STATS 500 – Homework 3

Due in class Thursday, February 11

Part A: Using the `sat` data (see `help(sat)` for the description of variables):

1. Fit a model with `total` sat score as the response and `takers, ratio` and `salary` as predictors. Comment on the goodness of fit. Using this model, answer the questions 2-6:
2. Suppose you wish to claim that teachers' salary has a positive effect on the SAT scores. State the appropriate null and alternative hypothesis, the test statistic, the p-value, and your conclusion at significance level $\alpha = 0.01$.
3. Test the hypothesis that the variable `ratio` has an effect on the SAT scores in the full model.
4. Test the hypothesis $\beta_{takers} = \beta_{ratio} = \beta_{salary} = 0$. Explain in words what this hypothesis means.
5. Compute the 95% and 99% CIs for the parameter associated with `salary`. Using just these intervals, what can we deduce about the p-value for `salary` in the regression summary?
6. Compute and display a 95% joint confidence region for the parameters associated with `ratio` and `salary`. Add the origin to the plot. The location of the origin on the plot tells us the outcome of a certain hypothesis test. State that test and its outcome.
7. Now add `expend` (current expenditure per pupil) to the model and comment on the coefficients, their significance and the goodness of fit as compared to the model in question 1.
8. In the model of question 7, test the hypothesis $\beta_{salary} = \beta_{expend} = \beta_{ratio} = 0$. Based on your entire analysis, do you feel any of these predictors have an effect on the response?

   Solutions to Part A should be no longer than 3 pages.

Part B: Using the `teengamb` dataset:

Fit a model to predict gambling expenditure from all other available variables. Perform regression diagnostics on this model to answer the following questions. Display **only** those plots that are relevant to the questions below. Present your diagnostics in a logical order.

- Check the constant variance assumption for the errors. Modify the model if necessary (see below).
- Check the normality assumption.
- Check for large leverage points.
- Check for outliers.
- Check for influential points.
- Check the structure of the relationship between the predictors and the response.

Hints: You should start with a linear regression of `gamble` on `sex, status, income` and `verbal`. A diagnostic plot will reveal heteroscedasticity in residuals. A standard solution for the type of heteroscedasticity that you will see is to take the log or the square root of the response. Since in this case the `gamble` variable has some zero values, take the square root of `gamble` and fit a new model for this response. Then do all your diagnostic analysis for the new model.
Solutions to Part B should be no longer than 6 pages.

Part C: Confidence regions:

Consider a regression in which we regress Y on two predictor variables, A and B. For simplicity, assume that A and B are both standardized (mean 0, variance 1). Suppose we draw a joint confidence region for β_A and β_B. Explain, in your own words, why the confidence region will "lean to the left" if A and B are positively correlated, and will "lean to the right" if A and B are anti-correlated.
Hint: In each case (correlated and anti-correlated), think about whether it is easier to estimate $\beta_A+\beta_B$ or $\beta_A-\beta_B$. It might help to think about the case in which A and B are *perfectly* correlated or anti-correlated. Then think about what a confidence region would like like if $\beta_A + \beta_B$ is easy to estimate but $\beta_A - \beta_B$ is difficult to estimate, and vice versa.