Biost 536: Categorical Data Analysis in Epidemiology
Emerson, Fall 2014

Homework #2
October 5, 2014

Written problems: To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 5:30 pm on Sunday, October 12, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

> *On this (as all homeworks) Stata / R code and unedited Stata / R output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*
>
> ***In all problems requesting "statistical analyses" (either descriptive or inferential), you should present both***
>
> - ***Methods:** A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.*
> - ***Inference**: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to "Reporting Associations" for details.*

Questions 1-5 refer to analyses of at the data in the file `mri.txt` that is located on the class webpages. In those questions we are interested in associations between 5 year mortality and prevalence of atherosclerotic cardiovascular disease (ASCVD) as defined by history of prior angina, myocardial infarction, transient ischemic attacks, or stroke. You will likely find it useful to create a new variable indicating ASCVD. This variable can be derived from the *chd* and *stroke* variables.

For these questions, we will be considering adjustment for age and sex using both stratified and regression analyses. For the stratified analyses, it will be necessary to use an appropriate categorization of age

***Instructions for grading**: On this key I place the total points to be awarded for a particular problem just prior to the answer for the problem. I also sometimes provide very specific criteria for awarding points. Remember the purpose of peer grading is to have the grader focus more closely on answers to the problems that were potentially different from those that he/she provided, and to identify areas where the answers on the paper being graded might not be correct. It is not of value to anyone to be overly "lenient" in the grading. So while it is not appropriate to capriciously deduct points, it is equally not appropriate to give points only because "Well, they tried" (unless, of course, such a criterion is specified in the grading instructions). We welcome students' questions regarding the appropriateness of answers during the grading process. And grades assigned by the peer grader can always be appealed.*

In providing answers below, the required answers are in* bold face type*. Further verbage that is in* regular italics type ***is additional information that you are in fact responsible for knowing on exams, but was not really required for the homework assignment. That additional information may help you decide whether the answers you are grading are valid, however.

In each problem requiring both description of methods and results, full credit should only be given when both aspects are appropriately addressed. The grader should be able to determine the exact analysis method from their description.

I note that a major point that I try to make is that there is great similarity among all types of regression. I used "cut-and-paste" extensively. In problem 2, I wrote out the answers completely, even when the descriptive statistics parts of the answer were similar. In problems 3 and 4, I cut and past the answers and highlight what is different from problem 2 in* blue font*. By doing this I am hoping that you will be able to abstract that commonality of approach across the methods.

At the end of this document, I provide Stata code that I used to produce the answers. This is for your later reference when trying to understand Stata. I did not expect (and in fact do not tolerate) the Stata output as solutions to the homework assignment. It should not have been included in a homework set turned in by a student.

1. We are interested in analyzing associations between 5 year mortality and prevalence of ASCVD at study enrollment using statistical methods appropriate for binary response variables. The observation time for death among these subjects is potentially subject to censoring. Provide a statistical analysis demonstrating that such methods as logistic regression can be used to answer this question.

Instructions for grading: This problem is worth 10 points.

***Methods:* The minimum of the observation time among patients still alive was compared to 5 years.**

***Results:* Of patients still alive at the time of data analysis, the shortest observation time was 1827 days, which is just over 5 years. Hence, we have at least 5 years follow-up on all patients, and we will have no incomplete data on a binary indicator of death within 5 years.**

2. Using the risk difference (RD) as a measure of association, provide statistical inference regarding an association between 5 year survival and baseline prevalence of ASCVD, adjusting for age and sex.
    a. Answer the question using a stratified analysis (e.g., using Stata command `cs` or an equivalent analysis in R).

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading:

- ***Description of the method of analysis should make clear that an adjusted RD is estimated by first estimating RD within strata defined by combinations of age and sex, and then combining maximum likelihood estimates across strata in a weighted analysis, with standard errors of the adjusted RD computed by combining the estimated SE from each stratum according to the chosen weights. P values and CI are computed using the approximate normal distribution.***
- ***In deciding upon an exact analysis method, the student would have had to decide how to categorize age. They must make clear the categorization used. Points to consider include:***
    - ***It is greatly to be preferred to categorize a variable in terms of scientific values (e.g., 5 year age ranges) rather than based on quantiles. Quantiles are not reproducible from study to study, and depending on the distribution of the data, it is often the case that the lowest and highest quantile categories cover an extremely broad range, while the middle intervals are extremely narrow.***

  - ***I would think that a minimum of 4 categories should be used. Otherwise the extremely important predictor of age is inadequately adjusted.***
  - ***I divided age into 5 year intervals up to age 85, and then combined all patients age 85 and above into a single group. Other choices are acceptable, but you should be able to reproduce the analysis from their description.***

- ***In deciding upon an exact analysis method, the student would have had to decide how to weight the individual strata. They must make clear the weighting used, and ideally they provide a scientific interpretation of the impact of their weighting as described below. Points to consider include:***
  - ***The strata can be weighted according to the sex-age distribution among the ASCVD subjects. This has the advantage of being interpretable as the change in the death rates that might be expected if ASCVD were truly causal, if there were no unmeasured confounding, and the ASCVD in these patients could somehow be eliminated. (These weights correspond to the "istandard", but that term is jargon specific to Stata and should be avoided—not everyone uses Stata.)***
  - ***The strata can be weighted according to the sex-age distribution among the non-ASCVD subjects. This is interpretable as the change in the death rates that might be expected if ASCVD were truly causal, if there were no unmeasured confounding, and the non-ASCVD patients all somehow developed ASCVD. (These weights correspond to the "estandard", but that term is jargon specific to Stata and should be avoided—not everyone uses Stata.)***
  - ***The strata can be weighted according to the sex-age distribution among the combined ASCVD and non-ASCVD populations. Motivation for this weighting is generally based on the idea that you have to choose something, and this is something. However, it is true that this is an average RD across individuals in the sample. That is, in randomly choosing a person from the sample, we can say that this is the average RD without respect to age and sex.***

- ***The answer for the results should include some unadjusted estimates of death rates in the ASCVD and non-ASCVD patients. It is pretty easy to also give the rates within sexes and some estimate of the trend across age. Then the age-sex adjusted point estimates of the association (RD), a confidence interval for that measure of association, and a p value should be included. It would be extremely good to provide an interpretation of the scientific importance of the adjusted estimates. NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).***

- **Ans:** ***Methods:*** *I provide description of methods based on my personal preference of standardizing to the sex-age distribution of the ASCVD group. See above for possible interpretations based on other interpretations.* **Subjects with and without prior history of clinical atherosclerotic cardiovascular disease (ASCVD: angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke) at the time of study enrollment were compared with respect to differences in the incidence of death within 5 years of study enrollment. No subjects' observations were censored prior to five years, so five year mortality rates within each ASCVD diagnostic group, both overall and within strata defined by age and sex, were estimated using sample proportions, with differences in mortality rates compared using 95% confidence intervals computed using the asymptotic normal distribution for Wald-type maximum likelihood methods. The two-sided p value testing the null hypothesis of no difference in the probability of 5 year survival across ASCVD diagnostic groups was based**

on the chi squared statistic. Age and sex adjusted differences between the ASCVD and non-ASCVD groups in the probabilities of 5 year mortality were compared using direct standardization of risk differences computed within strata defined by the combination of sex and age categorized within five age groups (65-69, 70-74, 75-79, 80-84, and 85-99), with standard errors of the standardized risk difference derived from the stratum specific standard errors. The weighting scheme was based on the age and sex distribution of the ASCVD subjects in the sample. Hence the resulting difference in probabilities is interpretable as the absolute percentage of mortality in the ASCVD subjects that is attributable to ASCVD (in the absence of unmeasured confounding). Inference (95% CI and two-sided p values) for the age and sex adjusted risk difference used the normal distribution derived from maximum likelihood theory.

Results: *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).* **Mortality within 5 years of study entry was recorded for 735 generally healthy subjects (369 females, 366 males) who were 65-99 years of age when recruited for the Cardiovascular Health Study from Medicare rolls. Table 1 contains the proportion of subjects observed to die within categories defined by age, sex, and prior history of clinical atherosclerotic cardiovascular disease (ASCVD defined as prior diagnosis of angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke). Overall, 68 of 217 patients with ASCVD (31.3%) were observed to die within 5 years of study enrolment, while 53 of 518 patients without an initial diagnosis of ASCVD (10.2%) died within 5 years. This absolute risk difference of 21.2% is highly statistically significant (95% CI 14.4% to 27.8%, two-sided P < 0.0001). The 217 patients with prior ASCVD were more likely to be male (130 / 217 = 59.9%) and tended to be very slightly older (mean 75.5 y; sd 5.50 y) than the patients without prior history of ASCVD (45.6% male, age mean 74.2 y; sd 5.39 y). Overall 43 of the 369 females (11.7%) and 78 of the 366 males (21.3%) were observed to die. For both sexes, the 5 year mortality probability was higher in the ASCVD patients (23.0% for females, 36.9% for males) than in the non ASCVD patients (8.2% for females, 12.7% for males), with mortality probabilities tending to be higher for older age groups in both sex groups and both diagnostic groups, as expected.**

Table 1: Sample size and number of patients dying within five (5) years from study entry according to sex, age, and prior history of clinical atherosclerotic cardiovascular disease (ASCVD – angina, myocardial infarction, transient ischemic attack, or stroke).

	Females		Males		Both Sexes	
Age (y)	**No ASCVD**	**ASCVD**	**No ASCVD**	**ASCVD**	**No ASCVD**	**ASCVD**
65-69	1 / 52 (1.9%)	0 / 9 (0.0%)	5 / 38 (13.2%)	7 / 18 (38.9%)	6 / 90 (6.7%)	7 / 27 (25.9%)
70-74	6 / 126 (4.8%)	7 / 28 (25.0%)	12 / 105 (11.4%)	17 / 46 (37.0%)	18 / 231 (7.8%)	24 / 74 (32.4%)
75-79	9 / 62 (14.5%)	6 / 31 (19.4%)	5 / 56 (8.9%)	11 / 38 (28.9%)	14 / 118 (11.9%)	17 / 69 (24.6%)
80-84	4 / 29 (13.8%)	4 / 14 (28.6%)	4 / 22 (18.2%)	7 / 16 (43.8%)	8 / 51 (15.7%)	11 / 30 (36.7%)
85-99	3 / 13 (23.1%)	3 / 5 (60.0%)	4 / 15 (26.7%)	6 / 12 (50.0%)	7 / 28 (25.0%)	9 / 17 (52.9%)
All Ages	23 / 282 (8.2%)	20 / 87 (23.0%)	30 / 236 (12.7%)	48 / 130 (36.9%)	53 / 518 (10.2%)	68 / 217 (31.3%)

After direct standardization of the mortality probabilities for each diagnostic group to the age and sex distribution of the ASCVD patients, ASCVD was associated with an absolute difference in five year mortality probabilities of 19.3%, suggesting that in the absence of unmeasured confounding, their mortality probability would be 19.3% lower if the ASCVD subjects had not had prior disease. Based on the 95% confidence interval, such an observation is consistent with the possibilities that the true difference in 5 year mortality associated with ASCVD were anywhere between a 12.5% to 26.0% absolute higher probability of death than would be expected for their age and sex distribution. Such an

observation was highly statistically significant (two sided P < 0.0001). *Estimates for the other weighting schemes I explored are:*

- *Standardization to the age-sex distribution of the non ASCVD sample: estimated RD 18.1%, 95% CI 11.5% to 24.6%, P < 0.0001.*
- *Standardization to the age-sex distribution of the combined samples: estimated RD 18.4%, 95% CI 11.9% to 24.9%, P < 0.0001.*

b. Answer the question using an appropriate regression model.

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading:**

- ***Description of the method of analysis should make clear the method of regression analysis, the variables included in the regression model, the form of those variables, and the use of asymptotic normal theory as the basis for CI and p values.***
- ***In deciding upon an exact analysis method, the student would have had to decide which regression model to use. There are basically two models that could have been considered:***
  - ***Linear regression, which is equivalent to a generalized linear model with the Gaussian family and identity link (this is much to be preferred by me for reasons I discuss in lecture), or***
  - ***A generalized linear model with the Binomial family and identity link (I do not recommend the use of this model, owing to the possibility that model misspecification might lead to an unintentional change to the definition of the inferential contrast).***
- ***In deciding upon an exact analysis method, the student would have had to decide which standard errors to use. There are basically two models that could have been considered:***
  - ***Classical methods based on assumptions of homoscedasticity in linear regression (this is not the best choice when using linear regression with a binary response, because the mean-variance relationship will lead to heteroscedasticity; while the null hypothesis might dictate homoscedasticity across ASCVD diagnostic groups, it will not protect against sex and age effects on mortality) or the model based mean-variance relationship when using the binomial family with the generalized linear model (but, again, I do not recommend the use of the binomial family with an identity link), or***
  - ***"Robust" standard errors derived from the Huber-White sandwich estimator (this is what I recommend when using linear regression with a binary endpoint).***
- ***In deciding upon an exact analysis method, the student would have had to decide how to model age. We will be discussing all of these issues in great detail as the quarter unfolds, so the following discussion is more in anticipation of that. However, for the purposes of this homework, they must make clear the exact form used, which might be an untransformed linear term, a univariate transformation (e.g., the log), a categorized variable fit as dummy variables, a polynomial (e.g., the linear term and a squared term to fit a U-shaped relationship), linear splines, among many others.*** *Points to consider include:*
  - *First and foremost, there is no requirement in any regression model anywhere that the covariates have some particular distribution (though we do worry about influential points sometimes—see below). Hence, any justification of transforming a variable so its distribution would look more "normal" is just plain wrong.*

- *Second, we want to avoid data-driven choices of the modeling. We should not fit several models and then choose what seems to fit the data best. Instead, if we are unsure of the relationship between a covariate and the response, we should recognize that*
  - *If we are only modeling the covariate to gain precision of inference about the predictor of interest (POI), it does not really matter too much that we model the "true" data generating relationship. We tend to gain the bulk of the precision from any attempt to include the covariate.*
  - *If we are trying to adjust for a confounder, we may have a greater interest in modeling the "true" relationship. But we should still not fit lots of models. Instead, fitting a fairly flexible model (polynomials or linear splines) should be adequate, though we can lose power if we are unnecessarily too flexible.*
  - *Sometimes our specific question about the relationship between the response and the POI (e.g., "dose-response") dictates a particular model. For instance, we might be interested in whether there was evidence against a linear trend, or whether an association "tailed off" at the highest levels, or whether there was a "U-shaped" relationship. In those cases, the form of modeling the covariate should definitely be pre-specified.*
- *When you are unsure of the "true" (data generating) relationship between the response and a continuous variable, I believe just including the continuous variable as an untransformed linear term is most often best. This is the approach I preferred for age.*
- *In some cases we have good science to suggest that a logarithmic transformation is appropriate. These are cases in which it is a multiplicative difference in the covariate (e.g., every doubling of serum creatinine) that confers the same degree of association. In such settings, it is often the case that the covariate has a somewhat skewed distribution, but it is not generally the case that skewness alone should be the justification for a log transformation. I do note, however, that we do sometimes want to downweight the influence that large observations might have on the modeling of a covariate, in which case we might consider a log transformation. In the case of modeling age, there is little justification for a log transformation.*
- *Categorization of a continuous variable for the purposes of dummy variable adjustment to gain precision or remove confounding is rarely justifiable. (And categorization of the data to then fit it as "grouped continuous" data seems just plain silly to me, though I do recognize we sometimes collect continuous data as categorized data such as was done for the cancer surveillance data of problem 6). Unless an extremely large number of categories are considered, it is unlikely that the resulting fit to a "step function" is better than a fit to the best linear trend.*
- *If you are going to categorize, the points made about categorization for problem 2a all hold here: Categorize on a scientific, rather than a quantile basis, and have multiple such categories.*
- *More flexible approaches based on polynomials (age and age squared would probably have been adequate here) or linear splines (with, say 4 or 5 "knots") would cover most relationships that we are likely to see. (I did not anticipate that any student would choose one of these method. sIn later lectures I will come down on the side of a quadratic fit or linear splines when flexibility is needed. I think that "fractional polynomials" are bad ideas, so I will justify that opinion later, as well.)*

- ***In deciding upon an exact analysis method, the student would have had to decide <u>whether to model any interactions</u>.***
    - ***A sex-age interaction is certainly an option, and inclusion of a multiplicative interaction term would be most common. I anticipate that a few students might have included such an interaction, and I could argue that such would be a smart thing to do when you had sufficient data: it might better adjust for confounding relationships that we want to remove. If an interaction is included, it must be explicitly stated in the methods. (In my answer to the question, I fit such an interaction. However, in an uncharacteristic streak of honesty, I will admit that the model I would typically fit for "sex and age adjustment" would only include the "main effects".)***
    - *A three-way interaction between ASCVD, sex, and age would be necessary if you wanted to mimic the stratified analysis that averages across possible effect modification in a prescribed fashion. Such an analysis would then entail a "multiple partial" test to assess a mortality-ASCVD association, and in order to quantify the association we would have to compute an adjusted RD user a pre-specified linear combination of the regression parameters. I certainly did not anticipate any student doing this, but I do illustrate the way that such an approach can reproduce the estimates (but not in general the SE) from a stratified analysis performed in problem 2a.*
- ***In deciding upon an exact analysis method, the student would have had to decide <u>how to weight the individual observations</u>. The most common approach would be to do an unweighted analysis, and this is likely what everyone did without even thinking about it. And in that case, it is standard that no mention need be made of weighting in the description of the methods. Instead, if a weighted regression is performed, then that would have to be described (and justified). They must make clear the weighting used, and ideally they provide a scientific interpretation of the impact of their weighting as described below.***
    - *Weighted regressions can be used to adjust for unusual sampling plans or to emphasize how the regression results might apply to other populations when unmodeled nonlinear effects or effect modification might be an issue. In those settings, the issues to consider when choosing weights are very similar to those that were discussed in problem 2a..*
- ***The answer for the results should include some unadjusted estimates of death rates in the ASCVD and non-ASCVD patients. It is pretty easy to also give the rates within sexes and some estimate of the trend across age. Then the age-sex adjusted point estimates of the association (RD), a confidence interval for that measure of association, and a p value should be included. It would be extremely good to provide an interpretation of the scientific importance of the adjusted estimates.*** *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).*

Ans: ***<u>Methods:</u>*** *I provide description of methods based on my personal preference of a linear regression model. Use of a GLM with binomial family is acceptable, as are alternative ways of modeling age and alternative ways of considering interactions. However, the student <u>must</u> describe the models they used in a way that would allow you to reproduce the analysis.(In future homeworks we will be scrutinizing the choice of regression model much more closely.).* **Subjects with and without prior history of clinical atherosclerotic cardiovascular disease (ASCVD: angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke) at the time of study enrollment were compared with respect to differences in the incidence of death within 5 years of study enrollment. No subjects' observations were censored prior to five years, so five year mortality rates within each ASCVD diagnostic group were estimated using sample proportions, with differences in mortality rates compared using 95% confidence intervals**

computed using Wald-type maximum likelihood methods. Two-sided p values testing the null hypothesis of no difference in the probability of 5 year survival across ASCVD diagnostic groups were based on the chi squared statistic. For purposes of descriptive analyses, risk differences computed within strata defined by the combination of sex and age categorized within five age groups (65-69, 70-74, 75-79, 80-84, and 85-99). Age and sex adjusted differences between the ASCVD and non-ASCVD groups in the probabilities of 5 year mortality were compared using a linear regression model of the binary indicator of 5 year mortality regressed on an indicator of ASCVD, an indicator of male sex, and a variable measuring age at study enrollment that was modeled as linear continuous variable. Also included was a multiplicative interaction term for sex and age. Estimation of the age and sex adjusted difference in 5 year mortality was based on the regression coefficient for the ASCVD variable, with standard errors computed using the Huber-White sandwich estimator in order to account for the nonconstant variances inherent in binary data. Two-sided p values and 95% confidence intervals were computed assuming the asymptotic normal distribution for the linear regression parameter estimates..

Results: *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).***Mortality within 5 years of study entry was recorded for 735 generally healthy subjects (369 females, 366 males) who were 65-99 years of age when recruited for the Cardiovascular Health Study from Medicare rolls. Table 1 contains the proportion of subjects observed to die within categories defined by age, sex, and prior history of clinical atherosclerotic cardiovascular disease (ASCVD defined as prior diagnosis of angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke). Overall, 68 of 217 patients with ASCVD (31.3%) were observed to die within 5 years of study enrolment, while 53 of 518 patients without an initial diagnosis of ASCVD (10.2%) died within 5 years. This absolute risk difference of 21.2% is highly statistically significant (95% CI 14.4% to 27.8%, two-sided P < 0.0001). The 217 patients with prior ASCVD were more likely to be male (130 / 217 = 59.9%) and tended to be very slightly older (mean 75.5 y; sd 5.50 y) than the patients without prior history of ASCVD (45.6% male, age mean 74.2 y; sd 5.39 y). Overall 43 of the 369 females (11.7%) and 78 of the 366 males (21.3%) were observed to die. For both sexes, the 5 year mortality probability was higher in the ASCVD patients (23.0% for females, 36.9% for males) than in the non ASCVD patients (8.2% for females, 12.7% for males), with mortality probabilities tending to be higher for older age groups in both sex groups and both diagnostic groups, as expected.**

After adjustment for age, sex, and the age-sex interaction in a linear regression model, we estimate that ASCVD was associated with an absolute difference in five year mortality probabilities of 18.7%,. This observed trend toward higher mortality in the ASCVD group was highly statically significant (two-sided P < 0.0002), and based on the 95% confidence interval, such an observation is consistent with the possibilities that the true difference in 5 year mortality associated with ASCVD were anywhere between a 12.0% to 25.4% absolute higher probability of death than for non-ASCVD subjects of the same age and sex distribution. *Estimates for a regression model that did not include the age-sex interaction were RD= 18.9%; 95% CI 12.2% to 25.7%; P < 0.0001.*

c. What is the difference in the statistical models you used? That is, how would you explain any differences between the two analysis approaches?

Instructions for grading: This part of the problem is worth 10 points. To get full credit, the student must address each of the relevant points given in the answer (the points can be weighted equally). Of course, their response to each of these points will differ depending upon the models they fit.

Ans: The methods differ in their handling of

- **The modeling of age: The stratified analysis had to consider age categories without borrowing information across those categories, while in the linear regression analysis I was able to model age continuously (thereby borrowing information about a trend across all ages).** *(If the student fit a categorized age with dummy variables in the linear regression model, there would be no need to mention this difference.)*
- **The modeling of an age-sex interaction: The stratified analysis had to adjust for the age-sex interaction (using categorized age). In the regression model, the age-sex interaction did not have to be included. Though I did include the age-sex interaction in my regression model, I modeled it continuously, rather than discretely.**
- **The modeling of an ASCVD-age-sex interaction: The stratified analysis allowed an estimate of the RD to be different for each stratum defined by age and sex. The regression analysis is modeling the RD as if it were constant across all age-sex combinations.**
- **The weighting of the individual combinations of age and sex: In the stratified analysis, a specific "importance" weighting based on the distribution of ASCVD patients across age-sex strata was chosen to "average" over possibly different RD in each stratum. In the linear regression model, if there is effect modification of the RD by age and/or sex, the resultant RD estimate will be some sort of weighted average where the weights are "efficiency weights" that would be optimal if there were no effect modification.**
- *In the Stata code, I do present an analysis based on fitting a linear regression model to the three way interaction of ASCVD, sex, and categorized age (a saturated model). I then use Stata's lincom command to compute the linear combination of regression parameter estimates that would correspond exactly to the standardized estimate that uses the ASCVD age-sex distribution. I note that the standard errors will not agree exactly, because the regression using the "robust" Huber-White sandwich estimates will not be the same as the stratum specific estimated standard errors. I urge you to look at this section (labeled "OF SPECIAL NOTE")*

3. Using the odds ratio (OR) as a measure of association, provide statistical inference regarding an association between 5 year survival and baseline prevalence of ASCVD, adjusting for age and sex.

    a. Answer the question using a stratified analysis (e.g., using Stata command `cs, cc, mhodds` or an equivalent analysis in R).

<u>Instructions for grading</u>: This part of the problem is worth 10 points. Key points to consider in your grading:

- ***Description of the method of analysis should make clear that an adjusted OR is estimated by first estimating RD within strata defined by combinations of age and sex, and then combining maximum likelihood estimates across strata in a weighted analysis, with standard errors of the adjusted RD computed by combining the estimated SE from each stratum according to the chosen weights. P values and CI are computed using the approximate normal distribution.***
- ***In deciding upon an exact analysis method, the student would have had to decide <u>how to categorize age</u>. See problem 2a for a discussion.***
- ***In deciding upon an exact analysis method, the student would have had to decide <u>how to weight the individual strata</u>. They must make clear the weighting used, and ideally they provide***

a scientific interpretation of the impact of their weighting as described below. Points to consider include:

- ***The strata can be weighted according to the Mante-Haenszel statistics. These are approximately "efficiency weights" when there is a common OR across strata and are the most common approach with OR. Empirical justification for being so quick to use the M-H statistic might be the observation that it is easier to have no effect modification on the OR scale than on the RD or RR scales. (In specifying the methods, it is sufficient to say that the Mantel-Haenszel statistic was used.)***
- ***The strata can be weighted according to the sex-age distribution among the ASCVD subjects who did not die. In epidemiologic parlance, these would correspond to the "exposed" in the "control" population. When investigating a rare disease in a case-control study, the presumption would be that the "control" population is more representative of the general population. (These weights correspond to the "istandard", but that term is jargon specific to Stata and should be avoided—not everyone uses Stata.)***
- ***The strata can be weighted according to the sex-age distribution among the non-ASCVD subjects who did not die. In epidemiologic parlance, these would correspond to the "unexposed" in the "control" population. When investigating a rare disease in a case-control study, the presumption would be that the "control" population is more representative of the general population. (These weights correspond to the "estandard", but that term is jargon specific to Stata and should be avoided—not everyone uses Stata.)***
- ***The strata can be weighted according to the sex-age distribution among the combined ASCVD and non-ASCVD populations.***

- ***The answer for the results should include some unadjusted estimates of death rates in the ASCVD and non-ASCVD patients. It is pretty easy to also give the rates within sexes and some estimate of the trend across age. Then the age-sex adjusted point estimates of the association (OR), a confidence interval for that measure of association, and a p value should be included. It would be extremely good to provide an interpretation of the scientific importance of the adjusted estimates.*** *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).*

- **Ans: *Methods:*** *I provide description of methods based on the Mantel-Haenszel weighting..*
  Subjects with and without prior history of clinical atherosclerotic cardiovascular disease (ASCVD: angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke) at the time of study enrollment were compared with respect to differences in the incidence of death within 5 years of study enrollment. No subjects' observations were censored prior to five years, so five year mortality rates within each ASCVD diagnostic group, both overall and within strata defined by age and sex, were estimated using sample proportions, with ratios of the odds of mortality compared using 95% confidence intervals computed using the asymptotic normal distribution for Wald-type maximum likelihood methods. The two-sided p value testing the null hypothesis of no difference in the probability of 5 year survival across ASCVD diagnostic groups was based on the chi squared statistic. Age and sex adjusted differences between the ASCVD and non-ASCVD groups in the odds of 5 year mortality were based on the Mantel-Haenszel statistic computed within strata defined by the combination of sex and age categorized within five age groups (65-69, 70-74, 75-79, 80-84, and 85-99), with standard errors of the standardized risk difference derived from the stratum specific standard errors. ~~The weighting scheme was based on the age and sex~~

~~distribution of the ASCVD subjects in the sample. Hence the resulting difference in probabilities is interpretable as the absolute percentage of mortality in the ASCVD subjects that is attributable to ASCVD (in the absence of unmeasured confounding). Inference (95% CI and two-sided p values) for the age and sex adjusted risk difference used the normal distribution derived from maximum likelihood theory~~**.**

Results: *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).***Mortality within 5 years of study entry was recorded for 735 generally healthy subjects (369 females, 366 males) who were 65-99 years of age when recruited for the Cardiovascular Health Study from Medicare rolls. Table 1** *(given in the answer to problem 2a)* **contains the proportion of subjects observed to die within categories defined by age, sex, and prior history of clinical atherosclerotic cardiovascular disease (ASCVD defined as prior diagnosis of angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke). Overall, 68 of 217 patients with ASCVD (31.3%) were observed to die within 5 years of study enrolment, while 53 of 518 patients without an initial diagnosis of ASCVD (10.2%) died within 5 years. This corresponds to an odds ratio of 4.00 and is highly statistically significant (95% CI 2.67 to 6.00, two-sided P < 0.0001). The 217 patients with prior ASCVD were more likely to be male (130 / 217 = 59.9%) and tended to be very slightly older (mean 75.5 y; sd 5.50 y) than the patients without prior history of ASCVD (45.6% male, age mean 74.2 y; sd 5.39 y). Overall 43 of the 369 females (11.7%) and 78 of the 366 males (21.3%) were observed to die. For both sexes, the 5 year mortality probability was higher in the ASCVD patients (23.0% for females, 36.9% for males) than in the non ASCVD patients (8.2% for females, 12.7% for males), with mortality probabilities tending to be higher for older age groups in both sex groups and both diagnostic groups, as expected.**

After adjustment for age and sex in a stratified Mantel-Haenszel analysis, ASCVD was associated with 3.50-fold higher odds of five year mortality~~, suggesting that in the absence of unmeasured confounding, their mortality probability would be 19.3% lower if the ASCVD subjects had not had prior disease~~**. Based on the 95% confidence interval, such an observation is consistent with the possibilities that the true odds of 5 year mortality associated with ASCVD were anywhere between a 2.35-fold to 5.29-fold higher odds of death than would be expected for their age and sex distribution. Such an observation was highly statistically significant (two sided P < 0.0001).** *(Note that the adjusted OR is closer to the null than is the unadjusted OR. This is not consistent with age and sex merely being precision variables. Instead, this is the behavior of the OR when the unadjusted analysis is confounded by the covariates in the adjusted model.)*

b. Answer the question using an appropriate regression model.

Instructions for grading: This part of the problem is worth 10 points. Key points to consider in your grading:

- ***Description of the method of analysis should make clear the method of regression analysis, the variables included in the regression model, the form of those variables, and the use of asymptotic normal theory as the basis for CI and p values.***
- ***In deciding upon an exact analysis method, the student would have had to decide which regression model to use. Logistic regression is the only logical choice here..***
- ***In deciding upon an exact analysis method, the student would have had to decide which standard errors to use. Generally there is not much advantage in using the "robust" standard errors with logistic regression.***

- ***In deciding upon an exact analysis method, the student would have had to decide how to model age. The issues with logistic regression are exactly the same as they are will other forms of regression.*** *(See the comments with the answers to problem 2b.)*
- ***In deciding upon an exact analysis method, the student would have had to decide whether to model any interactions. The issues with logistic regression are exactly the same as they are will other forms of regression.*** *(See the comments with the answers to problem 2b.)*
- ***In deciding upon an exact analysis method, the student would have had to decide how to weight the individual observations. The issues with logistic regression are exactly the same as they are will other forms of regression.*** *(See the comments with the answers to problem 2b.)*
- ***The answer for the results should include some unadjusted estimates of death rates in the ASCVD and non-ASCVD patients. It is pretty easy to also give the rates within sexes and some estimate of the trend across age. Then the age-sex adjusted point estimates of the association (OR), a confidence interval for that measure of association, and a p value should be included. It would be extremely good to provide an interpretation of the scientific importance of the adjusted estimates.*** *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).*

Ans: ***Methods:*** *I provide description of methods based on a logistic regression model modeling age, sex, and the age-sex interaction.. Alternative ways of modeling age and alternative ways of considering interactions are acceptable. However, the student must describe the models they used in a way that would allow you to reproduce the analysis.(In future homeworks we will be scrutinizing the choice of regression model much more closely.).* **Subjects with and without prior history of clinical atherosclerotic cardiovascular disease (ASCVD: angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke) at the time of study enrollment were compared with respect to differences in the incidence of death within 5 years of study enrollment. No subjects' observations were censored prior to five years, so five year mortality rates within each ASCVD diagnostic group were estimated using sample proportions, with ratios of odds of mortality compared using 95% confidence intervals computed using Wald-type maximum likelihood methods. Two-sided p values testing the null hypothesis of no difference in the odds of 5 year survival across ASCVD diagnostic groups were based on the chi squared statistic. For purposes of descriptive analyses, risk differences computed within strata defined by the combination of sex and age categorized within five age groups (65-69, 70-74, 75-79, 80-84, and 85-99). Age and sex adjusted comparisons between the ASCVD and non-ASCVD groups in the odds of 5 year mortality were compared using a logistic regression model of the binary indicator of 5 year mortality regressed on an indicator of ASCVD, an indicator of male sex, and a variable measuring age at study enrollment that was modeled as linear continuous variable. Also included was a multiplicative interaction term for sex and age. Estimation of the age and sex adjusted odds ratios for 5 year mortality was based on the regression coefficient for the ASCVD variable~~, with standard errors computed using the Huber-White sandwich estimator in order to account for the nonconstant variances inherent in binary data~~. Two-sided p values and 95% confidence intervals were computed assuming the asymptotic normal distribution for the logistic regression parameter estimates.**.

Results: *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).***Mortality within 5 years of study entry was recorded for 735 generally healthy subjects (369 females, 366 males) who were 65-99 years of age when recruited for the Cardiovascular Health Study from Medicare rolls. Table 1 contains the proportion of subjects observed to die within categories defined by age, sex, and prior history**

of clinical atherosclerotic cardiovascular disease (ASCVD defined as prior diagnosis of angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke). Overall, 68 of 217 patients with ASCVD (31.3%) were observed to die within 5 years of study enrolment, while 53 of 518 patients without an initial diagnosis of ASCVD (10.2%) died within 5 years. This corresponds to an odds ratio of 4.00 and is highly statistically significant (95% CI 2.67 to 6.00, two-sided P < 0.0001)). The 217 patients with prior ASCVD were more likely to be male (130 / 217 = 59.9%) and tended to be very slightly older (mean 75.5 y; sd 5.50 y) than the patients without prior history of ASCVD (45.6% male, age mean 74.2 y; sd 5.39 y). Overall 43 of the 369 females (11.7%) and 78 of the 366 males (21.3%) were observed to die. For both sexes, the 5 year mortality probability was higher in the ASCVD patients (23.0% for females, 36.9% for males) than in the non ASCVD patients (8.2% for females, 12.7% for males), with mortality probabilities tending to be higher for older age groups in both sex groups and both diagnostic groups, as expected.

After adjustment for age, sex, and the age-sex interaction in a logistic regression model, we estimate that ASCVD was associated with a 3.50-fold higher odds of five year mortality. This observed trend toward higher mortality in the ASCVD group was highly statically significant (two-sided P < 0.0001), and based on the 95% confidence interval, such an observation is consistent with the possibilities that the true odds of 5 year mortality associated with ASCVD were anywhere between a 2.32-fold to 5.29-fold absolute higher odds of death than for non-ASCVD subjects of the same age and sex distribution..

c. What is the difference in the statistical models you used? That is, how would you explain any differences between the two analysis approaches?

***Instructions for grading*: This part of the problem is worth 10 points. To get full credit, the student must address each of the relevant points given in the answer (the points can be weighted equally). Of course, their response to each of these points will differ depending upon the models they fit.**

Ans: The methods differ in their handling of

- **The modeling of age: The stratified analysis had to consider age categories without borrowing information across those categories, while in the logistic regression analysis I was able to model age continuously (thereby borrowing information about a trend across all ages).** *(If the student fit a categorized age with dummy variables in the logistic regression model, there would be no need to mention this difference.)*
- **The modeling of an age-sex interaction: The stratified analysis had to adjust for the age-sex interaction (using categorized age). In the regression model, the age-sex interaction did not have to be included. Though I did include the age-sex interaction in my regression model, I modeled it continuously, rather than discretely.**
- **The modeling of an ASCVD-age-sex interaction: The stratified analysis allowed an estimate of the OR to be different for each stratum defined by age and sex. The regression analysis is modeling the OR as if it were constant across all age-sex combinations.**
- **The weighting of the individual combinations of age and sex: In the stratified analysis, a specific "quasi-efficiency" weighting based on the Mantel-Haenszel statistic across age-sex strata was chosen to "average" over possibly different OR in each stratum. In the logistic regression model, if there is effect modification of the OR by age and/or sex, the resultant OR estimate will be some sort of weighted geometric mean where the weights are "efficiency weights" that would be optimal if there were no effect modification.**

4. Using the risk ratio (RR) as a measure of association, provide statistical inference regarding an association between 5 year survival and baseline prevalence of ASCVD, adjusting for age and sex.

    a. Answer the question using a stratified analysis (e.g., using Stata command `cs, ir` or an equivalent analysis in R).

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading:**

- ***Description of the method of analysis should make clear that an adjusted RR is estimated by first estimating risk within strata defined by combinations of age and sex, and then combining maximum likelihood estimates for the stratum specific risk across strata in a weighted analysis, with the RR defined as the ratio of the weighted risk for the ASCVD group divided by the weighted risk for the non-ASCVD group. Standard errors of the adjusted RR computed by combining the estimated SE from each stratum according to the chosen weights. P values and CI are computed using the approximate normal distribution.***
- ***In deciding upon an exact analysis method, the student would have had to decide how to categorize age. See problem 2a for a discussion.***
- ***In deciding upon an exact analysis method, the student would have had to decide how to weight the individual strata. They must make clear the weighting used, and ideally they provide a scientific interpretation of the impact of their weighting as described below. Points to consider are similar to aspects of the RD analysis and the OR analysis:***
    - ***The strata can be weighted according to the Mantel-Haenszel statistics. These are approximately "efficiency weights" when there is a common RR across strata. In justifying use of the M-H weights with rare diseases, we can invoke the close correspondence between OR and RR in that setting. (In specifying the methods, it is sufficient to say that the Mantel-Haenszel statistic was used.)***
    - ***The strata can be weighted according to the sex-age distribution among the ASCVD subjects (note that this is like RD rather than OR). (These weights correspond to the "istandard", but that term is jargon specific to Stata and should be avoided—not everyone uses Stata.)***
    - ***The strata can be weighted according to the sex-age distribution among the non-ASCVD subjects (note that this is like RD rather than OR) (These weights correspond to the "estandard", but that term is jargon specific to Stata and should be avoided—not everyone uses Stata.)***
    - ***The strata can be weighted according to the sex-age distribution among the combined ASCVD and non-ASCVD populations.***
- ***The answer for the results should include some unadjusted estimates of death rates in the ASCVD and non-ASCVD patients. It is pretty easy to also give the rates within sexes and some estimate of the trend across age. Then the age-sex adjusted point estimates of the association (RR), a confidence interval for that measure of association, and a p value should be included. It would be extremely good to provide an interpretation of the scientific importance of the adjusted estimates.*** *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).*
- **Ans:** ***Methods:*** *I provide description of methods based on the Mantel-Haenszel weighting..*
  Subjects with and without prior history of clinical atherosclerotic cardiovascular disease

(ASCVD: angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke) at the time of study enrollment were compared with respect to differences in the incidence of death within 5 years of study enrollment. No subjects' observations were censored prior to five years, so five year mortality rates within each ASCVD diagnostic group, both overall and within strata defined by age and sex, were estimated using sample proportions, with risk ratio compared using 95% confidence intervals computed using the asymptotic normal distribution for Wald-type maximum likelihood methods. The two-sided p value testing the null hypothesis of no difference in the probability of 5 year survival across ASCVD diagnostic groups was based on the chi squared statistic. Age and sex adjusted comparisons between the ASCVD and non-ASCVD groups in the risk of 5 year mortality were based on a Mantel-Haenszel weighting of the risk computed within strata defined by the combination of sex and age categorized within five age groups (65-69, 70-74, 75-79, 80-84, and 85-99), with the risk ratio defined as the ratio of the weighted average of the stratum specific risks for each diagnostic group (ASCVD divided by non ASCVD). Standard errors of the standardized risk difference derived from the stratum specific standard errors. ~~The weighting scheme was based on the age and sex distribution of the ASCVD subjects in the sample. Hence the resulting difference in probabilities is interpretable as the absolute percentage of mortality in the ASCVD subjects that is attributable to ASCVD (in the absence of unmeasured confounding). Inference (95% CI and two-sided p values) for the age and sex adjusted risk difference used the normal distribution derived from maximum likelihood theory~~**.**

Results: *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).* **Mortality within 5 years of study entry was recorded for 735 generally healthy subjects (369 females, 366 males) who were 65-99 years of age when recruited for the Cardiovascular Health Study from Medicare rolls. Table 1** *(given in the answer to problem 2a)* **contains the proportion of subjects observed to die within categories defined by age, sex, and prior history of clinical atherosclerotic cardiovascular disease (ASCVD defined as prior diagnosis of angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke). Overall, 68 of 217 patients with ASCVD (31.3%) were observed to die within 5 years of study enrolment, while 53 of 518 patients without an initial diagnosis of ASCVD (10.2%) died within 5 years. This corresponds to an risk ratio of 3.06 and is highly statistically significant (95% CI 2.22 to 4.23, two-sided P < 0.0001). The 217 patients with prior ASCVD were more likely to be male (130 / 217 = 59.9%) and tended to be very slightly older (mean 75.5 y; sd 5.50 y) than the patients without prior history of ASCVD (45.6% male, age mean 74.2 y; sd 5.39 y). Overall 43 of the 369 females (11.7%) and 78 of the 366 males (21.3%) were observed to die. For both sexes, the 5 year mortality probability was higher in the ASCVD patients (23.0% for females, 36.9% for males) than in the non ASCVD patients (8.2% for females, 12.7% for males), with mortality probabilities tending to be higher for older age groups in both sex groups and both diagnostic groups, as expected.**

After adjustment for age and sex in a stratified analysis using Mantel-Haenszel weights, ASCVD was associated with 2.63-fold higher probability of five year mortality~~, suggesting that in the absence of unmeasured confounding, their mortality probability would be 19.3% lower if the ASCVD subjects had not had prior disease~~**. Based on the 95% confidence interval, such an observation is consistent with the possibilities that the true probability of 5 year mortality associated with ASCVD were anywhere between a 1.92-fold to 3.62-fold higher probability of death than would be expected for their age and sex distribution. Such an observation was highly statistically significant (two sided P < 0.0001).**

b. Answer the question using an appropriate regression model.

***Instructions for grading*: This part of the problem is worth 10 points. Key points to consider in your grading:**

- ***Description of the method of analysis should make clear the method of regression analysis, the variables included in the regression model, the form of those variables, and the use of asymptotic normal theory as the basis for CI and p values.***
- ***In deciding upon an exact analysis method, the student would have had to decide which regression model to use. We could consider the GLM with Binomial family and a log link, but I recommend Poisson regression instead for reasons similar to those invoked for RD.***
- ***In deciding upon an exact analysis method, the student would have had to decide which standard errors to use. With Poisson regression, the "robust" standard errors are to be preferred, but if the probabilities are low, this will not make much difference.***
- ***In deciding upon an exact analysis method, the student would have had to decide how to model age. The issues with Poisson regression are exactly the same as they are will other forms of regression.*** *(See the comments with the answers to problem 2b.)*
- ***In deciding upon an exact analysis method, the student would have had to decide whether to model any interactions. The issues with Poisson regression are exactly the same as they are will other forms of regression.*** *(See the comments with the answers to problem 2b.)*
- ***In deciding upon an exact analysis method, the student would have had to decide how to weight the individual observations. The issues with Poisson regression are exactly the same as they are will other forms of regression.*** *(See the comments with the answers to problem 2b.)*
- ***The answer for the results should include some unadjusted estimates of death rates in the ASCVD and non-ASCVD patients. It is pretty easy to also give the rates within sexes and some estimate of the trend across age. Then the age-sex adjusted point estimates of the association (OR), a confidence interval for that measure of association, and a p value should be included. It would be extremely good to provide an interpretation of the scientific importance of the adjusted estimates.*** *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).*

Ans: ***Methods:*** *I provide description of methods based on a Poisson regression model modeling age, sex, and the age-sex interaction.. Alternative ways of modeling age and alternative ways of considering interactions are acceptable. However, the student must describe the models they used in a way that would allow you to reproduce the analysis.(In future homeworks we will be scrutinizing the choice of regression model much more closely.).* **Subjects with and without prior history of clinical atherosclerotic cardiovascular disease (ASCVD: angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke) at the time of study enrollment were compared with respect to differences in the incidence of death within 5 years of study enrollment. No subjects' observations were censored prior to five years, so five year mortality rates within each ASCVD diagnostic group were estimated using sample proportions, with risk ratio compared using 95% confidence intervals computed using Wald-type maximum likelihood methods. Two-sided p values testing the null hypothesis of no difference in the probability of 5 year survival across ASCVD diagnostic groups were based on the chi squared statistic. For purposes of descriptive analyses, risk differences computed within strata defined by the combination of sex and age categorized within five age groups (65-69, 70-74, 75-79, 80-84, and 85-99). Age and sex adjusted comparisons between the ASCVD and non-ASCVD groups in the probability of 5 year mortality were compared using a Poisson regression model of the binary indicator of 5 year mortality regressed on an indicator of**

ASCVD, an indicator of male sex, and a variable measuring age at study enrollment that was modeled as linear continuous variable. Also included was a multiplicative interaction term for sex and age. Estimation of the age and sex adjusted risk ratios for 5 year mortality was based on the regression coefficient for the ASCVD variable with standard errors computed using the Huber-White sandwich estimator in order to account for the nonconstant variances inherent in binary data. Two-sided p values and 95% confidence intervals were computed assuming the asymptotic normal distribution for the Poisson regression parameter estimates..

Results: *NOTE: The student's answers do NOT need to be as detailed as I provide. I just included detail to show how the descriptive statistics would not change very substantially for the different measures of association (RD, RR, or OR).***Mortality within 5 years of study entry was recorded for 735 generally healthy subjects (369 females, 366 males) who were 65-99 years of age when recruited for the Cardiovascular Health Study from Medicare rolls. Table 1 contains the proportion of subjects observed to die within categories defined by age, sex, and prior history of clinical atherosclerotic cardiovascular disease (ASCVD defined as prior diagnosis of angina, myocardial infarction (MI), transient ischemic attack (TIA), or stroke). Overall, 68 of 217 patients with ASCVD (31.3%) were observed to die within 5 years of study enrolment, while 53 of 518 patients without an initial diagnosis of ASCVD (10.2%) died within 5 years. This corresponds to a risk ratio of 3.06 and is highly statistically significant (95% CI 2.22 to 4.23, two-sided P < 0.0001)). The 217 patients with prior ASCVD were more likely to be male (130 / 217 = 59.9%) and tended to be very slightly older (mean 75.5 y; sd 5.50 y) than the patients without prior history of ASCVD (45.6% male, age mean 74.2 y; sd 5.39 y). Overall 43 of the 369 females (11.7%) and 78 of the 366 males (21.3%) were observed to die. For both sexes, the 5 year mortality probability was higher in the ASCVD patients (23.0% for females, 36.9% for males) than in the non ASCVD patients (8.2% for females, 12.7% for males), with mortality probabilities tending to be higher for older age groups in both sex groups and both diagnostic groups, as expected.**

After adjustment for age, sex, and the age-sex interaction in a Poisson regression model, we estimate that ASCVD was associated with a 2.65-fold higher odds of five year mortality. This observed trend toward higher mortality in the ASCVD group was highly statically significant (two-sided P < 0.0001), and based on the 95% confidence interval, such an observation is consistent with the possibilities that the true probability of 5 year mortality associated with ASCVD were anywhere between a 1.93-fold to 3.67-fold absolute higher probability of death than for non-ASCVD subjects of the same age and sex distribution..

c. What is the difference in the statistical models you used? That is, how would you explain any differences between the two analysis approaches?

Instructions for grading: This part of the problem is worth 10 points. To get full credit, the student must address each of the relevant points given in the answer (the points can be weighted equally). Of course, their response to each of these points will differ depending upon the models they fit.

Ans: The methods differ in their handling of

- **The modeling of age: The stratified analysis had to consider age categories without borrowing information across those categories, while in the Poisson regression analysis I was able to model age continuously (thereby borrowing information about a trend across all ages).** *(If the student fit a categorized age with dummy variables in the logistic regression model, there would be no need to mention this difference.)*
- **The modeling of an age-sex interaction: The stratified analysis had to adjust for the age-sex interaction (using categorized age). In the regression model, the age-sex interaction**

did not have to be included. Though I did include the age-sex interaction in my regression model, I modeled it continuously, rather than discretely.

- **The modeling of an ASCVD-age-sex interaction: The stratified analysis allowed an estimate of the RR to be different for each stratum defined by age and sex. The regression analysis is modeling the RR as if it were constant across all age-sex combinations.**
- **The weighting of the individual combinations of age and sex: In the stratified analysis, a specific "quasi-efficiency" weighting based on the Mantel-Haenszel statistic across age-sex strata was chosen to "average" over possibly different risks in each stratum. The RR was then estimated as a ratio of averages. In the Poisson regression model, if there is effect modification of the RR by age and/or sex, the resultant RR estimate will be some sort of weighted geometric mean of stratum specific RR where the weights are "efficiency weights" that would be optimal if there were no effect modification.**

5. Comment very briefly on the similarity or differences among the three approaches. Which would you tend to prefer in general? Why?

***Instructions for grading*: This problem is worth 10 points. The grader may use his/her judgement.**

Ans: The primary differences among the analyses all relate to the relative advantages and disadvantages of the RD vs RR vs OR, and those advantages and disadvantages would be the major factor in choosing among them:

- **RD is easiest to interpret, RR accentuates effects with low probabilities, OR has some invariance properties that make it nice, including invariance to case-control or cohort sampling. The OR might tend to have less effect modification.**
- **If special weighting is needed, the stratified analyses are more straightforward. (Obtaining the equivalent inference about RR from a regression analysis is quite involved.)**
- **Regression does not require categorization of continuous variables, so may provide greater precision.**
- **Correspondences between stratified RD and linear regression are quite straightforward. The correspondence between RR and Poisson regression is much weaker.**
- **Logistic regression is used quite widely in the setting of binary response, and thus readers will perhaps have greater familiarity**

Question 6 pertains to the analysis of colorectal cancer incidence for whites living in the U.S. as a function of birthplace (U.S. born vs foreign born) (see datafile `surveillance.txt` and documentation `surveillance.doc` on the class web pages).

6. Using the incidence ratio as a measure of association, provide inference for an association between incidence of colorectal cancer and birthplace, after adjustment for age, sex, and SEER.
    a. Answer the question using directly standardized rates, with standardization to the U.S. population.

***Instructions for grading**: This part of the problem is worth 10 points. The criteria for grading this problem should be very similar to the grading of problem 4a above. Additional key points to consider in your grading:*

- ***In describing the methods, we need to talk about the computation of the person-years of observation..***
- ***It would have been reasonable to alter the categorization of ages or to restrict to adults. If this is done, it should be made clear.***
- ***One could either provide a single analysis adjusting for sex and age, or one could produce analyses for each sex separately. The latter would be more the norm with cancer incidence statisics.***

Ans: ***Methods:*** *I provide description of methods based on the weighting by the distribution of SEER, age, and sex in the census data used to construct the denominators..* **Colorectal cancer incidence rates were compared between subjects known to be born in the US and those known to be foreign born using directly standardized rates. Numerator data comes from the SEER registries aggregated by 5 year age intervals and sex for each SEER site. Time period of surveillance was between 10 and 12 years depending on the SEER site. Denominator data was estimated from a sample of 1980 census data, and multiplied by the years of surveillance for each site. Incidence rates were standardized to the age, SEER, and sex distribution of the census data combined across birthplace groups. Incidence rates in each stratum were estimated by maximum likelihood estimates based on a Poisson distribution, with inference about the ratio of weighted incidence rates derived from using the asymptotic normal distribution for the stratum specific weight estimates. Analyses were performed for each sex separately.**

Results: **During the period of surveillance, a total of 62,668 incident cases of colorectal cancer in subjects known to be US born and 11,026 incident cases in subjects known to be foreign born were reported by the participating SEER sites. Over that period of time, the estimated time of observation was over 192 million person years and 12 million person years for US born and foreign born, respectively. Crude cumulative incidence rates were thus 32.6 and 89.3 cases per 100,000 person years across all SEERs, all ages, and both sexes. The unadjusted incidence of colorectal cancer (i.e., across all SEERs and all ages) for foreign born females was estimated to be 2.42 times that of US born females (95% CI 2.35 to 2.49, P < 0.0001). For males, the unadjusted incidence of colorectal cancer for foreign born males was 3.14 times higher than that for US born males (95% CI 3.05 to 3..23, P < 0.0001).**

After standardization to the sex, age and SEER distribution of the 1980 census data for the case catchment area, the incidence of colorectal cancer for foreign born females was estimated to be only 98.8% that of US born females (95% CI 94.9% to 103%, P = 0.59). For males, the age and SEER adjusted incidence of colorectal cancer for foreign born was estimated to be 1.047 times that for US born males (95% CI 1.004 to 1.091, two-sided P = 0.03). After adjusting for variations due to age and geographic location (SEER), we are thus unable to detect a significant difference in colorectal cancer incidence by birthplace in women, but do find a marginally significant increase in colorectal cancer among foreign born males compared to US born males.

b. Answer the question using an appropriate regression model.

***Instructions for grading**: This part of the problem is worth 10 points. The criteria for grading this problem should be very similar to the grading of problem 4b above. Additional key points to consider in your grading:*

- ***In describing the methods, we need to talk about the computation of the person-years of observation..***
- ***It would have been reasonable to alter the categorization of ages or to restrict to adults. If this is done, it should be made clear.***
- ***One could either provide a single analysis adjusting for sex and age, or one could produce analyses for each sex separately. The latter would be more the norm with cancer incidence statisics.***

Ans: *Methods:* Colorectal cancer incidence rates were compared between subjects known to be born in the US and those known to be foreign born using Poisson regression. Numerator data comes from the SEER registries aggregated by 5 year age intervals and sex for each SEER site. Time period of surveillance was between 10 and 12 years depending on the SEER site. Denominator data was estimated from a sample of 1980 census data, and multiplied by the years of surveillance for each site. A Poisson regression model was fit to the data for each sex separately. In each such analysis, the model included a binary indicator of foreign birth, a linear continuous term based on the midpoints of the age intervals, dummy variables for SEER sites, and the SEER-age interaction. Inference (95% confidence intervals and two-sided P values) about the ratio of incidence rates was derived using the asymptotic normal distribution for Wald type statistics using the Poisson regression parameter estimate for the indicator of foreign birth, along with standard error estimates based on the Huber-White sandwich estimator. *(Note that very different inference is obtained when the "robust SE" are not used. I believe it best to consider the robust SE here, because the Poisson assumption is pretty strong, as is the assumption that we have modeled the data well..)*

Results: **During the period of surveillance, a total of 62,668 incident cases of colorectal cancer in subjects known to be US born and 11,026 incident cases in subjects known to be foreign born were reported by the participating SEER sites. Over that period of time, the estimated time of observation was over 192 million person years and 12 million person years for US born and foreign born, respectively. Crude cumulative incidence rates were thus 32.6 and 89.3 cases per 100,000 person years across all SEERs, all ages, and both sexes. The unadjusted incidence of colorectal cancer (i.e., across all SEERs and all ages) for foreign born females was estimated to be 2.42 times that of US born females (95% CI 2.35 to 2.49, P < 0.0001). For males, the unadjusted incidence of colorectal cancer for foreign born males was 3.14 times higher than that for US born males (95% CI 3.05 to 3..23, P < 0.0001).**

After adjustment for age, SEER, and an age-SEER interaction, the incidence of colorectal cancer for foreign born females was estimated to be only 93.0% that of US born females (95% CI 81.5% to 163%, P = 0.28). For males, the age and SEER adjusted incidence of colorectal cancer for foreign born was estimated to be only 95.5% that for US born males (95% CI 82.3% to 111%, two-sided P = 0.540). After adjusting for variations due to age and geographic location (SEER), we are thus unable to detect a significant difference in colorectal cancer incidence by birthplace in either women or men.

c. What is the difference in the statistical models you used? That is, how would you explain any differences between the two analysis approaches?

Instructions for grading: This part of the problem is worth 10 points. To get full credit, the student must address each of the relevant points given in the answer (the points can be weighted equally). Of course, their response to each of these points will differ depending upon the models they fit.

Ans: We obtained quite different estimates for RR in males for the two analyses. The directly standardized rates estimated that foreign born males had (barely) significantly higher risk of

cancer, while the Poisson analysis estimated that the foreign born males had a lower risk of cancer, though the estimate was not statistically significantly different from 1. The methods differ in their handling of

- **The modeling of age: The stratified analysis had to consider age categories without borrowing information across those categories, while in the Poisson regression analysis I was able to model age continuously (thereby borrowing information about a trend across all ages).** *(If the student fit a categorized age with dummy variables in the Poisson regression model, there would be no need to mention this difference.)*
- **The modeling of an age-SEERinteraction: The stratified analysis had to adjust for the age-SEER interaction (using categorized age). In the regression model, the age-SEER interaction did not have to be included. Though I did include the age-SEER interaction in my regression model, I modeled it continuously, rather than discretely.**
- **The modeling of a birthplace-age-SEER interaction: The stratified analysis allowed an estimate of the RR to be different for each stratum defined by age and SEER. The regression analysis is modeling the RR as if it were constant across all age-SEER combinations.**
- **The weighting of the individual combinations of age and SEER: In the stratified analysis, a specific "importance" weighting based on the combined group SEER-age-sex distribution was chosen to "average" over possibly different risks in each stratum. The RR was then estimated as a ratio of averages. In the Poisson regression model, if there is effect modification of the RR by age and/or SEER, the resultant RR estimate will be some sort of weighted geometric mean of stratum specific RR where the weights are "efficiency weights" that would be optimal if there were no effect modification.**

It is likely the difference in the weighting and the difference between taking a ratio of averages vs a geometric mean of ratios that contributes the most to the observed differences in inference. The lack of substantial precision in either of these analyses suggests that we should be most cautious in attributing the differences in the results to any particular factor, however.

Appendix

Stata Code and Results

```
. *************************
. * PROBLEMS 1-5
. * Reading in the data
.
. clear

. quietly: infile ptid mridate age male race weight height packyrs yrsquit alcoh ///
>         physact chf chd stroke diabetes genhlth ldl alb crt plt sbp aai fev ///
>         dsst atrophy whgrd numinf volinf obstime death using ///
>         http://www.emersonstatistics.com/datasets/mri.txt

.
. *(The first line of the file was variable names, so we can drop it):
. drop in 1
(1 observation deleted)

.
. * Problem 1
. summ obstime if death==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     obstime |       602    1945.694    108.4126       1827       2159

. di 1827 / 365.25
5.0020534

.
. gen deadin5= 0

. replace deadin5=1 if obstime <= 5 * 365.25
(121 real changes made)

. replace deadin5=. if obstime==.
(0 real changes made)
```

```
.
. * Problem 2a
.
. * Categorizing age in 5 year categories (and verify categorization)
. egen agectg= cut(age), at(65, 70, 75, 80, 85, 100)

. tabstat age, by(agectg) stat(n mean sd min q max) col(stat)

Summary for variables: age
     by categories of: agectg

  agectg |         N      mean        sd       min       p25       p50       p75       max
---------+--------------------------------------------------------------------------------
      65 |       117  67.94017  .9311709        65        67        68        69        69
      70 |       305  71.89836  1.390581        70        71        72        73        74
      75 |       187  76.82353  1.464961        75        75        77        78        79
      80 |        81  81.60494  1.393548        80        80        81        83        84
      85 |        45  87.82222  2.971753        85        86        87        89        99
---------+--------------------------------------------------------------------------------
   Total |       735  74.56599  5.451364        65        71        74        78        99
--------------------------------------------------------------------------------------------


.
.
. * Creating indicator of ASCVD
. g ascvd= 1

. replace ascvd= 0 if chd==0 & stroke==0
(518 real changes made)

. replace ascvd= . if chd==. | stroke==.
(0 real changes made)


.
. * Tabulate agectg and male by ASCVD for reference when looking at standardizations
. table ascvd agectg male

----------------------------------------------------------------------
          |                       male and agectg
          | ------------ 0 ------------    ------------ 1 ------------
    ascvd |   65    70    75    80    85      65    70    75    80    85
```

```
----------+-------------------------------------------------------------
        0 |   52   126    62    29    13      38   105    56    22    15
        1 |    9    28    31    14     5      18    46    38    16    12
-------------------------------------------------------------------------
```

```
. table ascvd deadin5 agectg if male==0, col
```

```
-------------------------------------------------------------------------------------------
      |                                 agectg and deadin5
      | -------- 65 -------  -------- 70 -------  -------- 75 -------  -------- 80 -------
ascvd |     0      1  Total      0      1  Total      0      1  Total      0      1  Total
------+------------------------------------------------------------------------------------
    0 |    51      1     52    120      6    126     53      9     62     25      4     29
    1 |     9             9     21      7     28     25      6     31     10      4     14
-------------------------------------------------------------------------------------------
```

```
----------------------------
      | agectg and deadin5
      | -------- 85 -------
ascvd |     0      1  Total
------+---------------------
    0 |    10      3     13
    1 |     2      3      5
----------------------------
```

```
. table ascvd deadin5 agectg if male==1, col
```

```
-------------------------------------------------------------------------------------------
      |                                 agectg and deadin5
      | -------- 65 -------  -------- 70 -------  -------- 75 -------  -------- 80 -------
ascvd |     0      1  Total      0      1  Total      0      1  Total      0      1  Total
------+------------------------------------------------------------------------------------
    0 |    33      5     38     93     12    105     51      5     56     18      4     22
    1 |    11      7     18     29     17     46     27     11     38      9      7     16
-------------------------------------------------------------------------------------------
```

```
----------------------------
      | agectg and deadin5
      | -------- 85 -------
ascvd |     0      1  Total
------+---------------------
```

```
    0 |    11      4     15
    1 |     6      6     12
--------------------------
```

```
. tabstat age male, stat(n mean sd min q max) col(stat) by(ascvd) long
```

ascvd	variable	N	mean	sd	min	p25	p50
0	age	518	74.16988	5.389123	65	70	73
	male	518	.4555985	.498506	0	0	0
1	age	217	75.51152	5.49504	66	71	75
	male	217	.5990783	.4912183	0	0	1
Total	age	735	74.56599	5.451364	65	71	74
	male	735	.4979592	.5003363	0	0	0

ascvd	variable	p75	max
0	age	77	95
	male	1	1
1	age	79	99
	male	1	1
Total	age	78	99
	male	1	1

```
.
. * An unadjusted analysis
. cs deadin5 ascvd, or woolf
```

	ascvd		
	Exposed	Unexposed	Total
Cases	68	53	121
Noncases	149	465	614
Total	217	518	735

```
                |                         |
           Risk |  .3133641    .1023166   |   .1646259
                |                         |
                |      Point estimate     |    [95% Conf. Interval]
                |-------------------------+------------------------
 Risk difference |         .2110475        |    .1440389     .278056
      Risk ratio |          3.06269        |    2.218941    4.227274
 Attr. frac. ex. |         .6734897        |    .5493345    .7634409
 Attr. frac. pop |         .3784901        |
      Odds ratio |         4.004052        |    2.673947    5.995794 (Woolf)
                +-------------------------------------------------
                              chi2(1) =    49.53  Pr>chi2 = 0.0000

.
. * Stratified analysis
. * Note that cs does not provide a p value for RD
. * We can use a CI to judge the p value
. * Because the unadjusted p value < .0001, I will try 99.99% CI
.
. * First with "internal standard": the ASCVD group
. cs deadin5 ascvd, by(male agectg) rd istandard level(99.99)

     male agectg |       RD    [99.99% Conf. Interval]       Weight
-----------------+-------------------------------------------------
           0 65  |   -.0192308    -.0933269   .0548654            9
           0 70  |     .202381    -.1244371    .529199           28
           0 75  |    .0483871    -.2779712   .3747454           31
           0 80  |    .1477833    -.3839266   .6794931           14
           0 85  |    .3692308    -.5968196   1.335281            5
           1 65  |    .2573099    -.2380346   .7526545           18
           1 70  |    .2552795    -.0468112   .5573702           46
           1 75  |     .200188    -.1221594   .5225354           38
           1 80  |    .2556818    -.3232546   .8346183           16
           1 85  |    .2333333    -.4826875   .9493542           12
-----------------+-------------------------------------------------
           Crude |    .2110475     .0780333   .3440616
 I. Standardized |    .1925401     .0583795   .3267007

.
. * Analyses interpreting the internal standard
. * Weights times the stratum specific RD to give an estimated difference
```

```
. * in number of deaths "due to ASCVD"
. * (This purports to estimate fraction of lives saved if ASCVD were changed to non-ASCVD)
. di (9 * -.0192308 + 18 * .2573099 + 28 * .202381 + 46 * .2552795 + ///
>         31 * .0483871 + 38 * .200188 + 14 * .1477833 + 16 * .2556818 + ///
>         5 * .3692308 + 12 * .2333333)
41.781199

.
. * The difference in number of deaths as a fraction of the ASCVD group
. * (This agrees with the adjusted RD from age-sex stratified analyses
. di (41.781199 / 217)
.19254009

.
. * Alternatively with "external standard": the non-ASCVD group
. * (This purports to estimate fraction of additional deaths if non-ASCVD became ASCVD)
. cs deadin5 ascvd, by(male agectg) rd estandard level(99.99)

     male agectg |       RD    [99.99% Conf. Interval]       Weight
-----------------+-------------------------------------------------
            0 65 |   -.0192308    -.0933269   .0548654           52
            0 70 |     .202381    -.1244371    .529199          126
            0 75 |    .0483871    -.2779712   .3747454           62
            0 80 |    .1477833    -.3839266   .6794931           29
            0 85 |    .3692308    -.5968196   1.335281           13
            1 65 |    .2573099    -.2380346   .7526545           38
            1 70 |    .2552795    -.0468112   .5573702          105
            1 75 |     .200188    -.1221594   .5225354           56
            1 80 |    .2556818    -.3232546   .8346183           22
            1 85 |    .2333333    -.4826875   .9493542           15
-----------------+-------------------------------------------------
           Crude |    .2110475     .0780333   .3440616
 E. Standardized |    .1805084     .0513041   .3097128

.
. * Alternative create an ad hoc standard based on total age-sex distribution
. * (This corresponds to an average RD for a randomly chosen subject in the dataset)
. egen wts= count(agectg), by(male agectg)

. cs deadin5 ascvd, by(male agectg) rd standard(wts) level(99.99)
```

```
     male agectg |       RD    [99.99% Conf. Interval]       Weight
-----------------+-----------------------------------------------
            0 65 |   -.0192308    -.0933269   .0548654           61
            0 70 |     .202381    -.1244371    .529199          154
            0 75 |    .0483871    -.2779712   .3747454           93
            0 80 |    .1477833    -.3839266   .6794931           43
            0 85 |    .3692308    -.5968196   1.335281           18
            1 65 |    .2573099    -.2380346   .7526545           56
            1 70 |    .2552795    -.0468112   .5573702          151
            1 75 |     .200188    -.1221594   .5225354           94
            1 80 |    .2556818    -.3232546   .8346183           38
            1 85 |    .2333333    -.4826875   .9493542           27
-----------------+-----------------------------------------------
           Crude |    .2110475     .0780333   .3440616
    Standardized |    .1840606       .05552   .3126013

.
. * Problem 2b
.
. * The analysis I would have been more prone to use a priori (just out of habit)
. regress deadin5 ascvd age male, robust

Linear regression                               Number of obs =     735
                                                F(  3,   731) =   20.16
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0927
                                                Root MSE      =   .3542

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ascvd |   .1892352   .0343236     5.51   0.000     .1218505    .2566198
         age |   .0086417   .0027878     3.10   0.002     .0031686    .0141148
        male |   .0712176   .0262331     2.71   0.007     .0197165    .1227188
       _cons |  -.5710826   .2023432    -2.82   0.005    -.9683256   -.1738396
------------------------------------------------------------------------------

.
. * An analysis that adjusted for the age-sex interaction
. regress deadin5 ascvd i.male##c.age, robust
```

```
Linear regression                                       Number of obs =     735
                                                        F(  4,   730) =   18.02
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0961
                                                        Root MSE      =  .35379

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ascvd |    .186887   .0343053     5.45   0.000     .1195382    .2542359
      1.male |    .660529   .3957052     1.67   0.095     -.116327    1.437385
         age |   .0128757   .0036334     3.54   0.000     .0057425     .020009
             |
  male#c.age |
          1  |  -.0079006   .0053842    -1.47   0.143     -.018471    .0026698
             |
       _cons |  -.8855701   .2623387    -3.38   0.001    -1.400598   -.3705417
------------------------------------------------------------------------------

.
. * An analysis that adjusted for the age-sex interaction with weights
. regress deadin5 ascvd i.male##c.age [pweight=wts], robust
(sum of wgt is   7.5205e+04)

Linear regression                                       Number of obs =     735
                                                        F(  4,   730) =   15.42
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0881
                                                        Root MSE      =  .34055

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ascvd |   .1922566   .0378728     5.08   0.000      .117904    .2666093
      1.male |   .7697128   .4320423     1.78   0.075    -.0784809    1.617907
         age |     .01153   .0039461     2.92   0.004     .0037828    .0192771
             |
  male#c.age |
```

```
          1  |  -.0095355   .0058722    -1.62   0.105    -.0210639    .0019928
             |
       _cons |  -.7848995   .2842625    -2.76   0.006    -1.342969   -.2268299
------------------------------------------------------------------------------

.
. ************************
. **
. ** OF SPECIAL NOTE: Showing explicit correspondence between linear regression
. **                  and stratified analysis of RD
. **
. ************************
. * An analysis that will best mimic the stratified analysis
. * Note that this is a "saturated" model, so the fitted values from the
. * regression will agree exactly with the sample proportions from the strata
. regress deadin5 i.agectg##i.male##i.ascvd, robust

Linear regression                                      Number of obs =     735
                                                       F( 18,   715) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.1119
                                                       Root MSE      =  .35433

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agectg |
         70  |   .0283883   .0272554     1.04   0.298    -.0251219    .0818985
         75  |   .1259305   .0492979     2.55   0.011     .0291446    .2227165
         80  |   .1187003    .067733     1.75   0.080     -.014279    .2516796
         85  |   .2115385   .1200408     1.76   0.078    -.0241362    .4472131
             |
      1.male |   .1123482   .0588555     1.91   0.057     -.003202    .2278984
             |
 agectg#male |
       70 1  |  -.0456815    .069462    -0.66   0.511    -.1820555    .0906924
       75 1  |  -.1682238   .0837499    -2.01   0.045    -.3326488   -.0037987
       80 1  |   -.068461   .1209539    -0.57   0.572    -.3059283    .1690062
       85 1  |  -.0764507   .1757914    -0.43   0.664    -.4215798    .2686783
             |
```

```
     1.ascvd |  -.0192308   .0193095    -1.00   0.320    -.0571408    .0186793
             |
agectg#ascvd |
       70 1  |   .2216117   .0873304     2.54   0.011     .0501571    .3930664
       75 1  |   .0676179   .0872135     0.78   0.438    -.1036074    .2388431
       80 1  |    .167014   .1399027     1.19   0.233    -.1076552    .4416833
       85 1  |   .3884615   .2524925     1.54   0.124    -.1072538    .8841768
             |
  male#ascvd |
        1 1  |   .2765407   .1305232     2.12   0.034     .0202862    .5327952
             |
 agectg#male#|
       ascvd |
     70 1 1  |  -.2236422   .1746071    -1.28   0.201    -.5664462    .1191618
     75 1 1  |  -.1247398   .1769923    -0.70   0.481    -.4722267     .222747
     80 1 1  |  -.1686421   .2428956    -0.69   0.488    -.6455161    .3082318
     85 1 1  |  -.4124381   .3394609    -1.21   0.225    -1.078897    .2540211
             |
       _cons |   .0192308   .0193095     1.00   0.320    -.0186793    .0571408
------------------------------------------------------------------------------

.
. * Brute force computation of stratum specific RDs agree with stratified analysis
. di -.0192308 + .2216117                          // RD in 70-74 yo F
.2023809

. di -.0192308 + .0676179                          // RD in 75-79 yo F
.0483871

. di -.0192308 + .167014                           // RD in 80-84 yo F
.1477832

. di -.0192308 + .3884615                          // RD in 85-99 yo F
.3692307

. di -.0192308 + .2765407                          // RD in 65-69 yo M
.2573099

. di -.0192308 + .2765407 + .2216117 - .2236422   // RD in 70-74 yo M
.2552794
```

```
. di -.0192308 + .2765407 + .0676179 - .1247398   // RD in 75-79 yo M
.200188

. di -.0192308 + .2765407 + .167014 - .1686421    // RD in 80-84 yo M
.2556818

. di -.0192308 + .2765407 + .3884615- .4124381    // RD in 85-99 yo M
.2333333


.
. * The linear combination that mimics the "istandard" stratified analysis
. lincom  (                                                                    ///
>   9 * (1.ascvd) +                                                           ///
>  RD in 65-69 yo F
>  28 * (1.ascvd + 70.agectg#1.ascvd) +                                       ///
>  RD in 70-74 yo F
>  31 * (1.ascvd + 75.agectg#1.ascvd) +                                       ///
>  RD in 75-79 yo F
>  14 * (1.ascvd + 80.agectg#1.ascvd) +                                       ///
>  RD in 80-84 yo F
>   5 * (1.ascvd + 85.agectg#1.ascvd) +                                       ///
>  RD in 85-99 yo F
>  18 * (1.ascvd + 1.male#1.ascvd) +                                          ///
>  RD in 65-69 yo M
>  46 * (1.ascvd + 1.male#1.ascvd + 70.agectg#1.ascvd + 70.agectg#1.male#1.ascvd) +    ///
>  RD in 70-74 yo M
>  38 * (1.ascvd + 1.male#1.ascvd + 75.agectg#1.ascvd + 75.agectg#1.male#1.ascvd) +    ///
>  RD in 75-79 yo M
>  16 * (1.ascvd + 1.male#1.ascvd + 80.agectg#1.ascvd + 80.agectg#1.male#1.ascvd) +    ///
>  RD in 80-84 yo M
>  12 * (1.ascvd + 1.male#1.ascvd + 85.agectg#1.ascvd + 85.agectg#1.male#1.ascvd)) /   ///
>  RD in 85-99 yo M
>  217, level(99.99)                                                          //
>  Total ASCVD

 ( 1)  1.ascvd + .3410138*70.agectg#1.ascvd + .3179724*75.agectg#1.ascvd +
       .1382488*80.agectg#1.ascvd + .078341*85.agectg#1.ascvd + .5990783*1.male#1.ascvd +
       .2119816*70.agectg#1.male#1.ascvd + .1751152*75.agectg#1.male#1.ascvd +
       .0737327*80.agectg#1.male#1.ascvd + .0552995*85.agectg#1.male#1.ascvd = 0

------------------------------------------------------------------------------
```

```
     deadin5 |      Coef.   Std. Err.      t    P>|t|  [99.99% Conf. Interval]
-------------+------------------------------------------------------------------
         (1) |   .1925401   .0349623     5.51   0.000     .0557447    .3293355
------------------------------------------------------------------------------


. 
. * Test of ASCVD in saturated model
. regress deadin5 i.male##i.agectg##i.ascvd, robust

Linear regression                                      Number of obs =     735
                                                       F( 18,   715) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.1119
                                                       Root MSE      =  .35433

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.male |   .1123482   .0588555     1.91   0.057     -.003202    .2278984
             |
      agectg |
         70  |   .0283883   .0272554     1.04   0.298    -.0251219    .0818985
         75  |   .1259305   .0492979     2.55   0.011     .0291446    .2227165
         80  |   .1187003    .067733     1.75   0.080     -.014279    .2516796
         85  |   .2115385   .1200408     1.76   0.078    -.0241362    .4472131
             |
 male#agectg |
       1 70  |  -.0456815    .069462    -0.66   0.511    -.1820555    .0906924
       1 75  |  -.1682238   .0837499    -2.01   0.045    -.3326488   -.0037987
       1 80  |   -.068461   .1209539    -0.57   0.572    -.3059283    .1690062
       1 85  |  -.0764507   .1757914    -0.43   0.664    -.4215798    .2686783
             |
     1.ascvd |  -.0192308   .0193095    -1.00   0.320    -.0571408    .0186793
             |
  male#ascvd |
        1 1  |   .2765407   .1305232     2.12   0.034     .0202862    .5327952
             |
agectg#ascvd |
       70 1  |   .2216117   .0873304     2.54   0.011     .0501571    .3930664
       75 1  |   .0676179   .0872135     0.78   0.438    -.1036074    .2388431
```

```
       80 1  |    .167014   .1399027     1.19   0.233    -.1076552    .4416833
       85 1  |   .3884615   .2524925     1.54   0.124    -.1072538    .8841768
             |
 male#agectg#|
       ascvd |
     1 70 1  |  -.2236422   .1746071    -1.28   0.201    -.5664462    .1191618
     1 75 1  |  -.1247398   .1769923    -0.70   0.481    -.4722267     .222747
     1 80 1  |  -.1686421   .2428956    -0.69   0.488    -.6455161    .3082318
     1 85 1  |  -.4124381   .3394609    -1.21   0.225    -1.078897    .2540211
             |
       _cons |   .0192308   .0193095     1.00   0.320    -.0186793    .0571408
------------------------------------------------------------------------------

. testparm i.male#i.agectg#i.ascvd i.male#i.ascvd i.agectg#i.ascvd 1.ascvd

 ( 1)  1.ascvd = 0
 ( 2)  1.male#1.ascvd = 0
 ( 3)  70.agectg#1.ascvd = 0
 ( 4)  75.agectg#1.ascvd = 0
 ( 5)  80.agectg#1.ascvd = 0
 ( 6)  85.agectg#1.ascvd = 0
 ( 7)  1.male#70.agectg#1.ascvd = 0
 ( 8)  1.male#75.agectg#1.ascvd = 0
 ( 9)  1.male#80.agectg#1.ascvd = 0
 (10)  1.male#85.agectg#1.ascvd = 0

       F( 10,   715) =    3.49
            Prob > F =    0.0002

.
. * Some tests of interactions
. regress deadin5 i.male##i.agectg##i.ascvd, robust

Linear regression                                      Number of obs =     735
                                                       F( 18,   715) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.1119
                                                       Root MSE      =  .35433

------------------------------------------------------------------------------
             |               Robust
```

```
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.male |   .1123482   .0588555     1.91   0.057     -.003202    .2278984
             |
      agectg |
         70  |   .0283883   .0272554     1.04   0.298    -.0251219    .0818985
         75  |   .1259305   .0492979     2.55   0.011     .0291446    .2227165
         80  |   .1187003    .067733     1.75   0.080     -.014279    .2516796
         85  |   .2115385   .1200408     1.76   0.078    -.0241362    .4472131
             |
 male#agectg |
       1 70  |  -.0456815    .069462    -0.66   0.511    -.1820555    .0906924
       1 75  |  -.1682238   .0837499    -2.01   0.045    -.3326488   -.0037987
       1 80  |   -.068461   .1209539    -0.57   0.572    -.3059283    .1690062
       1 85  |  -.0764507   .1757914    -0.43   0.664    -.4215798    .2686783
             |
     1.ascvd |  -.0192308   .0193095    -1.00   0.320    -.0571408    .0186793
             |
  male#ascvd |
        1 1  |   .2765407   .1305232     2.12   0.034     .0202862    .5327952
             |
agectg#ascvd |
       70 1  |   .2216117   .0873304     2.54   0.011     .0501571    .3930664
       75 1  |   .0676179   .0872135     0.78   0.438    -.1036074    .2388431
       80 1  |    .167014   .1399027     1.19   0.233    -.1076552    .4416833
       85 1  |   .3884615   .2524925     1.54   0.124    -.1072538    .8841768
             |
 male#agectg#|
       ascvd |
     1 70 1  |  -.2236422   .1746071    -1.28   0.201    -.5664462    .1191618
     1 75 1  |  -.1247398   .1769923    -0.70   0.481    -.4722267     .222747
     1 80 1  |  -.1686421   .2428956    -0.69   0.488    -.6455161    .3082318
     1 85 1  |  -.4124381   .3394609    -1.21   0.225    -1.078897    .2540211
             |
       _cons |   .0192308   .0193095     1.00   0.320    -.0186793    .0571408
------------------------------------------------------------------------------

. testparm i.male#i.agectg#i.ascvd i.male#i.ascvd i.agectg#i.ascvd 1.ascvd

 ( 1)  1.ascvd = 0
 ( 2)  1.male#1.ascvd = 0
```

```
 ( 3)  70.agectg#1.ascvd = 0
 ( 4)  75.agectg#1.ascvd = 0
 ( 5)  80.agectg#1.ascvd = 0
 ( 6)  85.agectg#1.ascvd = 0
 ( 7)  1.male#70.agectg#1.ascvd = 0
 ( 8)  1.male#75.agectg#1.ascvd = 0
 ( 9)  1.male#80.agectg#1.ascvd = 0
 (10)  1.male#85.agectg#1.ascvd = 0

       F( 10,   715) =    3.49
            Prob > F =    0.0002

. testparm i.male#i.agectg#i.ascvd i.agectg#i.ascvd

 ( 1)  70.agectg#1.ascvd = 0
 ( 2)  75.agectg#1.ascvd = 0
 ( 3)  80.agectg#1.ascvd = 0
 ( 4)  85.agectg#1.ascvd = 0
 ( 5)  1.male#70.agectg#1.ascvd = 0
 ( 6)  1.male#75.agectg#1.ascvd = 0
 ( 7)  1.male#80.agectg#1.ascvd = 0
 ( 8)  1.male#85.agectg#1.ascvd = 0

       F(  8,   715) =    1.32
            Prob > F =    0.2317

. testparm i.male#i.agectg#i.ascvd i.male#i.ascvd

 ( 1)  1.male#1.ascvd = 0
 ( 2)  1.male#70.agectg#1.ascvd = 0
 ( 3)  1.male#75.agectg#1.ascvd = 0
 ( 4)  1.male#80.agectg#1.ascvd = 0
 ( 5)  1.male#85.agectg#1.ascvd = 0

       F(  5,   715) =    1.36
            Prob > F =    0.2393

.
. * Problem 3a
.
. cs deadin5 ascvd, by(male agectg) or
```

```
     male agectg |       OR       [95% Conf. Interval]   M-H Weight
-----------------+------------------------------------------------
            0 65 |           0            0          .      .147541 (Cornfield)
            0 70 |    6.666667     2.124776   20.94761     .8181818 (Cornfield)
            0 75 |    1.413333      .470773   4.270701     2.419355 (Cornfield)
            0 80 |         2.5     .5649769   11.16314     .9302326 (Cornfield)
            0 85 |           5      .638676    39.8223     .3333333 (Cornfield)
            1 65 |         4.2     1.152018   15.31843     .9821429 (Cornfield)
            1 70 |    4.543103     1.967126   10.49181     2.304636 (Cornfield)
            1 75 |    4.155556     1.353784   12.65875      1.43617 (Cornfield)
            1 80 |         3.5     .8453067   14.38831     .9473684 (Cornfield)
            1 85 |        2.75     .5774654   13.08275     .8888889 (Cornfield)
-----------------+------------------------------------------------
           Crude |    4.004052     2.677651   5.987529
    M-H combined |    3.496627     2.309146    5.29477
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(9) =    4.463  Pr>chi2 = 0.8784

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =     37.39
                                              Pr>chi2 =    0.0000


.
. * Problem 3b
.
. * The analysis I would have been more prone to use a priori (just out of habit)
. logistic deadin5 ascvd age male

Logistic regression                               Number of obs   =        735
                                                  LR chi2(3)      =      65.38
                                                  Prob > chi2     =     0.0000
Log likelihood = -296.04774                       Pseudo R2       =     0.0994

------------------------------------------------------------------------------
     deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ascvd |   3.566727   .7490352     6.06   0.000     2.363264     5.38304
         age |   1.065669   .0191898     3.53   0.000     1.028714    1.103952
        male |   1.803582   .3883253     2.74   0.006     1.182675    2.750466
------------------------------------------------------------------------------
```

```
.
. * An analysis that adjusted for the age-sex interaction
. logistic deadin5 ascvd i.male##c.age

Logistic regression                                     Number of obs   =        735
                                                        LR chi2(4)      =      72.01
                                                        Prob > chi2     =     0.0000
Log likelihood = -292.73116                             Pseudo R2       =     0.1095

------------------------------------------------------------------------------
     deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ascvd |   3.498582   .7365869     5.95   0.000     2.315694    5.285705
      1.male |   3114.393   9153.831     2.74   0.006     9.806624    989070.3
         age |   1.135452   .0350823     4.11   0.000     1.068732    1.206336
             |
  male#c.age |
          1  |   .9069276   .0347079    -2.55   0.011       .84139    .9775699
------------------------------------------------------------------------------


.
. * An analysis that will best mimic the stratified analysis
. logistic deadin5 i.agectg##i.male##i.ascvd
note: 65.agectg#0.male#1.ascvd != 0 predicts failure perfectly
      65.agectg#0.male#1.ascvd dropped and 9 obs not used

note: 85.agectg#1.male#1.ascvd omitted because of collinearity

Logistic regression                                     Number of obs   =        726
                                                        LR chi2(18)     =      76.84
                                                        Prob > chi2     =     0.0000
Log likelihood = -288.68701                             Pseudo R2       =     0.1175

------------------------------------------------------------------------------
     deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agectg |
         70  |   2.549998   2.787098     0.86   0.392     .2993594    21.72134
         75  |   8.660369   9.285539     2.01   0.044     1.058983    70.82458
         80  |   8.159992   9.338137     1.83   0.067     .8661463    76.87555
```

```
          85  |   15.29999   18.44229     2.26   0.024      1.44102    162.4471
              |
       1.male |   7.727265   8.639034     1.83   0.067     .8637496    69.12956
              |
 agectg#male |
       70 1  |   .3339662   .4116051    -0.89   0.374     .0298275    3.739276
       75 1  |   .0747149   .0944929    -2.05   0.040     .0062644     .891112
       80 1  |   .1797387   .2441713    -1.26   0.206     .0125401     2.57621
       85 1  |   .1568629   .2231731    -1.30   0.193     .0096492    2.550044
              |
     1.ascvd |   7.636364   11.84247     1.31   0.190     .3654634    159.5619
              |
agectg#ascvd |
       70 1  |   .8730159   1.453107    -0.08   0.935     .0334368    22.79398
       75 1  |   .1850794   .3064516    -1.02   0.308     .0072103    4.750778
       80 1  |    .327381   .5712754    -0.64   0.522     .0107082    10.00895
       85 1  |   .6547619   .6985765    -0.40   0.691     .0808951     5.29962
              |
  male#ascvd |
        1 1  |        .55   .7662408    -0.43   0.668     .0358508    8.437757
              |
 agectg#male#|
       ascvd |
     65 0 1  |  (empty)
     70 1 1  |   1.239028   1.956636     0.14   0.892     .0560905    27.36989
     75 1 1  |   5.345912   8.661547     1.03   0.301     .2233002    127.9836
     80 1 1  |   2.545455   4.510868     0.53   0.598     .0789457    82.07338
     85 1 1  |  (omitted)
------------------------------------------------------------------------------

.
. * Problem 4a
.
. * In the assignment, I suggested using Stata command ir
. * To do this we had to define a constant time variable, and to define a single
. * stratification variable
. g mAgectg= 100*male + agectg

. g time= 1

. ir deadin5 ascvd time, by(mAgectg)
```

```
         mAgectg |      IRR       [95% Conf. Interval]   M-H Weight
-----------------+------------------------------------------------
              65 |          0             0   225.3333      .147541 (exact)
              70 |       5.25      1.510783   18.90916     1.090909 (exact)
              75 |   1.333333      .3905268   4.194448            3 (exact)
              80 |   2.071429      .3858193   11.12131     1.302326 (exact)
              85 |        2.6      .3482377   19.41203     .8333333 (exact)
             165 |   2.955556      .8074879   11.80963     1.607143 (exact)
             170 |   3.233696      1.455468   7.420701     3.655629 (exact)
             175 |   3.242105      1.038475   11.90278     2.021277 (exact)
             180 |    2.40625      .6117207   11.20926     1.684211 (exact)
             185 |      1.875      .4446351   9.033643     1.777778 (exact)
-----------------+------------------------------------------------
           Crude |    3.06269      2.107235   4.472294              (exact)
    M-H combined |   2.634438      1.837129   3.777777
-------------------------------------------------------------------
 Test of homogeneity (M-H)    chi2(9) =      4.10  Pr>chi2 = 0.9046

.
. * The Stata command cs gets you the same thing more easily
. cs deadin5 ascvd, by(male agectg)

     male agectg |       RR       [95% Conf. Interval]   M-H Weight
-----------------+------------------------------------------------
            0 65 |          0            .          .      .147541
            0 70 |        5.25     1.910976   14.42326     1.090909
            0 75 |    1.333333     .5215043   3.408942            3
            0 80 |    2.071429     .6052232   7.089643     1.302326
            0 85 |         2.6     .7648441   8.838402     .8333333
            1 65 |    2.955556     1.085895   8.044341     1.607143
            1 70 |    3.233696     1.683615   6.210914     3.655629
            1 75 |    3.242105     1.224655   8.583029     2.021277
            1 80 |     2.40625      .845281   6.849839     1.684211
            1 85 |       1.875     .6814636   5.158933     1.777778
-----------------+------------------------------------------------
           Crude |     3.06269     2.218941   4.227274
    M-H combined |    2.634438     1.918642   3.617279
-------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(9) =    5.024  Pr>chi2 = 0.8322
```

```
. cs deadin5 ascvd, by(male agectg) istandard

     male agectg |       RR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
            0 65 |           0            .          .            9
            0 70 |        5.25     1.910976   14.42326           28
            0 75 |    1.333333     .5215043   3.408942           31
            0 80 |    2.071429     .6052232   7.089643           14
            0 85 |         2.6     .7648441   8.838402            5
            1 65 |    2.955556     1.085895   8.044341           18
            1 70 |    3.233696     1.683615   6.210914           46
            1 75 |    3.242105     1.224655   8.583029           38
            1 80 |     2.40625      .845281   6.849839           16
            1 85 |       1.875     .6814636   5.158933           12
-----------------+-------------------------------------------------
           Crude |     3.06269     2.218941   4.227274
 I. Standardized |    2.593559     1.878397   3.581004

. cs deadin5 ascvd, by(male agectg) estandard

     male agectg |       RR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
            0 65 |           0            .          .           52
            0 70 |        5.25     1.910976   14.42326          126
            0 75 |    1.333333     .5215043   3.408942           62
            0 80 |    2.071429     .6052232   7.089643           29
            0 85 |         2.6     .7648441   8.838402           13
            1 65 |    2.955556     1.085895   8.044341           38
            1 70 |    3.233696     1.683615   6.210914          105
            1 75 |    3.242105     1.224655   8.583029           56
            1 80 |     2.40625      .845281   6.849839           22
            1 85 |       1.875     .6814636   5.158933           15
-----------------+-------------------------------------------------
           Crude |     3.06269     2.218941   4.227274
 E. Standardized |    2.764214     1.991523   3.836702

. cs deadin5 ascvd, by(male agectg) standard(wts)

     male agectg |       RR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
            0 65 |           0            .          .           61
```

```
            0 70 |        5.25     1.910976   14.42326          154
            0 75 |    1.333333     .5215043   3.408942           93
            0 80 |    2.071429     .6052232   7.089643           43
            0 85 |         2.6     .7648441   8.838402           18
            1 65 |    2.955556     1.085895   8.044341           56
            1 70 |    3.233696     1.683615   6.210914          151
            1 75 |    3.242105     1.224655   8.583029           94
            1 80 |     2.40625      .845281   6.849839           38
            1 85 |       1.875     .6814636   5.158933           27
-----------------+------------------------------------------------
           Crude |     3.06269     2.218941   4.227274
    Standardized |    2.707733     1.962761   3.735462

.
. * Problem 4b
.
. * The analysis I would have been more prone to use a priori (just out of habit)
. poisson deadin5 ascvd age male, robust

Poisson regression                                Number of obs   =        735
                                                  Wald chi2(3)    =      76.46
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -313.17242                 Pseudo R2       =     0.0770

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ascvd |   .9993239   .1663382     6.01   0.000     .6733069    1.325341
         age |   .0463668    .013451     3.45   0.001     .0200033    .0727303
        male |   .4583881   .1706203     2.69   0.007     .1239784    .7927978
       _cons |  -5.991398     1.0062    -5.95   0.000    -7.963514   -4.019282
------------------------------------------------------------------------------

.
. * An analysis that adjusted for the age-sex interaction
. poisson deadin5 ascvd i.male##c.age, robust

Poisson regression                                Number of obs   =        735
                                                  Wald chi2(4)    =      75.71
                                                  Prob > chi2     =     0.0000
```

```
Log pseudolikelihood = -309.98212                 Pseudo R2       =     0.0864

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ascvd |   .9778957   .1641729     5.96   0.000     .6561227    1.299669
      1.male |   6.699502   2.108884     3.18   0.001     2.566165    10.83284
         age |   .1020999   .0215578     4.74   0.000     .0598475    .1443523
             |
  male#c.age |
          1  |  -.0812299   .0269918    -3.01   0.003    -.1341328    -.028327
             |
       _cons |  -10.27916   1.696492    -6.06   0.000    -13.60423   -6.954101
------------------------------------------------------------------------------

.
. * An analysis that is most analogous to the stratified analysis
. * (Note that to mimic the stratified RR we would have to use Stata command nlcom
. * to first produce average rates and then the ratio of those average rates. The
. * Poisson analysis takes geometric means of the rates and therefore the rate ratio)
. poisson deadin5 i.agectg##i.male##i.ascvd, robust

Poisson regression                                Number of obs   =        735
                                                  Wald chi2(19)   =    1527.38
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -306.48226                 Pseudo R2       =     0.0967

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agectg |
         70  |   .9065444   1.068064     0.85   0.396    -1.186822    2.999911
         75  |    2.02109   1.037755     1.95   0.051    -.0128716    4.055052
         80  |   1.970001   1.094374     1.80   0.072    -.1749335    4.114935
         85  |   2.484683   1.112925     2.23   0.026     .3033895    4.665976
             |
      1.male |   1.922857   1.075061     1.79   0.074    -.1842245    4.029938
             |
 agectg#male |
```

```
       70 1  |  -1.047437   1.178402    -0.89   0.374    -3.357063    1.262189
       75 1  |  -2.408808   1.197201    -2.01   0.044    -4.755279   -.0623368
       80 1  |  -1.646607   1.255575    -1.31   0.190    -4.107488    .8142746
       85 1  |  -1.778248   1.263389    -1.41   0.159    -4.254446    .6979492
             |
     1.ascvd |  -10.58728   1.045819   -10.12   0.000    -12.63705   -8.537517
             |
agectg#ascvd |
       70 1  |   12.24548   1.165982    10.50   0.000     9.960197    14.53076
       75 1  |   10.87497    1.15045     9.45   0.000     8.620128    13.12981
       80 1  |   11.31558   1.219961     9.28   0.000     8.924503    13.70666
       85 1  |   11.54318   1.218072     9.48   0.000     9.155805    13.93056
             |
  male#ascvd |
        1 1  |   11.67104   1.164207    10.02   0.000     9.389235    13.95284
             |
 agectg#male#|
       ascvd |
    70 1 1   |  -12.15554   1.316286    -9.23   0.000    -14.73541   -9.575666
    75 1 1   |  -10.78249   1.353523    -7.97   0.000    -13.43534    -8.12963
    80 1 1   |  -11.52125   1.426616    -8.08   0.000    -14.31736   -8.725131
    85 1 1   |  -11.99839   1.418479    -8.46   0.000    -14.77856   -9.218222
             |
       _cons |  -3.951091   .9908612    -3.99   0.000    -5.893144   -2.009039
------------------------------------------------------------------------------

.
. * The linear combination that mimics the "istandard" stratified analysis
. lincom  (                                                                   ///
>   9 * (1.ascvd) +                                                           ///
>  RR in 65-69 yo F
>  28 * (1.ascvd + 70.agectg#1.ascvd) +                                       ///
>  RR in 70-74 yo F
>  31 * (1.ascvd + 75.agectg#1.ascvd) +                                       ///
>  RR in 75-79 yo F
>  14 * (1.ascvd + 80.agectg#1.ascvd) +                                       ///
>  RR in 80-84 yo F
>   5 * (1.ascvd + 85.agectg#1.ascvd) +                                       ///
>  RR in 85-99 yo F
>  18 * (1.ascvd + 1.male#1.ascvd) +                                          ///
>  RR in 65-69 yo M
```

```
>  46 * (1.ascvd + 1.male#1.ascvd + 70.agectg#1.ascvd + 70.agectg#1.male#1.ascvd) +    ///
>  RR in 70-74 yo M
>  38 * (1.ascvd + 1.male#1.ascvd + 75.agectg#1.ascvd + 75.agectg#1.male#1.ascvd) +    ///
>  RR in 75-79 yo M
>  16 * (1.ascvd + 1.male#1.ascvd + 80.agectg#1.ascvd + 80.agectg#1.male#1.ascvd) +    ///
>  RR in 80-84 yo M
>  12 * (1.ascvd + 1.male#1.ascvd + 85.agectg#1.ascvd + 85.agectg#1.male#1.ascvd)) /   ///
>  RR in 85-99 yo M
>  217                                                                                 //
>  Total ASCVD

 ( 1)  [deadin5]1.ascvd + .3410138*[deadin5]70.agectg#1.ascvd +
       .3179724*[deadin5]75.agectg#1.ascvd + .1382488*[deadin5]80.agectg#1.ascvd +
       .078341*[deadin5]85.agectg#1.ascvd + .5990783*[deadin5]1.male#1.ascvd +
       .2119816*[deadin5]70.agectg#1.male#1.ascvd +
       .1751152*[deadin5]75.agectg#1.male#1.ascvd +
       .0737327*[deadin5]80.agectg#1.male#1.ascvd +
       .0552995*[deadin5]85.agectg#1.male#1.ascvd = 0

------------------------------------------------------------------------------
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .529144   .1728488     3.06   0.002     .1903666    .8679214
------------------------------------------------------------------------------


.
.
. **************************
. * PROBLEM 6
. * Reading in the data
. * Note the specification of bplace and SEER as strings
.
. clear

. infile str9 bplace male age str9 SEER startyr cases census80 ///
>   using http://www.emersonstatistics.com/datasets/surveillance.txt
'male' cannot be read as a number for male[1]
'age' cannot be read as a number for age[1]
'startyr' cannot be read as a number for startyr[1]
'cases' cannot be read as a number for cases[1]
'census80' cannot be read as a number for census80[1]
```

```
(973 observations read)

.
. *(The first line of the file was variable names, so we can drop it):
.
. drop in 1
(1 observation deleted)


.
. * For this assignment, we are going to ignore the unknown bplace
. * so it might be easiest just to drop all those lines
.
. drop if bplace=="Unknown"
(324 observations deleted)


.
. * We then would rather have an indicator variable of foreign born
.
. g foreign= 1

. replace foreign = 0 if bplace=="US"
(324 real changes made)

. table foreign bplace

------------------------
          |    bplace
  foreign |    US  nonUS
----------+-------------
        0 |   324
        1 |          324
------------------------


.
. * We now "encode" the character string SEER to be integers:
.
. encode SEER, generate(seercode)


.
. * Examine the integer coding by considering
. bysort SEER: summ seercode
```

```
-------------------------------------------------------------------------------------------
-> SEER = ATL

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    seercode |        72           1           0          1          1

-------------------------------------------------------------------------------------------
-> SEER = CT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    seercode |        72           2           0          2          2

-------------------------------------------------------------------------------------------
-> SEER = DET

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    seercode |        72           3           0          3          3

-------------------------------------------------------------------------------------------
-> SEER = HI

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    seercode |        72           4           0          4          4

-------------------------------------------------------------------------------------------
-> SEER = IA

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    seercode |        72           5           0          5          5

-------------------------------------------------------------------------------------------
-> SEER = NM

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
```

```
    seercode |        72           6           0          6          6

-------------------------------------------------------------------------------------------
-> SEER = SF

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    seercode |        72           7           0          7          7

-------------------------------------------------------------------------------------------
-> SEER = UT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    seercode |        72           8           0          8          8

-------------------------------------------------------------------------------------------
-> SEER = WWA

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    seercode |        72           9           0          9          9


.
. * Now it is easy to create a variable that has a unique value for
. * every stratum. For instance, age is a number between 2.5 and 87.5.
. * If we multiply seercode by 100 and add age, we will have a number in
. * which the first digit indicates SEER and the remaining digits
. * indicate age. Similarly, if we want to include sex, we can
. * add 1000 * male to that number, and all numbers above 1000
. * will be for males:
.
. g SEERage= 100* seercode + age

. bysort SEER: summ SEERage

-------------------------------------------------------------------------------------------
-> SEER = ATL

    Variable |       Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+--------------------------------------------------------
     SEERage |        72         145    26.12268      102.5      187.5

-------------------------------------------------------------------------------------------
-> SEER = CT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     SEERage |        72         245    26.12268      202.5      287.5

-------------------------------------------------------------------------------------------
-> SEER = DET

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     SEERage |        72         345    26.12268      302.5      387.5

-------------------------------------------------------------------------------------------
-> SEER = HI

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     SEERage |        72         445    26.12268      402.5      487.5

-------------------------------------------------------------------------------------------
-> SEER = IA

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     SEERage |        72         545    26.12268      502.5      587.5

-------------------------------------------------------------------------------------------
-> SEER = NM

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     SEERage |        72         645    26.12268      602.5      687.5

-------------------------------------------------------------------------------------------
-> SEER = SF
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     SEERage |        72         745    26.12268      702.5      787.5

-------------------------------------------------------------------------------
-> SEER = UT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     SEERage |        72         845    26.12268      802.5      887.5

-------------------------------------------------------------------------------
-> SEER = WWA

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     SEERage |        72         945    26.12268      902.5      987.5


. g mSEERage= 1000 * male + SEERage

. bysort male SEER: summ mSEERage

-------------------------------------------------------------------------------
-> male = 0, SEER = ATL

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36         145    26.30861      102.5      187.5

-------------------------------------------------------------------------------
-> male = 0, SEER = CT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36         245    26.30861      202.5      287.5

-------------------------------------------------------------------------------
-> male = 0, SEER = DET

    Variable |       Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+--------------------------------------------------------
    mSEERage |        36         345    26.30861      302.5      387.5

-------------------------------------------------------------------------------------------
-> male = 0, SEER = HI

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36         445    26.30861      402.5      487.5

-------------------------------------------------------------------------------------------
-> male = 0, SEER = IA

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36         545    26.30861      502.5      587.5

-------------------------------------------------------------------------------------------
-> male = 0, SEER = NM

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36         645    26.30861      602.5      687.5

-------------------------------------------------------------------------------------------
-> male = 0, SEER = SF

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36         745    26.30861      702.5      787.5

-------------------------------------------------------------------------------------------
-> male = 0, SEER = UT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36         845    26.30861      802.5      887.5

-------------------------------------------------------------------------------------------
-> male = 0, SEER = WWA
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36         945    26.30861      902.5      987.5

-------------------------------------------------------------------------------------------
-> male = 1, SEER = ATL

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36        1145    26.30861     1102.5     1187.5

-------------------------------------------------------------------------------------------
-> male = 1, SEER = CT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36        1245    26.30861     1202.5     1287.5

-------------------------------------------------------------------------------------------
-> male = 1, SEER = DET

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36        1345    26.30861     1302.5     1387.5

-------------------------------------------------------------------------------------------
-> male = 1, SEER = HI

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36        1445    26.30861     1402.5     1487.5

-------------------------------------------------------------------------------------------
-> male = 1, SEER = IA

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36        1545    26.30861     1502.5     1587.5

-------------------------------------------------------------------------------------------
-> male = 1, SEER = NM
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36        1645    26.30861     1602.5     1687.5

-------------------------------------------------------------------------------------------
-> male = 1, SEER = SF

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36        1745    26.30861     1702.5     1787.5

-------------------------------------------------------------------------------------------
-> male = 1, SEER = UT

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36        1845    26.30861     1802.5     1887.5

-------------------------------------------------------------------------------------------
-> male = 1, SEER = WWA

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    mSEERage |        36        1945    26.30861     1902.5     1987.5


.
. * We create a variable pyobs to record the person years of observation
.
. g pyobs = 12 * census80

. replace pyobs = 11 * census80 if startyr==1974
(72 real changes made)

. replace pyobs = 10 * census80 if startyr==1975
(72 real changes made)

.
. * Getting summary statistics for cases and pyobs by bplace
. bysort bplace: summ cases pyobs
```

```
-------------------------------------------------------------------------------------------
-> bplace = US

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       cases |       324    193.4198    298.1913          0       1361
       pyobs |       324    593580.4    433772.9       4488    1654524

-------------------------------------------------------------------------------------------
-> bplace = nonUS

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       cases |       324    34.03086    77.98689          0        404
       pyobs |       324    38122.76    37873.38       1180     143268


. di 324 * 193.4198, 324 * 593580.4
62668.015 1.923e+08

. di 324 * 34.03086, 324 * 38122.76
11025.999 12351774

.
. * Now we can use the Stata command egen to create a new variable that
. * will sum over the values of census80 within strata. We can list the
. * relevant variables for a single value of SEERage (SEER=="WWA", age 87.5)
. * to see how they relate to each other.
.
. egen MSAweights = sum(pyobs), by(mSEERage)

. list pyobs MSAweights mSEERage male SEER age if SEER=="WWA" & age==87.5

     +-------------------------------------------------+
     |  pyobs   MSAwei~s   mSEERage   male   SEER    age |
     |-------------------------------------------------|
 14. | 160171     194645      987.5      0    WWA   87.5 |
277. |  61292      83776     1987.5      1    WWA   87.5 |
486. |  22484      83776     1987.5      1    WWA   87.5 |
647. |  34474     194645      987.5      0    WWA   87.5 |
```

```
     +-------------------------------------------------+

.
. * If we are stratifying by sex, SEER, and age, we could then consider
.
. ir cases foreign pyobs, by(mSEERage) estandard

        mSEERage |      IRR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
           102.5 |          .             .          .       355470 (exact)
…
          1987.5 |   1.102965      .8871411    1.36433        61292 (exact)
-----------------+-------------------------------------------------
           Crude |   2.739475      2.684343   2.795539              (exact)
 E. Standardized |   1.015793        .98399   1.048625

. ir cases foreign pyobs, by(mSEERage) istandard

        mSEERage |      IRR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
           102.5 |          .             .          .         3730 (exact)
…
          1987.5 |   1.102965      .8871411    1.36433        22484 (exact)
-----------------+-------------------------------------------------
           Crude |   2.739475      2.684343   2.795539              (exact)
 I. Standardized |   1.017919      .9960834   1.040234

. ir cases foreign pyobs, by(mSEERage) standard(MSAweights)

        mSEERage |      IRR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
           102.5 |          .             .          .       359200 (exact)
…
          1987.5 |   1.102965      .8871411    1.36433        83776 (exact)
-----------------+-------------------------------------------------
           Crude |   2.739475      2.684343   2.795539              (exact)
    Standardized |   1.016107      .9869356    1.04614

.
. * If we are stratifying by SEER, and age for each sex separately, we
. * could then consider for females:
```

```
.
. ir cases foreign pyobs if male==0, by(mSEERage) estandard

        mSEERage |      IRR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
           102.5 |          .             .          .       355470 (exact)
…
           987.5 |    1.04679      .8586425   1.268264       160171 (exact)
-----------------+-------------------------------------------------
           Crude |   2.415259      2.347062   2.485077              (exact)
 E. Standardized |   .9911597      .9480163   1.036266

. ir cases foreign pyobs if male==0, by(mSEERage) istandard

        mSEERage |      IRR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
           102.5 |          .             .          .         3730 (exact)
…
           987.5 |    1.04679      .8586425   1.268264        34474 (exact)
-----------------+-------------------------------------------------
           Crude |   2.415259      2.347062   2.485077              (exact)
 I. Standardized |   .9750338      .9461601   1.004789

. ir cases foreign pyobs if male==0, by(mSEERage) standard(MSAweights)

        mSEERage |      IRR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
           102.5 |          .             .          .       359200 (exact)
…
           987.5 |    1.04679      .8586425   1.268264       194645 (exact)
-----------------+-------------------------------------------------
           Crude |   2.415259      2.347062   2.485077              (exact)
    Standardized |   .9888198      .9492069   1.030086

.
. * and for males:
.
. ir cases foreign pyobs if male==1, by(mSEERage) estandard

        mSEERage |      IRR       [95% Conf. Interval]       Weight
-----------------+-------------------------------------------------
```

```
          1102.5 |          .            .          .       377330 (exact)
…
          1987.5 |   1.102965      .8871411    1.36433        61292 (exact)
-----------------+------------------------------------------------
           Crude |    3.14215      3.052389   3.234092              (exact)
 E. Standardized |   1.043889      .9974866    1.09245

. ir cases foreign pyobs if male==1, by(mSEERage) istandard

        mSEERage |      IRR       [95% Conf. Interval]       Weight
-----------------+------------------------------------------------
          1102.5 |          .            .          .         3500 (exact)
…
          1987.5 |   1.102965      .8871411    1.36433        22484 (exact)
-----------------+------------------------------------------------
           Crude |    3.14215      3.052389   3.234092              (exact)
 I. Standardized |   1.064983      1.032104   1.098909

. ir cases foreign pyobs if male==1, by(mSEERage) standard(MSAweights)

        mSEERage |      IRR       [95% Conf. Interval]       Weight
-----------------+------------------------------------------------
          1102.5 |          .            .          .       380830 (exact)
…
          1987.5 |   1.102965      .8871411    1.36433        83776 (exact)
-----------------+------------------------------------------------
           Crude |    3.14215      3.052389   3.234092              (exact)
    Standardized |   1.047052       1.00449   1.091417

.
. * Poisson regression models
.
. poisson cases foreign i.seercode##c.age if male==0, exposure(pyobs) robust irr

Poisson regression                              Number of obs   =        324
                                                Wald chi2(18)   =    3428.27
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -2173.2883               Pseudo R2       =     0.9598

------------------------------------------------------------------------------
             |               Robust
```

```
       cases |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   .9298238   .0626936    -1.08   0.281     .8147196     1.06119
             |
    seercode |
           2 |   1.323019   .5689511     0.65   0.515     .5695286    3.073384
           3 |   .7845823   .3209213    -0.59   0.553     .3519392    1.749079
           4 |   .4089423   .2236066    -1.64   0.102     .1400333    1.194243
           5 |   1.092439   .4606888     0.21   0.834      .478013    2.496632
           6 |   .6998577   .2973378    -0.84   0.401     .3043497    1.609336
           7 |   .8896077   .3892627    -0.27   0.789     .3773471    2.097278
           8 |   .4904234   .2067036    -1.69   0.091     .2146876    1.120303
           9 |   1.380017   .6001429     0.74   0.459     .5884549    3.236351
             |
         age |   1.089075   .0048566    19.13   0.000     1.079598    1.098636
             |
    seercode#|
       c.age |
           2 |   .9990665   .0063945    -0.15   0.884     .9866117    1.011678
           3 |   1.002136   .0060811     0.35   0.725     .9902883    1.014126
           4 |   1.010089   .0082156     1.23   0.217     .9941142     1.02632
           5 |   .9980907   .0061407    -0.31   0.756     .9861275    1.010199
           6 |    1.00099   .0063206     0.16   0.875     .9886784    1.013455
           7 |   1.000442   .0064516     0.07   0.945     .9878766    1.013167
           8 |   1.001909   .0062202     0.31   0.759     .9897911    1.014175
           9 |   .9961563   .0064856    -0.59   0.554     .9835255    1.008949
       pyobs | (exposure)
------------------------------------------------------------------------------

. poisson cases foreign i.seercode##c.age if male==1, exposure(pyobs) robust irr

Poisson regression                                Number of obs   =        324
                                                  Wald chi2(18)   =    3585.35
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood =  -2279.588                 Pseudo R2       =     0.9587

------------------------------------------------------------------------------
             |               Robust
       cases |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   .9548106   .0721114    -0.61   0.540     .8234379    1.107143
```

```
             |
    seercode |
           2 |    1.33776   .6474204     0.60   0.548     .5181216    3.454017
           3 |   .8737164   .4040368    -0.29   0.770     .3529761    2.162697
           4 |    .668966   .3641983    -0.74   0.460     .2301405    1.944531
           5 |   1.190005   .5955212     0.35   0.728     .4462489    3.173369
           6 |   1.018839    .480583     0.04   0.968     .4042006    2.568115
           7 |     .98006    .458025    -0.04   0.966      .392147    2.449382
           8 |   .6810652   .3028958    -0.86   0.388     .2848569    1.628361
           9 |   1.186048   .5693461     0.36   0.722     .4629078    3.038855
             |
         age |   1.100458   .0062221    16.93   0.000      1.08833    1.112721
             |
    seercode#|
       c.age |
           2 |   .9988351   .0074822    -0.16   0.876     .9842774    1.013608
           3 |   1.000756   .0072416     0.10   0.917      .986663     1.01505
           4 |   1.003518   .0086349     0.41   0.683     .9867354    1.020585
           5 |   .9946222   .0076939    -0.70   0.486     .9796563    1.009817
           6 |    .993927   .0072857    -0.83   0.406     .9797494     1.00831
           7 |   .9985267   .0071784    -0.21   0.838     .9845559    1.012696
           8 |   .9966134   .0068849    -0.49   0.623     .9832101    1.010199
           9 |   .9972659   .0074354    -0.37   0.713     .9827987    1.011946
       pyobs | (exposure)
------------------------------------------------------------------------------

.
. poisson cases foreign i.seercode##c.age if male==0, exposure(pyobs) irr

Poisson regression                                Number of obs   =        324
                                                  LR chi2(18)     =  103759.83
                                                  Prob > chi2     =     0.0000
Log likelihood = -2173.2883                       Pseudo R2       =     0.9598

------------------------------------------------------------------------------
       cases |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   .9298238   .0141119    -4.79   0.000     .9025723    .9578982
             |
    seercode |
           2 |   1.323019   .1556738     2.38   0.017     1.050532    1.666185
```

```
           3  |   .7845823   .0962057    -1.98   0.048     .6169698    .9977301
           4  |   .4089423    .110997    -3.29   0.001     .2402288     .696144
           5  |   1.092439   .1320659     0.73   0.465     .8619749    1.384521
           6  |   .6998577   .1150971    -2.17   0.030     .5070176     .966043
           7  |   .8896077   .1113658    -0.93   0.350     .6960503     1.13699
           8  |   .4904234    .082868    -4.22   0.000     .3521614    .6829686
           9  |   1.380017   .1695029     2.62   0.009     1.084761    1.755637
              |
          age |   1.089075   .0015867    58.57   0.000      1.08597    1.092189
              |
    seercode#|
       c.age |
           2  |   .9990665    .001631    -0.57   0.567     .9958748    1.002268
           3  |   1.002136   .0017075     1.25   0.210     .9987952    1.005489
           4  |   1.010089   .0037935     2.67   0.008     1.002681    1.017551
           5  |   .9980907   .0016629    -1.15   0.251     .9948369    1.001355
           6  |    1.00099   .0022908     0.43   0.665     .9965103     1.00549
           7  |   1.000442   .0017293     0.26   0.798     .9970583    1.003837
           8  |   1.001909   .0023501     0.81   0.416     .9973131    1.006525
           9  |   .9961563   .0016993    -2.26   0.024     .9928313    .9994924
       pyobs  | (exposure)
------------------------------------------------------------------------------

. poisson cases foreign i.seercode##c.age if male==1, exposure(pyobs) irr

Poisson regression                              Number of obs   =        324
                                                LR chi2(18)     =  105742.94
                                                Prob > chi2     =     0.0000
Log likelihood =  -2279.588                     Pseudo R2       =     0.9587

------------------------------------------------------------------------------
       cases |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   .9548106   .0148911    -2.97   0.003     .9260661    .9844473
             |
    seercode |
           2  |    1.33776   .1672786     2.33   0.020     1.046985    1.709289
           3  |   .8737164   .1125963    -1.05   0.295     .6786965    1.124774
           4  |    .668966   .1638744    -1.64   0.101     .4138929    1.081235
           5  |   1.190005   .1537816     1.35   0.178     .9237406    1.533019
           6  |   1.018839   .1721772     0.11   0.912     .7315739    1.418905
```

```
           7  |     .98006   .1292317    -0.15   0.879     .7568544    1.269092
           8  |   .6810652    .116423    -2.25   0.025     .4871718    .9521278
           9  |   1.186048     .15563     1.30   0.193     .9170856    1.533891
              |
         age  |   1.100458    .001784    59.05   0.000     1.096967     1.10396
              |
    seercode#|
       c.age |
           2  |   .9988351   .0018027    -0.65   0.518     .9953081    1.002375
           3  |   1.000756   .0018608     0.41   0.684     .9971156     1.00441
           4  |   1.003518   .0035566     0.99   0.322      .996571    1.010513
           5  |   .9946222   .0018417    -2.91   0.004     .9910192    .9982384
           6  |    .993927   .0024221    -2.50   0.012      .989191    .9986856
           7  |   .9985267   .0018929    -0.78   0.437     .9948236    1.002244
           8  |   .9966134   .0024422    -1.38   0.166     .9918382    1.001411
           9  |   .9972659   .0018853    -1.45   0.148     .9935776    1.000968
       pyobs | (exposure)
------------------------------------------------------------------------------

.
. * Alternative models without the SEER-age interaction
. poisson cases foreign i.seercode age if male==0, exposure(pyobs) robust irr

Poisson regression                                Number of obs   =        324
                                                  Wald chi2(10)   =    2359.54
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -2193.2659                 Pseudo R2       =     0.9594

------------------------------------------------------------------------------
             |               Robust
       cases |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   .9329831   .0619816    -1.04   0.296     .8190781    1.062728
             |
    seercode |
           2  |   1.237432   .1195974     2.20   0.028     1.023889    1.495511
           3  |   .9129447   .0767056    -1.08   0.278     .7743308    1.076372
           4  |   .8210234   .0939528    -1.72   0.085     .6560687    1.027452
           5  |   .9505656   .0880113    -0.55   0.584      .792813    1.139708
           6  |    .750234   .0679398    -3.17   0.002     .6282226    .8959421
           7  |   .9193473   .0891643    -0.87   0.386     .7601942     1.11182
```

```
          8  |   .5606321   .0505182    -6.42   0.000     .4698689    .6689278
          9  |    1.04883   .1131266     0.44   0.658     .8489751    1.295733
             |
         age |   1.088513   .0020101    45.93   0.000      1.08458     1.09246
       pyobs | (exposure)
-------------------------------------------------------------------------------

. poisson cases foreign i.seercode age if male==1, exposure(pyobs) robust irr

Poisson regression                                Number of obs   =        324
                                                  Wald chi2(10)   =    2449.26
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -2297.3257                 Pseudo R2       =     0.9583

-------------------------------------------------------------------------------
             |               Robust
       cases |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
     foreign |   .9629698   .0705691    -0.51   0.607     .8341311    1.111709
             |
    seercode |
          2  |      1.239   .1392214     1.91   0.056     .9940887    1.544249
          3  |   .9241766   .0966682    -0.75   0.451     .7528699    1.134462
          4  |   .8452818   .1096172    -1.30   0.195     .6555664    1.089899
          5  |    .816159   .0989709    -1.68   0.094     .6435087     1.03513
          6  |   .6732063   .0765508    -3.48   0.001     .5387129     .841277
          7  |   .8883447   .0987852    -1.06   0.287     .7143757     1.10468
          8  |   .5411643   .0594274    -5.59   0.000     .4363704    .6711244
          9  |   .9849769   .1116449    -0.13   0.894      .788759    1.230008
             |
         age |   1.098247      .0022    46.78   0.000     1.093943    1.102567
       pyobs | (exposure)
-------------------------------------------------------------------------------

.
. poisson cases foreign i.seercode age if male==0, exposure(pyobs) irr

Poisson regression                                Number of obs   =        324
                                                  LR chi2(10)     =  103719.88
                                                  Prob > chi2     =     0.0000
Log likelihood = -2193.2659                       Pseudo R2       =     0.9594
```

```
------------------------------------------------------------------------------
       cases |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   .9329831   .0140759    -4.60   0.000     .9057987    .9609834
             |
    seercode |
           2 |   1.237432   .0313258     8.42   0.000     1.177533    1.300378
           3 |   .9129447   .0240646    -3.46   0.001     .8669766    .9613501
           4 |   .8210234   .0468577    -3.46   0.001     .7341343    .9181963
           5 |   .9505656   .0243412    -1.98   0.048     .9040351    .9994911
           6 |    .750234   .0261601    -8.24   0.000      .700674    .8032996
           7 |   .9193473   .0243689    -3.17   0.002     .8728046    .9683718
           8 |   .5606321    .020605   -15.75   0.000     .5216672    .6025074
           9 |    1.04883   .0279963     1.79   0.074     .9953694    1.105163
             |
         age |   1.088513   .0003824   241.44   0.000     1.087764    1.089262
       pyobs | (exposure)
------------------------------------------------------------------------------

. poisson cases foreign i.seercode age if male==1, exposure(pyobs) irr

Poisson regression                                Number of obs   =        324
                                                  LR chi2(10)     =  105707.47
                                                  Prob > chi2     =     0.0000
Log likelihood = -2297.3257                       Pseudo R2       =     0.9583

------------------------------------------------------------------------------
       cases |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   .9629698   .0148863    -2.44   0.015     .9342308    .9925929
             |
    seercode |
           2 |      1.239   .0340668     7.79   0.000     1.173997    1.307601
           3 |   .9241766   .0262271    -2.78   0.005     .8741759    .9770373
           4 |   .8452818   .0461471    -3.08   0.002     .7595061    .9407447
           5 |    .816159   .0230227    -7.20   0.000     .7722601    .8625533
           6 |   .6732063   .0249056   -10.70   0.000       .62612    .7238337
           7 |   .8883447   .0256136    -4.11   0.000     .8395351    .9399921
           8 |   .5411643   .0206165   -16.12   0.000     .5022285    .5831187
           9 |   .9849769   .0284767    -0.52   0.601     .9307155    1.042402
```

```
             |
         age |   1.098247   .0004171   246.78   0.000      1.09743    1.099064
       pyobs | (exposure)
------------------------------------------------------------------------------
```