Practice Final Questions

1) Answer the following questions. *Justify* your answers *briefly*.

a) In a multiple regression involving the dependent variable $Y$ and the independent variables $X_1, X_2, X_3$ and $X_4$, the p-value for the F-statistic was 0%. However, the p-values for the t-statistics associated with $X_3$ and $X_4$ were larger than 50%, while the p-values for the t-statistics associated with $X_1, X_2$ were 0%.
   (i) Possible (ii) Impossible (iii) Not enough information to answer the question

b) There are four possible explanatory variables $A, B, C$ and $D$ which are being considered in building a multiple regression model for a dependent variable $Y$. Suppose the following two models are fit based upon a sample of $n$ observations.
   Model I: $Y = \beta_0 + \beta_1 B + \beta_2 C + \varepsilon$ with an associated r-squared value denoted by $R^2(I)$
   Model II: $Y = \beta_0 + \beta_1 B + \beta_2 C + \beta_3 D + \varepsilon$ with an r-squared value denoted by $R^2(II)$
   Then,
   (i) $R^2(I) \geq R^2(II)$
   (ii) $R^2(I) \leq R^2(II)$
   (iii) Not enough information to answer the question.

c) Your friend fits a simple regression model of $Y$ on $X$. He makes sure that there are no violations of the regression assumptions and then examines the estimated slope, which he finds to be 0.0003. He says “Since this estimated slope coefficient is so small and close to zero, there is no evidence that $Y$ and $X$ have a linear relation”. Is your friend correct? *(Justify your answer briefly)*
   (i) Yes (ii) No

2) An analyst is interested in modeling the monthly bank charges that companies pay. He takes a random sample of 30 companies in a city and collects data on the following variables:
   - *charges* = monthly bank charges of the company (in $)
   - *sales* = yearly sales of the company (in millions of $)
   - *disburse* = average daily number of disbursements
   - *deposits* = average daily number of deposits
   - *banktype* = 1 if the company banks with a local bank
     = 0 if the company banks with a subsidiary of a foreign bank
   The relevant regression output follows:
The regression equation is
\[
\text{charges} = -762 + 13.4 \times \text{sales} + 0.828 \times \text{disbursement} + 3.41 \times \text{deposit} - 3606 \times \text{banktype}
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-762.3</td>
<td>426.0</td>
<td>-1.79</td>
<td>0.082</td>
<td></td>
</tr>
<tr>
<td>sales</td>
<td>13.3626</td>
<td>0.9527</td>
<td>2.3</td>
<td></td>
<td>2.3</td>
</tr>
<tr>
<td>disburse</td>
<td>0.8281</td>
<td>0.4048</td>
<td>1.8</td>
<td></td>
<td>1.8</td>
</tr>
<tr>
<td>deposit</td>
<td>3.4124</td>
<td>0.7685</td>
<td>3.4</td>
<td></td>
<td>3.4</td>
</tr>
<tr>
<td>banktype</td>
<td>-3606.3</td>
<td>647.5</td>
<td>1.0</td>
<td></td>
<td>1.0</td>
</tr>
</tbody>
</table>

S = 1114  R-Sq = 92.2%  R-Sq(adj) = 91.3%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>5</td>
<td>514454216</td>
<td>103.65</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Residual Error</td>
<td>43</td>
<td>43429481</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
<td>557883697</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Is there evidence that the model is useful?
b) The analyst feels that companies which do business with local banks pay a different amount on average in bank charges than companies doing business with subsidiaries of foreign banks, everything else being constant. Is there evidence to support the analyst’s suspicion? Justify your answer based on an appropriate test of hypothesis.
c) Is there a problem with multicollinearity in the X variables? Justify your answer in one sentence.
d) A company that has sales of $200 million, 1100 daily disbursements on average, 500 daily deposits on average and does business with a subsidiary of a foreign bank has monthly bank charges of $8,200. Based on the regression output, does this amount seem excessively high? Justify your answer briefly.

3) A clothes designer wants to introduce a new line of dress shirts for men. He wants to categorize the sizes of his shirts on the basis of the two following measurements:
   i) The size of the person’s collar measured in inches
   ii) The length of the right arm of the person, measured from shoulder tip to the wrist, in inches and wants to be able to predict the shirt’s other measurements, such as shoulder width, based on these two variables. He hires a consultant to help him. The consultant takes a random sample of 30 men and obtains data on the following three variables:
   Y = Shoulders=width in inches of the person’s shoulders from the left shoulder tip to the right shoulder tip
   X1 = Collar=The size of the person’s collar measured in inches
   X2 = RghtArm=The length of the right arm of the person, measured from shoulder tip to the wrist, in inches

On running the regression analysis, the consultant gets the following results:
The regression equation is
Shoulder = 1.90 + 0.748 Collar + 0.244 RghtArm

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.900</td>
<td>1.178</td>
<td>1.61</td>
<td>0.118</td>
</tr>
<tr>
<td>Collar</td>
<td>0.74840</td>
<td>0.05228</td>
<td>14.31</td>
<td></td>
</tr>
<tr>
<td>RghtSlee</td>
<td>0.24425</td>
<td>0.03771</td>
<td>6.48</td>
<td></td>
</tr>
<tr>
<td>S = 0.2826</td>
<td>R-Sq = 89.5%</td>
<td>R-Sq(adj) = 88.8%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Analysis of Variance
<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>18.4763</td>
<td>9.2382</td>
<td>115.64</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>27</td>
<td>2.1570</td>
<td>0.0799</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>20.6333</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Is there evidence that the model is useful?
b) Are each of the two explanatory variables useful in predicting shoulder width on an individual basis, in the presence of the other X variables?
c) The consultant explains the basic concepts of regression to the designer and of how her regression model works. The designer gets very excited and says “I get it! If I want to stitch shirts for men whose collar size is 15.5 inches and right arm length is 22 inches, such that these shirts will be a fairly comfortable fit for a large majority of men with such collar and sleeve measurements, then I should use a shoulder width measurement of 1.90 + 0.748*(15.5) + 0.244*(22) = 18.862 inches.

Based on the model, approximately what proportion of men with sizes 15.5 and 22 will have shoulder widths that are less than 18.862 (You may assume that the values of 15.5 and 22 are not extrapolations)? Do you thus think that the designer is using his model wisely for his intended purpose? Justify your answer. If you do not agree with the designer’s proposal, suggest an alternative calculation which you feel may be better in this situation (For your suggestion, you need not compute a number, you can just describe in words what you would do. You can answer this in a couple of lines)
e) By now, the designer is excited with the use of regression analysis to help him obtain the measurements of the shirts. He then says “Since the length of the right arm has proved to be useful in modeling the shoulder width, why do we not also include the length of the left arm in the model in order to obtain a model which will provide substantially better predictions?” Do you think the designer is right? Your answer must be either “yes” or “no”. Justify your answer clearly yet briefly.
Solutions

1)  
a) (i) Yes, $X_1$ and $X_2$ have useful information about $Y$ in the presence of the other $X$ variables, but $X_3$ and $X_4$ do not.

b) (ii) since adding an extra variable (in this example $D$) can never decrease the r-sq.

c) No. The estimated slope coefficient depends on the units of the measured data. Hence, one cannot determine whether the coefficient is different from zero by merely looking at the estimated coefficient but by looking at the t-statistic, which does not depend on what units the variables are measured in.

2)  
a) Since the p-value for the F-stat is 0, there is evidence that model is useful.

b) We need to test: $H_0: \beta = 0$ vs. $H_a: \beta \neq 0$

The t-stat is $-3606/647.5 = -5.56$ and thus there is evidence to reject the null.

c) No, since the VIF are all much less than 8 (or 9)

d) The predicted bank charge is $-762 + 13.4 \times 200 + 0.828 \times 1100 + 3.41 \times 500 - 3606 \times 0 = $4533.8

The upper end point of a 99.7% PI will be $4533.8 + 3 \times 1114 = $7875.8 and thus a charge of $8,200 is excessive. (It is OK if you answered with a 95% PI).

3)  
a) yes, since the p-value for the F-stat is zero

b) Yes, as the p-value that one gets for each of the t-stats on the two $X$ variables is zero.

c) The value of the fitted line is the estimated mean $Y$ value. Hence, only approximately 50% of men with collar size 15.5 and right sleeve length 22 inches will have shoulder widths less than 18.862. This is clearly not useful. I would find a value such that approximately say 90% of shoulder widths will be less than that value, so that 90% of men with collar size 15.5 and sleeve length 22 will fit into shirts with such a size. This can be computed as $18.862 + 1.28 \times (0.2826)$.

d) No, as left sleeve length will match right sleeve length. If you included it, you would just induce multi-collinearity into the model.