Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 10:
Assumptions; Diagnostics

November 18, 2014

1

Lecture Outline

- Model Diagnostics
  - Assessing distributional assumptions
  - Assessing model fit
- Case Diagnostics
  - Leverage
  - Influence
  - Outliers

2

Multiple Regression

- General notation for regression model

$$Y_i \mid X_i, W_{1i}, W_{2i} + \ldots \sim F(y; \theta_i) \text{ independent}$$
$$g(\theta_i) = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \beta_3 \times W_{2i} + \cdots$$

θ_i Summary measure for distn of $Y_i|X, W_1, W_2, \ldots$

$g(\ )$ "link" function used for modeling

β_0 "Intercept"

β_1 "Slope for Pred of Interest X"

β_j "Slope for covariate W_{j-1}"

- The link function is usually either none (means) or log (geom mean, odds, hazard)
  - But we also consider "complementary log-log"

3

Questions Addressed with Regression

- Associations between *Y* and predictor of interest *X*
  - First order trends vs exact relationship
  - Possibly modified by some other predictor *W*
  - Possibly adjusted for confounders, precision
- Estimates of *θ(Y|X,W,…)* within groups defined by *X, W, …*
  - Point estimates
    - (With PH we would just consider HR between specific groups)
  - Inference: Interval estimates, hypothesis testing
    - (Difficult to do with proportional hazards)
- Prediction: Distribution of *Y* within groups defined by *X, W, …*
  - Point estimates (usually use mean, median, etc.)
    - (Difficult to do with proportional hazards)
  - Inference: Interval estimates
    - (Difficult to do with proportional hazards

4

Maximal Assumptions

- Relevant sampling for scientific question
  - (or appropriate weighting of observations)
- Independence
  - (or correct modeling of dependent data within clusters)
- Sufficient sample sizes to approximate asymptotic distributions
- Variance appropriate to the model
  - (relaxed when using robust Huber-White SE)
- Regression model accurately describes summary measures across all groups
  - Linearity on appropriate scale; any effect modification of interest
  - Proportional hazards assumption holds with PH regression
- Shape of distribution same in each group
  - (or can be correctly derived using parameter estimates)

5

Inference About Linear Trend in g(θ)

- Relevant sampling for scientific question
  - (or appropriate weighting of observations)
- Independence
  - (or correct modeling of dependent data within clusters)
- Sufficient sample sizes to approximate asymptotic distributions
- Variance appropriate to the model
  - (relaxed when using robust Huber-White SE)
- Regression model accurately describes summary measures across all groups
  - Linearity on appropriate scale; any effect modification of interest
  - Proportional hazards assumption holds with PH regression
- Shape of distribution same in each group
  - (or can be correctly derived using parameter estimates)

6

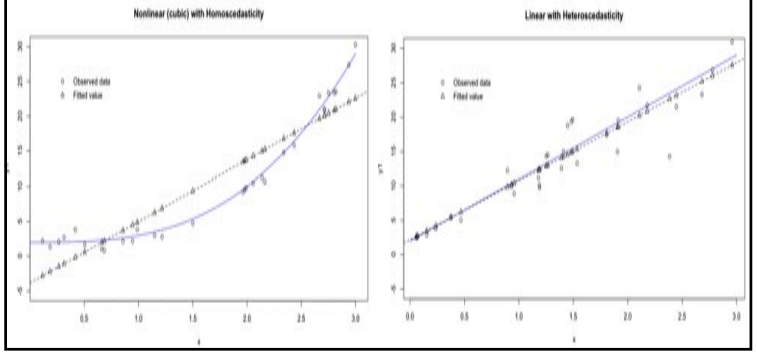
Estimating θ Within Groups

- Relevant sampling for scientific question
  - (or appropriate weighting of observations)
- Independence
  - (or correct modeling of dependent data within clusters)
- Sufficient sample sizes to approximate asymptotic distributions
- Variance appropriate to the model
  - (relaxed when using robust Huber-White SE)
- Regression model accurately describes summary measures across all groups
  - Linearity on appropriate scale; any effect modification of interest
  - Proportional hazards assumption holds with PH regression
- Shape of distribution same in each group
  - (or can be correctly derived using parameter estimates)

7

Estimating θ Within Groups

- Example: Linear regression
  - Nonlinearity matters, heteroscedasticity does not
- Other regressions: linearity of $g(\theta)$



Inference About θ Within Groups

- Relevant sampling for scientific question
  - (or appropriate weighting of observations)
- Independence
  - (or correct modeling of dependent data within clusters)
- Sufficient sample sizes to approximate asymptotic distributions
- Variance appropriate to the model
  - (relaxed when using robust Huber-White SE)
- Regression model accurately describes summary measures across all groups
  - Linearity on appropriate scale; any effect modification of interest
  - Proportional hazards assumption holds with PH regression
- Shape of distribution same in each group
  - (or can be correctly derived using parameter estimates)

9

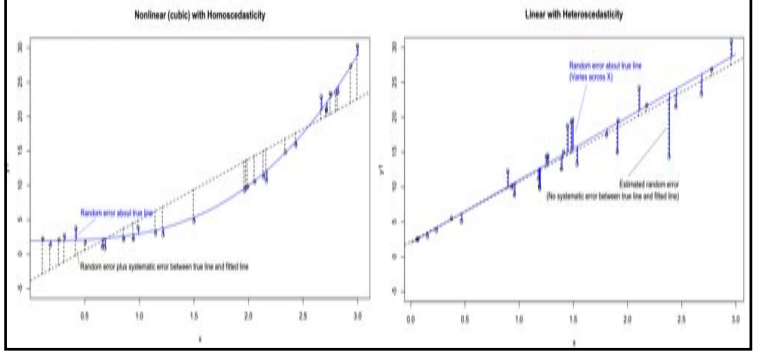
Estimating SD of *Y* Within Groups

- Relevant sampling for scientific question
  - (or appropriate weighting of observations)
- Independence
  - (or correct modeling of dependent data within clusters)
- Sufficient sample sizes to approximate asymptotic distributions
- Variance appropriate to the model (homoscedasticity for cts *Y*)
  - (relaxed when using robust Huber-White SE)
- Regression model accurately describes summary measures across all groups
  - Linearity on appropriate scale; any effect modification of interest
  - Proportional hazards assumption holds with PH regression
- Shape of distribution same in each group
  - (or can be correctly derived using parameter estimates)

10

Estimating SD Within Groups

- Example: Linear regression
  - Nonlinearity matters, heteroscedasticity matters
- Other regressions: Variance from mean-variance ➔ need linearity



Estimate *Y* Within Groups: Point Prediction

- Relevant sampling for scientific question
  - (or appropriate weighting of observations)
- Independence
  - (or correct modeling of dependent data within clusters)
- Sufficient sample sizes to approximate asymptotic distributions
- Variance appropriate to the model
  - (relaxed when using robust Huber-White SE)
- Regression model accurately describes summary measures across all groups
  - Linearity on appropriate scale; any effect modification of interest
  - Proportional hazards assumption holds with PH regression
- Shape of distribution same in each group
  - (or can be correctly derived using parameter estimates)

12

Inference for Distn of *Y* Within Groups

- Relevant sampling for scientific question
  - (or appropriate weighting of observations)
- Independence
  - (or correct modeling of dependent data within clusters)
- Sufficient sample sizes to approximate asymptotic distributions
- Variance appropriate to the model
  - (relaxed when using robust Huber-White SE)
- Regression model accurately describes summary measures across all groups
  - Linearity on appropriate scale; any effect modification of interest
  - Proportional hazards assumption holds with PH regression
- Shape of distribution same in each group
  - (or can be correctly derived using parameter estimates)

13

Role of Diagnostics

- Sometimes we want to assess whether
  - Regression model fits the bulk of the data well
    - Model diagnostics
      - Independence, link function, transformation of predictors, interactions, assumptions about variance
  - Individual cases might be different from the bulk of the data
    - Case diagnostics
      - Leverage, influence, outliers

14

What Can Go Wrong: General Case

- Errors we might make for the bulk of the data:
  - We could have specified *F* wrong
    - And if so, θ may not be the summary measure we think it is
  - We could have correlated observations
  - We could have the wrong link function *g( )*
  - We could have the wrong form in general for the specification of the covariates
    - (Maybe we should have transformed $X, W_1, W_2, \ldots$)
- Errors we might make for a few observations:
  - A very few observations might be very different from the bulk of the data

15

What Might Happen If We Go Wrong

- The interpretation of the regression parameters might be wrong
- The point estimates for the regression parameters might be
  - Biased (i.e., not tend to the truth on average across studies)
  - Inconsistent (i.e., not tend to the truth with large sample sizes)
- The confidence intervals for the regression parameters might be too narrow or too wide (i.e., not have the correct coverage probability)
- The reported P value might not be uniformly distributed when the null hypothesis is true

16

The Approach Used Here

- In this course, I have stressed a statistical analysis approach that minimizes our chances of going wrong.
  - Truly, I just stressed how to interpret analysis models robustly:
  - F is specified distribution-free manner (binary data is Bernoulli)
  - θ is just some summary measure (e.g., mean, median…) of a distribution (and we use distribution-free estimators)
  - $g(\ )$ is a link function chosen based on the comparison (difference or ratio) that I want to make across groups
  - X is a variable (or set of variables) modeled in a form (e.g. linear, quadratic, dummy variables,…) to answer my scientific question about the association between the POI and response
  - $W_1, W_2, \ldots$ are
    - confounders modeled in a way to best remove confounding based on our prior understanding, and
    - precision variables modeled sufficiently well to increase efficiency

17

Testing, Quantifying Associations

- We use our regression to model our question
- θ is the summary measure that is of greatest scientific interest
- $g(\ )$ is typically the identity function or log function according to whether the difference or ratio is scientifically most important
- β_1 (and any other coefficients for terms involving our POI) model appropriate linear contrasts of greatest scientific interest

18

Assumptions Needed: Associations

- The end result is that when answer questions about associations, we have very few assumptions that we need to verify ("diagnose problems")
  - *Independence*: Correct specification of any statistical independence or dependence among our observations
  - *Variance*: Robust specification of variance estimation techniques appropriate to the type of regression
  - *Distribution of Estimates*: Sufficiently large sample sizes for
    - parameter estimates to be approximately normally distributed about the true value, and
    - the estimated standard errors to be accurate.

19

Estimating θ Within Groups

- We are sometimes also interested in estimating the summary measure θ within groups defined by the predictor
- Having an approximate idea of the value is especially important when associations are quantified using ratios
  - A doubling of the risk may be chosen because it highlights information about the existence of an association
  - However, the clinical importance of a doubled risk is obviously less when the baseline risk is very low
- Occasionally we are truly interested in precisely quantifying θ within some group

20

Assumptions Needed: Estimates of θ

- In order to obtain absoutely reliable inference about θ, we need ALL the previous assumptions, PLUS
  - *Linearity*: Correct specification of regression model:
    - The linear predictor $X\beta$ must correctly model $g(\theta)$
- Note, however, that we will often get very good approximations to θ where we have the most data even if our model for the linear predictor is wrong.
  - In those typical cases, it will be in the fringes of the data where our (extrapolated) estimates will be most unreliable.
- Occasionally, however, there are certain "highly leveraged" observations that turn out to be "influential"
  - In the presence of influential observations, we can be very wrong

21

Role of Diagnostics

- Depending on where we are in the scientific process, we may want to assess the degree to which our data set appears to satisfy the various assumptions
- In "confirmatory" analyses we will have prespecified our model in order to robustly answer questions about trends
  - But after providing such an answer, we may want to explore our data to provide insight for later studies
  - And sometimes, we find things in our data that were completely unanticipated and make us question the credibility of our analysis results
- In "exploratory" analyses we will explore analyses trying to get the best description of relationships in order to generate hypotheses

22

Important Caveats

- Such diagnostic methods are always approximate
- Using diagnostics to alter your analysis plan (and hence the question answered) should always lessen our confidence in our statistical evidence
- Unfortunately, we do not always have a good way to quantify that lessened confidence in the P value and confidence intervals

23

The Real Problem

"Blood suckers hide 'neath my bed"

- "Eyepennies", Mark Linkous (Sparklehorse)

24

Nonrepresentative Samples

- Problems often result because of data that we didn't sample
- Recall "3.69 over N Rule"
  - Given a sample of size N, the upper 97.5% confidence bound on the proportion of the population not represented at all is approximately 3.69/n
- There is nothing your data can tell you about whether the unsampled population might be different
  - Only your sampling scheme tells you this

25

Because You Can't Stop Me

- There are common problems with all the model diagnostics
- They may detect "problems" that do not really affect your statistical inference
- They may not detect problems that truly exist
  - When assumptions do not hold, some data sets appear like the assumptions might be reasonable
  - Lack of power to prove "equivalence": Need infinite sample size
- Tendency to overfit the data
  - Inflated type I errors, anti-conservative CI

26

Role of Model Checking

- "Model checking" does not have good statistical foundations in either frequentist or Bayesian inference
- Most often, violations of the assumptions can happen only under the alternative, so model checking is multiple testing
- Bayesians should formally model the uncertainty in their assumptions

27

Because You Can't Stop Me

- The best approach is to use methods that have the fewest assumptions
- Do not try to make strong statistical inference about questions that are far more detailed than your current state of knowledge
- But after making inference about reasonable questions, DO explore your data for
  - information to use when using regression models in the next study, and
  - new hypotheses

28

Model Diagnostics

29

Evaluating Independence

30

Assessing Independence

- We must have variables that identify clusters
  - That is, we must have some reason to suspect correlation
- Things to look for
  - Correlations in time
  - Correlations in location
  - Correlations within families, hospitals, etc.
  - Correlations within subjects
- But we are interested in correlations ***AFTER*** adjustment for predictors

31

Assessing Independence: Caveats

- We will be tempted to measure correlations within clusters, and assume data are independent if the estimated correlation is small
  - However, small correlations within a moderate size cluster can cause big problems in the estimation of a standard error
- Suppose we have
  - *k* clusters of size *m* yielding *n=km* total observations
  - correlation is ρ between measurements within the same cluster
- Sampling distribution of a sample mean from clustered data is

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\left(1+(m-1)\rho\right)\right)$$

32

Assessing Independence: Example

- Sampling distribution of a sample mean from clustered data is

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\left(1+(m-1)\rho\right)\right)$$

- Suppose we have
  - *k* clusters of size *m= 21* yielding *n=km* total observations
  - correlation is *ρ= 0.05* between measurements within each cluster
- We are quite likely to estimate a small correlation and incorrectly assume that the data are independent
  - The resulting estimated SE is only half of the true SE

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\left(1+(m-1)\rho\right)\right) = \begin{cases} \rho = 0.05 \Rightarrow N\left(\mu, \frac{2\sigma^2}{n}\right) \\ \rho = \ \ 0 \ \ \Rightarrow N\left(\mu, \frac{\sigma^2}{n}\right) \end{cases}$$

33

Recommended Practices

- The above results and recommendations apply equally to binary response data or continuous response data
- It does not really seem worthwhile to waste time looking for correlations
  - We are not that good at estimating correlations in all the settings that it would cause problems
- If it seems plausible the data might be correlated within clusters, use an analysis model that can handle the correlation
  - This is the subject matter of Biost 540

34

Evaluating Distribution of Estimates

35

Assessing Asymptotic Distribution

- We usually rely on an approximate normal distribution for regression parameters
- Generally true in large samples
- But, the definition of "large" depends on the shape of the distribution for the data
  - A sample size of 3 is sufficiently large if *Y* is normally distributed
  - As a rule, "heavier tails" of response distribution requires larger sample size
    - "heavy tails"= tendency to outliers

36

Rules of Thumb

- Linear regression is quite robust for tests of zero slope when n > 50 (Lumley, et al.)
- Logistic, Poisson, proportional hazards asymptotics will depend on the number of events observed
  - (Unconditional exact logistic regression methods do exist: exlogistic in Stata or standalone StatExact)
- But some sampling schemes purposely alter the distribution of common statistics

37

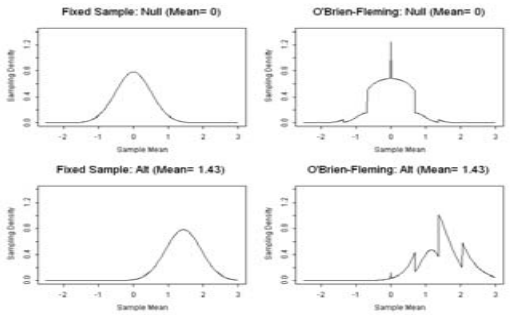
Asymptotic Distribution: Binary Data

- Binary data with *p* close to 0 or close to 1 is the ultimate in skewed data
- We typically define "large" by ensuring a reasonable number of events and non-events
- Classic rule-of-thumb: Large enough so we expect to see 5 events in every "cell" in contingency tables
  - A special variant of this issue arises with sparse data modeled with many dummy variables
  - We will discuss this further with conditional logistic regression for matched data

38

Fixed vs Sequential Sampling

- However: exceptions do exist
  - We sometimes purposely sample in a way that guarantees the sampling distribution will not be normal: Sequential sampling



39

Evaluating Variance Assumptions

40

Assessing Appropriate Variance

- Classic linear regression with continuous data: homoscedasticity
- Equality of variance across groups is most easily assessed by either
  - Stratified estimates of variances
    - Problem: Heterogeneity of means within strata can look like variability of response variables
    - Variance of residuals within strata
  - Scatterplots
    - Response versus predictors
    - Residuals versus fitted values
    - Residuals versus predictors

41

Linear Regression Residuals

Model :

$$E\left[Y_i \,\middle|\, \vec{X}_i\right] = \beta_0 + \beta_1 \times X_{1i} + \cdots + \beta_p \times X_{pi}$$

$$Y_i \,\middle|\, \vec{X}_i = \beta_0 + \beta_1 \times X_{1i} + \cdots + \beta_p \times X_{pi} + \varepsilon_i$$

Error ε_i is estimated by residual

$$\hat{e}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \times X_{1i} + \cdots + \hat{\beta}_p \times X_{pi}\right)$$

$$= Y_i - \hat{Y}_i$$

42

Stata: Estimation of Residuals

- Stata commands for estimation of residuals
- Obtain residuals from "`predict`" command
  - Following a linear regression
    - `predict` *`varname`*`, resid`
    - `predict` *`varname`*`, rstu` (studentized)
- Studentized residuals have been standardized to units of standard deviation
  - Often assumed to have t distn (approx normal)

43

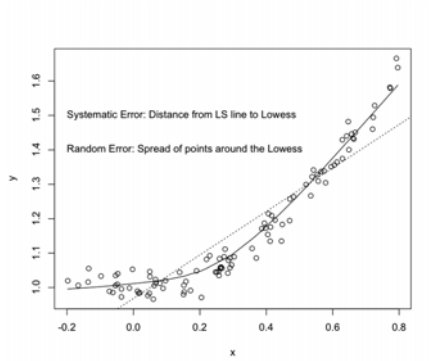
Linear Regr: Residual Analysis

- Assumptions in linear regression are primarily about the distribution of errors
- Thus we can examine the distribution of residuals
  - "Detrends" the data by subtracting off the estimated mean
  - Allows assessing the effect of multiple variables at once
  - Plots, stratified descriptive statistics, regression on squared residuals

44

Linear Regression Residuals

- Estimated residual variance also affected by having wrong model
- If our model is wrong, our variance estimates include both "systematic error" and "random error"



45

Linear Regr: Binary Data

- With binary data, plots of the residuals from linear regression are pretty boring
  - Each *Y* is either 0 or 1, so the residuals are similarly dichotomous for each combination of covariates
- Besides, we already know that there will be unequal variances due to the mean variance relationship
- We can use the robust SE to adjust for the heteroscedasticity when making inference
  - Note, however, that we generally need a larger sample size to get good behavior from the robust SE estimates than we need if we had homoscedasticity

46

Logistic, Poisson, PH Regr

- Assumptions about variance relate to mean variance relationships
  - The standard SE estimates use the estimated means

$$Y \sim (\mu, \quad V(\mu)) \quad \Rightarrow \quad \hat{V}(\mu) = V(\hat{\mu})$$

- Can be violated if
  - Data is not independent
    - "Overdispersed" or "underdispersed" binary or Poisson data
  - Model does not describe true relationship in g(θ) across groups
    - Wrong link function: e.g., multiplicative, additive, others
    - Wrong predictors and/or transformations
    - PH: nonproportional hazards (modeling of risk of event over time)

47

Logistic, Poisson, PH Regr

- We can use robust SE, which are instead based on something like a sample variance

$$Y \sim (\mu, \quad V(\mu)) \quad \Rightarrow \quad \text{Estimate} \sum V(\mu) = \sum (Y - \hat{\mu})^2$$

- With logistic regression, there is good agreement between the "model-based variance estimate" and the "sample variance"

$$Y \sim B(1, \quad p) \quad \Rightarrow \quad s^2 = \frac{n}{n\text{-}1} \hat{p}(1 - \hat{p})$$

  - Use of robust SE not as important
- With Poisson regression, there can be more differences between the standard (model based) SE and robust SE
  - I recommend using robust SE
  - But this will again mix "systematic" and "random" error

48

Evaluating Linearity

49

Assessing Model Fit

- The regression models we consider in this class are all based on "linear predictors"
- The summary of the response distribution is predicted to vary in some way across groups according to a linear function of the modeled predictors
- The modeled predictors may be transformations of the original measurements
  - E.g., log transformation of nadir PSA
  - E.g., dummy variables

50

Criteria

- Criteria for assessing model fit varies by type of regression
- Linear regression
  - Linearity of means
- Logistic regression
  - Linearity of log odds
- Poisson regression
  - Linearity of log rates
- Proportional hazards regression
  - Linearity of log hazards

51

General Methods

- Nonparametric description within strata
  - Strata generally not based on quantiles
- Graphical methods
  - Plots of data or residuals
    - Most useful with means (linear regression)
- Model based methods
  - Fit more flexible models and examine higher order terms
    - Plots of fitted values
  - With binary response data, I find this approach easiest

52

Modeling Based Diagnostics

- There is to my mind only a difference in perspective between
  - Addressing scientific questions about functional relationships between the POI and the response, and
  - Performing diagnostic analyses looking for violations of assumptions related to the "linearity" of effect
- My top choice: Pose scientific questions that can be answered rigorously
- But sometimes unanticipated problems lead to the need to explore the data

53

Ex: Mortality and Serum Cholesterol

- We consider the CHS data set related to inflammatory biomarkers
- We have 4 years of follow-up for survival on 5000 people
- We have serum cholesterol at study entry on 4953 people

54

As a Scientific Question

- Many reports in the scientific literature report increased risk of CVD in subjects with higher serum cholesterol
- Based on our prior understanding that the elderly with poor liver function will tend to have low cholesterol, we might expect a mixture of
  - People nearing death who have poorly functioning livers
  - People with bad lipid profiles that place them at risk for death
- We want to investigate whether the association between mortality and serum cholesterol is nonlinear
- We consider a function that models cholesterol as a linear term plus other terms (e.g., quadratic, splines, dummy variables)
  - We can test for the significance of all the terms that go beyond linearity with cholesterol

55

As Diagnostic Exploration

- Many reports in the scientific literature report increased risk of CVD in subjects with higher serum cholesterol
- Our analysis of the Cardiovascular Health Study data suggests a linear trend among 65-100 yo in which subjects with higher serum cholesterol trend toward higher probability of survival
- Given those surprising results, we want to explore what might be going on.
- We can consider more flexible models to differentiate between
  - Data that is not at all consistent with those prior studies, and
  - A curvilinear relationship that might give different trends according to the distribution of cholesterol levels (and other factors)

56

Hierarchy of Scientific Questions

- Linear trend in mortality over cholesterol
- Detect nonlinear trend: Alternative approaches
  - General tests for nonlinearity
    - Fit a quadratic model
      - A single parameter estimated across all data
    - Fit dummy variables
      - Multiple parameters that make no use of ordering
    - Linear splines
      - Multiple parameters that fit a somewhat smoother curve
  - Tests aimed at detecting whether there is increased risk at highest levels compared to moderate levels
    - Careful parameterization of a few linear splines

57

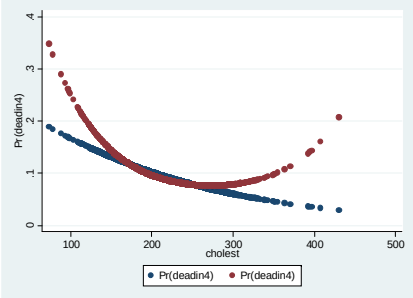
Quadratic Fit

```
. logistic deadin4 cholest
Logistic regression                               Number of obs   =       4953
                                                  LR chi2(1)      =      20.73
                                                  Prob > chi2     =     0.0000
Log likelihood = -1579.2394                       Pseudo R2       =     0.0065
------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholest |   .9943485   .0012533    -4.50   0.000     .9918951    .9968079
------------------------------------------------------------------------------

. g cholsqr= cholest^2
. logistic deadin4 cholest cholsqr
Logistic regression                               Number of obs   =       4953
                                                  LR chi2(2)      =      27.24
                                                  Prob > chi2     =     0.0000
Log likelihood = -1575.9847                       Pseudo R2       =     0.0086
------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholest |   .9748668   .0071995    -3.45   0.001     .9608578    .9890801
     cholsqr |   1.000047    .000017     2.74   0.006     1.000013    1.00008
------------------------------------------------------------------------------
```

58

Quadratic Fit

- Fitted values for linear fit and quadratic fit



59

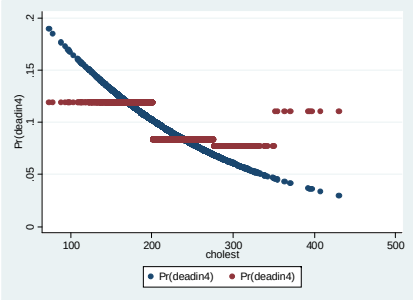
Dummy Variables

```
. logistic deadin4 i.cholctg
Logistic regression                               Number of obs   =       4953
                                                  LR chi2(3)      =      17.50
                                                  Prob > chi2     =     0.0006
Log likelihood = -1580.8527                       Pseudo R2       =     0.0055
------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholctg |
          2  |   .6749677    .066151    -4.01   0.000     .5570062    .8179109
          3  |   .6203559   .1473107    -2.01   0.044     .3895051    .9880266
          4  |   .9194561   .9772883    -0.08   0.937      .114496    7.38366
------------------------------------------------------------------------------

. test i4.cholctg=i3.cholctg
 ( 1)  - [deadin4]3.cholctg + [deadin4]4.cholctg = 0
           chi2(  1) =    0.13
         Prob > chi2 =    0.7168

. test i4.cholctg=i2.cholctg
 ( 1)  - [deadin4]2.cholctg + [deadin4]4.cholctg = 0
           chi2(  1) =    0.08
         Prob > chi2 =    0.7712
```

60

Dummy Variables

- Fitted values for linear fit and dummy variables



61

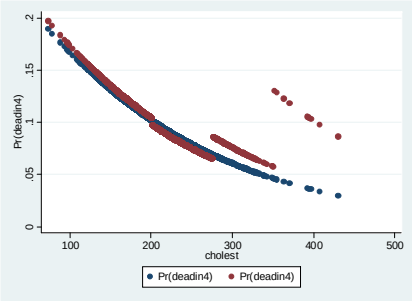
Dummy Variables (with Linear Term)

```
. recode cholest min/200=1 200/275=2 275/350=3 350/max=4, gen(cholctg)
. logistic deadin4 cholest i.cholctg
Logistic regression                               Number of obs   =       4953
                                                  LR chi2(4)      =      23.40
                                                  Prob > chi2     =     0.0001
Log likelihood =  -1577.904                       Pseudo R2       =     0.0074
------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholest |   .9941415    .002398    -2.44   0.015     .9894526    .9988526
             |
     cholctg |
           2 |   .9308006   .1530162    -0.44   0.663      .674413    1.284658
           3 |   1.264607   .4770091     0.62   0.534      .603786    2.648673
           4 |   3.129703   3.681596     0.97   0.332     .3120329    31.39105
------------------------------------------------------------------------------
.predict dummy
. test i2.cholctg i3.cholctg i4.cholctg
 ( 1)  [deadin4]2.cholctg = 0
 ( 2)  [deadin4]3.cholctg = 0
 ( 3)  [deadin4]4.cholctg = 0

           chi2(  3) =    3.00
         Prob > chi2 =    0.3922
```

62

Dummy Variables with Linear

- Fitted values for linear fit and dummy variables with linear



63

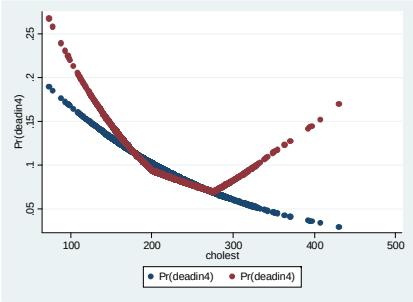
Linear Splines with Linear

```
. mkspline schol1 200 schol2 275 schol3 350 schol4= cholest
. logistic deadin4 cholest schol1 schol2 schol3 schol4
Logistic regression                               Number of obs   =       4953
                                                  LR chi2(4)      =      27.35
                                                  Prob > chi2     =     0.0000
Log likelihood = -1575.9303                       Pseudo R2       =     0.0086
------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholest |   1.005594   .0243539     0.23   0.818     .9589766    1.054478
      schol1 |   .9846408   .0240712    -0.63   0.527     .9385746    1.032968
      schol2 |    .990173   .0238553    -0.41   0.682     .9445041     1.03805
      schol3 |   1.001799   .0287906     0.06   0.950     .9469306    1.059848
      schol4 |  (omitted)
------------------------------------------------------------------------------

. test schol1 schol2 schol3

 ( 1)  [deadin4]schol1 = 0
 ( 2)  [deadin4]schol2 = 0
 ( 3)  [deadin4]schol3 = 0

           chi2(  3) =    7.67
         Prob > chi2 =    0.0533
```

64

Linear Splines

- Fitted values for linear fit and linear splines



65

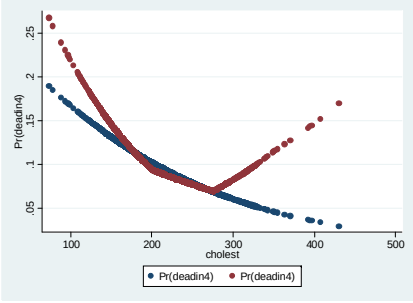
Linear Splines

```
. logistic deadin4 schol1 schol2 schol3 schol4

Logistic regression                               Number of obs   =       4953
                                                  LR chi2(4)      =      27.35
                                                  Prob > chi2     =     0.0000
Log likelihood = -1575.9303                       Pseudo R2       =     0.0086

------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      schol1 |   .9901491   .0026837    -3.65   0.000     .9849029    .9954231
      schol2 |   .9957122   .0026219    -1.63   0.103     .9905866    1.000864
      schol3 |   1.007404   .0083733     0.89   0.375     .9911252    1.023949
      schol4 |   1.005594   .0243539     0.23   0.818     .9589766    1.054478
------------------------------------------------------------------------------
```

66

Linear Splines

- Fitted values for linear fit and linear splines



67

Comments on Inference

- Only the quadratic fit had a significant test for nonlinearity (but the splines came close)
  - There was an efficiency advantage to the parsimony of having only a single parameter modeling “nonlinear”
  - It could not fit as flexible models, but it was flexible enough
- We need to distinguish between our confidence in a nonlinear trend and our confidence in there being a tendency for higher mortality at the highest cholesterol levels compared to medium levels
  - Estimates suggested the curve went back up
  - But neither the dummy variable based tests nor the spline based tests were statistically significant

68

Comments on Inference

- If I really wanted to trust the P values for tests of nonlinearity or the true U-shape, I would have needed to pre-specify a single test
- However, displaying the more flexible models does have descriptive value
- I can easily believe that the scientific community would have been satisfied learning that
  - There was a significant downward trend
  - There was a significant test of nonlinearity
  - Descriptively: mortality risk tended to go back up at highest levels
    - But there was not statistically significant evidence of that

69

Comments on Diagnostics

- Had I not anticipated the need to test for nonlinearity, I still would have done these sorts of analyses to examine my belief in a linear trend.
- Certainly, after exploring the data I would caution against using my regression model to estimate precise mortality at any particular cholesterol level

70

Department of Redundancy Department

- Bottom line for inferential questions about nonlinearity
  - I would use quadratic for its parsimony
- Bottom line for inferential questions about the mortality increasing again at high cholesterol levels
  - I would use the linear splines for their flexibility and ability to hone in on my question
- Bottom line for diagnostic exploration
  - I would probably just use the linear splines for plotting
  - I might even report the P values, noting that they could not be trusted due to multiple comparison issues in *ad hoc* analyses

71

Case Diagnostics

72

Detecting Unusual Cases

- When using regression models to explore associations between variables, we are always very interested in whether there are individual cases that behave somewhat differently than the bulk of the data

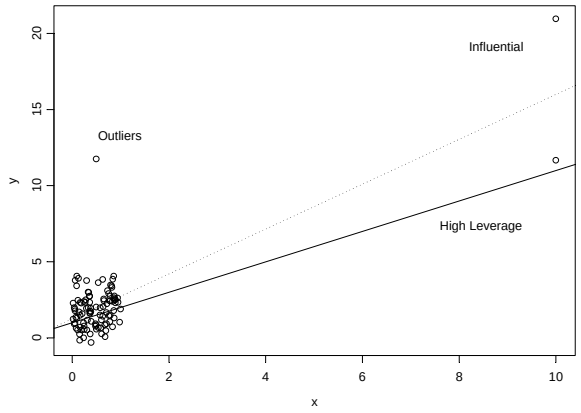
73

Detecting Unusual Cases

- Some cases may be poorly described by the overall regression model
  - "Outliers"
- Some cases may be overly influential in fitting the regression model
  - "Influential cases" affect estimates
  - "Highly leveraged cases" affect statistical significance

74

Outliers, Leverage, Influence



75

Outliers

- "Outliers" are cases whose response is far from that predicted by the model as judged by the residual
- Well developed for linear regression, providing you assume normally distributed data
- Consider how many SD a single case is from its group mean relative to the sample size of the data set
  - The expected magnitude of the largest residual is a function of n
- (Lacking anything else, still probably reasonable)

76

Multiple Comparisons

- We must consider the fact that we are looking at the largest and smallest residuals
- Essentially looking at all n residuals
- The residuals are nearly independent, so we can adjust for multiple comparisons using that approximation

77

Adjustments

- Compute a "p value" for each residual based on the t distribution
- Bonferroni: Compare the P value associated with the absolute value of each outlier to α / (2n)
- Modified Bonferroni: Use kα / (2n) as the threshold for the k-th largest residual (in absolute value)
- Assume independence: Use inverse binomial distribution to find threshold
  - In Stata: invbinomial (n, k, α / 2)

78

Outliers: Binary Data

- Binary response can be either 0 or 1
  - Residuals are similarly dichotomized
- If *0* < p < 1, there must sometimes be 0's and sometimes 1's
- This argues that looking at residuals is not very useful
  - Consider model fit instead

79

Detecting Influential Cases

- "Influential" cases are those cases which affect our inference too much
- Such cases can affect our inference by
  - Changing the scientific estimate of association markedly from what it would be if the case were not in the data set
  - Changing the strength of statistical evidence (e.g., P value) markedly from what it would be if the case were not in the data set

80

Conceptual Method

- Finding influential cases is conceptually quite easy
- In turn, leave each case out and see what happens
- There can, of course, be influential pairs (triples, etc.) of cases, but trying to detect these is hampered by the "curse of dimensionality"

81

Actual Methods

- In linear regression, influence of individual cases on the scientific estimates can be computed without fitting all the additional regressions
- In other forms of regressions, "one-step" approximations are often used to assess the approximate influence of a case
  - Do a single iteration away from the estimates from the full data

82

Stata: Linear regression

- In Stata, "predict" can be used to obtain statistics related to the influence of a case on the scientific estimate of association
- Linear regression:
  - dfbeta: the change in a slope parameter divided by the standard error of the slope
  - After performing a "`regress`" command
    - "`predict varname, dfbeta(pred)`"
  - Alternative form to produce dfbetas for every variable
    - "`dfbeta`"

83

Stata: Logistic

- After logistic regression, Stata will compute an omnibus statistic measuring the influence of a case
- After "logit" or "logistic"
  - "`predict varname, dbeta`"
- Pregibon's influence statistic
  - Large absolute values for dbetas suggests that deleting a case would affect the linear predictor
    - Affects the linear combination of the covariates

84

Stata: Logistic

- A more useful group of statistics is to get a user-written function ldfbeta

```
*
* Use user-written function ldfbeta to compute delta-betas
*
* If it is not installed on the machine you are using,
* connect to the internet and type "findit ldftbeta".
* Then click on the highlighted line that says
*
* ldfbeta from http://www.ats.ucla.edu/stat/stata/ado/analysis
*
* On the screen that opens, click on
*
* (click here to install)
```

85

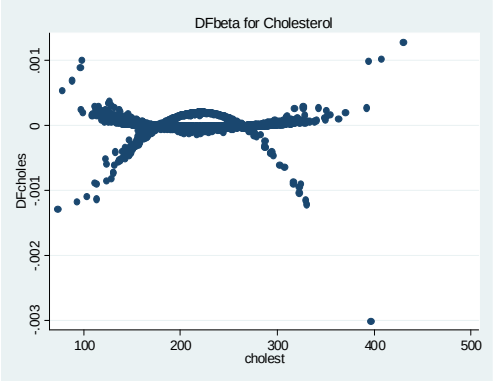
Example: Influence Using DFbeta

```
. logistic deadin4 cholest cholsqr age

Logistic regression                               Number of obs   =       4953
                                                  LR chi2(3)      =     198.21
                                                  Prob > chi2     =     0.0000
Log likelihood = -1490.4989                       Pseudo R2       =     0.0623

------------------------------------------------------------------------------
     deadin4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cholest |   .9808592   .0073151    -2.59   0.010     .9666263    .9953018
     cholsqr |   1.000035   .0000171     2.06   0.039     1.000002    1.000069
         age |   1.108365   .0086364    13.20   0.000     1.091567    1.125422
------------------------------------------------------------------------------

. ldfbeta
DFcholes:  DFbeta(cholest)
DFcholsq:  DFbeta(cholsqr)
DFage:     DFbeta(age)
```

86

DFbetas for Cholest vs Cholest

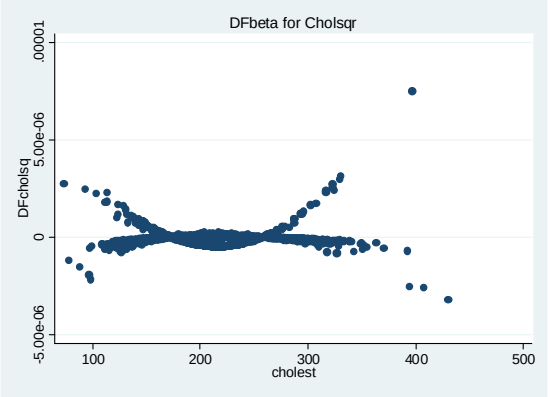
- Note a single case with cholesterol near 390 that seems to have a lot of influence



87

DFbetas for Cholsqr vs Cholest

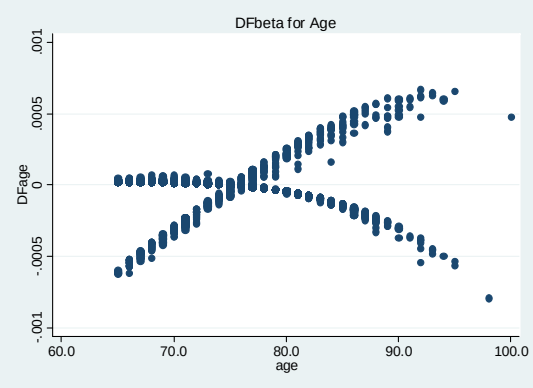
- Note a single case having cholesterol about 390 that seems to have more influence than other points



88

DFbetas for Age vs Age

- Greatest influence with the extremes of age



89

Detecting Influential Cases

- The influential case noted for cholest and cholsqr was an 80 yo male who had a cholesterol of 396 and who died within 4 years
- I would not delete this case, but it might be useful to consider how that case drives any inference we report
  - Does it "drive" the estimate? (influential)
  - Does it "drive" any statistical significance?(highly leveraged)

90

Detecting Influential Cases

- The DFbetas are looking at the change in the regression parameter estimate as each case is deleted
- Other approaches standardize this to look at the change in the Z statistic
- Personally, I would rather separate the scientific measures of influence from the statistical measures of influence
  - Scientific: Slope when each case is deleted
  - Statistical:P value when each case is deleted
- This generally requires programming
  - Unless there are just a few cases you want to consider

91

Influential Cases with Interactions

- Interactions can often appear statistically significant when some outlier is present in the data
- Interactions are often able to make a model fit the outlier better
  - There is often a very small sample size that is influencing a multiplicative interaction with dummy variables in particular
- But, I am very loathe to introduce an interaction into a model just to fit an outlier
- I often examine influence of cases whenever I consider interactions

92

Comments

- When you detect influential cases, you should check that there was not an error in measuring the case
  - If so, fix it
- If there is no error, then report both the estimates / inference when using all the data and the estimates / inference when omitting the case
  - The bottom line analysis is the one including all data
  - By reporting both estimates, however, you point out to your readers the lessened confidence you should have in the results

93