Biost 536: Categorical Data Analysis in Epidemiology
Emerson, Fall 2013

Homework #2 Stata and Discussion
October 29, 2013

This document presents Stata code that might have been used to solve Homework #2. However, the main focus of this document is to highlight the differences and similarities among the many alternative analyses that could have been considered. Hence, the following disclaimers should be noted:

- I present analyses that I do not routinely recommend. My point is to demonstrate where they might not be answering the question in the most scientifically or statistically rigorous manner.
- I certainly do not recommend that when performing an analysis, an analyst would perform multiple analyses and then choose among them. Instead, by learning the properties of the analyses from this discussion, it would be hoped that any problems could be anticipated and the most appropriate analysis models chosen *a priori*.

This (quite lengthy) file is produced from Stata output that was obtained by executing the commented code in the file hw2key.do. Comments from the "do" file are displayed in lime (and can be safely ignored), and the Stata commands themselves are displayed in blue. `The Stata output is displayed in Courier font.` To that output, I have added **discussion of general properties of the analysis methods presented in boldface**. They are divided into

- "**Stata Comments**" that discuss the vagaries of producing and interpreting the analyses using Stata,
- "**Statistical Comments**" that discuss the general properties of the statistical analyses, and
- "**Scientific Comments**" that present the interpretation of the analyses.

Creating log, reading in data, restricting to pertinent cases

```
. log using z:documents/teach/courses/b536/hw2Stata.log, replace
-------------------------------------------------------------------------------
      name:  <unnamed>
       log:  z:documents/teach/courses/b536/hw2Stata.log
  log type:  text
 opened on:  29 Oct 2013, 02:59:18

. * The log file will echo all commands and then store the tabular output
. * Graphical output will need to be separately copied to, for instance,
. * a Word document
.
. * Note that I can input the file by reading directly from the course web pages.
. * I use the "quietly" prefix to suppress the many warnings that arise from
. * the use of "NA" as an indicator for missing data. (Stata would not have
. * given a warning if I had used the Stata code of "." for missing data.)
. quietly: infile id site age male bkrace smoker estrogen prevdis diab2 bmi ///
>      systBP aai cholest crp fib ttodth death cvddth using                  ///
>          http://www.emersonstatistics.com/datasets/inflamm.txt

. * First line in datafile had variable names
. list in 1

     +-------------------------------------------------------------------------+
  1. | id | site | age | male | bkrace | smoker | estrogen | prevdis | diab2 |
     |  . |    . |   . |    . |      . |      . |        . |       . |     . |
     |-----------------------------------------+---------------------------|
     | bmi | systBP | aai | cholest | crp | fib | ttodth  | death  | cvddth  |
     |   . |      . |   . |       . |   . |   . |      .  |     .  |      .  |
     +-------------------------------------------------------------------------+

. drop in 1
(1 observation deleted)

. * Restrict all analyses to females
. drop if male==1
(2096 observations deleted)

. * Remove cases with missing values for estrogen
. * (This is for the purposes of the homework. In real life, we would want to
. * carefully think through what we would do with cases having missing data.)
. tabulate estrogen, missing
```

```
   estrogen |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,559       88.12       88.12
          1 |        340       11.71       99.83
          . |          5        0.17      100.00
------------+-----------------------------------
      Total |      2,904      100.00

. drop if estrogen==.
(5 observations deleted)

. * Create indicator of death attributed to cardiovascular disease within 4 years
. * (Again, this is for the purposes of the homework. In real life, we would
. * definitely think through what it means to die of other causes prior to
. * dying from cardiovascular disease-- a situation termed a "competing risk".)
. * This code relies on my prior knowledge that no subjects were missing data
. * for either ttodth or cvddth and that no subjects were censored prior to 4 yrs
. * If I did not know those facts, I would have to decide how to handle any
. * censored data and add code to check for missing data.
. g cvddeath4=0

. replace cvddeath4=1 if ttodth<= 4*365.25 & cvddth==1
(91 real changes made)

. * Create a variable modeling interaction between estrogen and prevdis
. g estr_prev= estrogen * prevdis

. * Formatting variables for descriptive statistics output
. format age %8.2f

. format estrogen prevdis cvddeath4 %8.3f

. bysort estrogen: tabstat age prevdis cvddeath4, col(stat) ///
>       stat(n mean sd min q max) format

--------------------------------------------------------------------------------------------
-> estrogen = 0.000

    variable |         N      mean        sd       min       p25       p50       p75       max
-------------+------------------------------------------------------------------------------
         age |   2559.00     72.82      5.61     65.00     68.00     72.00     76.00    100.00
     prevdis |  2559.000     0.201     0.401     0.000     0.000     0.000     0.000     1.000
   cvddeath4 |  2559.000     0.034     0.182     0.000     0.000     0.000     0.000     1.000
--------------------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------------------
-> estrogen = 1.000

    variable |         N      mean        sd       min       p25       p50       p75       max
-------------+--------------------------------------------------------------------------------
         age |    340.00     70.57      4.30     65.00     68.00     69.00     73.00     87.00
     prevdis |   340.000     0.088     0.284     0.000     0.000     0.000     0.000     1.000
   cvddeath4 |   340.000     0.009     0.094     0.000     0.000     0.000     0.000     1.000
-------------------------------------------------------------------------------------------
```

Statistical comments

- **As a rule, I look at the descriptive statistics listed above.**
- **Variables prevdis and cvddeath4 are binary variables. My use of the above command to get relevant descriptive statistics is indicative of how lazy I can be: The sample mean is interpretable as the sample proportion, but the rest of the statistics are largely not of interest.**
- **As noted in class, the sample means are of the greatest interest to me as I try to consider whether differences among the estrogen exposed and unexposed groups might lead to confounding (I will of course have to also consider my beliefs about the other variables' association with death.)**

Scientific comments

- **In our sample, 340 subjects had prior exposure to estrogen, while 2,559 were unexposed.**
- **Subjects exposed to estrogen averaged 2.25 years younger than those who were unexposed.**
- **Prevalence of prior CVD was 8.8% among subjects exposed to estrogen, while the prevalence of prior CVD was 20.1% among those who were unexposed.**
- **Among 340 subjects exposed to estrogen, 3 (0.9%) were observed to die of CVD within 4 years of study accrual, while among the 2,559 subjects unexposed to estrogen, 88 (3.4%) were observed to die of CVD within 4 years of study.**

PROBLEM #1: Analyses based on risk difference (RD)

Statistical comments

- **In a two sample test of binomial proportions, analyses based on RD are probably the most common.**
  - **When adjusting for additional covariates, however, it is more common to base analyses on OR.**
- **In Problem #1, we focus on a number of tests that can be used to analyze RD.**
  - **Unadjusted analyses based on**
    - **Fisher's exact test (sometimes used in small samples but tends to be overly conservative; only provides tests, no estimates or CI; variant using unconditional exact approach available in some software (but not Stata))**
    - **chi square test (based on asymptotic results; may be anti-conservative in small samples; not typically used to provide CI; variant using unconditional exact approach available in some software (but not Stata))**
    - **t tests (nonstandard approach, but valid based on large sample theory; variants presume equal variance or allow for the possibility of unequal variance)**
    - **unweighted simple linear regression (variants presume homoscedasticity (equal variances) or allow for the possibility of heteroscedasticity (unequal variances); correspond exactly to the t test that presumes equal variances or approximately to the t test that allows for unequal variances, approximately)**
    - **weighted simple linear regression (weights estimated using the mean-variance relationship of the binomial distribution; variants allowing for use of "robust" standard errors)**
  - **Adjusted analyses based on**
    - **standardized rates**
      - **adjustment for main effects and interactions among confounders / precision variables averaging across any effect modification**
    - **unweighted multiple linear regression (variants presume homoscedasticity or allow for the possibility of heteroscedasticity)**
      - **adjustment for main effects (in various forms possibly including interactions among confounders / precision variables) with or without inclusion of interactions with predictor of interest (POI)**
    - **weighted multiple linear regression (weights estimated using the mean-variance relationship of the binomial distribution; variants allowing for use of "robust" standard errors)**
      - **adjustment for main effects (in various forms possibly including interactions among confounders / precision variables) with or without inclusion of interactions with predictor of interest (POI)**

```
. * PROBLEM #1: analyses based on RD
. *
. * OVERVIEW:
. * Problem 1a: unadjusted analyses of cvddeath4 - estrogen association
. *               Fisher's exact test: tabulate; cs
```

```
. *               chi-square test: tabulate; cs
. *               t-test: ttest
. *                       - equal variances
. *                       - unequal variances
. *               ordinary linear regression: regress
. *                       - presuming homoscedasticity
. *                       - allowing for possible heteroscedasticity
. *               weighted linear regression: binreg; glm
. *                       - estimated binomial weights with standard SE
. *                       - estimated binomial weights with robust SE
. *
. * Problem 1b: evidence of effect modification by prevdis
. *               descriptive statistics
. *               inferential statistics
. *
. * Problem 1c: evidence of confounding by prevdis
. *               Association of estrogen - prevdis in sample
. *               Association of prevdis - cvddeath4 after adjustment for estrogen
. *
. * Problem 1d: analyses of cvddeath4 - estrogen association adjusted for prevdis
. *               Adjustment for main effect
. *                       ordinary linear regression: regress
. *                               - presuming homoscedasticity
. *                               - allowing for possible heteroscedasticity
. *                       weighted linear regression: binreg; glm
. *                               - estimated binomial weights with standard SE
. *                               - estimated binomial weights with robust SE
. *               Adjustment for main effect and estrogen - prevdis interaction
. *                       stratified analysis: cs, ir
. *                       ordinary linear regression: regress
. *                               - presuming homoscedasticity
. *                               - allowing for possible heteroscedasticity
. *                       weighted linear regression: binreg; glm
. *                               - estimated binomial weights with standard SE
. *                               - estimated binomial weights with robust SE
. *
. * Problem 1e: evidence of further confounding by age
. *               Association of age - estrogen beyond adjustment for prevdis
. *                       based on regression
. *                       graphical assessment
. *               Association of age - cvddeath4 after adjustment for estrogen, prevdis
. *
. * Problem 1f: association of cvddeath4 - estrogen adjusted for prevdis, age
. *               Adjustment for main effect (unweighted vs weighted, classical vs robust)
```

```
. *                       Dichotomized
. *                       Dummy variables
. *                              Quintiles
. *                              Scientific intervals (5 year)
. *                       Grouped linear
. *                       Continuous linear
. *                       Quadratic
. *                       Piecewise linear
. *                       Splines
. *                 Adjustment for main effect and age - prevdis interaction
. *                        (Note could mix and match above models for main effect with
. *                         alternative models for interactions)
. *
.
```

Problem #1a: Analyses of RD in 2 by 2 contingency tables (two sample tests of binomial proportions)

```
. * EXECUTION OF CODE FOR PROBLEM 1
. * Problem 1a: unadjusted analyses of cvddeath4 - estrogen association
. *               Fisher's exact test: tabulate; cs
. tabulate estrogen cvddeath4, row exact

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |       cvddeath4
  estrogen |     0.000      1.000 |     Total
-----------+----------------------+----------
     0.000 |     2,471         88 |     2,559
           |     96.56       3.44 |    100.00
-----------+----------------------+----------
     1.000 |       337          3 |       340
           |     99.12       0.88 |    100.00
-----------+----------------------+----------
     Total |     2,808         91 |     2,899
           |     96.86       3.14 |    100.00

           Fisher's exact =                 0.007
   1-sided Fisher's exact =                 0.004

. cs cvddeath4 estrogen, exact

                 | estrogen                |
                 |   Exposed   Unexposed   |      Total
-----------------+-------------------------+------------
           Cases |         3          88   |         91
        Noncases |       337        2471   |       2808
-----------------+-------------------------+------------
           Total |       340        2559   |       2899
                 |                         |
            Risk |  .0088235    .0343884   |   .0313901
                 |                         |
                 |      Point estimate     |    [95% Conf. Interval]
                 |-------------------------+------------------------
 Risk difference |        -.0255649        |   -.0377575   -.0133723
```

```
      Risk ratio |         .2565842       |      .08164    .8064117
 Prev. frac. ex. |         .7434158       |    .1935883      .91836
 Prev. frac. pop |         .0871892       |
                 +-------------------------------------------------
                                   1-sided Fisher's exact P = 0.0039
                                   2-sided Fisher's exact P = 0.0073

. *               chi-square test: tabulate; cs
. tabulate estrogen cvddeath4, row
+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+
           |       cvddeath4
  estrogen |     0.000      1.000 |     Total
-----------+----------------------+----------
     0.000 |     2,471         88 |     2,559
           |     96.56       3.44 |    100.00
-----------+----------------------+----------
     1.000 |       337          3 |       340
           |     99.12       0.88 |    100.00
-----------+----------------------+----------
     Total |     2,808         91 |     2,899
           |     96.86       3.14 |    100.00

. cs cvddeath4 estrogen
                 | estrogen               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |         3          88  |         91
        Noncases |       337        2471  |       2808
-----------------+------------------------+------------
           Total |       340        2559  |       2899
                 |                        |
            Risk |  .0088235    .0343884  |   .0313901
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0255649       |   -.0377575   -.0133723
      Risk ratio |         .2565842       |      .08164    .8064117
 Prev. frac. ex. |         .7434158       |    .1935883      .91836
 Prev. frac. pop |         .0871892       |
```

```
+-------------------------------------------------
             chi2(1) =     6.45  Pr>chi2 = 0.0111
```

Stata Comments:

- **tabulate and cs can each be used to obtain**
  - **estimated risks within groups, and**
  - **either chi-squared or Fisher's exact test**
- **labeling of output in cs is a little obscure, if you ask me**
  - **the first argument variable is presumed to be an indicator of "cases" (cf: y in regression)**
  - **the second argument variable is presumed to be an indicator of "exposure" (cf: x in regression)**
- **cs (but not tabulate) provides point estimates and confidence intervals for risk difference (among others)**
  - **The CI for the risk difference correspond to those obtained from a weighted regression in which the weights are based on the inverse variance using estimated proportions (see comments re binreg and glm below)**

Statistics Comments:

- **In testing binomial proportions across two groups, we generally have a parameter of interest ($\theta = p_1 - p_0$) and a nuisance parameter (we can consider p_0 or p_1 or some average)**
  - **One issue we have to address is the mean-variance relationship: the variability of our estimated difference depends on both p_0 and p_1:**

$$Var(\hat{p}_1 - \hat{p}_0) = \frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}$$

  - **Under a null hypothesis $H_0: \theta = p_1 - p_0 = \theta_0$, we just have to estimate the variability for some value of p_0 as:**

$$Var(\hat{p}_1 - \hat{p}_0 \mid \theta = \theta_0) = \frac{(p_0 + \theta_0)(1 - p_0 - \theta_0)}{n_1} + \frac{p_0(1-p_0)}{n_0}$$

  - **However, we do not know p_0, and so we consider it a "nuisance".**
  - **Under the typical null hypothesis of no difference in proportions ($H_0: \theta = p_1 - p_0 = \theta_0 = 0$), we might consider an estimate of the nuisance parameter based on the combined samples (by inspecting the formula for the average $\bar{p}$, we see that the numerator is just the total number of cases, and the denominator the total sample size):**

$$\bar{p} = \frac{n_1\hat{p}_1 + n_0\hat{p}_0}{n_1 + n_0} \quad \Rightarrow \quad Var(\hat{p}_1 - \hat{p}_0 \mid \theta = \theta_0 = 0) = \frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_0}$$

- **When samples sizes are sufficiently large, we can use the chi square test which just assumes the above variance based on $\bar{p}$**
  - **The asymptotic distribution of the Z statistic computed using that variance is normally distributed.**

- **However, if sample sizes are too small, the chi squared test might be anti-conservative (have a true type I error larger than we want) for some values of p_0**
- **We can correct for this potential anti-conservatism using an unconditional exact test. (I did this using some programs I have written for R. I will make them available.)**
  - **In the data for this problem, the chi square test reports a two-sided P value of 0.01109. The unconditional exact test based on the chi square test reports a P value of 0.01174.**

- **Many people recommend the use of Fisher's exact test when sample sizes are small. This "conditions" on the sum of events observed across the groups (so it "conditions" on $\bar{p}$), and computes the probability of observing more extreme differences between groups among all possible study outcomes that have the same value of $\bar{p}$.**
  - **Usual rule of thumb: "small sample sizes" means expected counts are less than 5 in some cell**
  - **In this case, the expected counts for deaths within the estrogen group would be computed as**
    (Tot N) * (Prop cvddeath4) * (Prop estrogen) = 2899 * (91 / 2899) * (340 / 2899) = 10.7
  - **When expected counts in each cell are large, there is very little difference between the chi square test and Fisher's exact test**
  - **However, as noted in class, when expected counts are small, Fisher's exact test can be extremely conservative owing to the problem of discrete outcomes: the counts have to be integers.**
    - **Fisher's exact test is only exact if we are willing to sometimes flip a coin to decide whether to reject the null hypothesis or not. This is termed a "randomized test", and people are generally unwilling to do that.**
    - **Improvements to the Fisher's exact test are often possible using "unconditional exact tests" that consider the worst case p value as we vary p_0.**
    - **In the data for this problem, the two-sided Fisher's exact test reports a P value of 0.00735. The unconditional exact test based on the Fisher's exact statistic would report a P value of 0.00701.**
      - **The very small difference between the unconditional exact test and the exact test is because our sample sizes in each group are relatively large, so discreteness is not such a problem.**

Problem #1a: Analyses of RD in two samples using t tests (somewhat unorthodox but absolutely valid)

```
. *                t-test: ttest
. *                        - equal variances
. ttest cvddeath4, by(estrogen)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    2559    .0343884    .0036029    .1822604    .0273235    .0414534
       1 |     340    .0088235    .0050792    .0936561   -.0011672    .0188143
---------+--------------------------------------------------------------------
combined |    2899    .0313901    .0032391    .1743998     .025039    .0377413
---------+--------------------------------------------------------------------
    diff |            .0255649    .0100574                .0058445    .0452853
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.5419
Ho: diff = 0                     degrees of freedom =     2897

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9945         Pr(|T| > |t|) = 0.0111          Pr(T > t) = 0.0055
```

Statistical Comments

- **The standard approach ("99 out of 100 data analysts agree…") to the analysis of a 2 by 2 contingency table is to use the chi square test or Fisher's exact test.**
- **However, the t test is a perfectly good nonparametric test of means, and proportions are means. Hence, in reasonable sample sizes, we can use the t test.**
  - **If you do so, you should be prepared to have to make the above comment.**
- **In fact, there is truly little difference between using the chi square test and the t test.**
  - **The chi square test computes the variance using the average $\bar{p}$, while the t test computes the average either using a pooled variance or the estimated proportions from each group (see below).**
  - **The chi square test uses the standard normal distribution to compute critical values, while the t test uses the t distribution.**
    - **A t distribution with an infinite number of degrees of freedom is the standard normal.**
- **The above analysis presumed equal variances in both groups. Under the null hypothesis, this is reasonable, but under an alternative hypothesis it is not reasonable.**
  - **With a binary response, there is a mean-variance relationship.**
  - **If the means (proportions) are equal, the variances must also be equal.**
  - **If the means (proportions) are not equal, the variance must also be unequal.**

- **Hence, if the null hypothesis is true, the t test that presumes equal variances will provide valid inference (in large samples0 in the sense that the type I error will be correct.**

- **Of course, when choosing an "optimal" statistical test, we also need to consider the statistical power of the test (Recall that we have the highest positive predictive value of a test when the power is high and the type I error is low, with the "Bayes factor" in a two sample test that considers only simple hypotheses (single values for null and for alternative) being the ratio of power to type I error.**
- **We thus might want to consider the t test that allows for the possibility of unequal variances in each group.**
  - **Because it estimates the variance in each group independently, this test will not "borrow information" across the groups. This might lead to a less precise estimate of the unknown variance when the variances are in fact equal, but when the variances are equal <u>or</u> when the variances are not equal it will estimate the variance for each of the groups in an unbiased manner.**
  - **We can compare the t test that presumes equal variances, the t test that allows for unequal variances, and the chi squared test with respect to how the standard errors are estimated:**

$$\textit{Chi-squared test} \qquad \bar{p} = \frac{n_1\hat{p}_1 + n_0\hat{p}_0}{n_1 + n_0} \qquad \Rightarrow \qquad Var\left(\hat{p}_1 - \hat{p}_0 \mid \theta = \theta_0 = 0\right) = \frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_0}$$

$$t \text{ test: equal vrnc} \qquad s_p^2 = \frac{n_1\hat{p}_1(1-\hat{p}_1) + n_0\hat{p}_0(1-\hat{p}_0)}{n_1 + n_0 - 2} \qquad \Rightarrow \qquad Var\left(\hat{p}_1 - \hat{p}_0 \mid \theta = \theta_0 = 0\right) = \frac{s_p^2}{n_1} + \frac{s_p^2}{n_0}$$

$$= \frac{n_1 + n_0}{n_1 + n_0 - 2}\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_0} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_1}\right)$$

$$t \text{ test: unequal vrnc} \qquad \Rightarrow \qquad Var\left(\hat{p}_1 - \hat{p}_0 \mid \theta = \theta_0 = 0\right) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}$$

- **Using the results from the t test that presumes equal variances, we can anticipate whether a higher or lower p value will be obtained from the t test that allows for the possibility of unequal variances.**
- **Absent a mean-variance relationship:**
  - **If the group with the higher sample size also has larger variance, the P value from the t test that allows for unequal variances will be lower: the conservative nature of the t test that presumes equal variances will be corrected.**

  - **If the group with the higher sample size also has smaller variance, the P value from the t test that allows for unequal variances will be higher: the anti-conservative nature of the t test that presumes equal variances will be corrected.**
- **In the presence of the binomial mean-variance relationship, the two P values will behave in a similar fashion:**
  - **If the group with the higher sample size has a mean closer to 0.5 (and hence a larger variance), the P value from the t test that allows for unequal variances will be lower than the P value from the t test that presumes equal variances.**
  - **If the group with the higher sample size has a mean further away from 0.5 (and hence a smaller variance), the P value from the t test that allows for unequal variances will be higher than the P value from the t test that presumes equal variances.**
- **Because both tests are (asymptotically) valid in testing binary outcomes, we are free to choose either test (or the chi-squared test), so long as we do so before looking at the data. Hence, if we can anticipate which group would have a mean closer to 0.5 in advance, and providing we know the sample sizes for each group, it is okay to choose the test that will tend toward the smaller sample size. (Of course, if we guess wrong in advance <u>it is not okay to change our minds after looking at the data</u>.**
- **In this problem, there were more unexposed subjects. Based on prior studies that suggested lower CVD in women who took hormone therapy, we might anticipate that the death rate will be higher in the unexposed. We do not, however, anticipate as high as 50% mortality in the unexposed group over 4 years, so we anticipate the t test that presumes unequal variances will have a lower P value.**
  - **The following analysis bears this out.**

```
. *                       - unequal variances
. ttest cvddeath4, by(estrogen) unequal
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    2559    .0343884    .0036029    .1822604    .0273235    .0414534
       1 |     340    .0088235    .0050792    .0936561   -.0011672    .0188143
---------+--------------------------------------------------------------------
combined |    2899    .0313901    .0032391    .1743998     .025039    .0377413
---------+--------------------------------------------------------------------
    diff |            .0255649    .0062273                .0133396    .0377902
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   4.1053
Ho: diff = 0                     Satterthwaite's degrees of freedom =  741.117

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Problem #1a: Analyses of RD in two samples using classical unweighted simple linear regression (somewhat unorthodox but absolutely valid)

```
. *               ordinary linear regression: regress; glm
. *                        - presuming homoscedasticity
. regress cvddeath4 estrogen

      Source |       SS       df       MS              Number of obs =    2899
-------------+------------------------------           F(  1,  2897) =    6.46
       Model |  .196150448     1  .196150448           Prob > F      =  0.0111
    Residual |  87.9473473  2897  .030358076           R-squared     =  0.0022
-------------+------------------------------           Adj R-squared =  0.0019
       Total |  88.1434978  2898  .030415286           Root MSE      =  .17424

------------------------------------------------------------------------------
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0255649   .0100574    -2.54   0.011    -.0452853   -.0058445
       _cons |   .0343884   .0034443     9.98   0.000     .0276349     .041142
------------------------------------------------------------------------------
```

Statistical Comments

- **We are fitting a linear regression model in which**

$$E[cvddeath4 \mid estrogen] = \beta_0 + \beta_1 \times estrogen$$

- **Interpretation of the regression parameters:**
  - **The intercept β_0 is the mean (proportion) death within 4 years after study accrual among subjects unexposed to estrogen (*estrogen=0)*.**
  - **The slope β_1 is the difference in mean (proportion) death within 4 years after study accrual between subjects exposed to estrogen (*estrogen=1)* and subjects unexposed to estrogen (*estrogen=0)*.**
- **Properties of ordinary least squares regression: In simple linear regression (one predictor variable) having a binary predictor (a 0-1 variable),**
  - **the predicted values will be equal to the sample mean in each group:**
    - **$\hat{\beta}_0$ will be equal to the sample mean among subjects having a predictor value equal to 0**
    - **$\hat{\beta}_1$ will be equal to the difference in sample means: group 1 minus group 0**
  - **the estimated within group standard deviation (labeled "Root MSE" in Stata) will be the square root of the pooled variance, and**
  - **statistical inference will correspond <u>exactly</u> to the t test that presumes equal variances.**

- **Ordinary least squares regression corresponds to a "generalized linear model" (GLM) with no mean-variance relationship (so "family" is "Gaussian") and an identity link.**
  - **In most statistical software implementing a GLM (including Stata), the Gaussian family and identity link are the default values.**
  - **Hence, the following code will produce the same estimates and nearly the same inference**
    - **In most GLM programs, CI and P values are computed using an asymptotic normal distribution instead of the t distribution. With a large sample size, this makes no difference: the normal distribution is the t distribution with infinite degrees of freedom.**
    - **(My opinion: using the t distribution would make more sense for all GLM—not just Gaussian family.)**

```
. glm cvddeath4 estrogen, family(gaussian) link(identity)

Iteration 0:   log likelihood =  953.05459

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2897
                                                  Scale parameter =  .0303581
Deviance         =  87.94734731                   (1/df) Deviance =  .0303581
Pearson          =  87.94734731                   (1/df) Pearson  =  .0303581

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  -.656126
Log likelihood   =  953.0545892                   BIC             = -23007.29

------------------------------------------------------------------------------
             |                 OIM
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0255649   .0100574    -2.54   0.011    -.0452771   -.0058527
       _cons |   .0343884   .0034443     9.98   0.000     .0276377    .0411392
------------------------------------------------------------------------------
```

Problem #1a: Analyses of RD in two samples unweighted simple linear regression with "robust" SE (somewhat unorthodox but absolutely valid)

```
. *                        - allowing for possible heteroscedasticity
. regress cvddeath4 estrogen, robust

Linear regression                                      Number of obs =    2899
                                                       F(  1,  2897) =   16.88
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0022
                                                       Root MSE      =  .17424

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0255649    .006223    -4.11   0.000    -.0377668    -.013363
       _cons |   .0343884   .0036035     9.54   0.000     .0273228    .0414541
------------------------------------------------------------------------------

. glm cvddeath4 estrogen, robust

Iteration 0:   log pseudolikelihood =  953.05459

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2897
                                                  Scale parameter =  .0303581
Deviance         =  87.94734731                   (1/df) Deviance =  .0303581
Pearson          =  87.94734731                   (1/df) Pearson  =  .0303581

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  -.656126
Log pseudolikelihood =  953.0545892               BIC             = -23007.29

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0255649   .0062219    -4.11   0.000    -.0377596   -.0133702
       _cons |   .0343884   .0036029     9.54   0.000      .027327    .0414499
------------------------------------------------------------------------------
```

Statistical Comments

- **Regression parameter interpretation and estimates are unchanged by the use of the "robust SE".**
  - **(We are not weighting our estimates by any differences in variation, we are only changing our SE estimates.)**
- **We use the Huber-White sandwich estimator to compute "robust" standard errors.**
  - **This allows for the possibility of different variances across the different groups defined by the predictors.**
- **When used with a single binary predictor, the above analysis should mimic the t test that allows for the possibility of unequal variances**
  - **The major difference will be due to the use of the t distribution in the t test and (typically) the use of the normal distribution with the robust SE.**

Problem #1a: Analyses of RD in two samples using weighted simple linear regression with both standard and robust SE (somewhat unorthodox but absolutely valid)

```
. *               weighted linear regression: binreg; glm
. *                       - estimated binomial weights with standard SE
. binreg cvddeath4 estrogen, rd

Iteration 1:   deviance =  800.4201
Iteration 2:   deviance =  800.4201

Generalized linear models                         No. of obs      =      2899
Optimization     : MQL Fisher scoring             Residual df     =      2897
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  800.4201353                   (1/df) Deviance =  .2762928
Pearson          =         2899                   (1/df) Pearson  =   1.00069

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = u                       [Identity]

                                                  BIC             = -22294.81

------------------------------------------------------------------------------
             |                 EIM
   cvddeath4 | Risk Diff.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0255649   .0062208    -4.11   0.000    -.0377575   -.0133723
       _cons |   .0343884   .0036022     9.55   0.000     .0273282    .0414487
------------------------------------------------------------------------------

. glm cvddeath4 estrogen, family(binomial) link(identity)

Iteration 0:   log likelihood = -400.21007
Iteration 1:   log likelihood = -400.21007

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2897
                                                  Scale parameter =         1
Deviance         =  800.4201353                   (1/df) Deviance =  .2762928
Pearson          =         2899                   (1/df) Pearson  =   1.00069

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  .2774819
```

```
Log likelihood   = -400.2100677                     BIC             = -22294.81

------------------------------------------------------------------------------
             |                 OIM
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0255649   .0062208    -4.11   0.000    -.0377575   -.0133723
       _cons |   .0343884   .0036022     9.55   0.000     .0273282    .0414487
------------------------------------------------------------------------------
Coefficients are the risk differences.

. *                       - estimated binomial weights with robust SE
. binreg cvddeath4 estrogen, rd vce(robust)

Iteration 1:   deviance =  800.4201
Iteration 2:   deviance =  800.4201

Generalized linear models                         No. of obs      =      2899
Optimization     : MQL Fisher scoring             Residual df     =      2897
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  800.4201353                   (1/df) Deviance =  .2762928
Pearson          =         2899                   (1/df) Pearson  =   1.00069

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = u                       [Identity]

                                                  BIC             = -22294.81

------------------------------------------------------------------------------
             |             Semirobust
   cvddeath4 | Risk Diff.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0255649   .0062219    -4.11   0.000    -.0377596   -.0133702
       _cons |   .0343884   .0036029     9.54   0.000      .027327    .0414499
------------------------------------------------------------------------------

. glm cvddeath4 estrogen, family(binomial) link(identity) vce(robust)

Iteration 0:   log pseudolikelihood = -400.21007
Iteration 1:   log pseudolikelihood = -400.21007

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2897
                                                  Scale parameter =         1
```

```
Deviance         =  800.4201353                  (1/df) Deviance =  .2762928
Pearson          =         2899                  (1/df) Pearson  =   1.00069

Variance function: V(u) = u*(1-u)                [Bernoulli]
Link function    : g(u) = u                      [Identity]

                                                 AIC             =  .2774819
Log pseudolikelihood = -400.2100677              BIC             = -22294.81

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0255649   .0062219    -4.11   0.000    -.0377596   -.0133702
       _cons |   .0343884   .0036029     9.54   0.000      .027327    .0414499
------------------------------------------------------------------------------
Coefficients are the risk differences.
```

Stata Comments

- **Note that the interval estimates from the GLM model implementing the weighted regression are the same as produced by the Stata command cs.**
- **The Stata command binreg is just an alias for the GLM model**
  - **Owing to the numerical methods used, there may be slight differences between the two commands, but these are inconsequential**
- **Stata tells us about how many iterations were necessary to find the appropriate weights (see below).**
  - **Almost always, I am uninterested in this information about the number of iterations, so I deleted that output from this file (I do not know how to get Stata to stop telling me about it).**

Statistical Comments

- **In doing a weighted regression, our regression model has not changed, so the interpretation of our parameters has similarly not changed when the straight line relationship holds (and in the two sample problem, the straight line relationship always holds)**
  - **By weighting our data differently, the interpretation of what we mean by a "linear trend" might be different when the straight line relationship does not hold**
- **Statistical optimality theory tells us that more efficient estimates can be obtained when weighting observations by the inverse of their within group variance**
  - **This "optimality" is specified by the Gauss-Markov theorem which defines the linear combination of the data that provides an unbiased estimator with the lowest variance**
  - **<u>However</u>, this presumes that you do have the correct model for the mean, i.e., the straight line model is correct**
    - **(Note that the straight line model has to be correct with a binary predictor).**

- **Weighted least squares regression historically presumes you know the weights in advance.**
  - **In our problem, however, we want to weight be the variance derived from the mean-variance relationship.**
- **In the GLM, we iteratively fit our model, compute new weights, re-fit our model, etc.**
  - **This will then converge on the estimator obtained had we known the estimated weights in advance.**
- **In the most general case, the weighted regression estimates will be different from the ordinary least squares estimates**
  - **<u>However</u>, with a binary predictor there will be no difference in the estimates when compared to ordinary least squares.**
- **When using GLM, the SE estimates for the regression parameters will very closely mimic the "robust SE" computed with the sandwich estimator.**
  - **The iterative search was adjusting for differences in the variability across predictor groups.**
  - **Hence, using the "robust" option makes little difference in the binomial family.**
- **As will be noted in later analyses, we will need to exercise care when using the identity link with the binomial or Poisson families' mean-variance relationship. This can be traced to the form of the estimating equation.**
  - **The estimating equations for the GLM are actually derived from "maximum likelihood theory" (ML).**
    - **Likelihood theory considers the formula for the density of the probability distribution.**
    - **We find the estimates that maximize the likelihood (probability density) of the data.**
      - **We actually maximize the log likelihood, because that tends to be easier to work with.**
  - **For the very special case of Gaussian (normally distributed) data, maximizing the likelihood is equivalent to minimizing the squared difference between our data and the estimated mean: an intuitively appealing approach called "least squares estimation" (LSE)**
    - **The only difference is the variance appearing in the denominator of the ML estimating equation**
    - **We use calculus to tell us that the maximum will occur where the derivative is zero**

Sum of squares $$SS(\beta)=\sum_{i=1}^{n}(Y_i-\mu_i)^2=\sum_{i=1}^{n}(Y_i-X_i\beta)^2 \quad\Rightarrow\quad U(\hat{\beta})=\frac{\partial}{\partial\beta}SS(\beta)=\sum_{i=1}^{n}(Y_i-X_i\hat{\beta})X_i=0$$

Gaussian ML $$\log(L(\beta))\propto\sum_{i=1}^{n}\frac{(Y_i-\mu_i)^2}{\sigma^2}=\sum_{i=1}^{n}\frac{(Y_i-X_i\beta)^2}{\sigma^2} \quad\Rightarrow\quad U(\hat{\beta})=\frac{\partial}{\partial\beta}\log(L(\beta))=\sum_{i=1}^{n}\frac{(Y_i-X_i\hat{\beta})}{\sigma^2}X_i=0$$

  - **For several other commonly used probability distributions (e.g., binomial, Poisson, exponential), the estimating equations derived from ML look just like the estimating equation from the Gaussian model, except the families' mean-variance relationship is substituted for the variance.**
  - **Because of this, we often consider general approaches ("quasi-likelihood" (QL) and "generalized estimating equations" (GEE)) that use the estimating equations without presuming that we know the true probability distribution.**

- **This can then be thought of as just a form of "weighted least squares" (WLS): We are minimizing the weighted sums of squares, where the weights are the inverse of the variance defined by the mean-variance relationship**
  - *(Very technical note for the biostatistically inclined: In thinking about this as WLS, we figure out what we will do by pretending we know the weights we want to use and take the derivative. Then when it comes to actually doing the analysis, we estimate the weights from our regression model, using an iterative search.)*

Weighted sum of squares $$WSS(\beta)=\sum_{i=1}^{n} w_i\left(Y_i-\mu_i\right)^2=\sum_{i=1}^{n}\frac{\left(Y_i-\mu_i\right)^2}{V\left(\mu_i\right)} \quad \Rightarrow \quad U\left(\hat{\beta}\right)=\frac{\partial}{\partial\beta}SS(\beta)\propto\sum_{i=1}^{n}\frac{\left(Y_i-\hat{\mu}_i\right)}{V\left(\mu_i\right)}\frac{\partial\hat{\mu}_i}{\partial\beta}=0$$

- **When fitting the GLM models with the binomial or Poisson families' mean-variance relationships, we try to minimize the "weighted sum of squares", which is equivalent to finding estimates that make the following estimating equations equal to zero**

	Binomial		Poisson	
Identity	$U(\beta)=\sum_{i=1}^{n}\frac{Y_i-X_i\beta}{X_i\hat{\beta}\left(1-X_i\hat{\beta}\right)}X_i$	$\Rightarrow$ *RD*	$U(\beta)=\sum_{i=1}^{n}\frac{Y_i-X_i\beta}{X_i\hat{\beta}}X_i$	$\Rightarrow$ *RD*
Log	$U(\beta)=\sum_{i=1}^{n}\frac{Y_i-e^{X_i\beta}}{\left(1-e^{X_i\hat{\beta}}\right)}X_i$	$\Rightarrow$ *RR*	$U(\beta)=\sum_{i=1}^{n}\left(Y_i-e^{X_i\beta}\right)X_i$	$\Rightarrow$ *RR*
Logit	$U(\beta)=\sum_{i=1}^{n}\left(Y_i-\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right)X_i$	$\Rightarrow$ *OR*		

- **Violation of the straight line model (or random variability with small sample size) may lead in some cases to negative weights and failure to converge. In particular, we worry about**
  - **The identity link with binomial family mean-variance relationships: a problem often shows up with very low probabilities (near 0) or very high probabilities (near 1)**
  - **The log link with binomial family mean-variance relationships: a problem shows up with very high probabilities (near 1)**
  - **The identity link with Poisson family mean-variance relationships: a problem often shows up with very low probabilities (near 0)**

- **In this "saturated regression model" that fits two groups (estrogen exposed and unexposed) with two parameters (intercept and slope), the weighted regression does not produce different estimates, only different SE.**

Scientific Comments

- **The following table summarizes the results obtained from the various analyses of risk difference (RD) conducted with the binary response and the binary predictor**

- **The analysis I would have chosen *a priori* would have been**
  - **(If I were only doing the unadjusted analysis) chi square test statistic and CI from Stata's cs command**
  - **(If I were going to do additional adjustment) Wald test statistic and CI from Stata's regress with robust SE**
- **We find an association between estrogen use and four year CVD mortality, estimating 2.56% lower 4 year CVD mortality among estrogen exposed than estrogen unexposed (95% CI: 3.78% to 1.34% lower; two-sided P < .0001).**
  - **Proper CI interpretation: If the true RD were between -3.78% and -1.34%, it is not all that unusual to observe an estimated RD of -2.56%.**

Method	Variant	Stata	Point Estimate	Lower 95% CI	Upper 95% CI	P value
Fisher's Exact	Nominal	`tabulate, row exact`	(-0.0255649)			0.00735
		`cs, exact`	-0.0255649	-.0377575	-.0133723	0.00735
	Uncond exact	--				0.00701
Chi square	Nominal	`tabulate, row`	(-0.0255649)			0.01109
		`cs`	-0.0255649	-.0377575	-.0133723	0.01109
	Uncond exact	--				0.01174
T test	Equal vrnc	`ttest`	-0.0255649	-.0452853	-.0058445	0.01108
	Unequal vrnc	`ttest, unequal`	-0.0255649	-.0377902	-.0133396	0.00004
OLS Regression	Classical SE	`regress`	-0.0255649	-.0452853	-.0058445	0.01108
		`glm`	-0.0255649	-.0452771	-.0058527	0.01102
	Robust SE (Huber-White sandwich est)	`regress, robust`	-0.0255649	-.0377668	-.013363	0.00004
		`glm, robust`	-0.0255649	-.0377596	-.0133702	0.00004
WLS Regression (estimate wts from binomial mean-variance)	Model SE glm	`binreg, rd`	-0.0255649	-.0377575	-.0133723	0.00004
		`glm, family(binomial) link(identity)`	-0.0255649	-.0377575	-.0133723	0.00004
	Robust SE (Huber-White sandwich est)	`binreg, rd vce(robust)`	-0.0255649	-.0377596	-.0133702	0.00004
		`glm, family(binomial) link(identity) robust`	-0.0255649	-.0377596	-.0133702	0.00004

Problem #1b: Analyses to detect effect modification with RD

```
. * Problem 1b: evidence of effect modification by prevdis
. *               descriptive statistics
. bysort prevdis: tabstat cvddeath4, col(stat) stat(n mean) by(estrogen)

------------------------------------------------------------------------------------------------
-> prevdis = 0.000

Summary for variables: cvddeath4
     by categories of: estrogen

estrogen |         N      mean
---------+--------------------
       0 |      2045  .0180929
       1 |       310  .0064516
---------+--------------------
   Total |      2355  .0165605
------------------------------

------------------------------------------------------------------------------------------------
-> prevdis = 1.000

Summary for variables: cvddeath4
     by categories of: estrogen

estrogen |         N      mean
---------+--------------------
       0 |       514  .0992218
       1 |        30  .0333333
---------+--------------------
   Total |       544  .0955882
------------------------------
```

Scientific Comments

- **The descriptive statistics seem to suggest effect modification**
  - **In subjects without prior history of CVD (*prevdis*=0):**
    - **Exposed subjects have 1.81% CVD mortality within 4 years, and unexposed subjects have 0.65% mortality within 4 years for a risk difference of 1.16%**
  - **In subjects with prior history of CVD (*prevdis*=1):**
    - **Exposed subjects have 9.92% CVD mortality within 4 years, and unexposed subjects have 3.33% mortality within 4 years for a risk difference of 6.59%%**
  - **That difference between 1.16% and 6.59% mortality is suggestive of effect modification**

- **We can also consider whether the observed data is suggestive of effect modification in the population**
- **To effect this we consider a linear regression fitting main effects for *estrogen* and *prevdis* along with the multiplicative interaction *estr_prev***

```
. *               inferential statistics
. regress cvddeath4 estrogen prevdis estr_prev, robust

Linear regression                                      Number of obs =    2899
                                                       F(  3,  2895) =   14.84
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0331
                                                       Root MSE      =  .17158

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0116413   .0054227    -2.15   0.032     -.022274   -.0010086
     prevdis |   .0811289   .0135213     6.00   0.000     .0546166    .1076411
   estr_prev |  -.0542472   .0357643    -1.52   0.129    -.1243733     .015879
       _cons |   .0180929   .0029495     6.13   0.000     .0123097    .0238762
------------------------------------------------------------------------------
```

Statistical Comments

- **We are fitting a linear regression model in which**

$$E[cvddeath4 \mid estrogen] = \beta_0 + \beta_1 \times estrogen + \beta_2 \times prevdis + \beta_3 \times estr\_prev$$

- **Interpretation of the regression parameters:**
  - **The intercept β_0 is the mean (proportion) death within 4 years after study accrual among subjects unexposed to estrogen (*estrogen=0)* and without prior history of CVD *(prevdis=0)*.**
    - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the sample mean for that group having both *estrogen* and *prevdis* equal to 0.**
  - **The slope β_1 is the difference in mean (proportion) death within 4 years after study accrual between subjects exposed to estrogen (*estrogen=1)* and subjects unexposed to estrogen (*estrogen=0)* <u>while holding all other modeled variables constant</u>. This last condition is only possible when *prevdis*=0, so the interpretation of this slope is the association between estrogen and 4 year CVD mortality among those without prior history of CVD.**

- **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the difference in sample means for the group having *estrogen=1* and *prevdis=0* minus the sample mean for the group having *estrogen=0* and *prevdis=0*.**

- **The slope β_2 is the difference in mean (proportion) death within 4 years after study accrual between subjects with prior history of CVD (*prevdis=1)* and subjects without prior history of CVD (*prevdis=0)* <u>while holding all other modeled variables constant</u>. This last condition is only possible when *estrogen*=0, so the interpretation of this slope is the association between prior CVD and 4 year CVD mortality among those without exposure to estrogen.**
  - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the difference in sample means for the group having *estrogen=0* and *prevdis=1* minus the sample mean for the group having *estrogen=0* and *prevdis=0*.**
- **The slope β_3 is not as easily interpreted using the idea of "holding all other modeled variables constant", because either one of the other variables <u>must</u> be different if the value of *estr_prev* is different. Instead we note that according to the regression model**
  - **Expected mortality in the group having having *estrogen=0* and *prevdis=0* is β_0.**
  - **Expected mortality in the group having having *estrogen=1* and *prevdis=0* is $\beta_0 + \beta_1$.**
  - **Expected mortality in the group having having *estrogen=0* and *prevdis=1* is $\beta_0 + \beta_2$.**
  - **Expected mortality in the group having having *estrogen=1* and *prevdis=1* is $\beta_0 + \beta_1 + \beta_2 + \beta_3$.**
  - **Hence**
    - **β_1 measures the association between estrogen and 4 year CVD mortality in subjects without prior history of CVD, and**
    - **$\beta_1 + \beta_3$ measures the association between estrogen and 4 year CVD mortality in subjects with prior history of CVD, so**
    - **β_3 measures the difference between the "effect" of estrogen in subjects with prior CVD and the "effect" of estrogen in subjects without prior CVD (a "difference of differences").**
    - **Note that in this "saturated" model, the estimate corresponds exactly to the corresponding difference of difference in sample means.**
  - **(We could have expressed the interaction parameter as the difference in the effect of prior CVD across the estrogen exposure strata. The interpretation of the interaction is symmetric in this sense.)**
- **The CI and P value given in the regression output is interpretable as statistical tests for the existence of effect modification in the population.**

- **Similar models could have been examined using GLM, and the estimates would have been identical for the saturated model.**
  - **Note the similarity between the inference from OLS and the robust SE and the inference using the weighted regression without the robust SE (though use of robust SE would not make much difference).**

```
. glm cvddeath4 estrogen prevdis estr_prev, family(binomial) link(identity)

Iteration 0:   log likelihood =  -367.7927
Iteration 1:   log likelihood =  -367.7927

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2895
                                                  Scale parameter =         1
Deviance         =   735.585394                   (1/df) Deviance =  .2540882
Pearson          =         2899                   (1/df) Pearson  =  1.001382

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  .2564972
Log likelihood   =  -367.792697                   BIC             = -22343.71

------------------------------------------------------------------------------
             |                 OIM
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0116413   .0054189    -2.15   0.032    -.0222622   -.0010204
     prevdis |   .0811289   .0135119     6.00   0.000      .054646    .1076118
   estr_prev |  -.0542472   .0357397    -1.52   0.129    -.1242956    .0158013
       _cons |   .0180929   .0029474     6.14   0.000     .0123161    .0238698
------------------------------------------------------------------------------
Coefficients are the risk differences.
```

Stata Comments

- **Stata will compute the "fitted" or "predicted" means for each case using the observed values of the predictors and the regression parameter estimates**

Statistical Comments

- **We demonstrate that the "fitted" or "predicted" values match up with the sample means**

```
. predict fit0
(option mu assumed; predicted mean cvddeath4)

. bysort estrogen prevdis: summ cvddeath4 fit0

-------------------------------------------------------------------------------------------------
-> estrogen = 0.000, prevdis = 0.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   cvddeath4 |      2045    .0180929    .1333201          0          1
        fit0 |      2045    .0180929           0   .0180929   .0180929

-------------------------------------------------------------------------------------------------
-> estrogen = 0.000, prevdis = 1.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   cvddeath4 |       514    .0992218    .2992508          0          1
        fit0 |       514    .0992218           0   .0992218   .0992218

-------------------------------------------------------------------------------------------------
-> estrogen = 1.000, prevdis = 0.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   cvddeath4 |       310    .0064516    .0801919          0          1
        fit0 |       310    .0064516           0   .0064516   .0064516

-------------------------------------------------------------------------------------------------
-> estrogen = 1.000, prevdis = 1.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   cvddeath4 |        30    .0333333    .1825742          0          1
        fit0 |        30    .0333333           0   .0333333   .0333333
```

Problem #1c: Analyses to detect confounding with RD

```
. * Problem 1c: evidence of confounding by prevdis
. *               Association of estrogen - prevdis in sample
. tabulate prevdis estrogen, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |       estrogen
   prevdis |     0.000      1.000 |     Total
-----------+----------------------+----------
     0.000 |     2,045        310 |     2,355
           |     86.84      13.16 |    100.00
-----------+----------------------+----------
     1.000 |       514         30 |       544
           |     94.49       5.51 |    100.00
-----------+----------------------+----------
     Total |     2,559        340 |     2,899
           |     88.27      11.73 |    100.00


. *               Association of prevdis - cvddeath4 after adjustment for estrogen
. bysort estrogen: regress cvddeath4 prevdis, robust

------------------------------------------------------------------------------------------------
-> estrogen = 0.000

Linear regression                                      Number of obs =    2559
                                                       F(  1,  2557) =   36.02
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0318
                                                       Root MSE      =  .17937

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prevdis |   .0811289   .0135172     6.00   0.000     .0546231    .1076347
       _cons |   .0180929   .0029486     6.14   0.000     .0123111    .0238748
------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------------------
-> estrogen = 1.000

Linear regression                                       Number of obs =     340
                                                        F(  1,   338) =    0.66
                                                        Prob > F      =  0.4185
                                                        R-squared     =  0.0066
                                                        Root MSE      =  .09348

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prevdis |   .0268817   .0331848     0.81   0.418     -.038393    .0921564
       _cons |   .0064516   .0045607     1.41   0.158    -.0025193    .0154225
------------------------------------------------------------------------------
```

Scientific Comments

- **We find that there is an association between estrogen use and prior CVD in the sample**
  - **Among subjects without prior CVD 13.16% have exposure to estrogen, while among subjects with prior CVD only 5.51% have exposure to estrogen.**
- **Based on our prior belief, history of prior CVD is likely causally associated with 4 year CVD mortality independent of estrogen exposure.**
  - **We can examine the data to see if those beliefs are supported by our data:**
    - **Among subjects without estrogen exposure, 4 year CVD mortality is 8.11% higher in subjects with prior CVD than it is in subjects no history of prior CVD.**
    - **(Among subjects with estrogen exposure, 4 year CVD mortality is 2.69% higher in subjects with prior CVD than it is in subjects no history of prior CVD. This difference may not seem as substantial, but we would regard the condition met if it is met in either estrogen exposure stratum.)**

Statistical Comments

- **Note that because we are basing inference on RD, our data is "collapsible" if there is no confounding**
- **We can thus compare the unadjusted estimate of RD to the estimate that is adjusted for *prevdis* on the next page**
  - **The unadjusted estimated RD was -2.56%, while the estimated RD adjusted for *prevdis* is -1.68%**
  - **I regard such a difference to be of a magnitude that would be consistent with confounding**

Problem #1d: Analyses to adjust for *prevdis*: unweighted regression analysis adjusting for main effect

```
. * Problem 1d: analyses of cvddeath4 - estrogen association adjusted for prevdis
. *               Adjustment for main effect
. *                       ordinary linear regression: regress
. *                                - allowing for possible heteroscedasticity
. regress cvddeath4 estrogen prevdis, robust
Linear regression                                      Number of obs =    2899
                                                       F(  2,  2896) =   21.67
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0323
                                                       Root MSE      =  .17162
------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0168093   .0060086    -2.80   0.005    -.0285909   -.0050276
     prevdis |    .077742   .0128537     6.05   0.000     .0525387    .1029453
       _cons |   .0187732    .002948     6.37   0.000     .0129928    .0245536
------------------------------------------------------------------------------
```

Scientific Comments

- **After adjustment for prior history of CVD, we find a statistically significant association between estrogen exposure and lower 4 year CVD mortality. A risk difference regression model adjusting for prior history of CVD finds that subjects with estrogen exposure have an absolute 1.68% lower risk of 4 year CVD mortality than subjects without estrogen exposure but who have similar history of CVD (95% CI: 2.86% to 0.503% lower; two-sided P = 0.005).**

Statistical Comments

- **We did not fit a saturated model, hence the "adjusted" estrogen effect may be averaging across different effects in different *prevdis* strata.**
- **In the case of two binary predictors and ordinary least squares regression, we can anticipate how the adjusted estimate will average the effects observed in the two strata.**
- **Basically, the "adjusted" effect will be a weighted average of the two stratum specific effects (the RD in the stratum without prior history of CVD (*prevdis=0*) and the RD in the stratum with prior history of CVD (*prevdis=1)*)**
    - **Technical note for the biostatistically interested: The weight given to each stratum will be proportional to the harmonic mean of the sample sizes for the estrogen groups in that stratum. (Recall that the harmonic mean of the sample sizes would be proportional to the reciprocal of the sum of the reciprocals of the sample sizes.)**
        - **Weight for stratum without prior history of CVD: 1 / ((1 / 2045) + (1 / 310)) = 269.19**
        - **Weight for stratum without prior history of CVD: 1 / ((1 / 514) + (1 / 30)) = 28.346**
- **If that is the case, we will obtain different weighted estimates of the strata when we use GLM with the mean-variance relationship of the binomial family:**

Problem #1d: Analyses to adjust for *prevdis*: weighted regression analysis adjusting for main effect

```
. *                        weighted linear regression: binreg; glm
. *                                - estimated binomial weights with standard SE
. glm cvddeath4 estrogen prevdis, family(binomial) link(identity)

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2896
                                                  Scale parameter =         1
Deviance         =  736.9232735                   (1/df) Deviance =  .2544625
Pearson          =         2899                   (1/df) Pearson  =  1.001036

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  .2562688
Log likelihood   = -368.4616367                   BIC             = -22350.34

------------------------------------------------------------------------------
             |                 OIM
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0121974   .0051992    -2.35   0.019    -.0223876   -.0020072
     prevdis |   .0776328   .0129369     6.00   0.000      .052277    .1029886
       _cons |   .0182657   .0029728     6.14   0.000     .0124391    .0240922
------------------------------------------------------------------------------
Coefficients are the risk differences.
```

Scientific Comments

- **After adjustment for prior history of CVD, we find a statistically significant association between estrogen exposure and lower 4 year CVD mortality. A weighted GLM risk difference regression model adjusting for prior history of CVD finds that subjects with estrogen exposure have an absolute 1.22% lower risk of 4 year CVD mortality than subjects without estrogen exposure but who have similar history of CVD (95% CI: 2.24% to 0.201% lower; two-sided P = 0.019).**

Problem #1d: Analyses to adjust for *prevdis*: using standardized rates to potentially average over effect modification

Statistical Comments

- **An alternative to the regression model is to use a stratified analysis**
  - **Suppose we have strata labeled $s = 1, 2, \ldots, S$**
  - **We estimate incidences p_{s1} (in the estrogen exposed) and p_{s0} (in the estrogen unexposed) in each stratum**
  - **We then average across the stratum specific incidences using some predefined weights $w_1, w_2, \ldots, w_S$**
    - **(We often choose weights that sum to 1 in order to be able to ignore the denominator)**

$$p_1 = \frac{\sum_{s=1}^{S} w_s p_{s1}}{\sum_{s=1}^{S} w_s} \qquad p_0 = \frac{\sum_{s=1}^{S} w_s p_{s0}}{\sum_{s=1}^{S} w_s}$$

  - **The standardized RD is then $\theta = p_1 - p_0$ is then interpretable as the difference of incidence in a population where the weights represent the prevalence of the different strata.**
    - **For this reason, we often standardize age-adjusted rates according to some reference population that we care about**
  - **Because we are considering RD (a difference of means) we could have equivalently estimated an effect within each stratum $\theta_s = p_{s1} - p_{s0}$, and then calculated a weighted average across strata**

$$p_1 = \frac{\sum_{s=1}^{S} w_s p_{s1}}{\sum_{s=1}^{S} w_s} \qquad p_0 = \frac{\sum_{s=1}^{S} w_s p_{s0}}{\sum_{s=1}^{S} w_s} \qquad \theta = p_1 - p_0 \qquad \Leftrightarrow \qquad \theta_s = p_{s1} - p_{s0} \qquad \theta = \frac{\sum_{s=1}^{S} w_s \theta_s}{\sum_{s=1}^{S} w_s}$$

- **Note that this approach does not have to assume that the effect in each stratum is the same**
  - **If it is the same, then the stratified rate is adjusted for confounding and/or precision**
  - **If it is not the same, then the stratified rate is some sort of average effect.**
- **Because of this ambiguity in how the "adjusted" effect is estimated, we might want to explicitly decide on the weights assigned to each stratum. Choices include the following:**

- **Weighting according to the distribution of the estrogen exposed across *prevdis* strata (Stata's internal standards)**

```
. *               Adjustment for main effect and estrogen - prevdis interaction
. *                        stratified analysis: cs
. cs cvddeath4 estrogen, rd by(prevdis) istandard

         prevdis |       RD       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
               0 |   -.0116413    -.0222622  -.0010204          310
               1 |   -.0658885    -.1351271   .0033501           30
-----------------+-------------------------------------------------
           Crude |   -.0255649    -.0377575  -.0133723
 I. Standardized |   -.0164278    -.0278776   -.004978
```

- **Weighting according to the distribution of the estrogen unexposed across *prevdis* strata (Stata's external standards)**

```
. cs cvddeath4 estrogen, rd by(prevdis) estandard

         prevdis |       RD       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
               0 |   -.0116413    -.0222622  -.0010204         2045
               1 |   -.0658885    -.1351271   .0033501          514
-----------------+-------------------------------------------------
           Crude |   -.0255649    -.0377575  -.0133723
 E. Standardized |   -.0225374      -.03883  -.0062447
```

- **Weighting according to some distribution that might matter to us**
  - **I give the example of using weights according to the distribution of *prevdis* in the entire sample**

```
. g std= 544 / 2899

. replace std= 1 - std if prevdis==0
(2355 real changes made)

. cs cvddeath4 estrogen, rd by(prevdis) standard(std)

         prevdis |       RD       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
               0 |   -.0116413    -.0222622  -.0010204     .8123491
               1 |   -.0658885    -.1351271   .0033501     .1876509
-----------------+-------------------------------------------------
           Crude |   -.0255649    -.0377575  -.0133723
    Standardized |   -.0218208    -.0374173  -.0062244
```

- **Weighting proportional to the harmonic means of the sample sizes in each group (see above)**
  - **This is the weighting that would be used in an unweighted regression analysis with identity link**
  - **This is the "efficient" weighting to use when there is no effect modification**
  - **Note that this point estimate should agree exactly with the results from the linear regression that adjusted for *prevdis***
    - **The CI will differ very slightly due to the difference in how the Huber-White sandwich estimator approximates the variances that might differ across groups. In general the CI will be very close.**

```
. drop std

. g std= 1 / ( (1 / 2045) + (1 / 310) )

. replace std = 1 / ( (1 / 514) + (1 / 30) ) if prevdis==1
(544 real changes made)

. cs cvddeath4 estrogen, rd by(prevdis) standard(std)

         prevdis |       RD       [95% Conf. Interval]       Weight
-----------------+------------------------------------------------
               0 |   -.0116413    -.0222622  -.0010204     269.1932
               1 |   -.0658885    -.1351271   .0033501     28.34559
-----------------+------------------------------------------------
           Crude |   -.0255649    -.0377575  -.0133723
    Standardized |   -.0168093    -.0284644  -.0051541
```

Problem #1d: Analyses to adjust for *prevdis*: weighted regression analysis adjusting for main effect and interaction between *estrogen* and *prevdis*

Statistical Comments

- **In the above stratified analyses, we have no opportunity to estimate the effect of *prevdis***
  - **This is not of too much concern, because that was not our question**
- **If we did care, we could fit a model with effect modification**
  - **A multiplicative effect is all that makes sense when we have a binary POI and a binary confounder**
  - **In other cases, we could consider more complicated interactions, but a multiplicative interaction is still the most commonly used**

```
. *                      ordinary linear regression: regress
. *                                - allowing for possible heteroscedasticity
. regress cvddeath4 estrogen prevdis estr_prev, robust
Linear regression                                      Number of obs =    2899
                                                       F(  3,  2895) =   14.84
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0331
                                                       Root MSE      =  .17158
------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0116413   .0054227    -2.15   0.032     -.022274   -.0010086
     prevdis |   .0811289   .0135213     6.00   0.000     .0546166    .1076411
   estr_prev |  -.0542472   .0357643    -1.52   0.129    -.1243733     .015879
       _cons |   .0180929   .0029495     6.13   0.000     .0123097    .0238762
------------------------------------------------------------------------------
```

Stata Comments

- **Stata provides a P value and CI for each parameter estimate**
  - **The interpretation of the parameters has changed due to inclusion of the interaction (see above)**

Statistical Comments

- **We now have two parameters modeling an association between 4 year CVD mortality and estrogen exposure**
  - **If either *estrogen* or *estr_prev* parameters is nonzero, there is an association**
- **Note that we now have a multiple comparison issue if we look at both the *estrogen* and *estr_prev* slopes' P values**
  - **A simple Bonferroni correction would say that because we are looking at two parameters, we would compare each P value to $\alpha / 2$**
  - **Such an analysis would judge that there was not enough evidence to declare an association exists when $\alpha = 0.05$**

Statistical Comments

- **The better approach is to instead test all terms that involve our POI *estrogen***
  - **Such a test addresses the question of any association**
- **Conceptually, in linear regression it may be more powerful to adjust for an interaction than just using simple adjustment for the main effect of the confounder, if adjusting for effect modification greatly reduces our variabililty**
  - **With binary outcomes and the robust SE, however, we will not expect much improvement**
- **On the other hand, the fact that we have to test two parameters may tend to decrease our statistical power**
  - **In this case, we had less statistical significance (P = 0.017) than we did in the regression adjusting for *prevdis* alone when using robust SE (P = 0.005)**

```
. test estrogen estr_prev

 ( 1)  estrogen = 0
 ( 2)  estr_prev = 0

       F(  2,  2895) =    4.04
            Prob > F =    0.0177
```

Statistical Comments

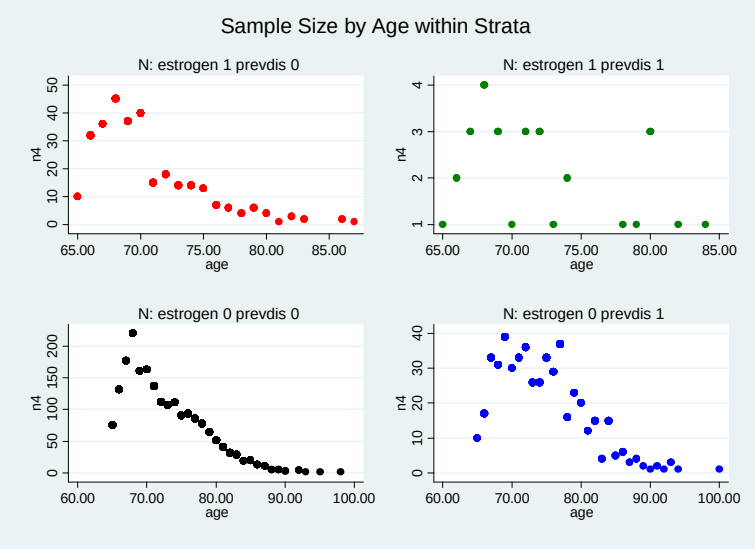
- **Summarizing the effect overall will have to use some other technique**
  - **We could use the above analysis to estimate effects within each subgroup, but our question was about an overall effect**
- **We could use Stata's lincom command to get standardized estimates**
  - **If we use the same weights as in a stratified analysis, we will get almost the same inference**
  - **Here I used weights proportional to the harmonic means of the sample size in each *prevdis* stratum in order to be able to compare the inference to the inference when we adjusted for only the main effect of *prevdis***
  - **The differences are negligible**

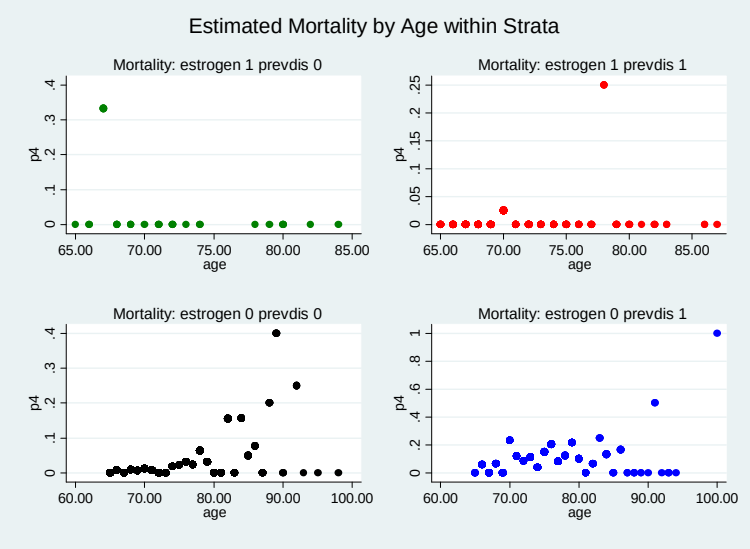
```
. lincom (269.1932 / (269.1932 + 28.34559)) * estrogen + (28.34559 / (269.1932 + 28.34559)) * (estrogen +
estr_prev)

 ( 1)  estrogen + .0952669*estr_prev = 0

------------------------------------------------------------------------------
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0168093   .0059507    -2.82   0.005    -.0284774   -.0051411
------------------------------------------------------------------------------
```

Problem #1e: Analyses to detect further confounding by age

```
. * Problem 1e: evidence of further confounding by age
. *               Association of age - estrogen beyond adjustment for prevdis
. *                        based on regression
. bysort prevdis: regress estrogen age, robust

-------------------------------------------------------------------------------------------
-> prevdis = 0.000

Linear regression                                      Number of obs =    2355
                                                       F(  1,  2353) =   53.01
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0163
                                                       Root MSE      =  .33547

------------------------------------------------------------------------------
             |               Robust
    estrogen |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0080881   .0011109    -7.28   0.000    -.0102665   -.0059096
       _cons |   .7154242   .0828228     8.64   0.000      .553011    .8778374
------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------
-> prevdis = 1.000

Linear regression                                      Number of obs =     544
                                                       F(  1,   542) =    4.76
                                                       Prob > F      =  0.0296
                                                       R-squared     =  0.0078
                                                       Root MSE      =   .2278

------------------------------------------------------------------------------
             |               Robust
    estrogen |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0033723   .0015459    -2.18   0.030    -.0064089   -.0003356
       _cons |   .3053106   .1185853     2.57   0.010     .0723674    .5382537
------------------------------------------------------------------------------
```

```
. *                Association of age - estrogen beyond adjustment for prevdis
. *                        graphical display of sample size, mortality by age within estr_prev strata
```





Scientific Comments

- **Estrogen use clearly associated with age in the regressions adjusting for *prevdis***
  - **Note that the prevalence of estrogen exposure is estimated to be 0.809% lower for each year difference in age among subjects with no prior history of CVD and to be 0.337% lower for each year difference in age among subjects with prior history of CVD**
  - **Our data set has ages ranging between 65 yo and 100 yo. Hence, those estimates might be better judged against a 10 year difference in age: 8.09% and 3.37% lower prevalence per decade difference in age**
- **Age associated with mortality within each stratum**

Problem #1f: Analyses to adjust for age

Stata comments

- **Stata has commands that allow categorizing data (xtile) and making splines (mkspline) easy**
- **Stata's facility for fitting "dummy" ("factor" or "class") variables and multiplicative interactions are often very useful**

```
. * Problem 1f: association of cvddeath4 - estrogen adjusted for prevdis, age
. *         Adjustment for main effect (unweighted vs weighted, classical vs robust)

. * Create variables to facilitate more flexible modeling of age
. *       age squared to be used in a quadratic fit
. g agesqr= age ^2

. *       splitting at median
. xtile ageQ2 = age, nq(2)

. *       coding within quintiles to be used as dummny variables
. xtile ageQ5 = age, nq(5)

. *       coding within (approx) 5 year age groups to be used as dummy variables
. g age5yr = age

. recode age5yr 65/69=1 70/74=2 75/79=3 80/84=4 85/max=5
(age5yr: 2899 changes made)

. *               I prefer to code variables by mean value within interval
. tabstat age, by(age5yr) col(stat) stat(n mean sd min q max) format
Summary for variables: age
     by categories of: age5yr
  age5yr |         N      mean        sd       min       p25       p50       p75       max
---------+--------------------------------------------------------------------------------
      67 |   1069.00     67.38      1.25     65.00     66.00     68.00     68.00     69.00
      72 |    892.00     71.77      1.44     70.00     70.00     72.00     73.00     74.00
      77 |    589.00     76.80      1.39     75.00     76.00     77.00     78.00     79.00
      82 |    253.00     81.58      1.40     80.00     80.00     81.00     83.00     84.00
      88 |     96.00     87.63      3.01     85.00     85.00     87.00     89.00    100.00
---------+--------------------------------------------------------------------------------
   Total |   2899.00     72.56      5.52     65.00     68.00     71.00     76.00    100.00
--------------------------------------------------------------------------------------------
. recode age5yr 1=67 2=72 3=77 4=82 5=88
(age5yr: 2899 changes made)

. *       creating linear splines within (approx) 5 year age groups
. mkspline age65 69.5 age70 74.5 age75 79.5 age80 84.5 age85 = age
```

Problem #1f: Analyses to adjust for age : dichotomization at median

```
. *                Dichotomized
. regress cvddeath4 estrogen prevdis ageQ2, robust

Linear regression                                      Number of obs =    2899
                                                       F(  3,  2895) =   17.98
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0372
                                                       Root MSE      =  .17122

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0125033   .0058917    -2.12   0.034    -.0240557   -.0009509
     prevdis |   .0741651   .0128437     5.77   0.000     .0489814    .0993488
       ageQ2 |   .0247358   .0063539     3.89   0.000     .0122771    .0371944
       _cons |  -.0178101   .0084728    -2.10   0.036    -.0344235   -.0011967
------------------------------------------------------------------------------
```

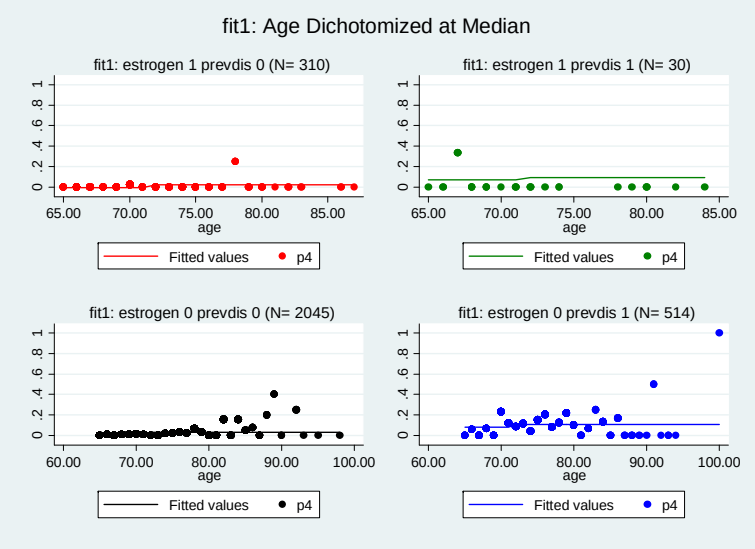
Statistical Comments

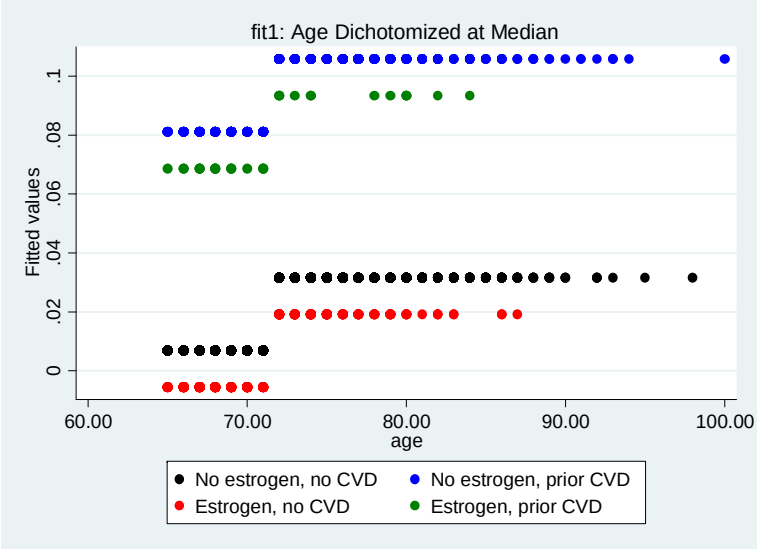
- **Splitting the data at the median ends up fitting all the older ages as if they had constant risk, as displayed in the following graph**
- **This seems silly to me. I care more about differences in age on the more scientific scale of years, rather than in percentiles of the population (and especially not the percentiles of the sample)**
- **Because I only modeled main effects, the predicted lines for the various strata must all be parallel.**
    - **This will carry forward to all models in which no interaction is fit.**

Scientific Comments

- **The interpretation of the estrogen coefficient is now "the RD comparing subjects exposed to estrogen to subjects who were not exposed, but who had the same prior history of CVD and were in the same half of the age distribution"**

```
. predict fit1
(option xb assumed; fitted values)

. * (see do file for code used to produce graphs)
```


fit1: Age Dichotomized at Median
fit1: estrogen 1 prevdis 0 (N= 310)
fit1: estrogen 1 prevdis 1 (N= 30)
fit1: estrogen 0 prevdis 0 (N= 2045)
fit1: estrogen 0 prevdis 1 (N= 514)
0 .2 .4 .6 .8 1
65.00 70.00 75.00 80.00 85.00
60.00 70.00 80.00 90.00 100.00
age
Fitted values
p4


fit1: Age Dichotomized at Median
Fitted values
0 .02 .04 .06 .08 .1
60.00 70.00 80.00 90.00 100.00
age
No estrogen, no CVD
No estrogen, prior CVD
Estrogen, no CVD
Estrogen, prior CVD

Problem #1f: Analyses to adjust for age : categorization by quintiles; modeling with dummy variables

```
. *                    Dummy variables : Quintiles
. regress cvddeath4 estrogen prevdis i.ageQ5, robust

Linear regression                                       Number of obs =    2899
                                                        F(  6,  2892) =   10.57
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0428
                                                        Root MSE      =   .1708
------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0107076   .0058563    -1.83   0.068    -.0221905    .0007752
     prevdis |   .0726238   .0128373     5.66   0.000     .0474526     .097795
       ageQ5 |
          2  |   .0125381   .0074469     1.68   0.092    -.0020637    .0271399
          3  |   .0069448   .0068327     1.02   0.310    -.0064527    .0203423
          4  |   .0264212   .0088937     2.97   0.003     .0089825    .0438599
          5  |   .0504967   .0113501     4.45   0.000     .0282417    .0727517
       _cons |   .0013138   .0033625     0.39   0.696    -.0052793    .0079068
------------------------------------------------------------------------------
```

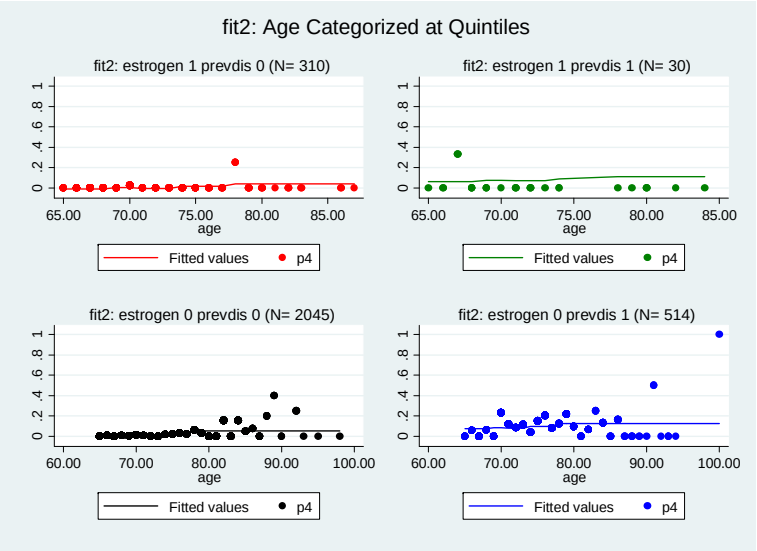
Statistical Comments

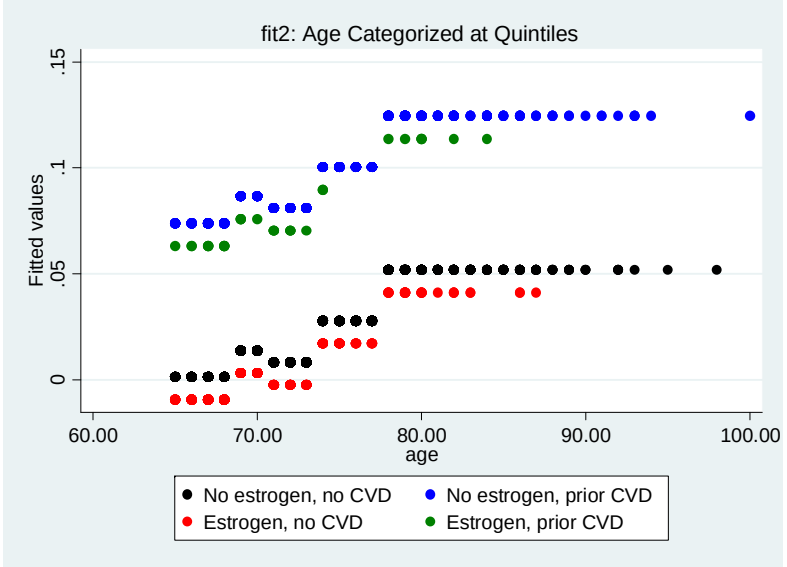
- **Splitting the data at quintiles divides the sample more finely, but it tends to result in very short age intervals for some groups, as displayed in the following graph**
- **This again seems silly to me. I care more about differences in age on the more scientific scale of years, rather than in percentiles of the population (and especially not the percentiles of the sample)**

Scientific Comments

- **The interpretation of the estrogen coefficient is now "the RD comparing subjects exposed to estrogen to subjects who were not exposed, but who had the same prior history of CVD and were in the same quintile of the age distribution"**

```
. predict fit2
(option xb assumed; fitted values)
. * (see do file for code used to produce graphs)
```


fit2: Age Categorized at Quintiles
fit2: estrogen 1 prevdis 0 (N= 310)
fit2: estrogen 1 prevdis 1 (N= 30)
fit2: estrogen 0 prevdis 0 (N= 2045)
fit2: estrogen 0 prevdis 1 (N= 514)
0 .2 .4 .6 .8 1
65.00 70.00 75.00 80.00 85.00
60.00 70.00 80.00 90.00 100.00
age
Fitted values
p4


fit2: Age Categorized at Quintiles
Fitted values
0 .05 .1 .15
60.00 70.00 80.00 90.00 100.00
age
No estrogen, no CVD
No estrogen, prior CVD
Estrogen, no CVD
Estrogen, prior CVD

Problem #1f: Analyses to adjust for age : categorizing in (approx) 5 year intervals; modeling with dummy variables

```
. *               Dummy variables : Scientific intervals (5 year)
. regress cvddeath4 estrogen prevdis i.age5yr, robust

Linear regression                                      Number of obs =    2899
                                                       F(  6,  2892) =   10.88
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0450
                                                       Root MSE      =  .17061
------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0093431   .0058526    -1.60   0.111    -.0208188    .0021326
     prevdis |    .071772   .0127901     5.61   0.000     .0466933    .0968506
      age5yr |
         72  |   .0157194   .0057321     2.74   0.006     .0044799    .0269588
         77  |    .045269   .0097963     4.62   0.000     .0260605    .0644776
         82  |   .0363536   .0145961     2.49   0.013     .0077338    .0649735
         88  |   .0729661   .0300742     2.43   0.015     .0139971    .1319351
       _cons |  -.0006053   .0030667    -0.20   0.844    -.0066183    .0054078
------------------------------------------------------------------------------
```

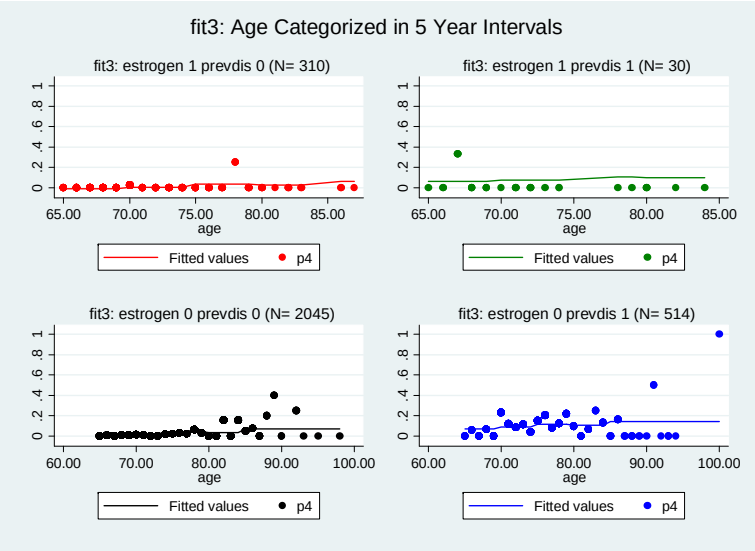
Statistical Comments

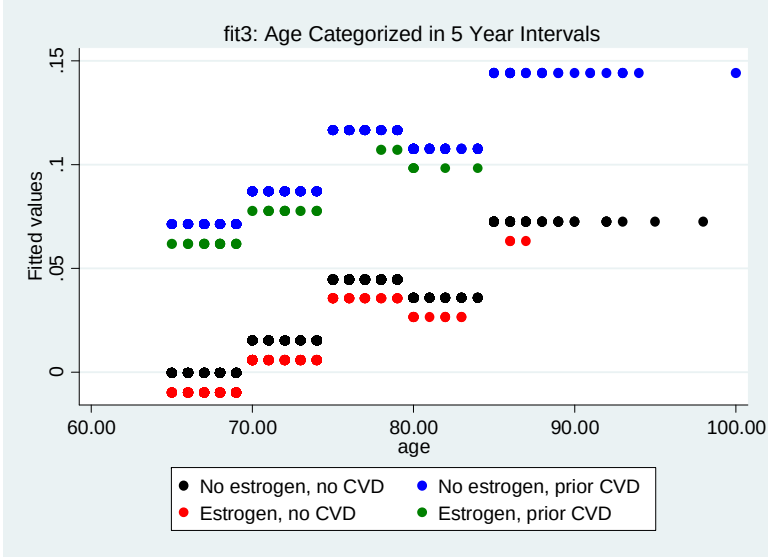
- **Splitting the data at 5 year intervals considers the scientific scale of years. Note however that owing to relatively sparse data at the oldest ages, my last interval covers the 15 years from 85-100**
- **Note that with dummy variables, we do not force continuous or "monotonic" relationships**
  - **The estimates can go up or down for each successive age group**

Scientific Comments

- **The interpretation of the estrogen coefficient is now "the RD comparing subjects exposed to estrogen to subjects who were not exposed, but who had the same prior history of CVD and were in the same 5 year age group (15 years for over 85)"**

```
. predict fit3
(option xb assumed; fitted values)
. * (see do file for code used to produce graphs)
```


fit3: Age Categorized in 5 Year Intervals
fit3: estrogen 1 prevdis 0 (N= 310)
fit3: estrogen 1 prevdis 1 (N= 30)
fit3: estrogen 0 prevdis 0 (N= 2045)
fit3: estrogen 0 prevdis 1 (N= 514)
0 .2 .4 .6 .8 1
65.00 70.00 75.00 80.00 85.00
60.00 70.00 80.00 90.00 100.00
age
Fitted values
p4


fit3: Age Categorized in 5 Year Intervals
Fitted values
0 .05 .1 .15
60.00 70.00 80.00 90.00 100.00
age
No estrogen, no CVD
No estrogen, prior CVD
Estrogen, no CVD
Estrogen, prior CVD

Problem #1f: Analyses to adjust for age : categorizing in (approx) 5 year intervals; modeling continuously

```
. *                        Grouped linear
. regress cvddeath4 estrogen prevdis age5yr, robust

Linear regression                                      Number of obs =    2899
                                                       F(  3,  2895) =   20.11
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0435
                                                       Root MSE      =  .17065

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0099337   .0058642    -1.69   0.090    -.0214322    .0015648
     prevdis |     .07182   .0128227     5.60   0.000     .0466775    .0969626
      age5yr |   .0033666   .0006979     4.82   0.000     .0019981    .0047351
       _cons |  -.2252545   .0490502    -4.59   0.000    -.3214312   -.1290777
------------------------------------------------------------------------------
```

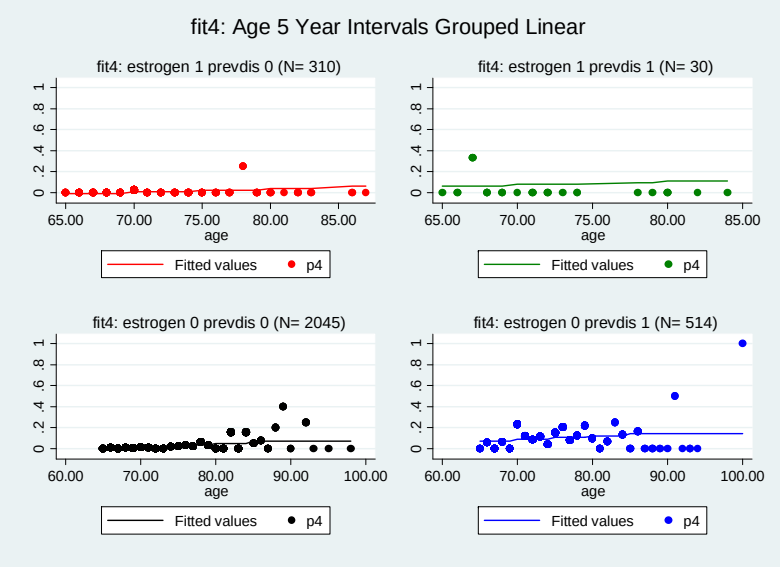
Statistical Comments

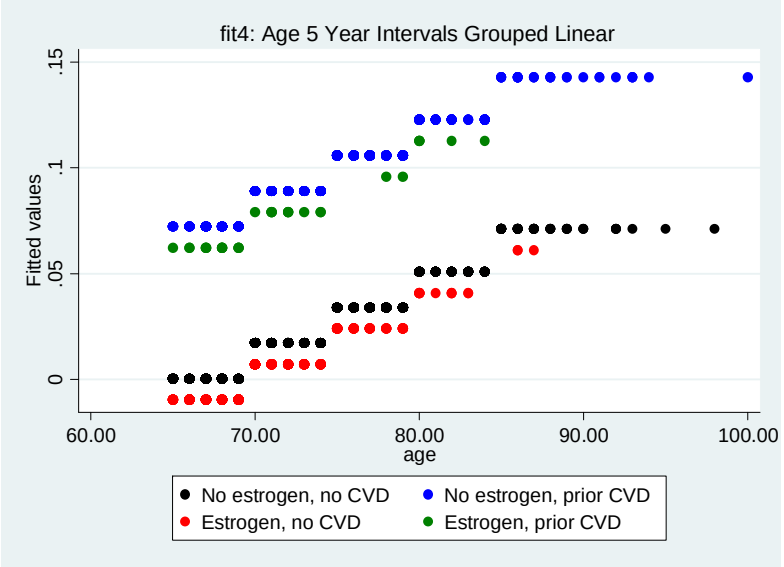
- **I can take the 5 year grouped data and fit it as a linear term**
  - **Doing so is made more reasonable by the fact that I coded my grouped variable with the average age in each group.**
- **This forces monotonic, but not continuous relationships**
- **This seems silly to me, unless you do not have the continuously measured ages.**

Scientific Comments

- **The interpretation of the estrogen coefficient is now "the RD comparing subjects exposed to estrogen to subjects who were not exposed, but who had the same prior history of CVD and were in the same 5 year age group (15 years for over 85)"**
  - **Note how I do not use different words, even though I fit a different model**
  - **Such is the lack of precision of natural language (and the reason we invented mathematical notation)**

```
. predict fit4
(option xb assumed; fitted values)

. * (see do file for code used to produce graphs)
```


fit4: Age 5 Year Intervals Grouped Linear
fit4: estrogen 1 prevdis 0 (N= 310)
fit4: estrogen 1 prevdis 1 (N= 30)
fit4: estrogen 0 prevdis 0 (N= 2045)
fit4: estrogen 0 prevdis 1 (N= 514)
0 .2 .4 .6 .8 1
65.00 70.00 75.00 80.00 85.00
60.00 70.00 80.00 90.00 100.00
age
Fitted values
p4


fit4: Age 5 Year Intervals Grouped Linear
Fitted values
0 .05 .1 .15
60.00 70.00 80.00 90.00 100.00
age
No estrogen, no CVD
No estrogen, prior CVD
Estrogen, no CVD
Estrogen, prior CVD

Problem #1f: Analyses to adjust for age : modeling continuous linear relationship across age

```
. *                        Continuous linear
. regress cvddeath4 estrogen prevdis age, robust

Linear regression                                      Number of obs =    2899
                                                       F(  3,  2895) =   19.77
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0443
                                                       Root MSE      =  .17058

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0095863   .0058756    -1.63   0.103     -.021107    .0019345
     prevdis |   .0712153   .0128743     5.53   0.000     .0459715     .096459
         age |   .0035347   .0007764     4.55   0.000     .0020123     .005057
       _cons |  -.2373083   .0547612    -4.33   0.000    -.3446831   -.1299334
------------------------------------------------------------------------------
```

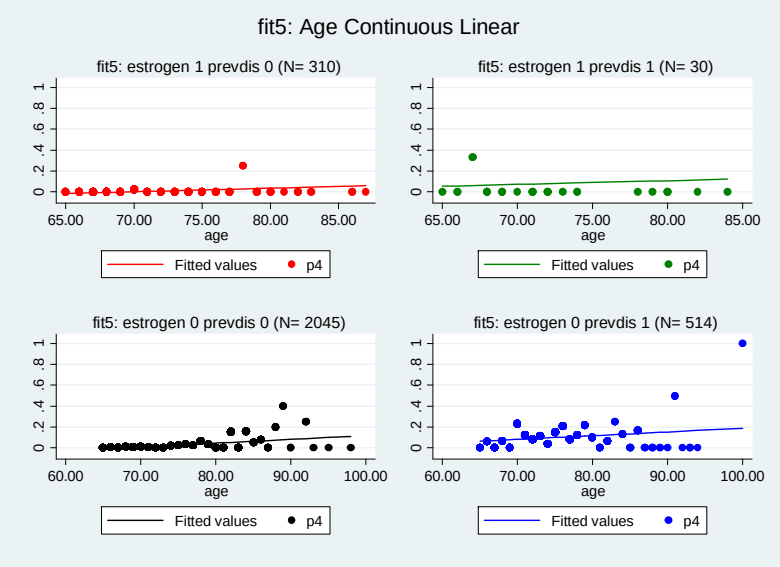
Statistical Comments

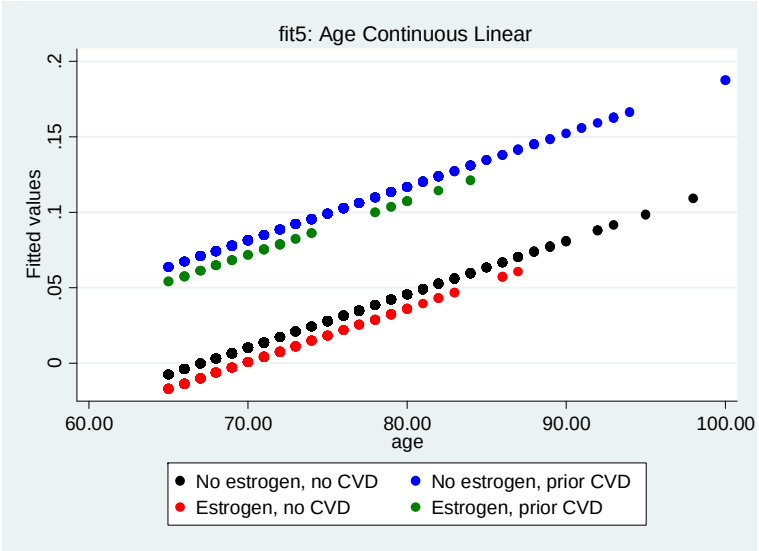
- **Fitting the linear continuous age forces monotonic and continuous estimated relationships**
- **Personally, this is generally my first choice unless I *a priori* know about expected "changepoints" that might be induced by, say, adolescence or menopause in the case of age.**

Scientific Comments

- **The interpretation of the estrogen coefficient is now "the RD comparing subjects exposed to estrogen to subjects who were not exposed, but who had the same prior history of CVD and were of the same age"**
  - **Truly we are just adjusting for the linear trend in age, but I would likely just make that clear in the statistical methods**
  - **If it were a major point, however, I might use the term "adjusted for linear trends in age"**

```
. predict fit5
(option xb assumed; fitted values)

. * (see do file for code used to produce graphs)
```


fit5: Age Continuous Linear
fit5: estrogen 1 prevdis 0 (N= 310)
fit5: estrogen 1 prevdis 1 (N= 30)
fit5: estrogen 0 prevdis 0 (N= 2045)
fit5: estrogen 0 prevdis 1 (N= 514)
age
Fitted values
p4


fit5: Age Continuous Linear
Fitted values
age
No estrogen, no CVD
No estrogen, prior CVD
Estrogen, no CVD
Estrogen, prior CVD

Problem #1f: Analyses to adjust for age : modeling continuous quadratic relationship across age

```
. *                        Quadratic
. regress cvddeath4 estrogen prevdis age agesqr, robust

Linear regression                                       Number of obs =    2899
                                                        F(  4,  2894) =   15.61
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0451
                                                        Root MSE      =  .17054

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0098691   .0058444    -1.69   0.091    -.0213288    .0015906
     prevdis |   .0712635   .0128798     5.53   0.000      .046009     .096518
         age |  -.0138405   .0204054    -0.68   0.498    -.0538511      .02617
      agesqr |   .0001156   .0001387     0.83   0.404    -.0001562    .0003875
       _cons |   .4111158   .7473654     0.55   0.582    -1.054306    1.876538
------------------------------------------------------------------------------
```

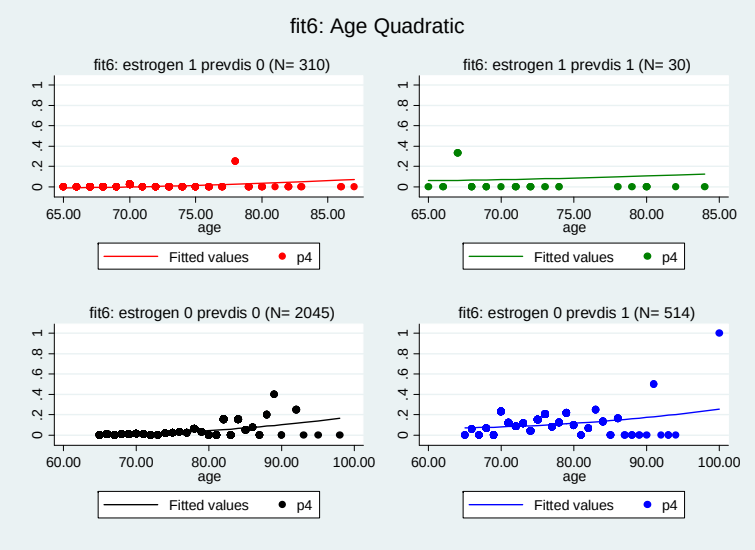
Statistical Comments

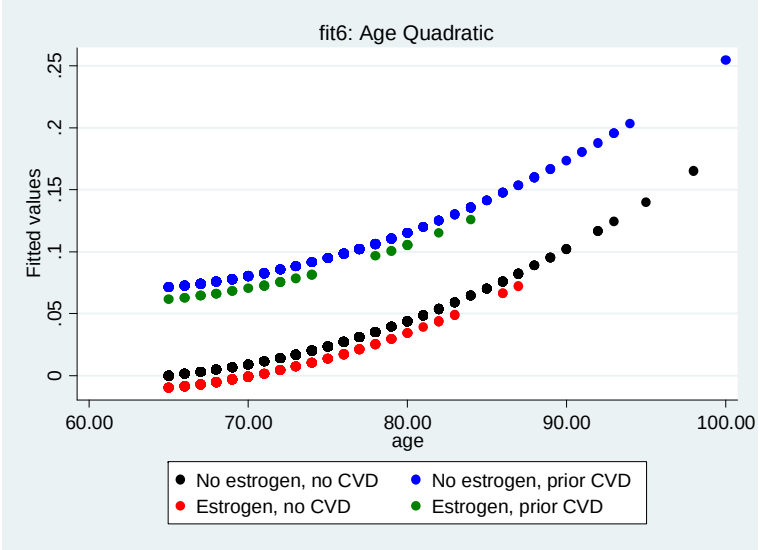
- **Fitting a quadratic continuous age allows greater flexibility**
- **Personally, this is generally my first choice when I want to allow a bit more flexibility.**
  - **It allows curvature, but still provides better efficiency by borrowing information**
- **Graphs really are the best way to interpret the predicted models here**

Scientific Comments

- **I would term the interpretation of the estrogen coefficient as "the RD comparing subjects exposed to estrogen to subjects who were not exposed, but who had the same prior history of CVD and adjusting for any quadratic trends in age"**

```
. predict fit6
(option xb assumed; fitted values)

. * (see do file for code used to produce graphs)
```


fit6: Age Quadratic
fit6: estrogen 1 prevdis 0 (N= 310)
fit6: estrogen 1 prevdis 1 (N= 30)
fit6: estrogen 0 prevdis 0 (N= 2045)
fit6: estrogen 0 prevdis 1 (N= 514)
0 .2 .4 .6 .8 1
65.00 70.00 75.00 80.00 85.00
60.00 70.00 80.00 90.00 100.00
age
Fitted values
p4


fit6: Age Quadratic
Fitted values
0 .05 .1 .15 .2 .25
60.00 70.00 80.00 90.00 100.00
age
No estrogen, no CVD
No estrogen, prior CVD
Estrogen, no CVD
Estrogen, prior CVD

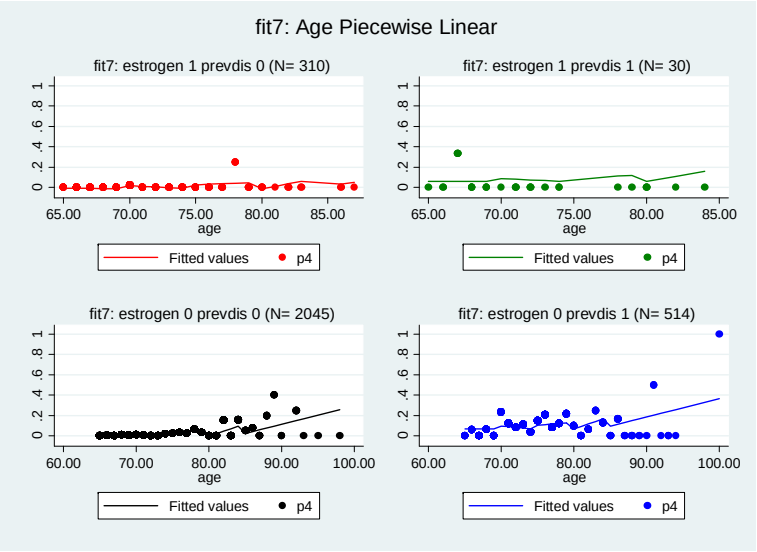
Problem #1f: Analyses to adjust for age : modeling piecewise linear relationship across age (within approx 5 year intervals)

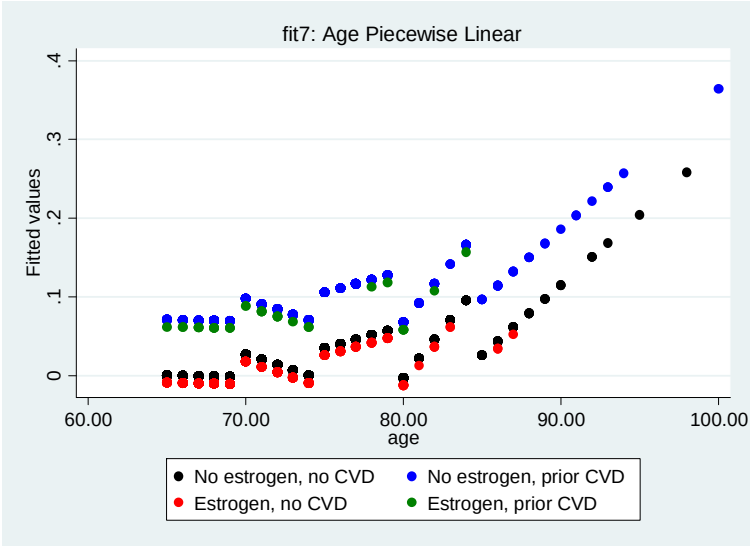
```
. *                     Piecewise linear
. regress cvddeath4 estrogen prevdis i.age5yr##c.age, robust

Linear regression                                       Number of obs =    2899
                                                        F( 11,  2887) =    6.19
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0528
                                                        Root MSE      =  .17005

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0093233   .0058342    -1.60   0.110     -.020763    .0021164
     prevdis |   .0705833   .0127969     5.52   0.000     .0454912    .0956753
             |
      age5yr |
         72  |   .4717309   .3022323     1.56   0.119     -.120882    1.064344
         77  |  -.4050118   .5646485    -0.72   0.473    -1.512167    .7021431
         82  |  -1.996646   .9443038    -2.11   0.035    -3.848224   -.1450684
         88  |  -1.517935   1.109883    -1.37   0.172    -3.694178    .6583081
             |
         age |  -.0003834   .0017959    -0.21   0.831    -.0039049    .0031381
             |
age5yr#c.age |
         72  |  -.0063292    .004239    -1.49   0.136    -.0146411    .0019827
         77  |   .0059116   .0074184     0.80   0.426    -.0086343    .0204575
         82  |   .0249877   .0117111     2.13   0.033     .0020248    .0479507
         88  |   .0182467   .0128299     1.42   0.155      -.00691    .0434034
             |
       _cons |    .025383   .1203647     0.21   0.833    -.2106264    .2613924
------------------------------------------------------------------------------
```

Statistical Comments

- **Fitting piecewise linear age allows greater flexibility but makes interpretability much more difficult**
  - **The "discontinuities" in the predicted values are not desirable, nor is the rapidly changing slopes (which presumably arise from sparse data and influential points)**
- **Again, graphs really are the best way to interpret the predicted models here**

```
. predict fit7
(option xb assumed; fitted values)

. * (see do file for code used to produce graphs)
```



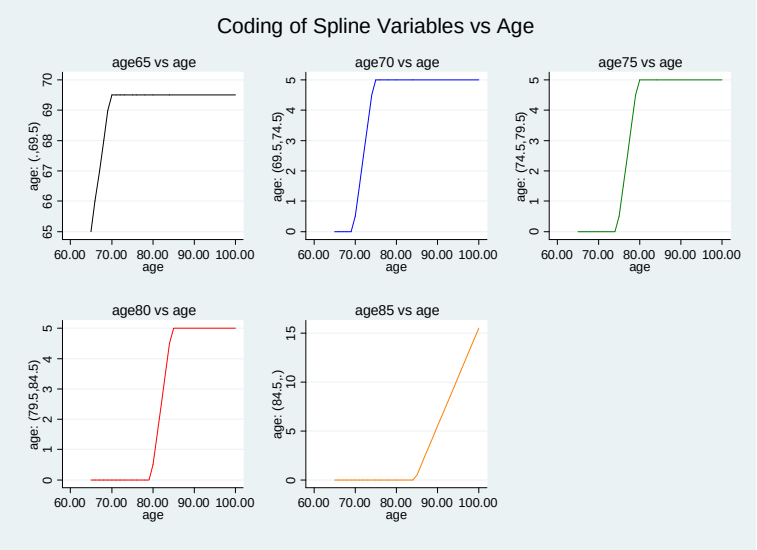


Problem #1f: Analyses to adjust for age : modeling linear splines across age (knots at approx 5 year intervals)

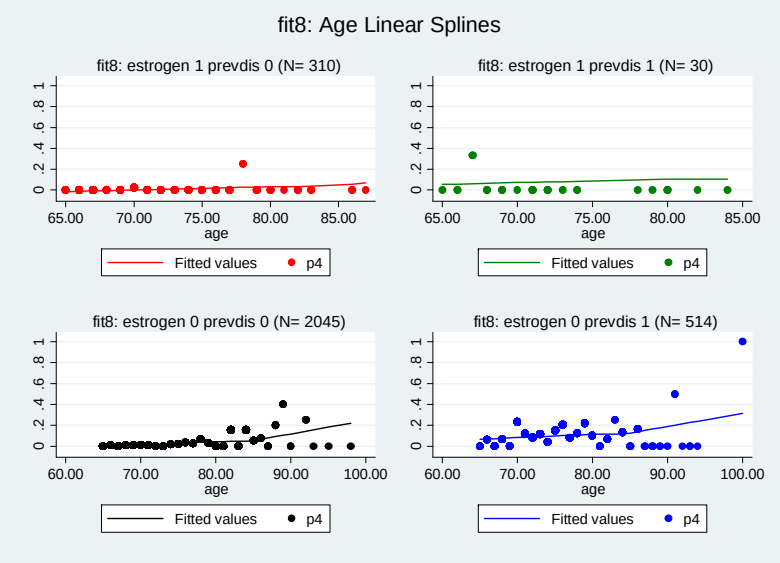
```
. *                        Splines
. regress cvddeath4 estrogen prevdis age65 age70 age75 age80 age85, robust

Linear regression                                       Number of obs =    2899
                                                        F(  7,  2891) =    9.31
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0456
                                                        Root MSE      =  .17058

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0099385   .0058545    -1.70   0.090    -.0214179    .0015408
     prevdis |   .0709798   .0128473     5.52   0.000      .045789    .0961707
       age65 |   .0035669   .0022312     1.60   0.110    -.0008081    .0079419
       age70 |   .0025654   .0024597     1.04   0.297    -.0022575    .0073883
       age75 |    .004109   .0037135     1.11   0.269    -.0031724    .0113904
       age80 |   .0007946   .0071635     0.11   0.912    -.0132514    .0148407
       age85 |   .0127429   .0111594     1.14   0.254    -.0091383     .034624
       _cons |  -.2378035   .1498665    -1.59   0.113    -.5316595    .0560524
------------------------------------------------------------------------------
```

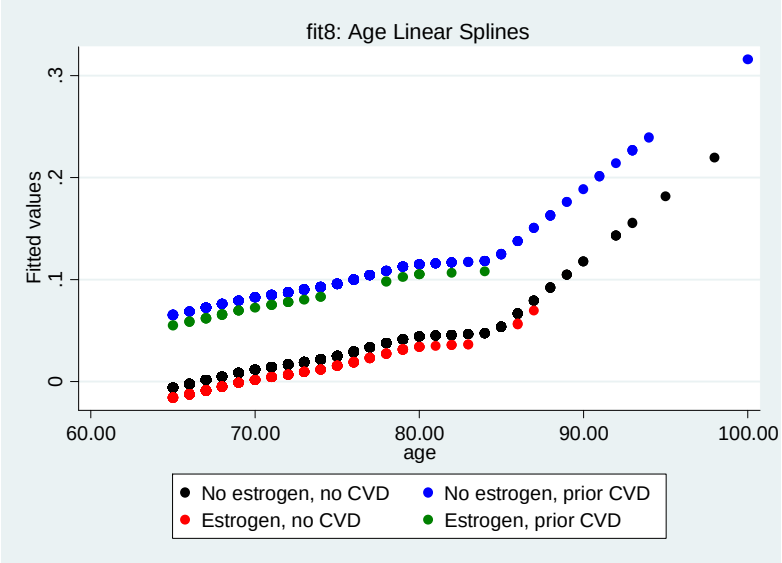
Statistical Comments

- **Linear splines allow greater flexibility but makes interpretability much more difficult**
  - **We have at least gotten rid of the discontinuities**
- **Interpretation of the coefficients is very difficult unless you understand how each of the modeled covariates relate to age**
  - **This is shown in the following graph**
- **Again, graphs really are the best way to interpret the predicted models here**
  - **And to my eye, the splines just reproduce something like a quadratic fit, though they do a bit better fitting the highest age groups**
  - **(Or perhaps they overfit the highest age groups where we have very sparse data)**




```
. predict fit8
(option xb assumed; fitted values)

. * (see do file for code used to produce graphs)
```





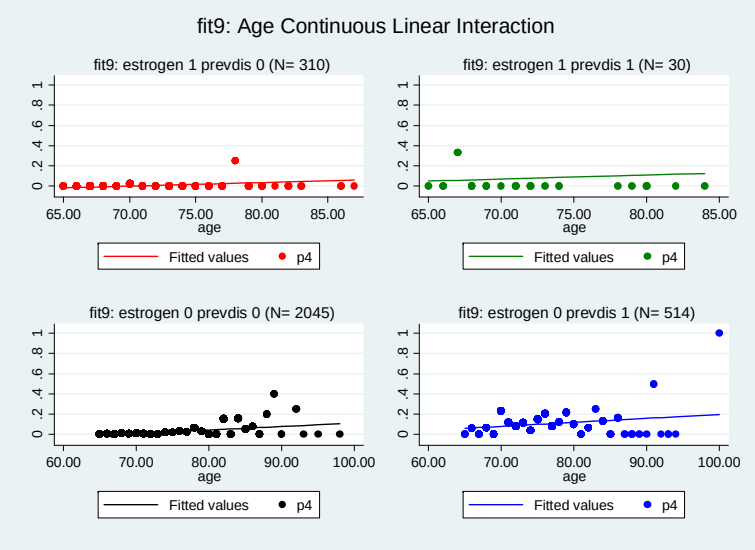
Problem #1f: Analyses to adjust for age : modeling linear age relationship with interaction with *prevdis*

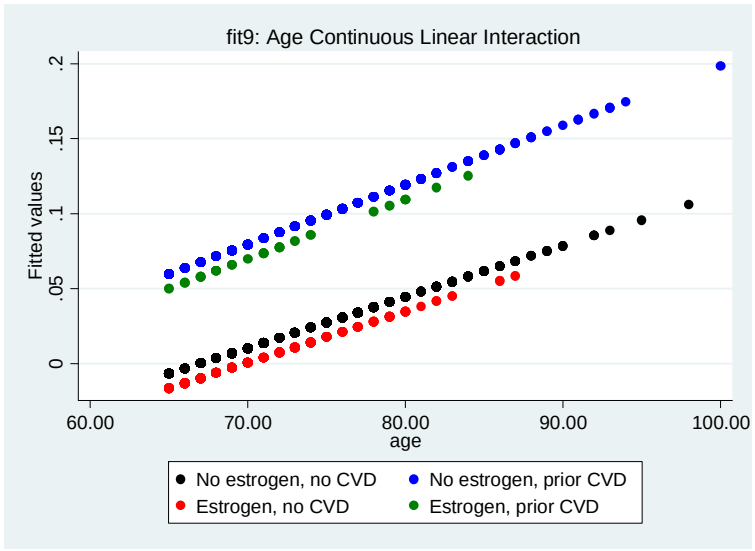
```
. *               Adjustment for main effect and age - prevdis interaction
. *                       (Note could mix and match above models for main effect with
. *                        alternative models for interactions)
. *                       Continuous linear and interaction
. regress cvddeath4 estrogen i.prevdis##c.age, robust

Linear regression                                       Number of obs =    2899
                                                        F(  4,  2894) =   16.87
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0444
                                                        Root MSE      =   .1706

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0097198   .0058796    -1.65   0.098    -.0212485    .0018088
   1.prevdis |   .0305219   .1668584     0.18   0.855    -.2966515    .3576953
         age |   .0034099   .0007829     4.36   0.000     .0018748     .004945
             |
     prevdis#|
       c.age |
          1  |   .0005518    .002293     0.24   0.810    -.0039442    .0050478
             |
       _cons |  -.2282845    .054835    -4.16   0.000    -.3358041   -.1207648
------------------------------------------------------------------------------
```

Statistical Comments

- **Fitting interactions among the potential confounders would allow greater flexibility**
  - **The curves across strata do not have to be parallel (though for the linear fit they do not seem very different)**
- **Interpretation of the parameters is more difficult, but not overly so**
  - **We just have a different slope and intercept in each stratum**

```
. predict fit9
(option xb assumed; fitted values)

. * (see do file for code used to produce graphs)
```


fit9: Age Continuous Linear Interaction
fit9: estrogen 1 prevdis 0 (N= 310)
fit9: estrogen 1 prevdis 1 (N= 30)
fit9: estrogen 0 prevdis 0 (N= 2045)
fit9: estrogen 0 prevdis 1 (N= 514)
0 .2 .4 .6 .8 1
65.00 70.00 75.00 80.00 85.00
60.00 70.00 80.00 90.00 100.00
age
Fitted values
p4


fit9: Age Continuous Linear Interaction
Fitted values
0 .05 .1 .15 .2
60.00 70.00 80.00 90.00 100.00
age
No estrogen, no CVD
No estrogen, prior CVD
Estrogen, no CVD
Estrogen, prior CVD

Problem #1f: Analyses to adjust for age : modeling quadratic age relationship with interaction with *prevdis*

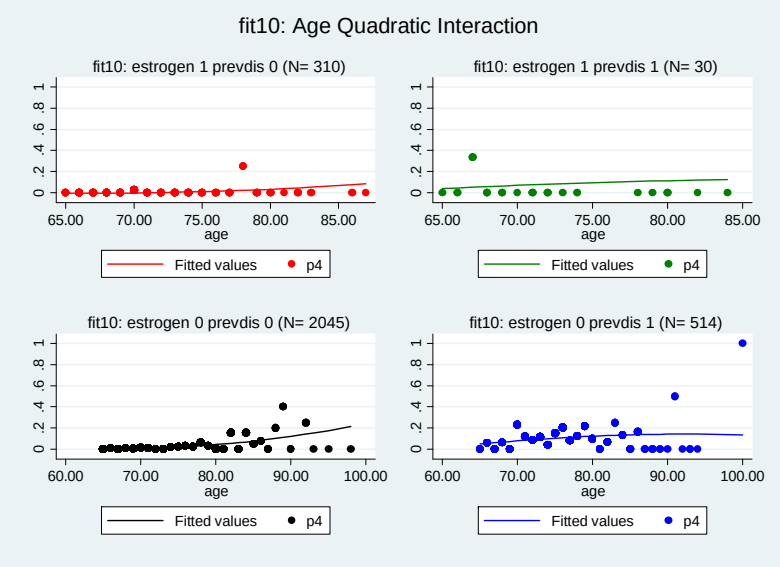
```
. *                     Quadratic and interaction
. regress cvddeath4 estrogen i.prevdis##c.age i.prevdis##c.agesqr, robust

Linear regression                                       Number of obs =    2899
                                                        F(  6,  2892) =   11.95
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0463
                                                        Root MSE      =  .17049

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -.0098138   .0058777    -1.67   0.095    -.0213386     .001711
   1.prevdis |  -1.861262     2.2912    -0.81   0.417    -6.353811    2.631287
         age |  -.0283475   .0191684    -1.48   0.139    -.0659326    .0092376
             |
     prevdis#|
       c.age |
          1  |    .051068    .061399     0.83   0.406    -.0693222    .1714582
             |
      agesqr |   .0002125   .0001316     1.61   0.106    -.0000456    .0004706
             |
     prevdis#|
    c.agesqr |
          1  |  -.0003351   .0004091    -0.82   0.413    -.0011372    .0004671
             |
       _cons |   .9505706   .6951954     1.37   0.172    -.4125579    2.313699
------------------------------------------------------------------------------
```

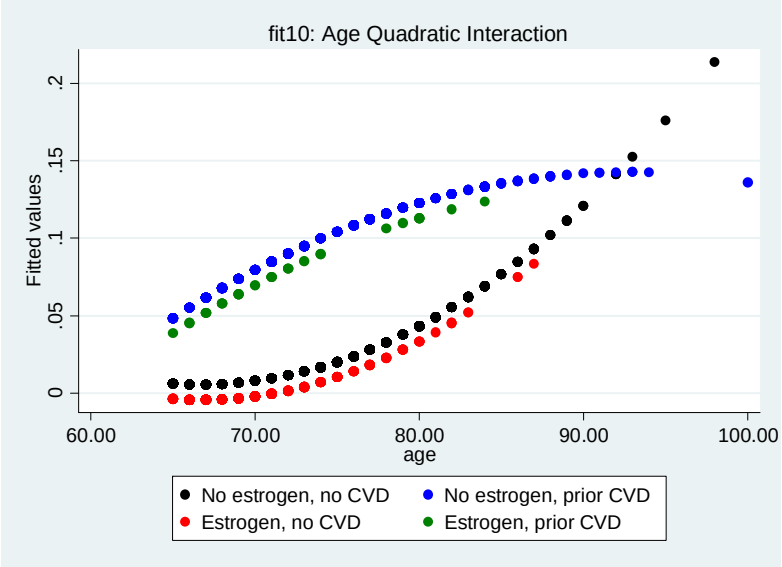
Statistical Comments

- **Fitting interactions among the potential confounders with a quadratic form for age is still greater flexibility**
  - **The curves across strata do not have to be parallel (and for the quadratic fit they seem very different)**
- **Interpretation of the parameters is fairly difficult, the graph shows it best**
- **My fear is that this analysis is prone to overfitting the data.**

```
. predict fit10
(option xb assumed; fitted values)
```

```
. * (see do file for code used to produce graphs)
```





Problem #1f: Analyses to adjust for age : modeling quadratic age relationship with interaction with *prevdis* and *estrogen*

```
. *               Adjustment for three-way interaction
. regress cvddeath4 i.estrogen##prevdis##c.age i.estrogen##prevdis##c.agesqr, robust

Linear regression                               Number of obs =    2899
                                                F( 11,  2887) =    8.36
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0487
                                                Root MSE      =  .17042

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  1.estrogen |  -1.538786   .8676814    -1.77   0.076    -3.240123    .1625518
   1.prevdis |  -2.176157   2.365643    -0.92   0.358    -6.814677    2.462363
             |
    estrogen#|
     prevdis |
        1 1  |   8.378526   6.273068     1.34   0.182    -3.921618    20.67867
             |
         age |   -.029608   .0203429    -1.46   0.146    -.0694962    .0102801
             |
    estrogen#|
       c.age |
          1  |   .0435968   .0238044     1.83   0.067    -.0030784    .0902721
             |
     prevdis#|
       c.age |
          1  |   .0593127   .0633587     0.94   0.349    -.0649202    .1835456
             |
    estrogen#|
     prevdis#|
       c.age |
        1 1  |  -.2202408   .1643311    -1.34   0.180    -.5424591    .1019774
             |
      agesqr |   .0002221   .0001395     1.59   0.111    -.0000514    .0004956
             |
    estrogen#|
    c.agesqr |
          1  |  -.0003086   .0001627    -1.90   0.058    -.0006277    .0000104
             |
     prevdis#|
    c.agesqr |
```

```
          1  |  -.0003881   .0004219    -0.92   0.358    -.0012154    .0004391
             |
    estrogen#|
     prevdis#|
    c.agesqr |
        1 1  |   .0014262   .0010725     1.33   0.184    -.0006767     .003529
             |
       _cons |    .990803   .7388898     1.34   0.180    -.4580017    2.439608
------------------------------------------------------------------------------

. test 1.estrogen 1.estrogen#1.prevdis 1.estrogen#c.age 1.estrogen#c.agesqr 1.estrogen#1.prevdis#c.age
1.estrogen#1.prevdis#c.agesqr

 ( 1)  1.estrogen = 0
 ( 2)  1.estrogen#1.prevdis = 0
 ( 3)  1.estrogen#c.age = 0
 ( 4)  1.estrogen#c.agesqr = 0
 ( 5)  1.estrogen#1.prevdis#c.age = 0
 ( 6)  1.estrogen#1.prevdis#c.agesqr = 0

       F(  6,  2887) =    8.24
            Prob > F =    0.0000
```

Stata Comments

- **Stata allows you to specify the inclusion of main effects and interactions by prefixing terms**
  - **A single variable prefixed with "i." will be fit as dummy variables**
  - **Interactions can be created by joining two variables with"#"**
  - **If you want one of those variables treated continuously rather than as dummy variables, prefix it with "c."**
- **When testing for an association between response and a variable that might also be included in interactions, you must simultaneously test that all terms containing that parameter have a zero coefficient**
  - **The Stata commands "test", "testparm", and "lrtest" are post-estimation commands that can be used to do this**
  - **Note the naming conventions for the automatically generated variables**

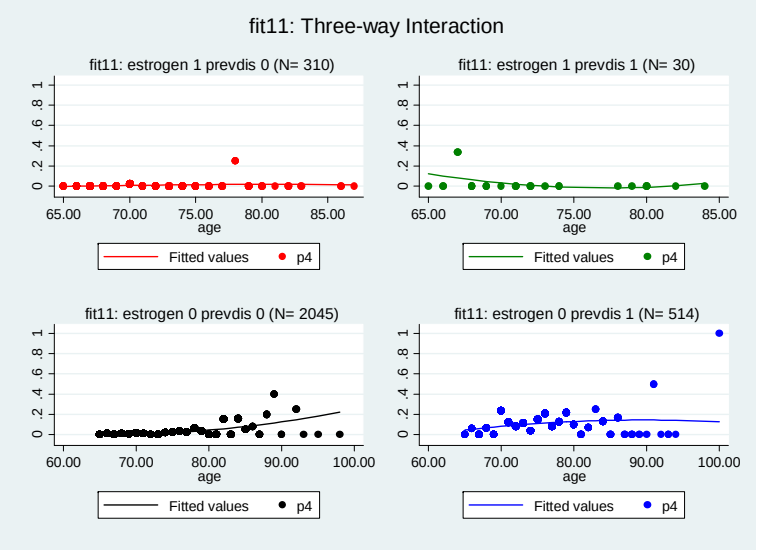
Statistical Comments

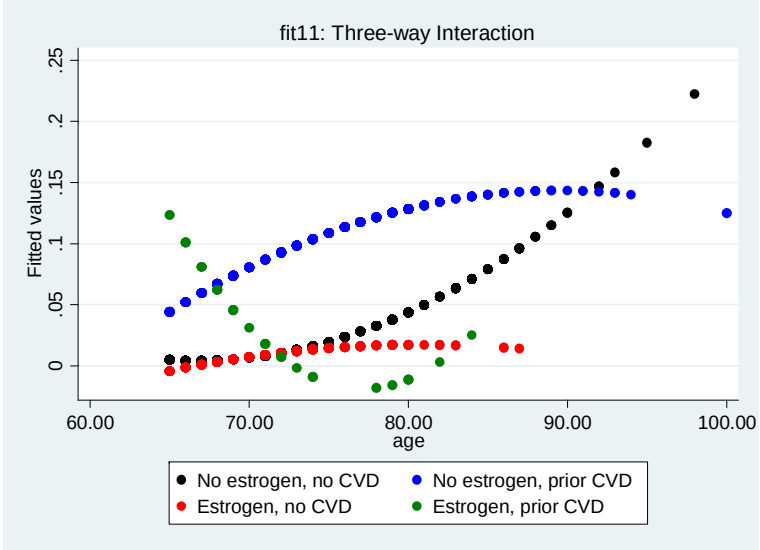
- **According to the multiple partial test of all parameters involving *estrogen*, there is a highly statistically significant association between 4 year CVD mortality and estrogen use.**
  - **It is difficult to assess whether this is a protective or deleterious effect from the coefficients**
  - **The graphs do not help too much here, either**
  - **Besides, we have very sparse data in some strata, so I do not trust the "large sample" assumptions**

Scientific Comments

- **I would not have *a priori* decided to fit this model. I would recommend against such complicated models**

```
. predict fit11
(option xb assumed; fitted values)

. * (see do file for code used to produce graphs)
```





PROBLEM #2: Analyses based on risk ratio (relative risk, RR)

Statistical comments

- **In a two sample test of binomial proportions, analyses based on RR are not as common as RD or OR.**
  - **However, it is not uncommon to describe hypotheses as a relative decrease or increase in risk even when we do not analyze the RR**
  - **When adjusting for additional covariates, however, it is more common to base analyses on OR.**
- **In Problem #2, we mirror the analyses explored for problem #1 insofar as possible. Notable differences include**
  - **t tests will not apply**
  - **regression models will consider the log link**
  - **stratified models will have more difficult correspondence with regression models**

Problem #2a: Analyses of RR in 2 by 2 contingency tables (two sample tests of binomial proportions)

```
. * EXECUTION OF CODE FOR PROBLEM 2
. * Problem 2a: unadjusted analyses of cvddeath4 - estrogen association
. *               chi-square test: cs
. cs cvddeath4 estrogen

                 | estrogen                |
                 |   Exposed   Unexposed   |      Total
-----------------+-------------------------+-----------
           Cases |         3          88   |         91
        Noncases |       337        2471   |       2808
-----------------+-------------------------+-----------
           Total |       340        2559   |       2899
                 |                         |
            Risk |  .0088235    .0343884   |   .0313901
                 |                         |
                 |      Point estimate     |    [95% Conf. Interval]
                 |-------------------------+-------------------------
 Risk difference |        -.0255649        |   -.0377575   -.0133723
      Risk ratio |         .2565842        |      .08164    .8064117
 Prev. frac. ex. |         .7434158        |    .1935883      .91836
 Prev. frac. pop |         .0871892        |
                 +-----------------------------------------------------
                              chi2(1) =     6.45  Pr>chi2 = 0.0111

. di  (3 / 340)   /    (88 / 2559)
.25658422
```

Stata Comments

- **Note that Stata command cs provides the estimated risk ratio, which is just the ratio of the sample proportions**
- **The CI is computed by the same methods as used in the Stata command glm with the binomial family and log link**
- **The P value can be provided by either the chi square test or Fisher's exact test**
  - **This is the same test as used for the risk difference, which makes sense: In a two sample problem, the risk difference is nonzero if and only if the risk ratio is not equal to 1. (Similarly, the risk difference is nonzero if and only if the odds ratio is not equal to 1.)**

Scientific Comments

- **I would most often summarize the association between a binary response and a binary POI with the risk difference**
- **However: An advantage of RR over RD is that RR is unaffected by "contaminating" a sample with subjects who would not have an event if they were exposed or unexposed**
    - **For instance, suppose in some "target population" the event rate is 20% in exposed and 10% in unexposed.**
    - **But imagine there is some "non-target population" having event rate of 0% in both the exposed and unexposed.**
    - **Now imagine we were unlucky and sampled in such a way that 40% of our subjects were from the "target" population and 60% were from the "nontarget" population. In our sample we would thus expect**
        - **8% incidence of events in the exposed (.4 times 20% plus .6 times 0%) and 4% incidence in the unexposed (.4 times 10% plus .6 times 0%)**
        - **RD is expected to be 4% rather than the 10% in our target population**
        - **RR is expected to be 2 (8% divided by 4%), just like in our target population (which had 20% divided by 10%)**
- **A disadvantage of RR is that we get very different answers if we, for instance, talk about survival instead of mortality**
    - **The RR in our sample for mortality is (3 / 340) / (88 / 2559) = .25658422**
    - **The RR is our sample for survival is (337 / 340) / (2471 / 2559) = 1.0265, which is not 1 / 0.256584 = 3.8974**
    - **Note that when using the OR, we have better agreement when talking about odds of events or odds of nonevents:**
        - **OR for mortality is (3 / 337) / (88 / 2471) = 0.24997**
        - **OR for survival is (337 / 3) / (2471 / 88) = 4.0005, which is 1 / 0.24997**
- **In any case, standard reporting practice would likely be to report the chi squared or Fisher's exact test when the two sample comparison is the primary interest**
    - **CI for RD, RR, or OR may not agree with the P value decision owing to the failure to fully account for the mean-variance relationship and the nuisance parameter**

Problem #2a: Analyses of RR using generalized linear model (GLM) with Gaussian mean-variance; standard SE

```
. *               generalized linear model: glm
. *                     - presuming homoscedasticity
. glm cvddeath4 estrogen, family(gaussian) link(log)

Iteration 0:   log likelihood = -4377.0276
Iteration 1:   log likelihood = -2964.6117  (not concave)
Iteration 2:   log likelihood = -1928.6162
Iteration 3:   log likelihood =  323.20504
Iteration 4:   log likelihood =  945.91524
Iteration 5:   log likelihood =  952.81919
Iteration 6:   log likelihood =  953.04683
Iteration 7:   log likelihood =  953.05456
Iteration 8:   log likelihood =  953.05459

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2897
                                                  Scale parameter =  .0303581
Deviance         =  87.94734731                   (1/df) Deviance =  .0303581
Pearson          =  87.94734731                   (1/df) Pearson  =  .0303581
Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  -.656126
Log likelihood   =  953.0545892                   BIC             = -23007.29
------------------------------------------------------------------------------
             |                 OIM
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -1.360267   1.075538    -1.26   0.206    -3.468282    .7477486
       _cons |  -3.370034   .1001588   -33.65   0.000    -3.566342   -3.173727
------------------------------------------------------------------------------
. di   exp(-3.370034), exp( -1.360267)
.03438847 .25659226

. di   -1.360267 / 1.075538
-1.2647317

. di   -1.360267  -  1.96 * 1.075538,         -1.360267    +    1.96 * 1.075538
-3.4683215 .74778748

. di   exp( -1.360267  -  1.96 * 1.075538),         exp(-1.360267    +    1.96 * 1.075538)
.03116931 2.1123213
```

Stata Comments

- **We can use the Stata glm command to estimate the mean when using a multiplicative model (log link)**
- **When we use the Gaussian family, we do not weight the estimating equation by estimated variance**
- **However, the fact that the log link is nonlinear (and thus has a term from the derivative of the inverse link function) means that we still have to use an iterative search**
  - **Consider the various possible estimating equations**

	Normal	Binomial	Poisson
Identity	$U(\beta)=\sum_{i=1}^{n}\frac{Y_i-X_i\beta}{\sigma^2}X_i \Rightarrow RD$	$U(\beta)=\sum_{i=1}^{n}\frac{Y_i-X_i\beta}{X_i\hat{\beta}(1-X_i\hat{\beta})}X_i \Rightarrow RD$	$U(\beta)=\sum_{i=1}^{n}\frac{Y_i-X_i\beta}{X_i\hat{\beta}}X_i \Rightarrow RD$
Log	$U(\beta)=\sum_{i=1}^{n}\frac{Y_i-e^{X_i\beta}}{\sigma^2}X_i e^{X_i\beta} \Rightarrow RR$	$U(\beta)=\sum_{i=1}^{n}\frac{Y_i-e^{X_i\beta}}{(1-e^{X_i\hat{\beta}})}X_i \Rightarrow RR$	$U(\beta)=\sum_{i=1}^{n}(Y_i-e^{X_i\beta})X_i \Rightarrow RR$
Logit		$U(\beta)=\sum_{i=1}^{n}\left(Y_i-\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right)X_i \Rightarrow OR$	

- **Stata tells you about each of iterative steps**
  - **Almost always, I do not care to see this output, but I do not know how to turn it off**
  - **Very rarely, it is of interest from an extremely technical standpoint to help understand why the estimation procedure is not converging**
    - **Sometimes we are trying to estimate infinity or negative infinity**
      - **log(0) is negative infinity, so in a group with no events we will run into problems if that group is modeled exactlly**
      - **logit (1) is positive inifinity, so in a group with all events we will run into problems if that group is modeled exactly**
    - **Sometimes our starting estimates are not very good**
    - **Sometimes we have negative weights owing to model misspecification (nonlinearities; effect modification)**
  - **In all additional output, I will delete the information about iterations.**
- **By default, glm with the Gaussian family reports parameter estimates on the untransformed scale**
  - **So with a log link, we may be more interested in using the eform option to get exponentiated slopes**

- **It is useful, however, to first consider the coefficient table on the untransformed scale**
  - **"Coefficients" are the estimated regression parameters**
  - **"Std. Err." are the estimated standard errors**
  - **"z" is the Z statistic testing a null hypothesis that the true regression parameter is 0: Z = coeff / std error.**
  - **"P > |z|" is the two-sided P value based on the asymptotic (large sample) distribution using the normal distribution.**
  - **"Conf Interval" is based on large sample results: (coeff) ± (crit value) × (std err)**
    - **For a 95% CI, the critical value is 1.96**

Statistical Comments

- **We are fitting a simple regression model in which**

$$\log\left(E\left[cvddeath4 \mid estrogen\right]\right) = \beta_0 + \beta_1 \times estrogen$$

- **Interpretation of the regression parameters:**
  - **The intercept β_0 is the log mean (proportion) death within 4 years after study accrual among subjects unexposed to estrogen (*estrogen=0)*.**
    - **The exponentiated intercept e^{β_0} is the mean (proportion) death within 4 years after study accrual among subjects unexposed to estrogen (*estrogen=0)*.**
  - **The slope β_1 is log ratio mean (proportion) death within 4 years after study accrual between subjects exposed to estrogen (*estrogen=1)* and subjects unexposed to estrogen (*estrogen=0)*.**
    - **The exponentiated slope e^{β_1} is the relative risk (RR) of the mean (proportion) death within 4 years after study accrual comparing subjects exposed to estrogen (*estrogen=1)* to subjects unexposed to estrogen (*estrogen=0)*.**
- **In this simple regression model (one predictor variable) having a binary predictor (a 0-1 variable), the predicted values will be (nearly) equal to the sample mean in each group**
  - **$\hat{\beta}_0$ will be equal to the log sample mean among subjects having a predictor value equal to 0**
  - **$\hat{\beta}_1$ will be equal to the log ratio of sample means: group 1 divided by group 0**
  - **(Minor differences in the estimates are due to the method of iterative search: IRLS vs MLE—see later)**

```
. glm cvddeath4 estrogen, family(gaussian) link(log) eform

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2897
                                                  Scale parameter =  .0303581
Deviance         =  87.94734731                   (1/df) Deviance =  .0303581
Pearson          =  87.94734731                   (1/df) Pearson  =  .0303581

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  -.656126
Log likelihood   =  953.0545892                   BIC             = -23007.29

------------------------------------------------------------------------------
             |                 OIM
   cvddeath4 |      exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .2565924   .2759748    -1.26   0.206     .0311705    2.112239
------------------------------------------------------------------------------

. di .2565924 -  1.96 * .2759748,  .2565924 +  1.96 * .2759748
-.28431821 .79750301
```

Stata Comments

- **When Stata presents exponentiated forms of the regression coefficients (i.e., when option eform is included)**
  - **The intercept term is suppressed**
    - **Note that the exponentiated intercept would be the risk in the reference group, while all other exponentiated coefficients would be a relative risk**
  - **The Z statistic and P value will be taken directly from the table in the untransformed report (no eform)**
  - **The CI will be the exponentiated CI from the untransformed report**
  - **The standard error will be computed for the RR using the "delta method"**
    - **These are asymptotically correct, but behave less well in small samples**
    - **It is better to do inference on the log scale and then backtransform, which is what Stata does**

- **Note that the estimate given above does not match the sample risk and RR exactly**
  - **This is due to the use of MLE rather than IRLS (iteratively reweighted least squares) in the search as below**
  - **In very large samples they will agree**
  - **In small samples, the differences are generally of no consequence**

```
. glm cvddeath4 estrogen, family(gaussian) link(log) eform irls

Generalized linear models                         No. of obs      =      2899
Optimization     : MQL Fisher scoring             Residual df     =      2897
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  87.94734731                   (1/df) Deviance =  .0303581
Pearson          =  87.94734731                   (1/df) Pearson  =  .0303581

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  BIC             = -23007.29

------------------------------------------------------------------------------
             |                 EIM
   cvddeath4 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .2565843   .2758539    -1.27   0.206     .0311963    2.110363
------------------------------------------------------------------------------
```

Problem #2a: Analyses of RR using generalized linear model (GLM) with Gaussian mean-variance; robust SE

```
. *                    - allowing for possible heteroscedasticity
. glm cvddeath4 estrogen, family(gaussian) link(log) robust

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2897
                                                  Scale parameter =  .0303581
Deviance         =  87.94734731                   (1/df) Deviance =  .0303581
Pearson          =  87.94734731                   (1/df) Pearson  =  .0303581

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  -.656126
Log pseudolikelihood =  953.0545892               BIC             = -23007.29

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -1.360267   .5843287    -2.33   0.020     -2.50553   -.2150033
       _cons |  -3.370034   .1047694   -32.17   0.000    -3.575379    -3.16469
------------------------------------------------------------------------------

. glm cvddeath4 estrogen, family(gaussian) link(log) robust eform

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2897
                                                  Scale parameter =  .0303581
Deviance         =  87.94734731                   (1/df) Deviance =  .0303581
Pearson          =  87.94734731                   (1/df) Pearson  =  .0303581

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  -.656126
Log pseudolikelihood =  953.0545892               BIC             = -23007.29

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .2565924   .1499343    -2.33   0.020     .0816323    .8065388
------------------------------------------------------------------------------
```

Statistical Comments

- **Owing to the mean-variance relationship, we should use the Huber-White sandwich estimator if we have not used weighted estimation techniques**
- **Specification of the "robust" option does not change the parameter estimates, it only changes the standard error estimates.**
- **We can also compare regression models estimated using weights derived from the binomial family: In Stata we can use binreg with the rr option (IRLS is the default) or glm (MLE is the default)**

Problem #2a: Analyses of RR using generalized linear model (GLM) with binomial mean-variance; standard SE

```
. *               weighted generalized linear model: glm
. *                     - estimated binomial weights with standard SE
. binreg cvddeath4 estrogen, rr

Generalized linear models                         No. of obs      =      2899
Optimization     : MQL Fisher scoring             Residual df     =      2897
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  800.4201353                   (1/df) Deviance =  .2762928
Pearson          =  2898.856367                   (1/df) Pearson  =  1.000641

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u)                   [Log]

                                                  BIC             = -22294.81

------------------------------------------------------------------------------
             |                 EIM
   cvddeath4 | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .2565842   .1498819    -2.33   0.020     .0816595    .8062198
------------------------------------------------------------------------------

. glm cvddeath4 estrogen, family(binomial) link(log) eform

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2897
                                                  Scale parameter =         1
Deviance         =  800.4201353                   (1/df) Deviance =  .2762928
Pearson          =         2899                   (1/df) Pearson  =   1.00069

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  .2774819
Log likelihood   = -400.2100677                   BIC             = -22294.81

------------------------------------------------------------------------------
             |                 OIM
   cvddeath4 | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .2565842   .1499131    -2.33   0.020       .08164    .8064117
------------------------------------------------------------------------------
```

```
. *                       - estimated binomial weights with robust SE
. binreg cvddeath4 estrogen, rr vce(robust)

Generalized linear models                         No. of obs      =      2899
Optimization     : MQL Fisher scoring             Residual df     =      2897
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  800.4201353                   (1/df) Deviance =  .2762928
Pearson          =  2898.856367                   (1/df) Pearson  =  1.000641

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u)                   [Log]

                                                  BIC             = -22294.81

------------------------------------------------------------------------------
             |             Semirobust
   cvddeath4 | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .2565842   .1499389    -2.33   0.020     .0816239     .806571
------------------------------------------------------------------------------

. glm cvddeath4 estrogen, family(binomial) link(log) vce(robust) eform

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2897
                                                  Scale parameter =         1
Deviance         =  800.4201353                   (1/df) Deviance =  .2762928
Pearson          =         2899                   (1/df) Pearson  =   1.00069

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  .2774819
Log pseudolikelihood = -400.2100677               BIC             = -22294.81

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .2565842   .1499389    -2.33   0.020     .0816239     .806571
------------------------------------------------------------------------------
```

Statistical Comments

- **When we use the binomial family mean-variance relationship as weights in this saturated model, there is little difference between model-based SE and robust SE.**

Scientific Comments

- **As with the inference based on RD, my usual habit would be to use the chi squared statistic to test for an association,**
- ***A priori* I would want to use inference based on the unweighted analysis and robust SE estimates for CI.**
  - **My reasoning here is more related to the possibility of model misspecification leading to improper weights**
  - **This matters less with a log link unless some groups' risk is very near 1.**
    - **Of course, we find RR most useful when the risk is very low, rather than very high.**
- **We find that there is an association between estrogen use and four year CVD mortality in the sample**
  - **Estimate risk of 4 year CVD mortality among estrogen exposed is only 0.257 as large as that among estrogen unexposed (95% CI: .082 to .807 as large; chi squared two-sided P = 0.011).**
    - **Note that I made clear that I was using a test that was not directly linked to the CI. Sometimes I would just let the Statistical Methods make that clear, and not provide this greater detail.**

Problem #2a: Analyses of RR using generalized linear model (GLM) with binomial mean-variance; standard SE

```
. poisson cvddeath4 estrogen, irr

Poisson regression                                Number of obs   =       2899
                                                  LR chi2(1)      =       8.44
                                                  Prob > chi2     =     0.0037
Log likelihood = -401.75408                       Pseudo R2       =     0.0104

------------------------------------------------------------------------------
   cvddeath4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .2565842   .1506429    -2.32   0.021     .0811861      .81092
------------------------------------------------------------------------------
```

Stata Comments

- **Poisson regression is a GLM with a Poisson family mean-variance relationship (variance equals mean) and, by default, a log link**
- **When using the Stata command poisson, you get exponentiated estimates by using option irr**

Statistical Comments

- **We might consider that use of the Poisson family mean-variance is to be preferred when using the log link, because that leads to an unweighted estimating equation.**
- **We see that there is very little difference between this model and the binomial or Gaussian families in this (saturated) two sample problem.**

Problem #2b: Analyses to detect effect modification with RR

Scientific Comments

- **Effect modification is a function of both the summary measure (proportion or odds with binary data) and the comparison (difference or ratio) used to quantify an association**
- **Hence, we cannot just rely on the decision re effect modification in problem #1**
- **We must evaluate the evidence for effect modification for the RR**

```
. * Problem 2b: evidence of effect modification by prevdis
. *               descriptive statistics
. cs cvddeath4 estrogen, by(prevdis)

         prevdis |       RR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               0 |    .3565824      .086377   1.472048     4.870488
               1 |    .3359477      .048058   2.348428       2.8125
-----------------+-------------------------------------------------
           Crude |    .2565842       .08164   .8064117
    M-H combined |    .3490287     .1109699   1.097784
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    0.002  Pr>chi2 = 0.9613
```

Scientific Comments

- **The estimated RR is quite similar in the two strata defined by *prevdis*: 0.357 vs 0.336**
  - **It should be noted that RR less than 1.0 will always seem closer together than if we took the inverse RR. But in this case, 1 / 0.357 = 2.8011204 and 1/ 0.336 = 2.9761905 would be judged fairly similar, as well**
- **Furthermore, tests for effect modification in the population were not significant (P = .96 from the Mantel-Haenszel test for homogeneity)**
- **We can also examine a comparable analysis based on the generalized linear model**

```
. *               inferential statistics
. glm cvddeath4 estrogen prevdis estr_prev, family(gaussian) link(log) robust

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2895
                                                  Scale parameter =  .0294383
Deviance         =   85.2240145                   (1/df) Deviance =  .0294383
Pearson          =   85.2240145                   (1/df) Pearson  =  .0294383

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             = -.6862012
Log pseudolikelihood =  998.6486912               BIC             = -22994.07

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   -1.03119    .723528    -1.43   0.154    -2.449279    .3868985
     prevdis |   1.701837    .210275     8.09   0.000     1.289706    2.113969
   estr_prev |  -.0594486   1.227819    -0.05   0.961    -2.465929    2.347032
       _cons |  -4.012235    .162933   -24.63   0.000    -4.331578   -3.692892
------------------------------------------------------------------------------
```

Stata Comments

- **I did not specify the eform option, so the interpretation of the regression parameters is in terms of the log incidence**

Scientific Comments

- **We are fitting a linear regression model in which**

$$\log\left(E\left[cvddeath4 \mid estrogen\right]\right) = \beta_0 + \beta_1 \times estrogen + \beta_2 \times prevdis + \beta_3 \times estr\_prev$$

- **Interpretation of the regression parameters is similar to the interpretation of regression parameters in the analogous RD regression model that included the interaction, except we are now modeling the log risk of death. (Compare the wording of this description with the analogous wording for problem #1 on page 26-27: all I did was add the word "log" in key places. I highlighted the introduced wording in red (which I can hardly detect) in order to emphasize (to those of you with "normal" color vision) the minimal way the wording changes):**
  - **The intercept β_0 is the log mean (proportion) death within 4 years after study accrual among subjects unexposed to estrogen (*estrogen=0)* and without prior history of CVD *(prevdis=0)*.**
    - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the log sample mean for the group with both *estrogen* and *prevdis* equal to 0.**
      - **log(0.0180929) = -4.0122**

- **The slope β_1 is the difference in log mean (proportion) death within 4 years after study accrual between subjects exposed to estrogen (*estrogen=1)* and subjects unexposed to estrogen (*estrogen=0)* while holding all other modeled variables constant. This last condition is only possible when *prevdis*=0, so the interpretation of this slope is the association between estrogen and 4 year CVD mortality among those without prior history of CVD.**
  - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the difference in log sample means for the group having *estrogen=1* and *prevdis=0* minus the log sample mean for the group having *estrogen=0* and *prevdis=0*.**
    - **log(0.0064516) – log(0.0180929) = -1.03119**
- **The slope β_2 is the difference in log mean (proportion) death within 4 years after study accrual between subjects with prior history of CVD (*prevdis=1)* and subjects without prior history of CVD (*prevdis=0)* while holding all other modeled variables constant. This last condition is only possible when *estrogen*=0, so the interpretation of this slope is the association between prior CVD and 4 year CVD mortality among those without exposure to estrogen.**
  - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the difference in log sample means for the group having *estrogen=0* and *prevdis=1* minus the log sample mean for the group having *estrogen=0* and *prevdis=0*.**
    - **log(0.0992218) – log(0.0180929) = 1.70184**
- **The slope β_3 is not as easily interpreted using the idea of "holding all other modeled variables constant", because either one of the other variables must be different if the value of *estr_prev* is different. Instead we note that according to the regression model**
  - **Expected log mortality in the group having having *estrogen=0* and *prevdis=0* is β_0.**
  - **Expected log mortality in the group having having *estrogen=1* and *prevdis=0* is $\beta_0 + \beta_1$.**
  - **Expected log mortality in the group having having *estrogen=0* and *prevdis=1* is $\beta_0 + \beta_2$.**
  - **Expected log mortality in the group having having *estrogen=1* and *prevdis=1* is $\beta_0 + \beta_1 + \beta_2 + \beta_3$.**
  - **Hence**
    - **β_1 measures the association (log RR) between estrogen and 4 year CVD mortality in subjects without prior history of CVD, and**
    - **$\beta_1 + \beta_3$ measures the association (log RR) between estrogen and 4 year CVD mortality in subjects with prior history of CVD, so**
    - **β_3 measures the difference between the "effect" (log RR) of estrogen in subjects with prior CVD and the "effect" (log RR) of estrogen in subjects without prior CVD (a "difference of differences").**
    - **Note that in this "saturated" model, the estimate corresponds exactly to the corresponding difference of difference in log sample means.**
      - **(log(0.03333333) - log(0.0992218)) – (log(0.0064516) – log(0.0180929)) = -0.5960**
  - **(We could have expressed the interaction parameter as the difference in the effect (log RR) of prior CVD across the estrogen exposure strata. The interpretation of the interaction is symmetric in this sense.)**

- **We can further examine the predictions from this saturated model using Stata's predict command**
  - **Note that by default, the predicted mean, rather than predicted log mean, is returned:**

```
. predict fit0
(option mu assumed; predicted mean cvddeath4)

. bysort estrogen prevdis: summ fit0 cvddeath4

-> estrogen = 0.000, prevdis = 0.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fit0 |      2045    .0180929           0   .0180929   .0180929
   cvddeath4 |      2045    .0180929    .1333201          0          1

-> estrogen = 0.000, prevdis = 1.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fit0 |       514    .0992218           0   .0992218   .0992218
   cvddeath4 |       514    .0992218    .2992508          0          1

-> estrogen = 1.000, prevdis = 0.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fit0 |       310    .0064516           0   .0064516   .0064516
   cvddeath4 |       310    .0064516    .0801919          0          1

-> estrogen = 1.000, prevdis = 1.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fit0 |        30    .0333387           0   .0333387   .0333387
   cvddeath4 |        30    .0333333    .1825742          0          1
```

  - **The CI and P value for slope β_3 given in the regression output is interpretable as statistical tests for the existence of effect modification in the population.**
    - **Neither the small magnitude of the slope estimate nor the high P value (P = 0.96) are suggestive that prior history of CVD modifies the association between estrogen exposure and 4 year CVD mortality**
- **We can also examine the interpretation of output when using Stata's eform option**

```
. glm cvddeath4 estrogen prevdis estr_prev, family(gaussian) link(log) eform robust

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2895
                                                  Scale parameter =  .0294383
Deviance         =   85.2240145                   (1/df) Deviance =  .0294383
Pearson          =   85.2240145                   (1/df) Pearson  =  .0294383

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             = -.6862012
Log pseudolikelihood =  998.6486912               BIC             = -22994.07

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .3565823   .2579973    -1.43   0.154     .0863558    1.472407
     prevdis |   5.484013   1.153151     8.09   0.000     3.631718     8.28104
   estr_prev |    .942284   1.156954    -0.05   0.961     .0849299     10.4545
------------------------------------------------------------------------------
```

Stata Comments

- **Use of the eform option in Stata merely asks for printout that uses exponentiated parameter estimates**
  - **In this model using a log link, exponentiated intercept is the estimated risk and the exponentiated slopes are the estimated risk ratios**
  - **Stata suppresses printing of the intercept**
  - **Note again that the standard error in this output is <u>NOT</u> used to compute the Z statistic, P value, or CI: those are derived on the scale of the log risk**

Statistical Comments

- **Interpretation of the regression parameters is similar to the interpretation of regression parameters in the analogous RD regression model that included the interaction, except we are now modeling the risk of death. (Compare the wording of this description with the analogous wording for problem #1 on page 26-27: all I did was add the word "exponentiated" or change "difference" to "ratio" in key places. I highlighted the introduced wording in red (which I can hardly detect) in order to emphasize (to those of you with "normal" color vision) the minimal way the wording changes):**
  - **The exponentiated intercept e^{β_0} is the mean (proportion) death within 4 years after study accrual among subjects unexposed to estrogen (*estrogen=0)* and without prior history of CVD *(prevdis=0)*.**
    - **(The Stata output suppressed printing of this value.)**

- **The exponentiated slope e^{β_1} is the ratio of mean (proportion) death within 4 years after study accrual between subjects exposed to estrogen (*estrogen=1)* and subjects unexposed to estrogen (*estrogen=0)* while holding all other modeled variables constant. This last condition is only possible when *prevdis*=0, so the interpretation of this slope is the association between estrogen and 4 year CVD mortality among those without prior history of CVD.**
  - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the ratio of sample means for the group having *estrogen=1* and *prevdis=0* divided by the sample mean for the group having *estrogen=0* and *prevdis=0*.**
    - **0.0064516 / 0.0180929 = 0.3566**
- **The exponentiated slope e^{β_2} is the ratio of mean (proportion) death within 4 years after study accrual between subjects with prior history of CVD (*prevdis=1)* and subjects without prior history of CVD (*prevdis=0)* while holding all other modeled variables constant. This last condition is only possible when *estrogen*=0, so the interpretation of this slope is the association between prior CVD and 4 year CVD mortality among those without exposure to estrogen.**
  - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the ratio of sample means for the group having *estrogen=0* and *prevdis=1* divided by the sample mean for the group having *estrogen=0* and *prevdis=0*.**
    - **0.0992218 / 0.0180929 = 5.484**
- **The slope β_3 is not as easily interpreted using the idea of "holding all other modeled variables constant", because either one of the other variables must be different if the value of *estr_prev* is different. Instead we note that according to the regression model**
  - **Expected mortality in the group having having *estrogen=0* and *prevdis=0* is e^{β_0}.**
  - **Expected mortality in the group having having *estrogen=1* and *prevdis=0* is $e^{(\beta_0+\beta_1)}$.**
  - **Expected mortality in the group having having *estrogen=0* and *prevdis=1* is $e^{(\beta_0+\beta_2)}$.**
  - **Expected mortality in the group having having *estrogen=1* and *prevdis=1* is $e^{(\beta_0+\beta_1+\beta_2+\beta_3)}$.**
  - **Hence**
    - **e^{β_1} measures the association (RR) between estrogen and 4 year CVD mortality in subjects without prior history of CVD, and**
    - **$e^{(\beta_1+\beta_3)}$. measures the association (RR) between estrogen and 4 year CVD mortality in subjects with prior history of CVD, so**
    - **e^{β_3} measures the ratio between the "effect" (RR) of estrogen in subjects with prior CVD and the "effect" (RR) of estrogen in subjects without prior CVD (a "ratio of ratios").**
    - **Note that in this "saturated" model, the estimate corresponds exactly to the corresponding ratio of ratios in sample means.**
      - **(0.03333333 / 0.0992218) / (0.0064516 / 0.0180929) = 0.9421**

Problem #2c: Analyses to detect confounding by prevdis with RR

```
. * Problem 2c: evidence of confounding by prevdis
. * See problem 1
```

Statistical comments:

- **The fact that we have switched from RD to RR does not materially affect our analyses exploring associations between estrogen exposure and prior history of CVD.**
- **In the case of a binary outcome and a binary POI, switching from RD and RR will not materially affect our analyses exploring associations between 4 year CVD mortality and prior history of CVD after adjusting for estrogen exposure**
  - **Associations in the RR (an RR different from 1) must correspond to an association in the RD (a nonzero RD)**
- **If we were examining continuous covariates having a U-shaped association with the response, it is possible that there would be no average linear trend on one scale, though there would be a nonzero average linear trend on the other scale**
  - **That having been said, I would likely regard that evidence of confounding on one scale for a binary response (RD, RR, or OR) would motivate the adjustment for the covariate on all scales.**

Problem #2d: Analyses to adjust RR for confounding by *prevdis*: GLM with main effect

```
. * Problem 2d: analyses of cvddeath4 - estrogen association adjusted for prevdis
. *               Adjustment for main effect
. *                       unweighted generalized linear model: glm
. *                               - allowing for possible heteroscedasticity
. glm cvddeath4 estrogen prevdis, family(gaussian) link(log) robust eform

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2896
                                                  Scale parameter =  .0294282
Deviance         =  85.22404619                   (1/df) Deviance =  .0294282
Pearson          =  85.22404619                   (1/df) Pearson  =  .0294282

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             = -.6868908
Log pseudolikelihood =  998.6481523               BIC             = -23002.04

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .3411863   .2571784    -1.43   0.154      .077871    1.494885
     prevdis |    5.47916   1.138795     8.18   0.000     3.645853    8.234338
------------------------------------------------------------------------------
```

Stata Comments

- **There is no real advantage in looking at the intercept for our purposes here, so using the eform option makes reporting of our statistical analysis far easier**

Statistical Comments

- **If we had anticipated that there was no effect modification, the GLM based analysis is the obvious choice.**
- **We use the robust SE to handle the differences in variability of outcome induced by the mean-variance relationship.**
  - **To be either a confounder or a precision variable, difference in risk would be observed across groups, and thus there would similarly be differences in variances across groups**

Scientific Comments

- **After adjustment for difference in the history of prior CVD across groups defined by estrogen exposure, we estimate that risk of 4 year CVD mortality among the estrogen exposed is only 0.3412-fold as high as it is among the estrogen unexposed. Hence, when comparing two groups having similar prior history of CVD, we expect the risk of CVD death**

within 4 years to be associated with a 65.9% relative reduction in risk for subjects exposed to estrogen compared to the unexposed.

- **Such a relative reduction in risk would have to be interpreted to the underlying 1.81% estimated 4 year CVD mortality in the 2,355 subjects without prior history of CVD who had no estrogen exposure and the underlying 9.92% estimated 4 year CVD mortality in the 514 subjects with prior history of CVD who had no estrogen exposure. The 65.9% relative reduction in mortality is based on relatively sparse data: the observed 0.645% mortality in 310 estrogen-exposed subjects without prior CVD and the observed 3.3% estimated 4 year CVD mortality in the 30 estrogen-exposed subjects with prior CVD.**
- **Because of this lack of extensive data, the 95% confidence interval includes a wide range of relative risks having very different public health implications. The observed data suggestive of a 65.9% relative reduction in 4 year CVD mortality risk would not be judged unusual if the true association between estrogen exposure and CVD mortality were anywhere between a 92.2% relative reduction (so a risk only 0.0779-fold as high as the unexposed) and a 49.5% relative increase in risk (so a risk 1.495-fold higher in the estrogen exposed group compared to an unexposed group having the same prior history of CVD).**
- **The observed association is not statistically significant (two-sided P= .154).**

Problem #2d: Analyses to adjust RR for confounding by *prevdis*: GLM with main effect

```
. poisson cvddeath4 estrogen prevdis, irr

Poisson regression                                Number of obs   =       2899
                                                  LR chi2(2)      =      70.41
                                                  Prob > chi2     =     0.0000
Log likelihood = -370.77218                       Pseudo R2       =     0.0867

------------------------------------------------------------------------------
   cvddeath4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .3492807   .2057931    -1.79   0.074     .1100662    1.108396
     prevdis |   5.474078   1.163684     8.00   0.000     3.608792    8.303481
------------------------------------------------------------------------------
```

Statistical Comments

- **This Poisson analysis gives very similar results to the GLM with Gaussian family and log link.**
  - **Slight differences are due to the different weighting used in the estimating equation, but with the only very slight differences in the estimated effects across *prevdis* strata, this will not make too much difference**

Problem #2d: Analyses to adjust RR for confounding by *prevdis*: stratified analyses

Statistical Comments

- **As with analysis of RD, an alternative to the regression model for RR is to use a stratified analysis**
- **However, the use of ratios means that we will have to choose between**
  - **taking ratios of risks that were first averaged across strata,**
  - **taking arithmetic averages of stratum-specific ratios, or**
  - **taking geometric means of stratum-specific ratios.**
- **The following highlights these differences (parts of this are common to stratified rates with RD)**
  - **Suppose we have strata labeled $s = 1, 2, \ldots, S$**
  - **We estimate incidences p_{s1} (in the estrogen exposed) and p_{s0} (in the estrogen unexposed) in each stratum**
  - **We then average across the stratum specific incidences using some predefined weights $w_1, w_2, \ldots, w_S$**

$$p_1 = \frac{\sum_{s=1}^{S} w_s p_{s1}}{\sum_{s=1}^{S} w_s} \qquad p_0 = \frac{\sum_{s=1}^{S} w_s p_{s0}}{\sum_{s=1}^{S} w_s}$$

  - **The standardized RR is then $\theta = p_1 / p_0$ is then interpretable as the ratio of incidence in a population where the weights represent the prevalence of the different strata.**
    - **For this reason, we often standardize age-adjusted rates according to some specific population of interest**
  - **Now that we are considering RR (a ratio of means), we would only be estimating the same quantity with an average of stratum-specific risk ratios if <u>every</u> stratum has the same true risk ratio**
    - **Even then, the estimates for a RR based on ratios of average risk and a RR based on average stratum-specific risks might often be markedly different.**
  - **Suppose we have estimated an effect within each stratum $\theta_s = p_{s1} / p_{s0}$, and then calculated a weighted average (mean or geometric mean) across strata**

$$\left.\begin{array}{l} p_1 = \dfrac{\sum_{s=1}^{S} w_s p_{s1}}{\sum_{s=1}^{S} w_s} \quad p_0 = \dfrac{\sum_{s=1}^{S} w_s p_{s0}}{\sum_{s=1}^{S} w_s} \quad \theta = p_1 / p_0 \\ \theta_s = p_{s1} / p_{s0} \qquad \theta^{*} = \dfrac{\sum_{s=1}^{S} w_s \theta_s}{\sum_{s=1}^{S} w_s} \\ \theta_s = p_{s1} / p_{s0} \qquad \theta^{**} = \exp\left(\dfrac{\sum_{s=1}^{S} w_s \log(\theta_s)}{\sum_{s=1}^{S} w_s} \right) \end{array}\right\} \Longrightarrow \quad \theta = \theta^{*} = \theta^{**} \quad \text{only if} \quad \theta_s = \theta \quad \text{for } s = 1, 2, \ldots, S$$

- **As compared to standardized rates with RD, in the presence of effect modification, we not only have to choose suitable weights, but also how we will use them**

Stata Comments:

- **The standardized RR that first averages risk:**
  - **Easily computed using cs in Stata**
  - **More laboriously computed using GLM of model with interactions, log link, and post-estimation command nlcom**
- **A weighted arithmetic mean of stratum-specific RR:**
  - **Can be laboriously computed using GLM of model with interactions, log link, and post-estimation command nlcom**
- **A weighted geometric mean of stratum-specific RR:**
  - **Can be fairly easily computed using GLM of model with interactions, log link, and post-estimation command lincom**
  - **and then computes the ratio is implemented by cs in Stata**

- **Ratios of weighted average of risk in exposed to weighted average of risk in unexposed**
  - **Weighting according to the distribution of the estrogen exposed across *prevdis* strata (Stata's internal standards)**

```
. *               Adjustment for main effect and potential estrogen - prevdis interaction
. *                     stratified analysis: cs
. cs cvddeath4 estrogen, by(prevdis) istandard
         prevdis |       RR       [95% Conf. Interval]       Weight
-----------------+------------------------------------------------
               0 |    .3565824      .086377   1.472048          310
               1 |    .3359477      .048058   2.348428           30
-----------------+------------------------------------------------
           Crude |    .2565842       .08164   .8064117
 I. Standardized |    .3494282     .1111289   1.098725
```

  - **Weighting according to the distribution of the estrogen unexposed across *prevdis* strata (Stata's external standards)**

```
. cs cvddeath4 estrogen, by(prevdis) estandard
         prevdis |       RR       [95% Conf. Interval]       Weight
-----------------+------------------------------------------------
               0 |    .3565824      .086377   1.472048         2045
               1 |    .3359477      .048058   2.348428          514
-----------------+------------------------------------------------
           Crude |    .2565842       .08164   .8064117
 E. Standardized |    .3446237     .0977688   1.214758
```

  - **Weighting according to some distribution that might matter to us**
    - **I give the example of using weights according to the distribution of *prevdis* in the entire sample**

```
. g std= 544 / 2899
. replace std= 1 - std if prevdis==0
(2355 real changes made)

. cs cvddeath4 estrogen, by(prevdis) standard(std)

         prevdis |       RR       [95% Conf. Interval]       Weight
-----------------+------------------------------------------------
               0 |    .3565824      .086377   1.472048     .8123491
               1 |    .3359477      .048058   2.348428     .1876509
-----------------+------------------------------------------------
           Crude |    .2565842       .08164   .8064117
    Standardized |    .3450507     .0998757   1.192081
```

Statistical comments:

- **Because there was not strong evidence of effect modification, we get similar estimates with different weightings**
- **None of them, however, correspond with the estimate from the adjusted GLM model**

Problem #2d: Analyses to adjust for *prevdis*: weighted regression analysis adjusting for main effect and interaction between *estrogen* and *prevdis*

Statistical Comments

- **In the above stratified analyses, we have no opportunity to estimate the effect of *prevdis***
  - **This is not of too much concern, because that was not our question**
- **If we did care, we could fit a model with effect modification**
  - **A multiplicative effect is all that makes sense when we have a binary POI and a binary confounder**
  - **In other cases, we could consider more complicated interactions, but a multiplicative interaction is still the most commonly used**

```
. *                       generalized linear model with log link: glm
. *                               - allowing for possible heteroscedasticity
. glm cvddeath4 estrogen prevdis estr_prev, family(gaussian) link(log) robust eform

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2895
                                                  Scale parameter =  .0294383
Deviance         =   85.2240145                   (1/df) Deviance =  .0294383
Pearson          =   85.2240145                   (1/df) Pearson  =  .0294383

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             = -.6862012
Log pseudolikelihood =  998.6486912               BIC             = -22994.07

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .3565823   .2579973    -1.43   0.154     .0863558    1.472407
     prevdis |   5.484013   1.153151     8.09   0.000     3.631718     8.28104
   estr_prev |    .942284   1.156954    -0.05   0.961     .0849299     10.4545
------------------------------------------------------------------------------
```

Statistical Comments

- **As before, we will have to test both parameters that involve estrogen**
- **The same comments hold regarding the general relative merits of this model to that with only the main effect**
- **However, with no evidence of effect modification in our sample, this would have been an unnecessary choice for a model, providing we could have made that decision before looking at our data.**

```
. test estrogen estr_prev

 ( 1)  [cvddeath4]estrogen = 0
 ( 2)  [cvddeath4]estr_prev = 0

           chi2(  2) =    3.24
         Prob > chi2 =    0.1979
```

Scientific Comments

- **As before, we will need to use some other approach to summarize the overall association between estrogen exposure and 4 year CVD mortality**
- **We can use lincom to produce a weighted average as before**
  - **Even if we appropriately weight according to sample sizes, however, this will not tend to agree with the estimates derived from cs**

```
. lincom (1 - 544/2899) * estrogen + (544/2899) * (estrogen + estr_prev), eform

 ( 1)  [cvddeath4]estrogen + .1876509*[cvddeath4]estr_prev = 0

------------------------------------------------------------------------------
   cvddeath4 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3526159   .2173672    -1.69   0.091     .1053382     1.18037
------------------------------------------------------------------------------
```

- **We can laboriously compute the RR as the ratio of weighted stratum-specific risks, which will agree with cs**

```
. * Weight for no prior history stratum
. di (1 - 544/2899)
.81234909

. * Average risk in no estrogen stratum
. di .81234909 * exp(-4.012235) + (1-.81234909) * exp(-4.012235+1.701838)
.03331683

. * Average risk in estrogen exposure stratum
. di .81234909 * exp(-4.012235+  -1.03119) + (1-.81234909) * exp(-4.012235+1.701838+  -1.03119 + -.0596098)
.011496

. * Relative average risk
. di .011496/.03331683
.34505083
```

- **We can also use post-estimation command nlcom**
  - **nlcom can handle more complicated functions and (as we need in this case) ratios**
  - **We have to reference parameter estimates using: _b[*name*] (underscore b and square brackets)**

```
. * Inference using nlcom
. nlcom ( (1 - 544/2899) * exp(_b[_cons] + _b[estrogen]) + ///
>         (544/2899) * exp(_b[_cons] + _b[prevdis] + _b[estrogen] + _b[estr_prev]) ) ///
>         / ((1 - 544/2899) * exp(_b[_cons]) + (544/2899) * exp(_b[_cons] + _b[prevdis]))

       _nl_1:  ( (1 - 544/2899) * exp(_b[_cons] + _b[estrogen]) + (544/2899) * exp(_b[_cons] + _b[prevdis]
       + _b[estrogen] + _b[estr_prev]) ) / ((1 - 544/2899) * exp(_b[_cons]) + (544/2899) * exp(_b[_cons] +
       _b[prevdis]))

------------------------------------------------------------------------------
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .3450507     .21826     1.58   0.114    -.0827309    .7728324
------------------------------------------------------------------------------
```

Problem #2d: Analyses to adjust for *prevdis*: Poisson regression analysis adjusting for main effect and interaction between *estrogen* and *prevdis*

```
. poisson cvddeath4 estrogen prevdis estr_prev, irr

Poisson regression                                Number of obs   =       2899
                                                  LR chi2(3)      =      70.41
                                                  Prob > chi2     =     0.0000
Log likelihood = -370.77103                       Pseudo R2       =     0.0867

------------------------------------------------------------------------------
   cvddeath4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .3565824   .2588668    -1.42   0.155     .0859441    1.479461
     prevdis |   5.484015   1.184279     7.88   0.000      3.59154    8.373683
   estr_prev |   .9421321   1.171671    -0.05   0.962     .0823238    10.78197
------------------------------------------------------------------------------
```

Statistical Comments

As we might expect in this saturated model, we get pretty much the same results with the Gaussian family, Poisson family, or (not shown) the binomial family mean-variance relationships

Problem #2e: Analyses to detect further confounding by age with RR

```
. * Problem 2e: evidence of confounding by age
. * See problem 1
```

Statistical comments:

- **The fact that we have switched from RD to RR does not materially affect our analyses exploring associations between age and estrogen exposure after adjustment for prior history of CVD.**
- **In the case of a continuous covariates having a monotonically increasing association with the response, I would regard that my conclusions about the association with the response are also unchanged.**

Problem #2f: Analyses to adjust for age
Statistical comments

- **We can again consider all the different ways of modeling age**
- **However, I will restrict attention to just the models that I would regularly entertain**
  - **Adjustment for a linear effect with age**
  - **Adjustment for a quadratic effect with age**
  - **Adjustment with linear splines**

Problem #2f: Analyses to adjust for age : continuous linear adjustment (GLM Gaussian)

```
. glm cvddeath4 estrogen prevdis age, robust link(log) eform

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2895
                                                  Scale parameter =  .0291915
Deviance         =  84.50939886                   (1/df) Deviance =  .0291915
Pearson          =  84.50939886                   (1/df) Pearson  =  .0291915

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             = -.6946217
Log pseudolikelihood =  1010.854202               BIC             = -22994.78

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .3046138   .2005835    -1.81   0.071     .0838005    1.107268
     prevdis |   3.748734   1.134614     4.37   0.000     2.071357    6.784444
         age |   1.055203   .0242024     2.34   0.019     1.008818    1.103721
------------------------------------------------------------------------------
```

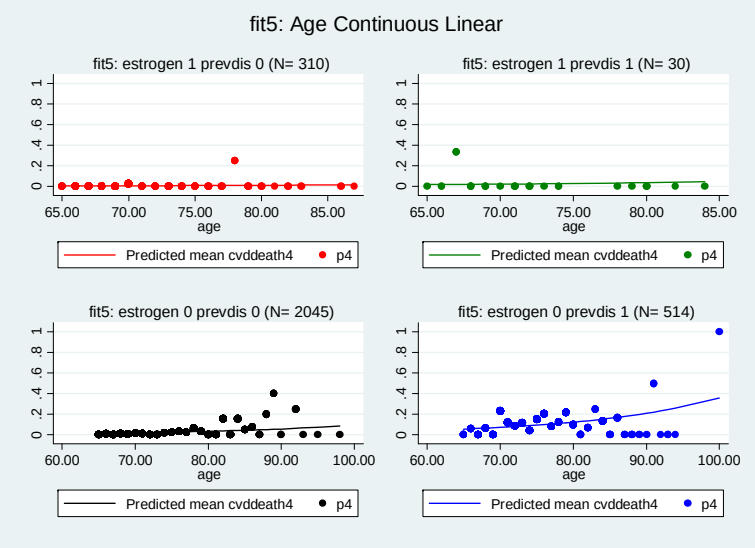
Scientific Comments

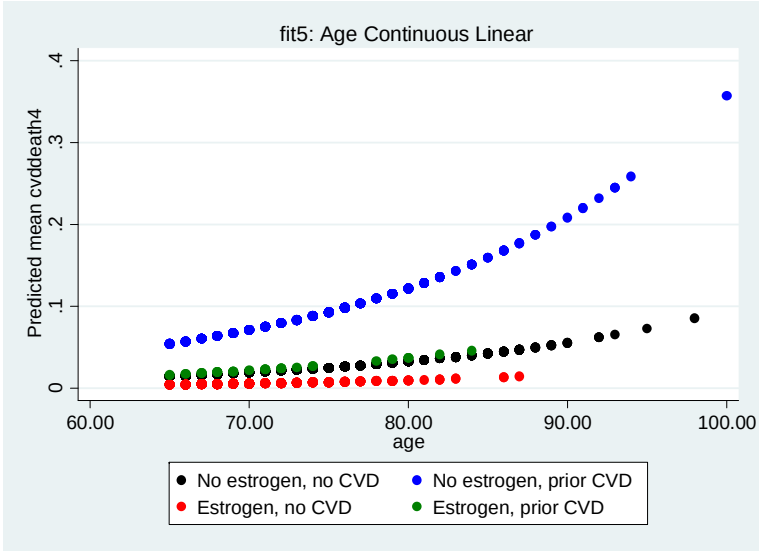
- **After adjustment for differences in the history of prior CVD and age across groups defined by estrogen exposure, we estimate that risk of 4 year CVD mortality among the estrogen exposed is only 0.3046-fold as high as it is among the estrogen unexposed. Hence, when comparing two groups having similar prior history of CVD and age, we expect the risk of CVD death within 4 years to be associated with a 69.5% relative reduction in risk for subjects exposed to estrogen compared to the unexposed.**
- **The 95% confidence interval includes a wide range of relative risks having very different public health implications. The observed data suggestive of a 69.5% relative reduction in 4 year CVD mortality risk would not be judged unusual**

if the true association between estrogen exposure and CVD mortality were anywhere between a 91.6% relative reduction (so a risk only 0.0838-fold as high as the unexposed) and a 10.7% relative increase in risk (so a risk 1.107-fold higher in the estrogen exposed group compared to an unexposed group having the same prior history of CVD).

- **The observed association is not statistically significant (two-sided P= .071).**

```
. predict fit5
(option mu assumed; predicted mean cvddeath4)

. * (see do file for code used to produce graphs)
```



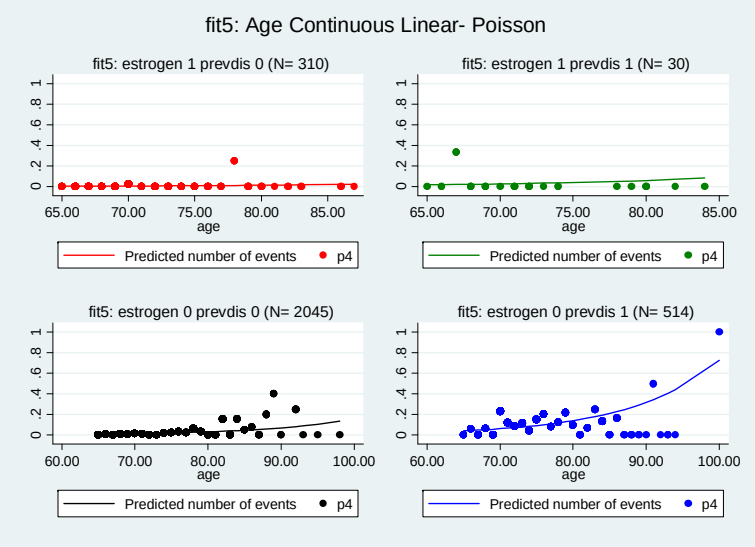


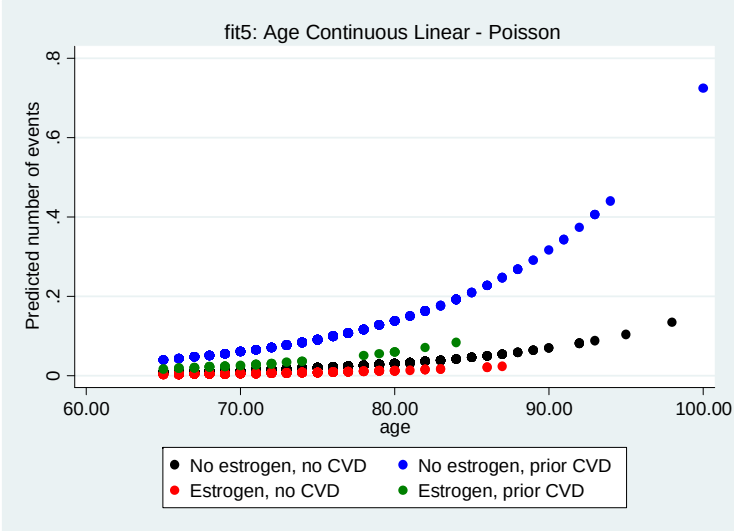
Statistical Comments

- **Note that we modeled a straight line relationship between age and log risk.**
- **Such a model predicts a curvilinear relationship between age and risk.**
  - **One could plot the risk on a log scale, but in this range of risk I would generally stick to the linear scale.**

Problem #2f: Analyses to adjust for age : continuous linear adjustment (Poisson)

```
. poisson cvddeath4 estrogen prevdis age, irr

Poisson regression                                Number of obs   =       2899
                                                  LR chi2(3)      =      96.87
                                                  Prob > chi2     =     0.0000
Log likelihood = -357.54163                       Pseudo R2       =     0.1193

------------------------------------------------------------------------------
   cvddeath4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .4345841   .2569713    -1.41   0.159     .1363813    1.384818
     prevdis |   4.577285   .9887876     7.04   0.000     2.997305    6.990126
         age |   1.086366   .0165222     5.45   0.000     1.054461    1.119237
------------------------------------------------------------------------------

. predict fit5
(option n assumed; predicted number of events)

. * (see do file for code used to produce graphs)
```





Statistical Comments

- **The Poisson family mean-variance relationship seems to be estimating a lesser protective effect of estrogen.**
- **My best guess: Not as much weight is being placed on subjects with high estimated means, thus removing some amount of influential points. (I probably trust the Poisson analysis a bit more due to the unweighted estimating equation.)**

Problem #2f: Analyses to adjust for age : continuous quadratic adjustment (GLM Gaussian)

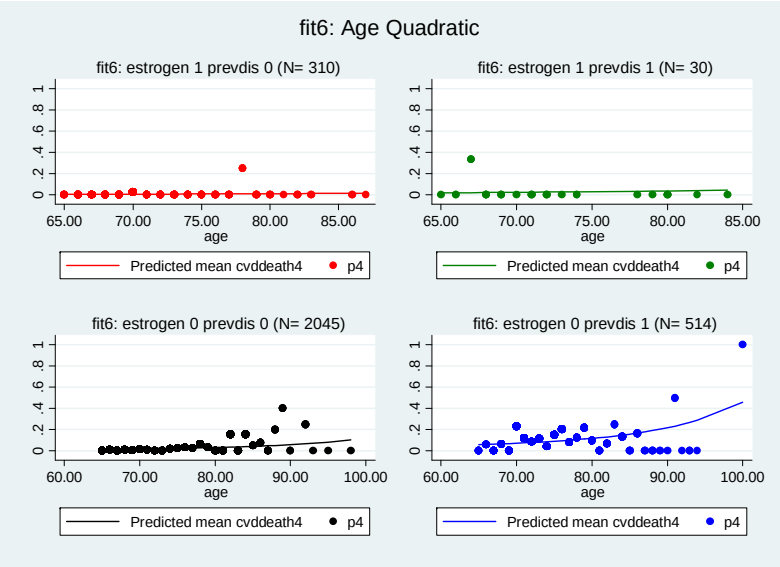
```
. glm cvddeath4 estrogen prevdis age agesqr, robust link(log) eform

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2894
                                                  Scale parameter =  .0291985
Deviance         =  84.50051681                   (1/df) Deviance =  .0291985
Pearson          =  84.50051681                   (1/df) Pearson  =  .0291985

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             = -.6940369
Log pseudolikelihood =  1011.006554               BIC             = -22986.82

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |    .313736   .2165283    -1.68   0.093     .0811145    1.213473
     prevdis |    3.76347   1.074513     4.64   0.000     2.150606    6.585914
         age |   .9528952   .5667561    -0.08   0.935     .2970118    3.057149
      agesqr |   1.000646   .0038334     0.17   0.866     .9931612    1.008188
------------------------------------------------------------------------------

. test age agesqr

 ( 1)  [cvddeath4]age = 0
 ( 2)  [cvddeath4]agesqr = 0

           chi2(  2) =   13.38
         Prob > chi2 =    0.0012
```

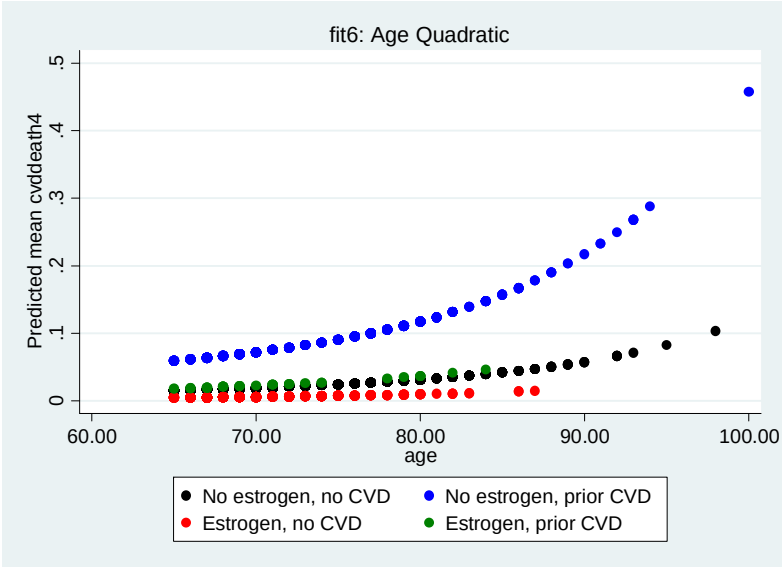
Statistical Comments

- **Note according to the coefficient table, neither the *age* nor the *agesqr* terms are statistically significant after adjusting for each other (and *prevdis* and *estrogen*).**
- **It would be an extremely bad thing to suggest that such results indicate no association between age and 4 year CVD mortality.**
    - **We have to look at the test of the two parameters simultaneously**

- **Though it was not our question, were we to have prespecified the question, the above analysis does suggest that the linear age term fits the data as well as the quadratic does.**
  - **The lack of statistical significance for the *agesqr* term suggests that we do not have strong evidence of a quadratic effect.**
  - **If we removed the quadratic term, the *age* coefficient then is statistically significant**
  - **I do not generally include a quadratic term without the linear term, so even though the P value is higher for the linear term, it is still the quadratic term I would most consider removing.**

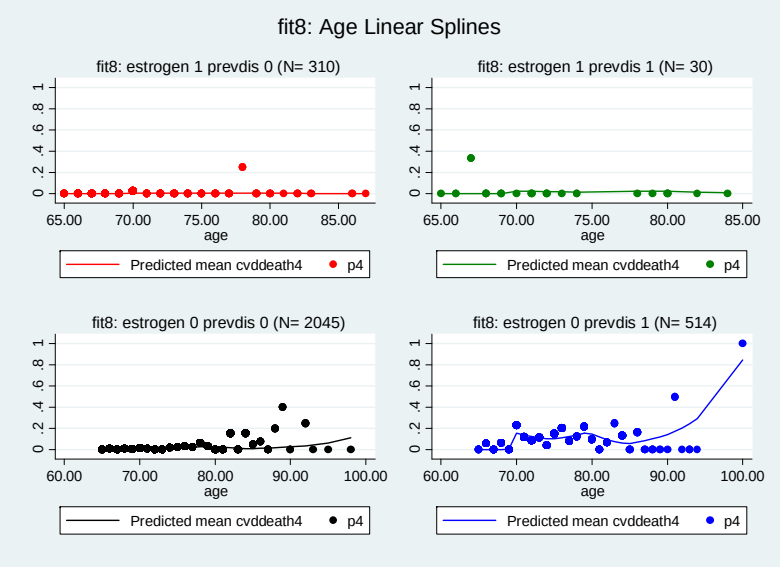
```
. predict fit6
(option mu assumed; predicted mean cvddeath4)

. * (see do file for code used to produce graphs)
```

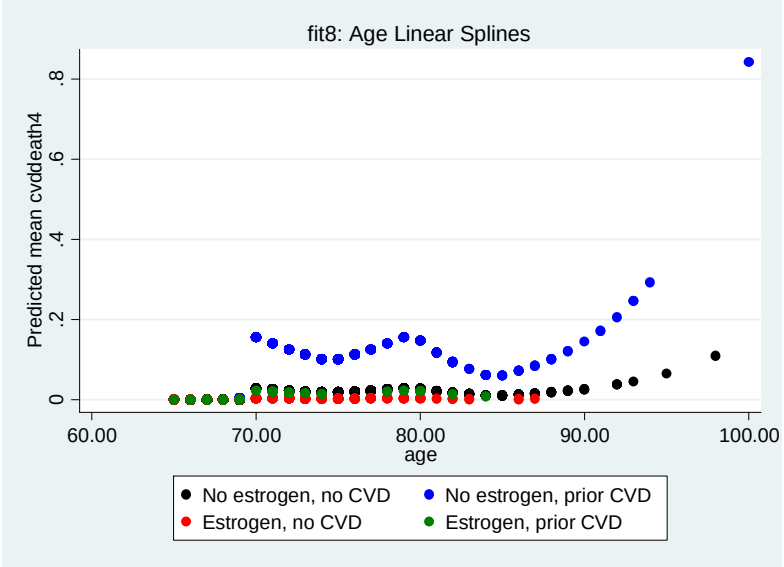




Problem #2f: Analyses to adjust for age : linear splines adjustment (Gaussian)

```
. *                        Splines
. glm cvddeath4 estrogen prevdis age65 age70 age75 age80 age85, robust link(log) eform

Generalized linear models                         No. of obs      =      2899
Optimization     : ML                             Residual df     =      2891
                                                  Scale parameter =  .0288953
Deviance         =  83.53623216                   (1/df) Deviance =  .0288953
Pearson          =  83.53623216                   (1/df) Pearson  =  .0288953

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             = -.7034445
Log pseudolikelihood =  1027.642748               BIC             = -22963.87

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .1465972    .117646    -2.39   0.017     .0304107    .7066839
     prevdis |   5.412527   2.150786     4.25   0.000     2.484028    11.79353
       age65 |    1482.35   3842.282     2.82   0.005     9.217622    238387.1
       age70 |   .8984226    .134875    -0.71   0.476     .6694139    1.205776
       age75 |   1.113403   .1195404     1.00   0.317     .9021177    1.374173
       age80 |     .80319   .1877442    -0.94   0.348     .5079857    1.269946
       age85 |   1.192671   .0877864     2.39   0.017     1.032447    1.377759
------------------------------------------------------------------------------

. predict fit8
(option mu assumed; predicted mean cvddeath4)

. * (see do file for code used to produce graphs)
```





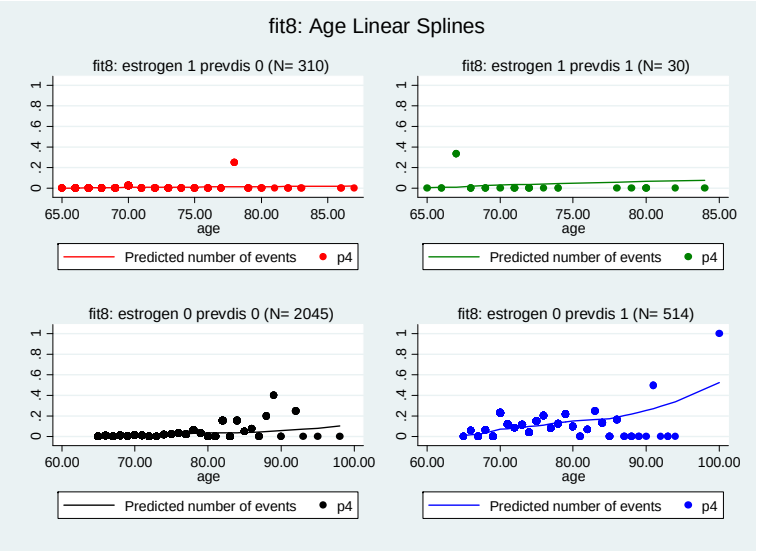
Statistical Comments

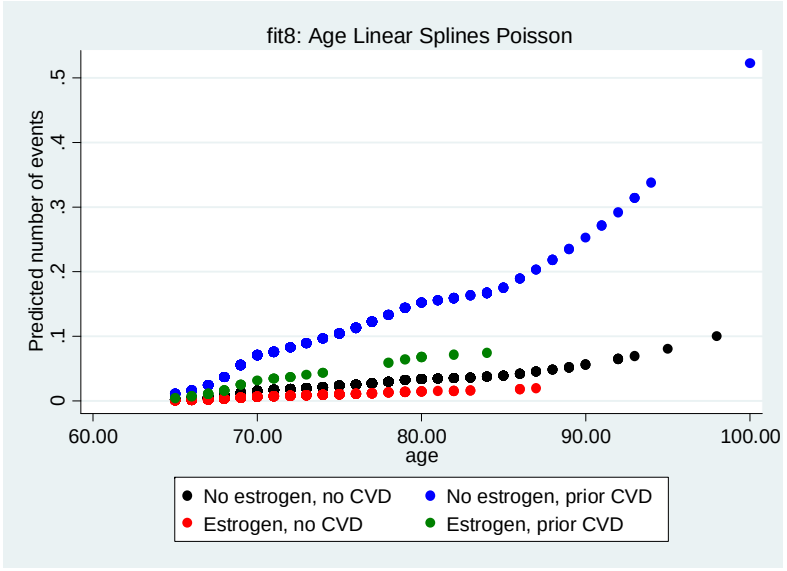
- **Note according to this adjusted analysis, after adjustment for *prevdis* and the age splines, there is a statistically significant association between estrogen use and 4 year CVD mortality (P = 0.017).**
- **However when we look at the plots, it sure does look to me like we have overfit the data.**
- **Had I prespecified the adjustment using splines, I would have to go with it, but I sure would mention in my discussion that the plots of the age effect looked most suspicious.**
- **This is the sort of thing that makes me greatly prefer just using the quadratic adjustment. It is less prone to influential points altering the estimates in such a big way. (The greatest harm done in statistical analysis is overfitting data.)**
- **We might anticipate that using Poisson regression would behave better, because its estimating equation is not weighted.**

Problem #2f: Analyses to adjust for age : linear splines adjustment (Poisson)

```
. poisson cvddeath4 estrogen prevdis age65 age70 age75 age80 age85, irr

Poisson regression                                Number of obs   =       2899
                                                  LR chi2(7)      =     101.81
                                                  Prob > chi2     =     0.0000
Log likelihood = -355.06777                       Pseudo R2       =     0.1254

------------------------------------------------------------------------------
   cvddeath4 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .4478631   .2649255    -1.36   0.174     .1404857    1.427771
     prevdis |   4.503398   .9692515     6.99   0.000     2.953514    6.866598
       age65 |   1.501907   .3195493     1.91   0.056     .9897814    2.279012
       age70 |   1.081051   .0991204     0.85   0.395     .9032343    1.293874
       age75 |   1.083755   .0936713     0.93   0.352     .9148708    1.283814
       age80 |   1.024482   .0997986     0.25   0.804     .8464199    1.240004
       age85 |   1.075444   .0731767     1.07   0.285      .941173    1.228871
------------------------------------------------------------------------------
```

```
. predict fit8
(option n assumed; predicted number of events)
```





Statistical Comments

- **As anticipated, the Poisson regression fit was less prone to influential points when using splines.**
- **It is noteworthy, however, that the estimated RR is markedly different in all Poisson regression models that I explored (results not shown) compared to the models using the Gaussian mean-variance relationship**
    - **The Gaussian models tended to estimate adjusted RR of 0.35, while the Poisson models tended to estimate adjusted RR of 0.45**
    - **These are just different weightings of the observations in a setting in which we do not really have a lot of data (recall the low numbers of events)**
    - **The P values were relatively similar, even when the estimates differed.**

PROBLEM #3: Analyses based on odds ratio (OR)

Statistical comments

- **In a two sample test of binomial proportions, analyses based on OR are relatively common, though RD probably predominates.**
  - **When adjusting for additional covariates, however, it is more common to base analyses on OR.**
- **In Problem #3, we mirror the analyses explored for problem #1 insofar as possible. Notable differences include**
  - **t tests will not apply**
  - **regression models will only consider logistic regression**
  - **stratified models will tend to behave nearly the same as regression models**
    - **We tend to use the Mantel-Haenszel (MH) statistic to do stratified analyses**
    - **This is not the same as "combining on the basis of the log odds ratio" (which would correspond to logistic regression), but the MH statistic is not all that different**

Problem #3a: Analyses of OR in 2 by 2 contingency tables (two sample tests of binomial proportions)

```
. * EXECUTION OF CODE FOR PROBLEM 3
. * Problem 3a: unadjusted analyses of cvddeath4 - estrogen association
. *               chi-square test: cc
. cc cvddeath4 estrogen
                                                        Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |         3          88  |         91      0.0330
        Controls |       337        2471  |       2808      0.1200
-----------------+------------------------+------------------------
           Total |       340        2559  |       2899      0.1173
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         .2499663       |     .050319    .7614012 (exact)
 Prev. frac. ex. |         .7500337       |    .2385988     .949681 (exact)
 Prev. frac. pop |         .0900147       |
                 +-----------------------------------------------
                              chi2(1) =     6.45  Pr>chi2 = 0.0111
. di  (3 / 337)   /    (88 / 2471)
.24996628
```

Stata Comments

- **Note that Stata command cc provides the estimated odds ratio, which is just the ratio of the sample odds**
- **The CI is computed by exact methods through inversion of Fisher's exact test, so will not agree with the logistic regression interval**
- **The P value can be provided by either the chi square test or Fisher's exact test**
  - **This is the same test as used for the risk difference, which makes sense: In a two sample problem, the risk difference is nonzero if and only if the risk ratio is not equal to 1. (Similarly, the risk difference is nonzero if and only if the odds ratio is not equal to 1.)**
  - **Of note, even when using the Fisher's exact test and the CI based on the inverted Fisher's exact test, there are some computational complexities due to discreteness that will sometimes lead to disagreement between the test and the CI.**

Scientific Comments

- **I would most often summarize the association between a binary response and a binary POI with the risk difference**
- **As noted previously, standard reporting practice would likely be to report the chi squared or Fisher's exact test when the two sample comparison is the primary interest**
  - **CI for RD, RR, or OR may not agree with the P value decision owing to the failure to fully account for the mean-variance relationship and the nuisance parameter as well as the discreteness of the data**

Problem #3a: Analyses of OR in logistic regression

```
. *               logistic regression
. logit cvddeath4 estrogen

Logistic regression                                     Number of obs   =       2899
                                                        LR chi2(1)      =       8.64
                                                        Prob > chi2     =     0.0033
Log likelihood = -400.21007                             Pseudo R2       =     0.0107

------------------------------------------------------------------------------
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -1.386429   .5899737    -2.35   0.019    -2.542757    -.230102
       _cons |  -3.335041   .1084819   -30.74   0.000    -3.547662   -3.122421
------------------------------------------------------------------------------

. logit cvddeath4 estrogen, robust

Logistic regression                                     Number of obs   =       2899
                                                        Wald chi2(1)    =       5.52
                                                        Prob > chi2     =     0.0188
Log pseudolikelihood = -400.21007                       Pseudo R2       =     0.0107

------------------------------------------------------------------------------
             |               Robust
   cvddeath4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |  -1.386429   .5900755    -2.35   0.019    -2.542956   -.2299025
       _cons |  -3.335041   .1085007   -30.74   0.000    -3.547699   -3.122384
------------------------------------------------------------------------------

. logistic cvddeath4 estrogen

Logistic regression                                     Number of obs   =       2899
                                                        LR chi2(1)      =       8.64
                                                        Prob > chi2     =     0.0033
Log likelihood = -400.21007                             Pseudo R2       =     0.0107

------------------------------------------------------------------------------
   cvddeath4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .2499663   .1474735    -2.35   0.019     .0786493    .7944526
------------------------------------------------------------------------------
```

Stata Comments

- **We can use either the Stata logit command, the Stata logistic command, the Stata binreg command, or the Stata glm command to provide regression based inference about the OR**
- **I highly recommend using the logit and logistic commands**
  - **The logit command reports inference based on the log odds (for the intercept) and the log odds ratio (for slopes)**
  - **The logistic command suppresses information about the intercept, and reports inference on the odds ratio (for slopes)**
    - **As described previously for the case of the log link, the SE from the logistic output is <u>NOT</u> the SE used to compute the Z statistic, the P values, and the CI—these come from the log odds scale**
- **However, the fact that the log link is nonlinear means that we still have to use an iterative search (I suppressed the output)**

Statistical Comments

- **We are fitting a simple logistic regression model in which**

$$\text{logit}\left(E\left[cvddeath4 \mid estrogen\right]\right) = \beta_0 + \beta_1 \times estrogen$$

- **Interpretation of the regression parameters:**
  - **The intercept β_0 is the log odds of death within 4 years after study accrual among subjects unexposed to estrogen (*estrogen=0)*.**
    - **The exponentiated intercept e^{β_0} is the odds of death within 4 years after study accrual among subjects unexposed to estrogen (*estrogen=0)*.**
    - **We can compute the probability of death within 4 years after study accrual as (odds / (odds + 1)), so the probability of an event is $p = e^{\beta_0} / \left(1 + e^{\beta_0}\right)$**
  - **The slope β_1 is log odds ratio for death within 4 years after study accrual between subjects exposed to estrogen (*estrogen=1)* and subjects unexposed to estrogen (*estrogen=0)*.**
    - **The exponentiated slope e^{β_1} is the odds ratio (OR) of the odds of death within 4 years after study accrual comparing subjects exposed to estrogen (*estrogen=1)* to subjects unexposed to estrogen (*estrogen=0)*.**
    - **To understand the scientific of a particular OR, we need to consider the baseline (reference) risk.**
- **In this simple logistic regression model (one predictor variable) having a binary predictor (a 0-1 variable), the predicted values will be (nearly) equal to the sample mean in each group**
  - **$\hat{\beta}_0$ will be equal to the log odds among subjects having a predictor value equal to 0**
  - **$\hat{\beta}_1$ will be equal to the sample log odds ratio: log of (odds for group 1 divided by odds for group 0)**
- **In the logistic regression model with binary data, there can be no real difference between the use of robust SE and standard SE.**
- **Note that the OR is very nearly the same as the RR in this setting of relatively rare events.**

Scientific Comments

- **As with the inference based on RD or RR, my usual habit would be to use the chi squared statistic to test for an association, which is all the more sensible when using logistic regression for other analyses: the chi squared test is the score test from logistic regression. I would tend to use CI from logistic regression in order to keep methods the same with adjusted analyses.**
- **We find that there is an association between estrogen use and four year CVD mortality in the sample**
    - **Estimate odds of 4 year CVD mortality among estrogen exposed is only 0.250 as large as that among estrogen unexposed (95% CI: .0786 to .794 as large; chi squared two-sided P = 0.011).**

Problem #3b: Analyses to detect effect modification with OR

Scientific Comments

- **Effect modification is a function of both the summary measure (proportion or odds with binary data) and the comparison (difference or ratio) used to quantify an association**
- **Hence, we cannot just rely on the decision re effect modification in problem #1**
- **We must evaluate the evidence for effect modification for the OR**
- **That having been said, the fact that OR is nearly the RR in this setting of rare events argues that we will likely observe the same type of evidence we did for RR**

```
.
. * Problem 3b: evidence of effect modification by prevdis
. *               descriptive statistics
. cc cvddeath4 estrogen, by(prevdis)

         prevdis |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               0 |   .3524044      .0409568   1.378646     4.839066 (exact)
               1 |   .3130494      .0075193   1.976695      2.71875 (exact)
-----------------+-------------------------------------------------
           Crude |   .2499663       .050319   .7614012              (exact)
    M-H combined |   .3382473      .1055259     1.0842
---------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.01  Pr>chi2 = 0.9251

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      3.66
                                              Pr>chi2 =    0.0559
```

Scientific Comments

- **The estimated OR is quite similar in the two strata defined by *prevdis*: 0.352 vs 0.313, though a little more different than were the RRs**
  - **It should be noted that RR less than 1.0 will always seem closer together than if we took the inverse OR. But in this case, 1 / 0.352 = 2.8409 and 1/ 0.313 = 3.194 would be judged fairly similar, as well**
- **Furthermore, tests for effect modification in the population were not significant (P = .93 from the Mantel-Haenszel test for homogeneity)**
- **We can also examine a comparable analysis based on the logistic regression model**

```
. *               inferential statistics
. logistic cvddeath4 estrogen prevdis estr_prev

Logistic regression                               Number of obs   =       2899
                                                  LR chi2(3)      =      73.48
                                                  Prob > chi2     =     0.0000
Log likelihood =  -367.7927                       Pseudo R2       =     0.0908

------------------------------------------------------------------------------
   cvddeath4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .3524043   .2567409    -1.43   0.152     .0845096    1.469523
     prevdis |   5.977933   1.327218     8.05   0.000     3.868723    9.237075
   estr_prev |   .8883244   1.119085    -0.09   0.925     .0752061    10.49277
------------------------------------------------------------------------------
```

Scientific Comments

- **We are fitting a linear regression model in which**

$$\text{logit}\left(E\left[cvddeath4 \mid estrogen\right]\right) = \beta_0 + \beta_1 \times estrogen + \beta_2 \times prevdis + \beta_3 \times estr\_prev$$

- **I only presented the output from the logistic command, and thus only exponentiated slopes are provided. Nonetheless, I provide the interpretation of the regression parameters as they would have been presented by the logit command.**
- **Interpretation of the regression parameters is similar to the interpretation of regression parameters in the analogous RD regression model that included the interaction, except we are now modeling the log risk of death. (Compare the wording of this description with the analogous wording for problem #1 on page 26-27: all I did was add the word "log odds" in key places. I highlighted the introduced wording in red (which I can hardly detect) in order to emphasize (to those of you with "normal" color vision) the minimal way the wording changes):**
  - **The intercept β_0 is the log odds of death within 4 years after study accrual among subjects unexposed to estrogen (*estrogen=0)* and without prior history of CVD *(prevdis=0)*.**
    - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the sample log odds for the group with both *estrogen* and *prevdis* equal to 0.**
  - **The slope β_1 is the logarithm of the ratio comparing odds of death within 4 years after study accrual between subjects exposed to estrogen (*estrogen=1)* and subjects unexposed to estrogen (*estrogen=0)* while holding all other modeled variables constant. This last condition is only possible when *prevdis*=0, so the interpretation of this slope is the association between estrogen and 4 year CVD mortality among those without prior history of CVD.**
    - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the difference in log odds for the group having *estrogen=1* and *prevdis=0* minus the log odds for the group having *estrogen=0* and *prevdis=0*.**

- **The slope β_2 is the difference in log odds of within 4 years after study accrual between subjects with prior history of CVD (*prevdis=1)* and subjects without prior history of CVD (*prevdis=0)* while holding all other modeled variables constant. This last condition is only possible when *estrogen*=0, so the interpretation of this slope is the association between prior CVD and 4 year CVD mortality among those without exposure to estrogen.**
  - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the difference in log odds for the group having *estrogen=0* and *prevdis=1* minus the log odds for the group having *estrogen=0* and *prevdis=0*.**
- **The slope β_3 is not as easily interpreted using the idea of "holding all other modeled variables constant", because either one of the other variables must be different if the value of *estr_prev* is different. Instead we note that according to the regression model**
  - **Expected log odds of mortality in the group having having *estrogen=0* and *prevdis=0* is β_0.**
  - **Expected log odds of mortality in the group having having *estrogen=1* and *prevdis=0* is $\beta_0+\beta_1$.**
  - **Expected log odds of mortality in the group having having *estrogen=0* and *prevdis=1* is $\beta_0+\beta_2$.**
  - **Expected log odds of mortality in the group having having *estrogen=1* and *prevdis=1* is $\beta_0+\beta_1+\beta_2+\beta_3$.**
  - **Hence**
    - **β_1 measures the association (log OR) between estrogen and 4 year CVD mortality in subjects without prior history of CVD, and**
    - **$\beta_1+\beta_3$ measures the association (log OR) between estrogen and 4 year CVD mortality in subjects with prior history of CVD, so**
    - **β_3 measures the difference between the "effect" (log OR) of estrogen in subjects with prior CVD and the "effect" (log OR) of estrogen in subjects without prior CVD (a "difference of differences").**
    - **Note that in this "saturated" model, the estimate corresponds exactly to the corresponding difference of difference in log odds.**
      - **(log(0.03333333) - log(0.0992218)) – (log(0.0064516) – log(0.0180929)) = -0.5960**
  - **(We could have expressed the interaction parameter as the difference in the effect (log OR) of prior CVD across the estrogen exposure strata. The interpretation of the interaction is symmetric in this sense.)**
- **We can further examine the predictions from this saturated model using Stata's predict command**
  - **Note that by default, the predicted mean, rather than predicted log odds or odds, is returned:**

```
. predict fit0
(option pr assumed; Pr(cvddeath4))

. bysort estrogen prevdis: summ fit0 cvddeath4
-> estrogen = 0.000, prevdis = 0.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fit0 |      2045    .0180929           0   .0180929   .0180929
   cvddeath4 |      2045    .0180929    .1333201          0          1

-> estrogen = 0.000, prevdis = 1.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fit0 |       514    .0992218           0   .0992218   .0992218
   cvddeath4 |       514    .0992218    .2992508          0          1

-> estrogen = 1.000, prevdis = 0.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fit0 |       310    .0064516           0   .0064516   .0064516
   cvddeath4 |       310    .0064516    .0801919          0          1

-> estrogen = 1.000, prevdis = 1.000

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fit0 |        30    .0333333           0   .0333333   .0333333
   cvddeath4 |        30    .0333333    .1825742          0          1
```

- **Interpretation of the regression parameters from the exponentiated output (Stata's logistic command) is similar to the interpretation of regression parameters in the analogous RD regression model that included the interaction, except we are now modeling the odds of death. (Compare the wording of this description with the analogous wording for problem #1 on page 26-27: all I did was add the word "exponentiated" or change "difference" to "ratio" in key places. I highlighted the introduced wording in red (which I can hardly detect) in order to emphasize (to those of you with "normal" color vision) the minimal way the wording changes):**
  - **The exponentiated intercept e^{β_0} is the odds of death within 4 years after study accrual among subjects unexposed to estrogen (*estrogen=0)* and without prior history of CVD *(prevdis=0)*.**
    - **(The Stata output suppressed printing of this value.)**

- **The exponentiated slope e^{β_1} is the ratio of odds of death within 4 years after study accrual between subjects exposed to estrogen (*estrogen=1)* and subjects unexposed to estrogen (*estrogen=0)* while holding all other modeled variables constant. This last condition is only possible when *prevdis*=0, so the interpretation of this slope is the association between estrogen and 4 year CVD mortality among those without prior history of CVD.**
  - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the ratio of odds of death for the group having *estrogen=1* and *prevdis=0* divided by the odds of death for the group having *estrogen=0* and *prevdis=0*.**
- **The exponentiated slope e^{β_2} is the ratio of odds of death within 4 years after study accrual between subjects with prior history of CVD (*prevdis=1)* and subjects without prior history of CVD (*prevdis=0)* while holding all other modeled variables constant. This last condition is only possible when *estrogen*=0, so the interpretation of this slope is the association between prior CVD and 4 year CVD mortality among those without exposure to estrogen.**
  - **Note that in this "saturated model" having four distinct groups and four regression parameters, the estimate is exactly equal to the ratio of odds of death for the group having *estrogen=0* and *prevdis=1* divided by the odds of death for the group having *estrogen=0* and *prevdis=0*.**
- **The slope β_3 is not as easily interpreted using the idea of "holding all other modeled variables constant", because either one of the other variables must be different if the value of *estr_prev* is different. Instead we note that according to the regression model**
  - **Expected odds of death in the group having having *estrogen=0* and *prevdis=0* is e^{β_0}.**
  - **Expected odds of death in the group having having *estrogen=1* and *prevdis=0* is $e^{(\beta_0+\beta_1)}$.**
  - **Expected odds of death in the group having having *estrogen=0* and *prevdis=1* is $e^{(\beta_0+\beta_2)}$.**
  - **Expected odds of death in the group having having *estrogen=1* and *prevdis=1* is $e^{(\beta_0+\beta_1+\beta_2+\beta_3)}$.**
  - **Hence**
    - **e^{β_1} measures the association (OR) between estrogen and 4 year CVD mortality in subjects without prior history of CVD, and**
    - **$e^{(\beta_1+\beta_3)}$. measures the association (OR) between estrogen and 4 year CVD mortality in subjects with prior history of CVD, so**
    - **e^{β_3} measures the ratio between the "effect" (OR) of estrogen in subjects with prior CVD and the "effect" (OR) of estrogen in subjects without prior CVD (a "ratio of ratios").**
    - **Note that in this "saturated" model, the estimate corresponds exactly to the corresponding ratio of ratios in sample odds of death.**
- **The CI and P value for slope β_3 given in the regression output is interpretable as statistical tests for the existence of effect modification in the population.**
  - **Neither the small magnitude of the slope estimate nor the high P value (P = 0.96) are suggestive that prior history of CVD modifies the association between estrogen exposure and 4 year CVD mortality**

Problem #3c: Analyses to detect confounding by prevdis with OR

```
. * Problem 3c: evidence of confounding by prevdis
. * See problem 1
```

Statistical comments:

- **The fact that we have switched from RD or RR to OR does not materially affect our analyses exploring associations between estrogen exposure and prior history of CVD.**
- **In the case of a binary outcome and a binary POI, switching from RD and RR will not materially affect our analyses exploring associations between 4 year CVD mortality and prior history of CVD after adjusting for estrogen exposure**
  - **Associations in the OR (an OR different from 1) must correspond to an association in the RD (a nonzero RD)**
- **If we were examining continuous covariates having a U-shaped association with the response, it is possible that there would be no average linear trend on one scale, though there would be a nonzero average linear trend on the other scale**
  - **That having been said, I would likely regard that evidence of confounding on one scale for a binary response (RD, RR, or OR) would motivate the adjustment for the covariate on all scales.**
- **I do note that if we were trying to compare unadjusted and adjusted analyses with respect to changes in the coefficient estimates, the OR is non-collapsible.**
  - **Hence, adjusting for precision variables can lead to changes in the coefficient for the POI.**
  - **In this case, however, the unadjusted OR was 0.250, and the OR adjusted for *prevdis* was 0.338—a value closer to the null OR of 1.0 than the unadjusted estimate. So in this case, we can use these findings to suggest confounding was present.**
    - **If *prevdis* were a precision variable, we would have expected the adjusted OR to be more extreme than the unadjusted OR.**

Problem #3d: Analyses to adjust OR for confounding by *prevdis*: logistic regression with main effect

```
. * Problem 3d: analyses of cvddeath4 - estrogen association adjusted for prevdis
. *                Adjustment for main effect
. *                        logistic regression
. logistic cvddeath4 estrogen prevdis

Logistic regression                                     Number of obs   =       2899
                                                        LR chi2(2)      =      73.47
                                                        Prob > chi2     =     0.0000
Log likelihood = -367.79717                             Pseudo R2       =     0.0908

------------------------------------------------------------------------------
   cvddeath4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .3382447   .2010262    -1.82   0.068     .1055221    1.084223
     prevdis |   5.955624   1.300542     8.17   0.000     3.881938    9.137049
------------------------------------------------------------------------------
```

Stata Comments

- **There is no real advantage in looking at the intercept for our purposes here, so using the logistic commandmakes reporting of our statistical analysis far easier**

Statistical Comments

- **If we had anticipated that there was no effect modification, the logistic regression based analysis is the obvious choice.**
- **There is no real advantage in using the robust SE in this model.**

Scientific Comments

- **After adjustment for difference in the history of prior CVD across groups defined by estrogen exposure, we estimate that odds of 4 year CVD mortality among the estrogen exposed is only 0.338-fold as high as it is among the estrogen unexposed. Hence, when comparing two groups having similar prior history of CVD, we expect the risk of CVD death within 4 years to be associated with a 66.2% relative reduction in odds of death for subjects exposed to estrogen compared to the unexposed.**
- **Such a relative reduction in odds of death would have to be interpreted to the underlying 1.81% estimated 4 year CVD mortality in the 2,355 subjects without prior history of CVD who had no estrogen exposure and the underlying 9.92% estimated 4 year CVD mortality in the 514 subjects with prior history of CVD who had no estrogen exposure. The 66.2% relative reduction in mortality is based on relatively sparse data: the observed 0.645% mortality in 310 estrogen-exposed subjects without prior CVD and the observed 3.3% estimated 4 year CVD mortality in the 30 estrogen-exposed subjects with prior CVD.**
- **Because of this lack of extensive data, the 95% confidence interval includes a wide range of relative risks having very different public health implications. The observed data suggestive of a 66.2% relative reduction in 4 year CVD**

mortality risk would not be judged unusual if the true association between estrogen exposure and CVD mortality were anywhere between a 89.45% relative reduction (so a risk only 0.1055-fold as high as the unexposed) and a 8.42% relative increase in risk (so a risk 1.0842-fold higher in the estrogen exposed group compared to an unexposed group having the same prior history of CVD).

- **The observed association is not statistically significant (two-sided P= .068).**

Problem #3d: Analyses to adjust OR for confounding by *prevdis*: stratified analyses

Statistical Comments

- **As with analysis of RD and RR, an alternative to the regression model for OR is to use a stratified analysis**
  - **Unlike those other measures, stratification models used with OR (Mantel-Haenszel statistics) are much closer to the regression models that we commonly use.**

```
. *              Adjustment for main effect and estrogen - prevdis interaction
. *                    stratified analysis: cs
```

- **Weighting according to the distribution of the estrogen exposed across *prevdis* strata (Stata's internal standards)**

```
. cc cvddeath4 estrogen, by(prevdis) istandard
         prevdis |       OR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
               0 |   .3524044      .0409568   1.378646          308 (exact)
               1 |   .3130494      .0075193   1.976695           29 (exact)
-----------------+-------------------------------------------------
           Crude |   .2499663       .050319   .7614012             (exact)
 I. Standardized |   .3382308      .1055106   1.084252
```

- **Weighting according to the distribution of the estrogen unexposed across *prevdis* strata (Stata's external standards)**

```
. cc cvddeath4 estrogen, by(prevdis) estandard
         prevdis |       OR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
               0 |   .3524044      .0409568   1.378646         2008 (exact)
               1 |   .3130494      .0075193   1.976695          463 (exact)
-----------------+-------------------------------------------------
           Crude |   .2499663       .050319   .7614012             (exact)
 E. Standardized |   .3295963      .0915301   1.186864
```

- **Weighting according to some distribution that might matter to us**
  - **I give the example of using weights according to the distribution of *prevdis* in the entire sample**

```
. g std= 544 / 2899
. replace std= 1 - std if prevdis==0
(2355 real changes made)
. cc cvddeath4 estrogen, by(prevdis) standard(std)

         prevdis |       OR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
               0 |   .3524044      .0409568   1.378646     .8123491 (exact)
               1 |   .3130494      .0075193   1.976695     .1876509 (exact)
-----------------+-------------------------------------------------
           Crude |   .2499663       .050319   .7614012             (exact)
    Standardized |   .3295789      .0914828   1.187351
```

Problem #3d: Analyses to adjust for *prevdis*: logistic regression analysis adjusting for main effect and interaction between *estrogen* and *prevdis*

Statistical Comments

- **In the above stratified analyses, we have no opportunity to estimate the effect of *prevdis***
  - **This is not of too much concern, because that was not our question**
- **If we did care, we could fit a model with effect modification**
  - **A multiplicative effect is all that makes sense when we have a binary POI and a binary confounder**
  - **In other cases, we could consider more complicated interactions, but a multiplicative interaction is still the most commonly used**

```
. *                       logistic regression
. logistic cvddeath4 estrogen prevdis estr_prev

Logistic regression                               Number of obs   =       2899
                                                  LR chi2(3)      =      73.48
                                                  Prob > chi2     =     0.0000
Log likelihood =  -367.7927                       Pseudo R2       =     0.0908

------------------------------------------------------------------------------
   cvddeath4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .3524043   .2567409    -1.43   0.152     .0845096    1.469523
     prevdis |   5.977933   1.327218     8.05   0.000     3.868723    9.237075
   estr_prev |   .8883244   1.119085    -0.09   0.925     .0752061    10.49277
------------------------------------------------------------------------------
```

Statistical Comments

- **As before, we will have to test both parameters that involve estrogen**
- **The same comments hold regarding the general relative merits of this model to that with only the main effect**
- **However, with no evidence of effect modification in our sample, this would have been an unnecessary choice for a model, providing we could have made that decision before looking at our data.**
- **The following test finds no statistically significant association between estrogen exposure and 4 year CVD mortality.**

```
. test estrogen estr_prev

 ( 1)  [cvddeath4]estrogen = 0
 ( 2)  [cvddeath4]estr_prev = 0

           chi2(  2) =    3.33
         Prob > chi2 =    0.1895
```

Scientific Comments

- **As before, we will need to use some other approach to summarize the overall association between estrogen exposure and 4 year CVD mortality**
- **We can use lincom to produce a weighted average as before**

```
. lincom (1 - 544/2899) * estrogen + (544/2899) * (estrogen + estr_prev), eform

 ( 1)  [cvddeath4]estrogen + .1876509*[cvddeath4]estr_prev = 0

------------------------------------------------------------------------------
   cvddeath4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3446598   .2145367    -1.71   0.087     .1017542    1.167425
------------------------------------------------------------------------------
```

Problem #3e: Analyses to detect further confounding by age with OR

```
. * Problem 3e: evidence of further confounding by age
. *               (see problem 1)
```

Statistical comments:

- **The fact that we have switched from RD to OR does not materially affect our analyses exploring associations between age and estrogen exposure after adjustment for prior history of CVD.**
- **In the case of a continuous covariates having a monotonically increasing association with the response, I would regard that my conclusions about the association with the response are also unchanged.**

Problem #3f: Analyses to adjust for age

Statistical comments

- **We can again consider all the different ways of modeling age in logistic regression**
- **However, I will restrict attention to just the models that I would regularly entertain**
  - **Adjustment for a linear effect with age**
  - **Adjustment for a quadratic effect with age**
  - **Adjustment with linear splines**

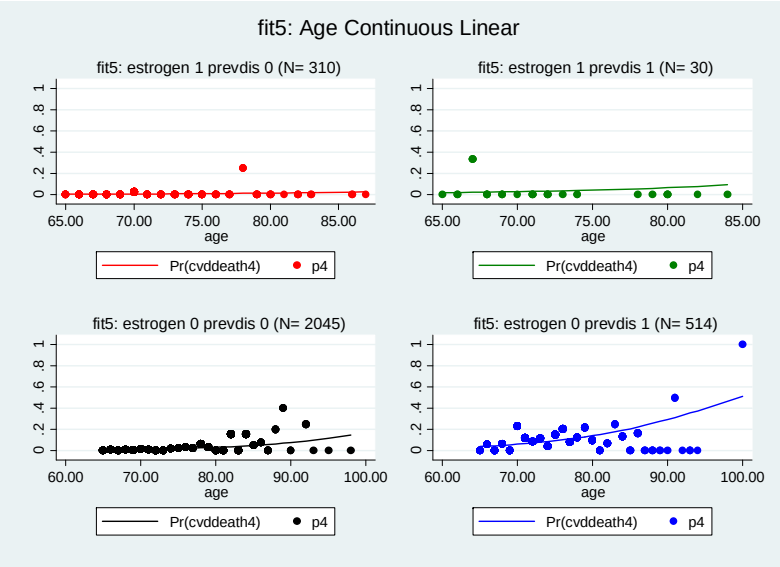
Problem #3f: Analyses to adjust for age : continuous linear adjustment

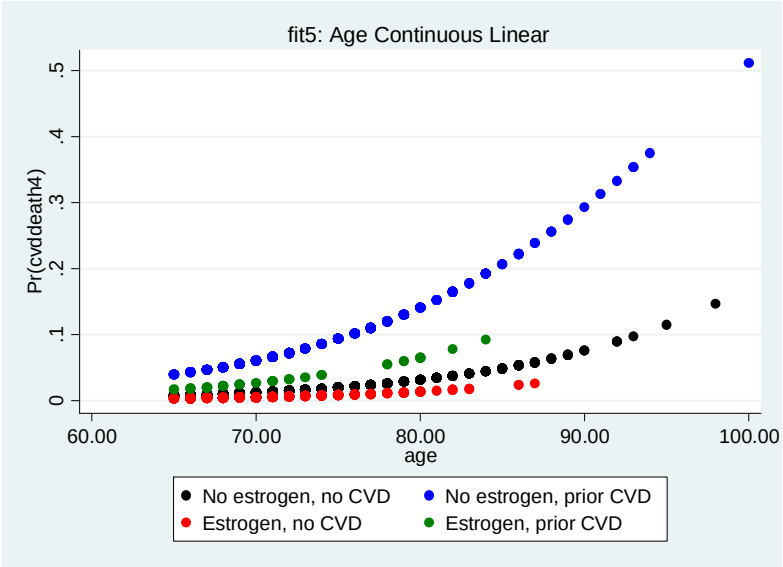
```
. * Problem 3f: association of cvddeath4 - estrogen adjusted for prevdis, age
. *                Adjustment for main effect
. *                        Continuous linear
. logistic cvddeath4 estrogen prevdis age

Logistic regression                                     Number of obs   =       2899
                                                        LR chi2(3)      =     102.52
                                                        Prob > chi2     =     0.0000
Log likelihood = -353.27188                             Pseudo R2       =     0.1267

------------------------------------------------------------------------------
   cvddeath4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .4271517   .2553941    -1.42   0.155      .132327    1.378846
     prevdis |   5.061125   1.123878     7.30   0.000     3.275129    7.821061
         age |   1.097151   .0183913     5.53   0.000     1.061691    1.133796
------------------------------------------------------------------------------
```

Scientific Comments

- **After adjustment for differences in the history of prior CVD and age across groups defined by estrogen exposure, we estimate that odds of 4 year CVD mortality among the estrogen exposed is only 0.427-fold as high as it is among the estrogen unexposed. Hence, when comparing two groups having similar prior history of CVD and age, we expect the risk of CVD death within 4 years to be associated with a 57.3% relative reduction in risk for subjects exposed to estrogen compared to the unexposed.**
- **The 95% confidence interval includes a wide range of odds ratios having very different public health implications. The observed data suggestive of a 57.3% relative reduction in odds of 4 year CVD mortality would not be judged unusual if the true association between estrogen exposure and CVD mortality were anywhere between a 86.8% relative reduction (so odds only 0..132-fold as high as the unexposed) and a 37.9% relative increase in risk (so odds 1.379-fold higher in the estrogen exposed group compared to an unexposed group having the same prior history of CVD).**
- **The observed association is not statistically significant (two-sided P= .155).**

```
. predict fit5
(option pr assumed; Pr(cvddeath4))

. * (see do file for code used to produce graphs)
```





Statistical Comments

- **Note that we modeled a straight line relationship between age and log odds of mortality.**
- **Such a model predicts a curvilinear relationship between age and risk.**
  - **One could plot the risk on a log odds scale, but I have next to never seen that done.**

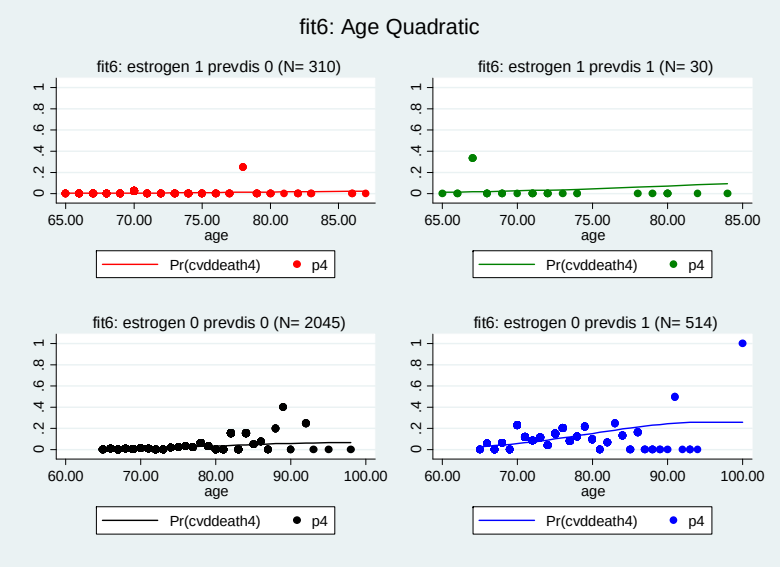
Problem #3f: Analyses to adjust for age : continuous quadratic adjustment

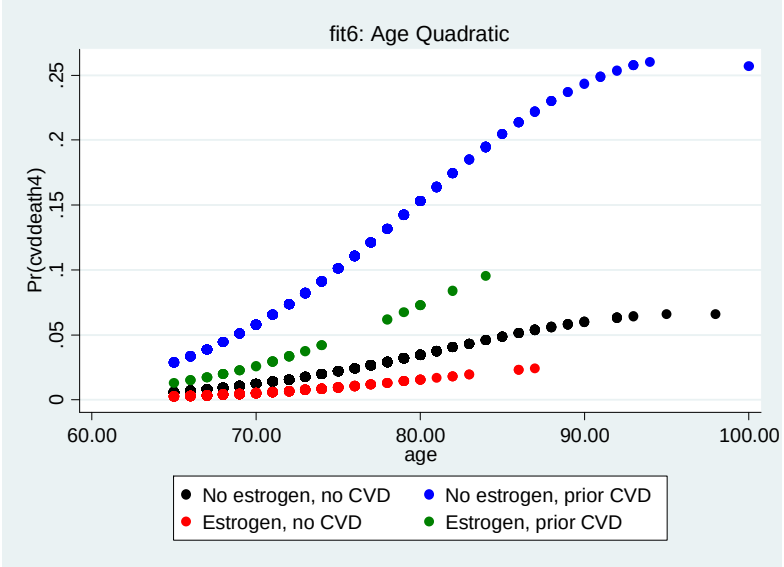
```
. *                        Quadratic
. logistic cvddeath4 estrogen prevdis age agesqr

Logistic regression                               Number of obs   =       2899
                                                  LR chi2(4)      =     104.08
                                                  Prob > chi2     =     0.0000
Log likelihood = -352.48968                       Pseudo R2       =     0.1286

------------------------------------------------------------------------------
   cvddeath4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .4348902   .2602475    -1.39   0.164     .1345868     1.40526
     prevdis |   5.031681   1.115348     7.29   0.000     3.258603     7.76953
         age |   1.625457   .5283819     1.49   0.135     .8595639     3.07378
      agesqr |   .9974845   .0020773    -1.21   0.226     .9934214    1.001564
------------------------------------------------------------------------------
```

Statistical Comments

- **Note according to the coefficient table, neither the *age* nor the *agesqr* terms are statistically significant after adjusting for each other (and *prevdis* and *estrogen*).**
- **See the comments for problem 2f regarding the assessment of an association with age, had that been part of our question.**
- **The statistical inference about the estrogen- mortality association is not different.**

```
. predict fit6
(option pr assumed; Pr(cvddeath4))

. * (see do file for code used to produce graphs)
```





Statistical Comments

- **The fit of these quadratic models shows some very mild, but strange, curvature owing to the fitting of a quadratic.**
- **This occurs in areas where there is no data for the estrogen-exposed groups, and thus it is not that disturbing.**

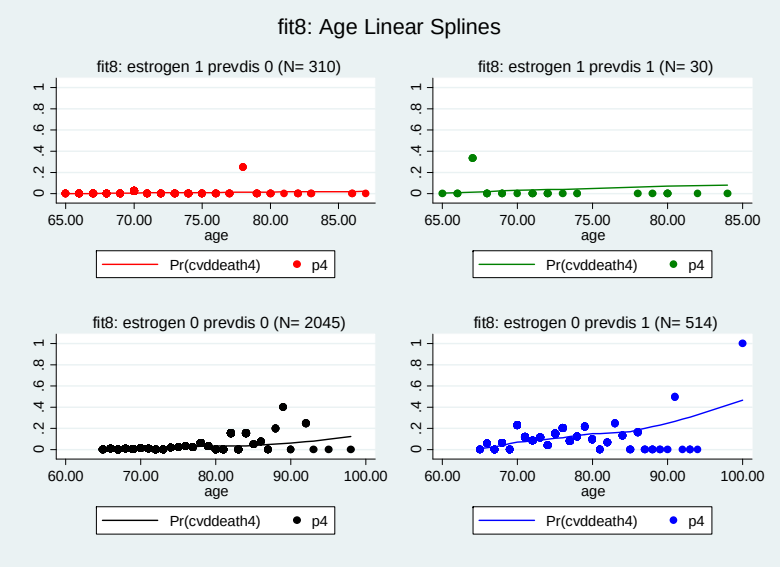
Problem #2f: Analyses to adjust for age : linear splines adjustment

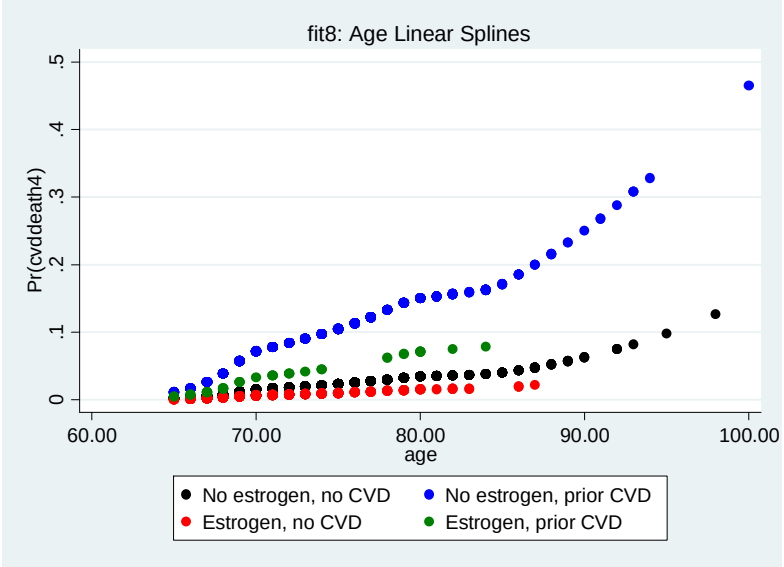
```
. *                        Splines
. logistic cvddeath4 estrogen prevdis age65 age70 age75 age80 age85

Logistic regression                                     Number of obs   =       2899
                                                        LR chi2(7)      =     106.99
                                                        Prob > chi2     =     0.0000
Log likelihood = -351.03814                             Pseudo R2       =     0.1322

------------------------------------------------------------------------------
   cvddeath4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    estrogen |   .4371173   .2617143    -1.38   0.167     .1351947    1.413306
     prevdis |   4.970239   1.102384     7.23   0.000     3.217979    7.676642
       age65 |   1.512439   .3255347     1.92   0.055     .9918978    2.306158
       age70 |   1.085221   .1022781     0.87   0.386     .9021849    1.305391
       age75 |   1.092258   .0986297     0.98   0.328      .915088    1.303731
       age80 |   1.024429   .1071291     0.23   0.817     .8345802    1.257465
       age85 |    1.10066   .0894965     1.18   0.238     .9385133     1.29082
------------------------------------------------------------------------------
```

Statistical Comments

- **The general fit of the splines was similar to that seen for the linear fit.**
- **The adjusted inference was quite similar to that obtained under other models**

```
. predict fit8
(option pr assumed; Pr(cvddeath4))

. * (see do file for code used to produce graphs)
```


fit8: Age Linear Splines
fit8: estrogen 1 prevdis 0 (N= 310)
fit8: estrogen 1 prevdis 1 (N= 30)
fit8: estrogen 0 prevdis 0 (N= 2045)
fit8: estrogen 0 prevdis 1 (N= 514)
0 .2 .4 .6 .8 1
65.00 70.00 75.00 80.00 85.00
60.00 70.00 80.00 90.00 100.00
age
Pr(cvddeath4)
p4


fit8: Age Linear Splines
Pr(cvddeath4)
0 .1 .2 .3 .4 .5
60.00 70.00 80.00 90.00 100.00
age
No estrogen, no CVD
No estrogen, prior CVD
Estrogen, no CVD
Estrogen, prior CVD