Quiz Practice Questions 4 (Attendance 8) for Statistics 512
Applied Regression Analysis
Material Covered: Chapter 9 Neter et al. and Kuhn

These are practice questions for the quiz. The quiz (not the practice questions) is worth 5% and marked out of 5 points. One or more questions is closely, but not necessarily exactly, related to one or more of these questions will appear on the quiz. These practice questions are *not* to be handed in. Quizzes are to be done *using Vista* on the Internet **before** 4am (West Lafayette time!) of the date of the quiz. Vista will *not* allow any quiz to be done late. It is *highly* recommended that you complete this practice quiz, by hand, *before* logging onto Vista. The quiz is an **individual** one which means that each student does this quiz by themselves without help from others.

Applied Linear Statistical Models
(Neter et al.) Questions.

Chapter	Problem(s)	hints
9, pages 392–399	9.6, 9.10, 9.16	Chemical Shipment data

(9.6) Chemical Shipment data, qz4-9-6-chem-partial

(a) *Partial Residual Plots.*
The partial residual plots are given in the SAS output.
One plot is $e(Y|X_1)$ vs $e(X_2|X_1)$ (TIME vs NUMBER)
The other plot is $e(Y|X_2)$ vs $e(X_1|X_2)$ (TIME vs WEIGHT)

(b) *Interpretation of Partial Residual Plots.*
The attached partial residual plots for the two regression coefficients (ignore the one against the intercept) both indicate a linear term in the second variable (either X_1 or X_2) may be a useful addition to the regression model already containing the first variable (either X_2 or X_1). Both plots also seem to reveal one outlying point.

(c) *Regression of Partial Residuals.*
The regression of Y on X_1 is $\hat{Y}(X_1) = -2.0558 + 8.10973X_1$
and so the partial residuals $e_i(Y|X_1) = Y_i - \hat{Y}_i(X_1)$ (which is plotted above)

The regression of X_2 on X_1 is $\hat{X}_2(X_1) = -1.05916 + 0.85472X_1$
and so $e_i(X_2|X_1) = Y_i - \hat{Y}_i(X_1)$

The regression of the residuals are $e(Y|X_1) = 5.07969e(X_2|X_1)$
This is the same partial residual plot as one of the partial residual plots given in part (a); it appears as the "TIME versus WEIGHT" plot.

(9.10) Chemical Shipment data, qz4-9-10-chem-leverage

(a) *Studentized Deleted Residuals For Y Outliers.*
The studentized deleted residuals (RStudent in SAS output) are
$0.4268, -0.8005, 0.4815, \ldots, -1.0258, -0.6141$.
The studentized deleted residual associated with observation 12, $t_{12} \approx 3.626$, appears to be a Y outlier because it is over three (studentized deleted residual) standard deviations (3.626 to be exact) in size. Use the Bonferroni outlier test procedure to investigate this.

1. *Statement.*
   H_0 : (largest) observation Y_{12} is *not* an outlier versus
   H_{12} : it is an outlier
2. *Test.*
   The test statistic is $|t_{12}| = |3.6262| = 3.6262$
   The Bonferroni critical value at $\alpha = 0.05$ is
   $t(1 - \alpha/2n; n - p - 1) = t(1 - 0.05/2(20); 20 - 3 - 1) = 3.58$
   (Use PRGM INVT ENTER 16 ENTER 0.99875 ENTER)
3. *Conclusion.*
   Since the test statistic, 3.6262, is larger than the critical value, 3.58, we (choose one) **accept** / **reject** the null hypothesis; that is, the data indicates observation Y_{12} *is* an outlying observation.

(b) *Rule of Thumb For X Outliers Using Leverage Values.*
If $h_{ii} > 2p/n = 2(3)/20 = 0.3$ (Hat Diag in SAS output),
then the corresponding X values are outliers.
Since $h_{77} = ? > 0.3$, X_7 appears to be an outlying X observation[1].

(c) *New X Outlier According To Leverage Values?*
Looking at the attached scatter plot of number versus weight, the observation $\mathbf{X}'_{new} = [1, 15, 8]$ does not look an outlier; the point $(15, 8)$ appears to within the scatter of points.
Also, from SAS,

$$h_{new,new} = \mathbf{X}'_{\mathbf{new}}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}_{\mathbf{new}} = 0.265$$

which, although a little bit large, is within the range of the other h_{ii} leverage values and so no extrapolation is involved in this case.

[1]Notice that this procedure identifies outlying X values, whereas the previous Bonferroni outlier procedure identifies outlying Y values.

(d) *Influential observations.*

i	$(DFFITS)_i$	D_i	$(DFBETAS)_{k(i)}, b_0$	$(DFBETAS)_{k(i)}, b_1$	$(DFBETAS)_{k(i)}, b_2$
7	0.179	0.0113	?	-0.057	0.112
12	0.844	0.1385	0.354	-0.122	?

According to all three criteria, neither observations 7 nor 12 are influential.

$(DFFITS)_i$:
Neither observation 7 nor 12 have absolute $(DFFITS)_i$ values larger than one (1) and so are not considered influential observations in this sense.

Cook's D_i:
Neither observation 7 nor 12 have Cook's D_i values with associated $F(p, n-p) = F(3, 20-3) = F(3, 17)$ percentiles larger than 0.50 ($P(F < 0.0113) = 0.0017$ and $P(F < 0.13385) = 0.061$) and so are not considered influential observations in this sense.

$(DFBETAS)_{k(i)}$:
Neither observations 7 nor 12 have absolute $(DFBETAS)_{k(i)}$ values for any of the three regression coefficients larger than one (1) and so are not considered influential observations in this sense.

(e) *Percent Difference Fitted Values With and Without Influential Observations*
From SAS, without observation 7,

$$\frac{\sum_{i=1}^{n} \left| \frac{\hat{Y}_{i(7)} - \hat{Y}_i}{\hat{Y}_i} \right| 100}{n} = 0.17\%$$

Without observation 12, it is ?%. Since both are smaller than 5%, both cases do not exercise undue influence on the regression.

(f) *Cook's Distance D_i.*
Observation 12 has the largest Cook's D_i value, but, as discussed above, this value is not large enough to be considered influential.

(9.16) Chemical Shipment data, qz4-9-16-chem-vif

(a) *Correlation matrix*

$$\begin{bmatrix} 1 & 0.97 & ? \\ 0.97 & ? & 0.93 \\ 0.97 & 0.93 & 1 \end{bmatrix}$$

which has large correlation values. This indicates that the response, Y, is highly correlated with both predictor variables, X_1 and X_2, which is "good". This also indicates the two predictor variables are highly correlated with one another, which is "bad".

(b) *Variance Inflation Factors.*
The variance inflation factors for the two regression coefficients in the model are both

$$VIF_k = 7.02753, \quad k = 1, 2$$

which indicates the multicollinearity, although it exists, is not significant, since $VIF_k = 7.02753 \leq 10$.