Stat 440/540. Regression Analysis. Name: ____________________

Fall 2006 Solutions

Midterm Exam (Solutions)

1. Table 1 in the output pages shows some data on total fat in grams and calories of breakfast sandwiches for a few fast food restaurants. There is an interest in understanding the relationship between calories (Y) and total fat (X).

   Here are a few summary statistics about this data, $\bar{X} = 24.875$, $\bar{Y} = 421.5$, $\sum(x_i - \bar{X})^2 = 688.875$, $\sum(y_i - \bar{Y})^2 = 111046$, $\sum X_i Y_i = 91924$, $SSE = 17081$.

   (a) Based on Figure 1 in the output pages, discuss on the relationship between calories and total fat. In particular, discuss about the appropiateness of a linear regression model to describe this data.

   As total fat increases, calories also increases. It seems that proposing a linear relationship to this data is appropiate except perhaps for data point associated to "Burger King" (43,620) which makes us think more of polynomial relationship between X and Y.

   (b) Find the least squares estimates of the two coefficients of the simple linear regression model

   Notice that $S_{xy} = \sum X_i Y_i - n\bar{X}\bar{Y} = 91924 - 8(24.875)(421.5) = 8045.5$. *Therefore* $\hat{\beta}_1 = 8045.5/688.875 = 11.679$ *and* $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 421.5 - (11.679)(24.875) = 130.9849$.

   (c) Find the value of the t-statistic for β_1 and used to test statistical significance of this intercept.

   $MSE = SSE/(n-2) = 17081/6 = 2846.83$ *and then* $SE(\hat{\beta}_1) = \sqrt{2846.83/688.875} = 2.033$. *The t-statistic value is* $T = \hat{\beta}_1 / SE(\hat{\beta}_1) = 11.679/2.033 = 5.744$ *which is significant at an* $\alpha = 0.05$ *level.*

   (d) For this data, determine the values of the analysis of variance table. Fill in the blanks.

   Realize that two of the entries in the table are already given $SST = 111046$, *and* $SSE = 17081$. *We have a total of* $n = 8$ *observations*

Source	df	Sum of Squares	Mean Squares	F
Regression	1	93965	93965	33.006
Error	6	17081	2846.833	
Total	7	111046		

(e) Now discuss on the appropiateness of the regression fit based on the residual plots that appear in Figures 2, 3 of the output pages. Discuss if there are any violations to usual model assumptions. Is there a need for variable transformations?

These residuals plots look ok given the fact that we only have 8 data points. There doesn't seem to be serious violation to normality or non-constant variance assumptions so no transformations are needed. Still, more data might be needed to understand better the relationship between these variables and detect problems with residuals

(f) Find a 95% prediction interval for Y given that $x = 26$. Give an interpretation of this interval in terms of the context of the data. Here are some quantiles of the t-distribution $t(.95, 8) = 1.86, t(.975, 8) = 2.31, t(.95, 7) = 1.89, t(.975, 6) = 2.45, t(.995, 6) = 3.71$

First find the point estimate of the prediction at $x = 26$*, this is* $\hat{Y} = 130.9849 + (11.679)(26) = 434.6389$. *Also,*

$$SE(pred) = \sqrt{2846.33(1 + 1/8 + (26 - 24.875)^2/688.875)} = 56.633$$

Since $n = 8$ *and we wish a .95 probability level, we use* $t(.975, 6) = 2.45$. *Therefore, the prediction interval is* $434.6389 \pm 2.45(56.633)$ *which is* $(295.89, 573.39)$. *With 95% probability the calories of a breakfast sandwich with* $x = 26$ *grams of fat is within these two limits.*

2. Suppose we have a sample of 12 discount department stores that advertize on television, radio, and in the newspapers. The variables X_1, X_2 and X_3 represent the respective amounts of money spent on these advertising activities during a certain month while y gives the store's revenues during that month. Regression output on this data is presented in the additional pages. The actual data is not included.

(a) Write down the fitted regression model equation.

From the output pages, we get that the estimated regression equation is

$$\hat{Y} = -15.3 + 2.620X_1 + 7.556X_2 + 1.901X_3$$

(b) What are the values of R^2 or $\bar{R}^2$? Give a brief interpretation of these values.

Again from the output pages, we have $R^2 = 0.32$ *and* $\bar{R}^2 = 0.065$. *This low values means that very low variability of the store's revenues is explained by a regression model including these 3 predictor variables.*

(c) Comment on the significance for each coefficient. Use $\alpha = 0.05$.

None of the coefficients are significant at an $\alpha = 0.05$ *level. From the table of coefficient estimates, we can see that the minimum p-value is 0.139 which is greater than 0.05*

(d) With the information given, can you test the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$? Why or why not? Justify your answer.

Yes, the F-statistic part of the model ANOVA table allows us to test this hypothesis. F=1.26, with a p-value of 0.353, so we do not reject the null hypothesis

(e) Find the appropiate statistic to test whether the reduced model $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$ is an adequate explanation of the data as compared to the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$.

$F = \frac{SSR(X_2|X_1)}{MSE(X_1,X_2)} = 1193.1/386.5 = 3.086$. *Also, notice that*

$MSE(X_1, X_2) = (SSR(X_3|X_1, X_2) + SSE(X_1, X_2, X_3))/(1+8) = (19.0+3459.7)/9 = 386.5$

(f) Why is observation no. 7 reported as an usual observation? Justify your answer.

Basically because it has a high standardized residual. $|r_7| > 2$. *You could have also argued this in terms of Cook's distances but not using leverages.*

(g) Figure 4 from the output shows the added variable plot for adding variable X_3 to a model that already contains variables X_1 and X_2. Interpret the plot.

After including variables X_1, X_2 *in the regression, the added variable plot does not show any relationship between* Y *and* X_3*. Therefore,* X_3 *would not contribute any additional information to a regression model including* X_1 *and* X_2

(h) With this information, is it possible compute the sample partial correlation of $r_{y3\cdot12}$. Why or why not? If yes, find this value.

Yes it is possible to compute this partial correlation,

$$r_{y3\cdot12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = 19.0/3478.7 = 0.00546$$