Homework 4
BIOST 515
Due: February 5, 2004 in lab

1. In this problem, we will explore how the distribution of the error term aeffects inferences. Assuming the regression model

$$y_i = 2 + 3x_i + \epsilon_i,$$

generate 500 samples of 100 x observations from a uniform(0,10) distribution. You will simulate the errors, ϵ_i, in the following 3 ways.

- From a skewed distribution. For this part, use a gamma with shape parameter = 1 and scale parameter = 2. Recenter your samples around zero (in other words, the mean should be zero).
- From a heavy-tailed distribution. For this part use a t distribution with 3 degrees of freedom.
- From a normal distribution with mean 0 and variance 4.

Do the following for each of the 3 error distributions:

(a) For each sample, calculate y and fit the least-squares estimates of the slope and intercept. Construct histograms of the sample values of $\hat{\beta}_0$ and sample values of $\hat{\beta}_1$. Discuss the shapes of these histograms.

(b) In what fraction of the samples does the joint confidence region for $\hat{\beta}$ contain the true values? What if you use the Bonferroni intervals? Comment.

(c) For the first sample, plot the residuals against the fitted values and create qq plots for the residuals. Comment.

Repeat the simulations with a sample size of 20. Comment on the differences, if any, in your results.

2. This is an extension of problem 1 from homework 3 (the patient satisfaction data).

(a) Obtain the jackknife residuals and indentify any potential outliers.

(b) Identify any high leverage points.

(c) Hospital management wishes to identify the mean patient satisfaction for patients who are 30 years old, whose index of illness severity is 59 and whose index of anxiety level is 2.0. How might you use leverage values to determine if estimating the fitted value will involve hidden extrapolation? Comment on whether you believe hidden extrapolation is present.

(d) Assess influence in this data set using Cook's D, DFFITS and DFBETAS. What do you conclude?

(e) Are there any obervations you might consider refitting the model without? If so, refit the model without the observation(s) and comment on the differnce in results.

3. The data KingCounty2001.data on the class web page contain a sample of data taken from birth certificates for children born in King County, WA in 2001. The data are restricted to singleton births (i.e. - no twins or triplets). A number of additional covariate measures are also available.

(a) The public program *First Steps* was implemented in the early 1990s to try and reduce the number of low birth weight infants. Babies that are born small are known to have a number of additional medical and developmental complications. Specifically, the goal of the *First Steps* program, authorized by the Maternity Care Access Act of 1989, was to provide "maternity care necessary to ensure healthy birth outcomes for low-income families." The legislation called for removal of unnecessary barriers to receiving prenatal care. Additional information about the program can be found in the report from DSHS on the web page. Use the King County data to summarize the evidence for the effectiveness of the program. Justify the methods that you use, summarize your conclusions, and state any limitations of your analysis. Also, include discussion of the potential covariates and why you chose the ones you did to inlcude in the analysis (for those of you who may be tempted, stepwise model selection is not a reason for including a predictor in this analysis). Examination of relationships between predictors, especially how they relate to your main predictor of interest may help with this analysis. Don't forget to perform the appropriate model diagnostics.

(b) (optional for extra credit) How would your analysis differ if the data were collected by oversampling women who are enrolled in the *First Steps* program? For example, what if women not in the program were sampled with probability 1/10 while women in the program were sampled with probability 1/3? You do not need to redo the anlaysis.

4. Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where the variance of ϵ_i is proportional to x_i^2. That is $\text{var}(\epsilon_i) = \sigma^2 x_i^2$.

   (a) Suppose we use the transformations $y^* = y/x$ and $x^* = 1/x$. Is this a variance-stabilizing transformation?

   (b) What are the relationships between the parameters in the original and transformed models?