Biost 518
Applied Biostatistics II

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

Lecture 10:
Diagnostics

March 3, 2006

Lecture Outline

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

Multiple Regression

- General notation for regression model

$$g(\theta_i) = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \beta_3 \times W_{2i} + \cdots$$

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)

Maximal Assumptions

- Independence
- Sufficient sample sizes for asymptotic distributions to be a good approximation
- Variance appropriate to the model
- Regression model accurately describes summary measures across groups
- Shape of distribution same in each group

Detecting Linear Trend in g(θ)

- Independence
  - (between identified clusters for robust SE)
- Sufficient sample sizes for asymptotic distributions to be a good approximation
- Variance appropriate to the model
  - (relaxed for robust SE)

Estimating θ in Groups (not PH)

- Independence
  - (between identified clusters for robust SE)
- Sufficient sample sizes for asymptotic distributions to be a good approximation
- Variance appropriate to the model
  - (relaxed for robust SE)
- Regression model accurately describes summary measures across groups

Predicting Range of Y in Groups

- Independence
  - (between identified clusters for robust SE)
- Sufficient sample sizes for asymptotic distributions to be a good approximation
- Variance appropriate to the model
  - (NOT relaxed for robust SE)
- Regression model accurately describes summary measures across groups
- Shape of distribution same in each group
  - (Normal distribution for standard PI)

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

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

The Real Problem

"Blood suckers hide 'neath my bed"

- "Eyepennies", Mark Linkous (Sparklehorse)

Nonrepresentative Samples

- Problems often result because of data that we didn't sample
  - Recall "3 over N Rule"
    - Given a sample of size N, the upper 95% confidence bound on the proportion of the population not represented at all is 3/n
- There is nothing your data can tell you about whether the unsampled population might be different
  - Only your sampling scheme tells you this

Model Diagnostics

Assessing Independence

- We must have variables that identify clusters
  - 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

Assessing Asymptotic Distribution

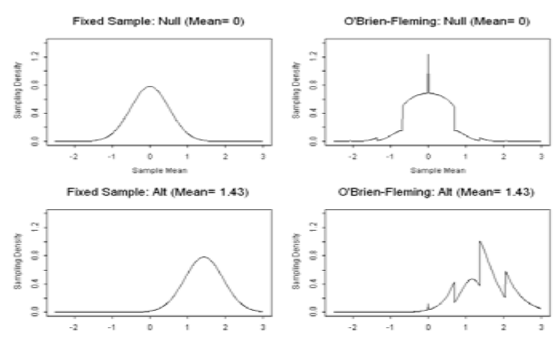
- We usually rely on an approximate normal distribution for regression parameters
  - Generally true in large samples
  - But, the definition of "large" depends on the shape of the distribution for the data
    - As a rule, "heavier tails" of response distribution requires larger sample size
      -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

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: StatExact)
- But some sampling schemes purposely alter the distribution of common statistics

Fixed vs Sequential Sampling

- Clinical trials often use a stopping rule



Assessing Appropriate Variance

- Classic linear regression: 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

Linear Regression Residuals

Model :

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

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

Error ε_i is estimated by residual

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

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

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)

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

Logistic, Poisson, PH Regr

- Assumptions about variance relate to mean variance relationships
  - 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)

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

Criteria

- Assess model fit by examining
  - Linear regression
    - Linearity of means
  - Logistic regression
    - Linearity of log odds
  - Poisson regression
    - Linearity of log rates
  - Proportional hazards regression
    - Linearity of log hazards

General Methods

- Nonparametric description within strata
  - Strata generally <u>not</u> 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

Ex: Hepatomegaly by Bili in PBC

- Examine log odds across strata

bilictg	N	Mn(bili)	Avail	Prop	Odds	Log odds
0.0 - 1.0	142	0.66	104	0.31	0.44	-0.51
1.0 - 2.0	107	1.34	77	0.44	0.79	-0.36
2.0 - 4.0	78	2.80	63	0.62	1.62	-0.21
4.0 - 8.0	48	5.69	38	0.76	3.22	-0.12
8.0 - 16.0	27	11.32	17	0.88	7.50	-0.05
16.0 - 32.0	16	19.54	13	0.85	5.50	-0.07

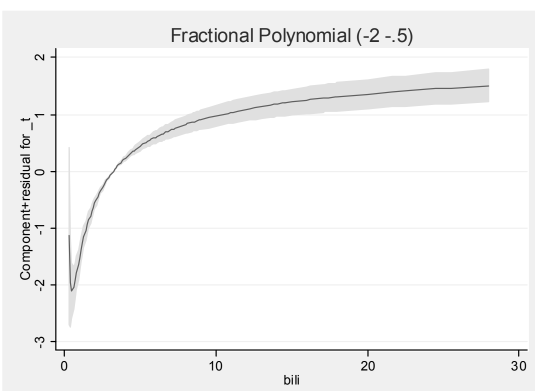
25

Ex: Survival and Bili in PBC

- Fit a flexible model
  - Examine pattern of fitted values versus predictor
- Using fractional polynomials in Stata
  - `. stset obstime status`
  - `. fracpoly stcox bili`
  - `. fracplot` (no longer documented)

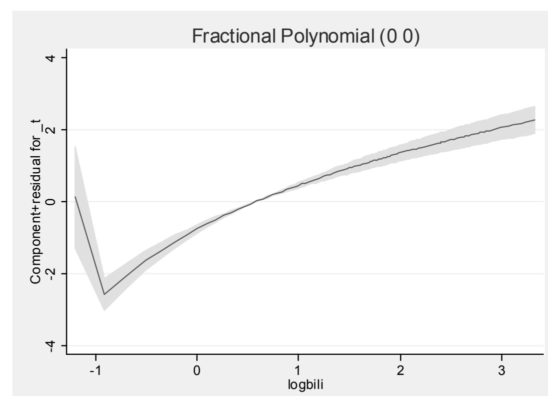
26

Ex: Survival and Bili in PBC



27

Ex: Survival and log(Bili) in PBC



28

Assessing Proportional Hazards

- Recall that in the proportional hazards model we use the regression model to
  - Borrow information across groups defined by the predictor
    - We assume the hazard ratio is linear in some modeled predictor(s)
  - Borrow information across time
    - We assume the hazard ratio is constant over time
  - The estimated standard errors in classical proportional hazards models depend on both of these assumptions

Graphical Method

- Log (- log) survival curves estimated for each stratum defined by levels of the predictor should look parallel
  - (And evenly spaced if linear in predictor)

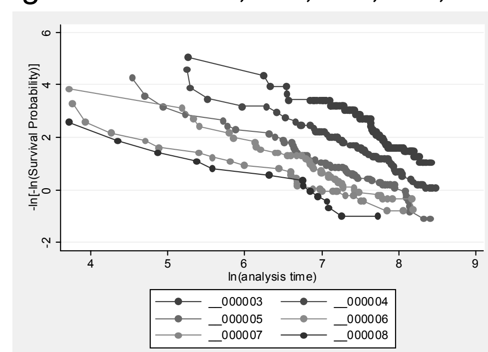
Stata Commands

```
. stset timvar eventind
. stphplot, by(stratvar)
```

- Produces a plot
  - log (– log (S(t))) vs log (t)
    - Why - log(-log (S(t)))?
      - Because. Why not?
    - Why log (t)?
      - If the survival times were truly Weibull distributed, then this plot would look like parallel straight lines

Ex: PBC Survival vs Bilirubin

- Categorized bili 0-1, 1-2, 2-4, 4-8, 8-16, 16+



Residuals Based Methods

- A number of methods for computing residuals have been described
    - Martingale residuals
    - Deviance residuals
    - Score residuals
    - Schoenfeld residuals
    - Cox-Snell residuals
  - The various forms of residuals differ somewhat in their ability to detect lack of linearity and/or nonproportional hazards

Stata: Schoenfeld Residuals

- Under proportional hazards, there should be no particular trend in the Schoenfeld residuals over time
  - Stata will produce plots and tests regressing these residuals over time

```
. stset timvar, fail(eventind)
. stcox pred1 pred2, scal(scalrsd) sch(schrsd)
. stphtest, detail
. stphtest, plot(pred1)
```


E: Survival vs log(Bili), Protime

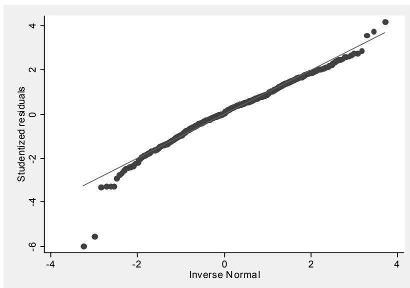
```
. stcox logbili protime, scal(scal*) sch(sch*)
. stphtest, detail
      Test of proportional hazards assumption
      Time:  Time
____________________________________________________
             |   rho     chi2   df        Prob>chi2
protime      | -0.449  14.77    1           0.0001
logbili      |  0.041   0.23    1           0.6317
global test  |         14.77    2           0.0006
```


Assessing Normality

- For normal based prediction intervals in linear regression, assess normality by looking at the residuals
  - Methods:
    - Histogram of residuals
    - QQ plot: Stata "qnorm"
      - Graph ordered residuals versus what we would expect from a normal distribution having the same mean and variance
      - Truly normal data approximates a straight line

Ex: log(FEV) vs age, loght

```
. predict stursd, rstu
. regress logfev smoker age loght if age>=9
. qnorm stursd
```



37

Because You Can't Stop Me

- The problem with all the model diagnostics
  - They may not detect problems that truly exist
    - Lack of power to prove "equivalence"
      - Need an infinite sample size
    - When assumptions do not hold, some data sets appear like the assumptions might be reasonable
  - Tendency to overfit the data
    - Inflated type I errors, anti-conservative CI

38

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)

39

Case Diagnostics

40

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

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

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)

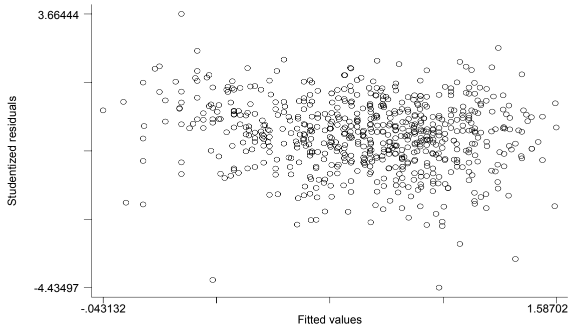
Multiple Regression Model

```
. regress logfev smoker age loght if age>=9
Number of obs =     439
Prob > F      =  0.0000
R-squared     =  0.6703
Root MSE      =  .14407

logfev |  Coef. StErr.      t   P>|t|    [95% CI]
smoker |  -.054  .0209  -2.56  0.011   -.095   -.012
   age |   .022  .0038   5.64  0.000    .014    .029
 loght |  2.870  .1301  22.06  0.000   2.614   3.125
 _cons |-11.095  .5201 -21.33  0.000 -12.117 -10.072
```

Example: FEV and Smoking

- Plot of residuals versus predicted values



Example: FEV and Smoking

- From residual plot we note extreme residuals
  - One large positive residual 3.664 standard deviations from 0
    - Based on the t distribution with 435 degrees of freedom, we would only expect 0.0139% of residuals to be this large if the log transformed FEV data were normally distributed within groups

Example: FEV and Smoking

- Large negative residuals -4.435, -4.215, and -3.593 standard deviations from 0
  - Based on the t distribution with 435 degrees of freedom, we would only expect 0.00058%, 0.00152% and 0.0182%, respectively, of studentized residuals to be this small if the log transformed FEV data were normally distributed within groups

Multiple Comparisons

- We must consider the fact that we are looking at the largest and smallest residuals
  - Essentially looking at all 439 residuals

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)

Example: FEV and Smoking

- Examples: Most extreme outliers of n=439 observations

		Adjusted Thresholds	
Extreme Residuals	Indiv P val	Worst Case Scenario	Independent Errors
-4.435	.0000058	.000057	.000058
-4.215	.000015	.000114	.000552
3.664	.000139	.000170	.001411
-3.593	.000182	.000228	.002488

FEV Example

- Applying the Bonferroni correction identifies four cases with extreme residuals, when we presume normally distributed residuals
  - But why do we think the FEV is lognormal within age, height, smoking groups?
  - Lack of effort would logically lead to skewed distribution of residuals

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

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”

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

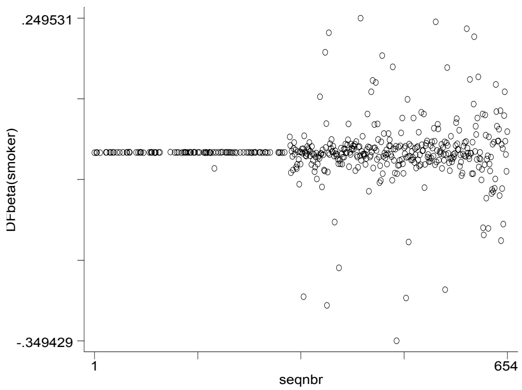
Stata: Linear regression

- 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`”

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

Ex: Influence in FEV Model



57

Ex: FEV data

- The dfbetas are the change in the t statistic associated with deleting a particular case
  - The t statistic for the smoking effect was -2.56 when using the entire dataset
    - The range of dfbetas from .25 to -.35 results in t statistics from -2.81 to -2.21 as individual cases are deleted
      - Critical value for a level .05 two-sided test based on t distribution with 434 degrees of freedom is 1.965

58

Detecting Influential Cases

- 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

59

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
  - But, I am very loathe to introduce an interaction into a model just to fit an outlier
  - I examine influence of cases whenever I consider interactions

60

FEV Example

- We could also consider sex, age, height interactions in the FEV data set
  - We find a statistically significant interaction between sex, age, and height
  - If we leave out the two cases with the large negative residuals, there is no statistically significant association
    - I choose to not model the interaction as it is likely driven largely by those outliers

Example: SEP "Normal Ranges"

- We consider the possibility of three way interactions between height, age, and sex
  - Osteoporosis affects women far more than men
    - Hence, we might expect the height - age interaction to be greatest in women and not so important in men

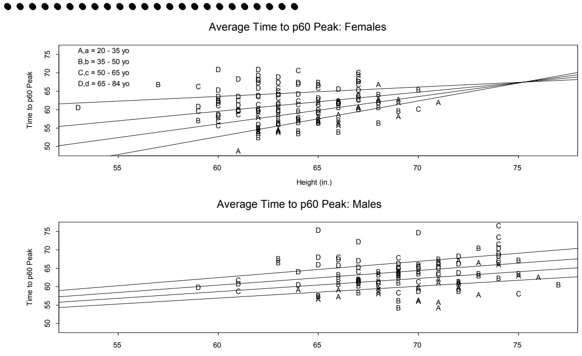
Example: SEP "Normal Ranges"

$$E(p60 \mid Ht, Age, Male) = \beta_0 + \beta_1 Ht + \beta_2 Age + \beta_3 Male$$
$$+ \beta_4 H.A + \beta_5 H.M + \beta_6 A.M + \beta_7 H.A.M$$

p60 - Height relationship for Age = a :

Sex	Intercept	Slope
F	$(\beta_0 + \beta_2 a)$	$(\beta_1 + \beta_4 a)$
M	$(\beta_0 + \beta_3 + (\beta_2 + \beta_6)a)$	$(\beta_1 + \beta_5 + (\beta_4 + \beta_7)a)$

Lines Predicted By Model



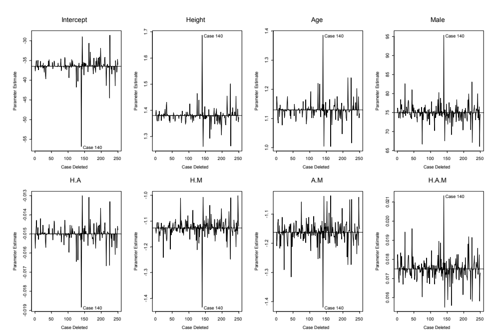
Example: SEP "Normal Ranges"

- From the inference, we find a statistically significant three way interaction
  - P= .0471

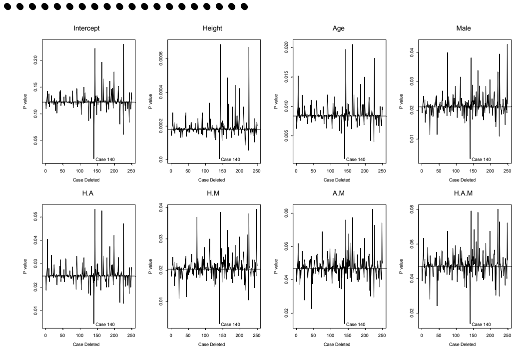
Example: SEP "Normal Ranges"

- I am now interested in ensuring that the evidence for an interaction is not based solely on a single person's observation
  - Hence, I consider 250 different regressions in which I leave out each case in turn
  - I plot the slope estimates and P values for each variable as a function of which case I left out
    - Case 0 corresponds to using the full data set

Influence on Estimated Parameters



Influence on P values

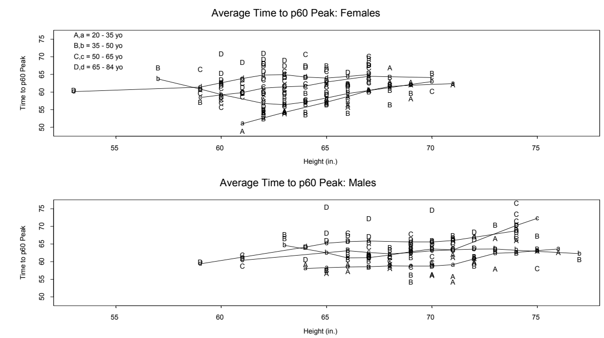


Example: SEP "Normal Ranges"

- Contrary to what I was afraid of, the only influential case actually lessened the evidence of an interaction
  - When Case 140 is removed from the data, the evidence for an interaction is a larger estimate and a lower P value
  - We can examine the scatterplot to see why Case 140 might be so influential

69

Stratified Scatterplots



70

Example: SEP "Normal Ranges"

- So now what do I do with Case 140
  - From the influence diagnostics, I now feel comfortable with the fact that the data really do suggest a three way interaction

71

Example: SEP "Normal Ranges"

- Personally, I do <u>NOT</u> remove the case from the dataset when making my prediction intervals
  - I do not know why Case 140 is so unusual
  - It is possible that people like her are actually more prevalent in the population than my sample would suggest
    - My best guess is that she represents 0.4% of the population, so leave her in

72