Suaray **11.21.06**

MATH 410/510: REGRESSION ANALYSIS- PRACTICE MIDTERM 2

You are allowed 1.25 hours to complete this exam. There are a total of 50 possible points. You must show all work to get credit. Print out all relevant SAS code and output. Look through the entire exam before you start. Please do not hesitate to raise your hand if you have a question.

1. a. Express the two sample problem as a specific instance of the multiple regression model. Clearly identify the design matrix, the response variables, and the parameters.
   b. Find the least squares estimates for the parameters using the usual method.

   SOLUTION

a. $\vec{y} = X\vec{\beta} + \vec{\varepsilon}$ is our model, where $\vec{y} = \begin{bmatrix} y_1 \\ . \\ y_m \\ \hline y_{m+1} \\ . \\ y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 & 0 \\ . & . \\ 1 & 0 \\ \hline 0 & 1 \\ . & . \\ 0 & 1 \end{bmatrix}$, $\vec{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$, and $\vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ . \\ \varepsilon_m \\ \hline \varepsilon_{m+1} \\ . \\ \varepsilon_n \end{bmatrix}$.

Please see notes.

b. Now solve $\vec{b} = (X'X)^{-1}X'\vec{y}$. You will get the sample mean from each group

2. The Federal Trade Commission (FTC) annually ranks varieties of domestic cigarettes according to their tar, nicotine, and carbon monoxide contents. The U.S. surgeon general considers each of these three substances hazardous to a smoker's health. Past studies have shown that increases in the tar and nicotine contents of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke. The data set (see http://www.amstat.org/publications/jse/v2n1/datasets.mcintyre.html) lists tar, nicotine, and carbon monoxide contents (in milligrams) and weight (in grams) for a simple of 25 (filter) brands tested in a recent year. Suppose we want to model carbon monoxide content, y, as a function of tar content, x_1, nicotine content, x_2, and weight ,x_3.
   a. Perform a regression analysis on the data using SAS. Are any of the variables significant?
   b. Calculate the variance inflation factors by finding the R_j values. What does this indicate about multicollinearity?
   c. What is your choice of a final model? Use forward selection to assist you in making your choice. Does a log transformation increase the coefficient of determination R^2 for your model?

SOLUTION

a. The REG Procedure

Model: MODEL1

Dependent Variable: y

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	495.25781	165.08594	78.98	<.0001
Error	21	43.89259	2.09012		
Corrected Total	24	539.15040			

Root MSE	1.44573	R-Square	0.9186
Dependent Mean	12.52800	Adj R-Sq	0.9070
Coeff Var	11.53996		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.20219	3.46175	0.93	0.3655	0
x1	1	0.96257	0.24224	3.97	0.0007	21.63071
x2	1	-2.63166	3.90056	-0.67	0.5072	21.89992
x3	1	-0.13048	3.88534	-0.03	0.9735	1.33386

The only variable that is significant is x1, while the overall F test indicates that the whole model is significant.

b. For example, to find VIF1, the code we need is

```
proc reg data=cig;
model x1=x2 x3;
run;
```

from which we get an R-squared value of 0.9537745. The VIFs larger than 10, along with the insignificant variables for an overall significant model point to multicollinearity.

c. Based on forward selection with SLE=0.1, we get

```
proc reg data=cig;
model y=x1 x2 x3/selection=f sle=.10;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	25
Number of Observations Used	25

Forward Selection: Step 1

Variable x1 Entered: R-Square = 0.9168 and C(p) = 0.4672

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	494.28131	494.28131	253.37	<.0001
Error	23	44.86909	1.95083		
Corrected Total	24	539.15040			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.74328	0.67521	32.20226	16.51	0.0005
x1	0.80098	0.05032	494.28131	253.37	<.0001

Bounds on condition number: 1, 1

No other variable met the 0.1000 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x1	1	0.9168	0.9168	0.4672	253.37	<.0001

We select the model with x1 alone. This corroborates the conclusion we would have reached had we looked at all 8 sub models and kept an eye on
1. VIFs
2. Significance of each regressor via partial t tests
3. Overall R-squared values.
The following code will allow us to do this:

```
proc reg data=cig;
model y=x1 x2 x3;
model y=x1 x2;
model y=x1 x3;
model y=x2 x3;
model y=x1;
model y=x2;
model y=x3;
run;
```

As for the log transformation, the following code will create a new transformed variable, which, after we perform a proc reg with it being the dependent variable, we see that the already high R squared does not increase by much at all.

```
data cigl;
set cig;
ly=log(y);
run;
```

3. a. There are a number of ways that we can detect outliers. Give a formula (or definition) of the following statistics, and describe the rules of thumb we use for outlier detection: i. Leverage ii. Cook's D iii. Studentized residuals

Discussed in class

b. Draw a sample graph illustrating a violation of the following assumptions concerning the residuals α : i. $Var(\varepsilon_i)=\sigma^2$, for all i (i.e. heteroscedasticity). ii. ε_i~N$(0,\sigma^2)$ for all i.

Discussed in class

4. Use stepwise regression to determine the best model for the asphalt data. Use an α level of SLE=SLS=0.15. Are there any potential outliers?

SOLUTION

4. Model: MODEL1

Dependent Variable: Y

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	994.00414	165.66736	10.67	<.0001
Error	24	372.51430	15.52143		
Corrected Total	30	1366.51844			

Root MSE	3.93972	R-Square	0.7274
Dependent Mean	6.50710	Adj R-Sq	0.6592
Coeff Var	60.54504		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-62.97048	36.11900	-1.74	0.0941
VISC	1	0.00307	0.00816	0.38	0.7100
__SURF	1	7.49803	3.96716	1.89	0.0709
__BASE	1	6.22582	4.81273	1.29	0.2081
__FINES	1	0.52221	1.17467	0.44	0.6606
__VOIDS	1	-0.24127	1.68496	-0.14	0.8873
RUN	1	-5.38630	0.98538	-5.47	<.0001

OUTLIER DETERMINATION

The SAS System   11:01 Thursday, May 10, 2007 35

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Output Statistics

<table>
<tr><th>Obs</th><th>Dependent Variable</th><th>Predicted Value</th><th>Std Error Mean Predict</th><th>Residual</th><th>Std Error Residual</th><th>Student Residual</th><th>-2-1 0 1 2</th><th>Cook's D</th></tr>
<tr><td>1</td><td>6.7500</td><td>11.0354</td><td>1.5319</td><td>-4.2854</td><td>3.630</td><td>-1.181</td><td>| **| |</td><td>0.035</td></tr>
<tr><td>2</td><td>13.0000</td><td>11.6445</td><td>1.7813</td><td>1.3555</td><td>3.514</td><td>0.386</td><td>| | |</td><td>0.005</td></tr>
<tr><td>3</td><td>14.7500</td><td>10.8504</td><td>1.4312</td><td>3.8996</td><td>3.671</td><td>1.062</td><td>| |** |</td><td>0.025</td></tr>
<tr><td>4</td><td>12.6000</td><td>11.5869</td><td>1.3975</td><td>1.0131</td><td>3.684</td><td>0.275</td><td>| | |</td><td>0.002</td></tr>
<tr><td>5</td><td>8.2500</td><td>13.1548</td><td>2.2482</td><td>-4.9048</td><td>3.235</td><td>-1.516</td><td>| ***| |</td><td>0.159</td></tr>
<tr><td>6</td><td>10.6700</td><td>11.6229</td><td>1.8463</td><td>-0.9529</td><td>3.480</td><td>-0.274</td><td>| | |</td><td>0.003</td></tr>
<tr><td>7</td><td>7.2800</td><td>12.3848</td><td>2.0724</td><td>-5.1048</td><td>3.351</td><td>-1.524</td><td>| ***| |</td><td>0.127</td></tr>
<tr><td>8</td><td>12.6700</td><td>11.4490</td><td>1.7889</td><td>1.2210</td><td>3.510</td><td>0.348</td><td>| | |</td><td>0.004</td></tr>
<tr><td>9</td><td>12.5800</td><td>10.5673</td><td>1.5427</td><td>2.0127</td><td>3.625</td><td>0.555</td><td>| |* |</td><td>0.008</td></tr>
<tr><td>10</td><td>20.6000</td><td>13.7904</td><td>1.5209</td><td>6.8096</td><td>3.634</td><td>1.874</td><td>| |*** |</td><td>0.088</td></tr>
<tr><td>11</td><td>3.5800</td><td>9.2581</td><td>1.7032</td><td>-5.6781</td><td>3.553</td><td>-1.598</td><td>| ***| |</td><td>0.084</td></tr>
<tr><td>12</td><td>7.0000</td><td>8.2729</td><td>1.7727</td><td>-1.2729</td><td>3.518</td><td>-0.362</td><td>| | |</td><td>0.005</td></tr>
<tr><td>13</td><td>26.2000</td><td>15.0738</td><td>2.0156</td><td>11.1262</td><td>3.385</td><td>3.287</td><td>| |******|</td><td>0.547</td></tr>
<tr><td>14</td><td>11.6700</td><td>12.0926</td><td>1.7542</td><td>-0.4226</td><td>3.528</td><td>-0.120</td><td>| | |</td><td>0.001</td></tr>
<tr><td>15</td><td>7.6700</td><td>12.6612</td><td>2.5469</td><td>-4.9912</td><td>3.006</td><td>-1.661</td><td>| ***| |</td><td>0.283</td></tr>
<tr><td>16</td><td>12.2500</td><td>12.0749</td><td>1.1568</td><td>0.1751</td><td>3.766</td><td>0.0465</td><td>| | |</td><td>0.000</td></tr>
<tr><td>17</td><td>0.7600</td><td>0.4695</td><td>1.4307</td><td>0.2905</td><td>3.671</td><td>0.0791</td><td>| | |</td><td>0.000</td></tr>
<tr><td>18</td><td>1.3500</td><td>1.7615</td><td>1.6323</td><td>-0.4115</td><td>3.586</td><td>-0.115</td><td>| | |</td><td>0.000</td></tr>
<tr><td>19</td><td>1.4400</td><td>2.0018</td><td>1.7859</td><td>-0.5618</td><td>3.512</td><td>-0.160</td><td>| | |</td><td>0.001</td></tr>
<tr><td>20</td><td>1.6000</td><td>3.1272</td><td>2.0065</td><td>-1.5272</td><td>3.390</td><td>-0.450</td><td>| | |</td><td>0.010</td></tr>
<tr><td>21</td><td>1.1000</td><td>-1.4451</td><td>2.1288</td><td>2.5451</td><td>3.315</td><td>0.768</td><td>| |* |</td><td>0.035</td></tr>
<tr><td>22</td><td>0.8500</td><td>0.6062</td><td>1.5241</td><td>0.2438</td><td>3.633</td><td>0.0671</td><td>| | |</td><td>0.000</td></tr>
<tr><td>23</td><td>1.2000</td><td>4.4006</td><td>2.0223</td><td>-3.2006</td><td>3.381</td><td>-0.947</td><td>| *| |</td><td>0.046</td></tr>
<tr><td>24</td><td>0.5600</td><td>2.6946</td><td>2.0039</td><td>-2.1346</td><td>3.392</td><td>-0.629</td><td>| *| |</td><td>0.020</td></tr>
<tr><td>25</td><td>0.7200</td><td>1.2446</td><td>2.3368</td><td>-0.5246</td><td>3.172</td><td>-0.165</td><td>| | |</td><td>0.002</td></tr>
<tr><td>26</td><td>0.4700</td><td>0.2119</td><td>2.7524</td><td>0.2581</td><td>2.819</td><td>0.0916</td><td>| | |</td><td>0.001</td></tr>
<tr><td>27</td><td>0.3300</td><td>-0.8277</td><td>1.9526</td><td>1.1577</td><td>3.422</td><td>0.338</td><td>| | |</td><td>0.005</td></tr>
<tr><td>28</td><td>0.2600</td><td>-2.4224</td><td>2.0234</td><td>2.6824</td><td>3.380</td><td>0.794</td><td>| |* |</td><td>0.032</td></tr>
<tr><td>29</td><td>0.7600</td><td>-0.8561</td><td>1.3419</td><td>1.6161</td><td>3.704</td><td>0.436</td><td>| | |</td><td>0.004</td></tr>
<tr><td>30</td><td>0.8000</td><td>3.7135</td><td>2.0680</td><td>-2.9135</td><td>3.353</td><td>-0.869</td><td>| *| |</td><td>0.041</td></tr>
<tr><td>31</td><td>2.0000</td><td>-0.4800</td><td>1.8875</td><td>2.4800</td><td>3.458</td><td>0.717</td><td>| |* |</td><td>0.022</td></tr>
</table>

Sum of Residuals   0
Sum of Squared Residuals   372.51430
Predicted Residual SS (PRESS)   670.16794

STEPWISE REGRESSION

```
proc reg data=asphalt;
model y=visc __surf __base __fines __voids run/selection=stepwise
sle=.15 sls=.15;
run;
```

The SAS System 11:01 Thursday, May 10, 2007 36

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read 31
Number of Observations Used 31

Stepwise Selection: Step 1

Variable RUN Entered: R-Square = 0.6576 and C(p) = 3.1488

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	898.56551	898.56551	55.69	<.0001
Error	29	467.95293	16.13631		
Corrected Total	30	1366.51844			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.33333	0.72185	1242.15054	76.98	<.0001
RUN	-5.38667	0.72185	898.56551	55.69	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable __SURF Entered: R-Square = 0.6947 and C(p) = 1.8789

Analysis of Variance

Sum of Mean

Source	DF	Squares	Square	F Value	Pr > F
Model	2	949.31954	474.65977	31.86	<.0001
Error	28	417.19890	14.89996		
Corrected Total	30	1366.51844			

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Stepwise Selection: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-23.00056	15.90890	31.14451	2.09	0.1593
__SURF	5.97498	3.23738	50.75403	3.41	0.0755
RUN	-5.40584	0.69372	904.76953	60.72	<.0001

Bounds on condition number: 1.0002, 4.0009

Stepwise Selection: Step 3

Variable __BASE Entered: R-Square = 0.7239 and C(p) = 1.3062

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	989.25205	329.75068	23.60	<.0001
Error	27	377.26639	13.97283		
Corrected Total	30	1366.51844			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-62.60050	28.03676	69.66015	4.99	0.0340
__SURF	7.42697	3.25057	72.94400	5.22	0.0304
__BASE	6.82824	4.03913	39.93251	2.86	0.1024
RUN	-5.26852	0.67669	847.00772	60.62	<.0001

Bounds on condition number: 1.0907, 9.5425

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

The SAS System 11:01 Thursday, May 10, 2007 38

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	RUN		1	0.6576	0.6576	3.1488	55.69	<.0001
2	__SURF		2	0.0371	0.6947	1.8789	3.41	0.0755
3	__BASE		3	0.0292	0.7239	1.306	2.86	0.1024

Thus the final model includes the variables RUN, __SURF, and __BASE (note: you have to place two consecutive underscores before SURF, etc for SAS to read the variables correctly).

As far as outliers, based on the criteria in 3a,

Average leverage is $(p+1)/n=7/31$. $2*(p+1)/n=0.452$. We can compare the leverage for individual observations to this value. If they are larger, they are considered x-outliers.

Based on Studentized Residuals, |d|>3 for observation 13.

Based on Cook's D, the appropriate F quantile to compare it to is $F(p+1, n\text{-}p\text{-}1, 0.50)= F(7,24,0.5)=0.93$. Since the Cook's D value for observation 13 is less than 0.93, we say that the observation is not an outlier.