STA 302f14 Assignment Eight[1]

This assignment assumes you are using the Formula sheet. There is a link on the course home page in case the one in this document does not work. The formula sheet (or part of it) will be provided with the quiz. **Bring your printouts for Question 12 to the quiz, including the plots.**

1. This question compares the error terms ϵ_i to the residuals $\widehat{\epsilon}_i$. Answer True or False to each statement. For statements about the residuals, show a calculation that proves your answer. You may use anything on the formula sheet.
   (a) $E(\epsilon_i) = 0$
   (b) $E(\widehat{\epsilon}_i) = 0$
   (c) $Var(\epsilon_i) = 0$
   (d) $Var(\widehat{\epsilon}_i) = 0$
   (e) ϵ_i has a normal distribution.
   (f) $\widehat{\epsilon}_i$ has a normal distribution.
   (g) $\epsilon_1, \ldots, \epsilon_n$ are independent.
   (h) $\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n$ are independent.

2. One of these statements is true, and the other is false. Pick one, and show it is true with a quick calculation. Start with something from the formula sheet.
   - $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\epsilon}}$
   - $\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\epsilon}}$
   - $\widehat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \widehat{\boldsymbol{\epsilon}}$

   As the saying goes, "Data equals fit plus residual."

3. The *deleted residual* is $\widehat{\epsilon}_{(i)} = Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{(i)}$, where $\widehat{\boldsymbol{\beta}}_{(i)}$ is defined as usual, but based on the $n-1$ observations with observation i deleted.
   (a) Guided by an expression on the formula sheet, write the formula for the Studentized deleted residual. You don't have to prove anything. You will need the symbols $\mathbf{X}_{(i)}$ and $MSE_{(i)}$, which are defined in the natural way.
   (b) If the model is correct, what is the distribution of the Studentized deleted residual? Make sure you have the degrees of freedom right.
   (c) Why are numerator and denominator independent?

4. For the general linear regression model, are $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ independent? Answer Yes or No and prove your answer.

5. For the general linear regression model, show that the squared sample correlation between $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ equals R^2. What does this imply about the plot of observed versus predicted values of the dependent variable?

[1] Copyright information is at the end of the last page.

6. For the general linear regression model, are $\widehat{\mathbf{Y}}$ and $\widehat{\boldsymbol{\epsilon}}$ independent?
   (a) Answer Yes or No and prove your answer.
   (b) What does this imply about the plot of predicted values against residuals?
7. For the general linear regression model, are $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ independent? Answer Yes or No and prove your answer.
8. For the general linear regression model, are $\mathbf{Y}$ and $\widehat{\boldsymbol{\epsilon}}$ independent? Answer Yes or No and prove your answer.
9. For the general linear regression model, calculate $\mathbf{X}'\widehat{\boldsymbol{\epsilon}}$. This will help with the next question.
10. For the general linear regression model,
    (a) Why does it not make sense to ask about independence of the independent variable values and the residuals?
    (b) Prove that the sample correlation between residuals and independent variable values must equal exactly zero.
    (c) Does this result depend on the correctness of the model?
    (d) What does the correlation between residuals and independent variable values imply about the corresponding plots?
11. In last week's analysis of the `Census Tract` data, you did a simultaneous test of `old`, `labor` and `income` controlling for the other variables. Here's a bit of my output.

```
> # Test old, labor and income
> redmodel = lm(crimerate ~ area+urban+docs+beds+hs)
> anova(redmodel,fullmodel)
Analysis of Variance Table

Model 1: crimerate ~ area + urban + docs + beds + hs
Model 2: crimerate ~ area + urban + old + docs + beds + hs + labor + income
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    135 19817
2    132 19792  3    25.683 0.0571  0.982
```

After controlling for other variables in the model, what proportion of the remaining variation is explained by `old`, `labor` and `income`? The answer is a number between zero and one that you can get with a calculator. Show some work.

12. Lecture slide set 7 used the `trees` data. Typing `help(trees)` at the R prompt gives more information. For this question, bring your R printouts to the quiz, *including the plots.*

    (a) Fit an ordinary model with two independent variables. How much of the variability in Volume is explained? You have to admit, that's pretty good.

    (b) Once you control for `Girth`, what proportion of the remaining variation in `Volume` is explained by `Height`? The answer is a number between zero and one that can be obtained from the default output (that is, the output of `summary`) using a calculator.

    (c) Once you control for `Height`, what proportion of the remaining variation in `Volume` is explained by `Girth`? The answer is a number between zero and one that can be obtained from the default output (that is, the output of `summary`) using a calculator.

    (d) Now let's look at the deleted Studentized residuals. One student made an excellent suggestion, which was to look at boxplots. Try `boxplot(varname)`, where `varname` is the name of the deleted Studentized residual. If you don't know what a boxplot is, look in the Wikipedia. This part is interesting, but it will not be on the quiz. Do you see one possible high outlier?

    (e) Now treat the deleted Studentized residuals as t-test statistics, with a Bonferroni correction to achieve a *joint* significance level of 0.05. What is the critical value? It's a number you get from R and display on your printout. This *could* be on the quiz.

    (f) Is there evidence of outliers? Answer yes or No.

    (g) Now plot predicted values against standardized residuals. Put a title on the plot. See `help(title)`. Do you see anything fishy, or perhaps wavy?

    (h) Now plot the independent variables in the model against the standardized residuals. It's a bit subejctive, but when I do this I see a curvilinear trend for one independent variable, but not for the other. Which one?

        Then I thought about it for a while. Finally, combining a bit of geometry with what little I know about trees, I came up with a model. This model has *one* independent variable, a function of Height and Girth, and it explains almost 98% of the variation in volume. The residual plots look pretty clean. Can you guess my model?
