STATS 500 - Homework 4

Due **Wednesday**, October 11, 2007

1. Based on Chapter 4, problem 1 (p. 97)

Using the `sat` dataset, fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. **Do not provide any plots about which you have nothing to say.** Present your diagnostics in a logical order, which may not match the order of the questions below.

- Check the constant variance assumption for the errors and for evidence of non-linearity via residual plots, and adjust model as appropriate
- Check the normality assumption.
- Check for large leverage points.
- Check for outliers.

Hints: You should start with a linear regression of `total` on `expend`, `salary`, `ratio` and `takers`. A diagnostic residual plot will reveal a non-linear relationship, which looks like a quadratic.

The next step is to discover which predictor has a non-linear relationship with the response. Notice that for this particular dataset, the plot of the residuals vs each predictor variable works the best for discovering the non-linear relationship.

Once you discover which predictor has a non-linear relationship with the response, you can add a quadratic term for that predictor to the model, and do all your diagnostic analysis for the new model.

Solutions to this problem should not exceed 5 pages.

2. (a) Verify rigorously that if the standard linear regression model holds, then

$$\mathrm{Var}(\hat{\epsilon}) = \sigma^2(I - H)$$

where H is the hat-matrix $X(X^\top X)^{-1}X^\top$.

(b) In your own works, explain why it may be (slightly) better to use the (internally) studentized residuals instead of the standard residuals for checking the constant variance assumption.