Stat 615/415 (Regression) Homework #5

Regression Diagnostics (chap. 3 [3.1-3.7, 3.11]).

1. (**3.12**) A student does not understand why the sum of squares SSPE is called a *pure error sum of squares* "since the formula looks like the one for an ordinary sum of squares". Explain.

2. (**3.19**) A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against responses Y_i and found positive relation. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relation.

   (a) How could the differences arise? Which is the more meaningful plot?

3. (Computer project, **3.3**) Refer to the GPA data from the previous h/w assignments.

   (a) Plot residuals e_i against the fitted values $\widehat{Y}_i$. What departures from the standard regression assumptions can be detected from this plot?

   (b) Prepare a Normal Q-Q plot of the residuals and use it to comment on whether the data passes or fails the assumption of normality. Conduct the Shapiro-Wilk test for normality.

   (c) Test whether residuals in this regression analysis have the same variance.

   (d) Conduct the lack-of-fit test and state your conclusion.

4. (Computer project,) **Crime rate** data set is available on our Blackboard site.

   A criminologist studies the relationship between level of education and crime rate in medium-sized U.S. counties. She collected data from a random sample of 84 counties; X is the percentage of individuals in the county having at least a high-school dipoma, and Y is the crime rate (crimes reported per 100,000 residents) last year.

i	1	2	3	4	5	6	7	8	$\cdots$
Y_i	8487	8179	8362	8220	6246	9100	6561	5873	$\cdots$
X_i	74	82	81	81	87	66	68	81	$\cdots$

   A linear regression of Y on X is then fit to these data. Test:

   (a) normal distribution of residuals;

   (b) constant variance of residuals;

   (c) presence of outliers;

   (d) lack of fit.

5. For the "toy" example, consider a small data set

X	0	0	1	2
Y	0	2	2	3

   Try to do as much as you can by hand, without the use of a computer. The numbers are quite simple!

(a) Plot these data and draw the least squares regression line, which has the expression $y = 1 + x$.

(b) Compute all the residuals.

(c) Compute all sums of squares by hand, from their definitions:

$$\begin{aligned}
\text{SSTot} &= \sum_i (Y_i - \bar{Y})^2 \\
\text{SSReg} &= \sum_i (\hat{Y}_i - \bar{Y})^2 \\
\text{SSErr} &= \text{SSTot} - \text{SSReg} = \sum_i (Y_i - \hat{Y}_i)^2 \\
\text{SS}\textsc{pe} &= \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 \\
\text{SS}\textsc{lof} &= \text{SSErr} - \text{SS}\textsc{pe} = \sum_j \sum_i (\bar{Y}_j - \hat{Y}_j)^2
\end{aligned}$$

Then conduct the lack-of-fit test. Explain the result.