STAT 3A03 Applied Regression With SAS

Fall 2017

Assignment 2 Solution Set

Q. 1 I will add subscripts relating to the question part to the parameters and their estimates as well as the errors and residuals.

a) The model is $y_i = \beta_{0a} + \beta_{1a}x_{1i} + \varepsilon_{a;i}$. From class notes the parameter estimates are

$$\hat{\beta}_{1a} = \frac{S_{1y}}{S_{11}} \qquad \hat{\beta}_{0a} = \overline{y} - \frac{S_{1y}}{S_{11}}\overline{x}_1.$$

Thus the residuals are

$$\begin{aligned} e_{a;i} &= y_i - \hat{y}_i \\ &= y_i - (\hat{\beta}_{0a} + \hat{\beta}_{1a}x_{1i}) \\ &= y_i - \left(\overline{y} - \frac{S_{1y}}{S_{11}}\overline{x}_1 + \frac{S_{1y}}{S_{11}}x_{1i}\right) \\ &= (y_i - \overline{y}) + \frac{S_{1y}}{S_{11}}(x_{1i} - \overline{x}_1) \end{aligned}$$

[4 marks]

b) Now we have the model $x_{2i} = \beta_{0b} + \beta_{1b}x_{1i} + \varepsilon_{b;i}$. The slope of the least squares line is then

$$\hat{\beta}_{1b} = \frac{S_{12}}{S_{11}}$$

Since we know that

$$r_{1,2} = \frac{S_{12}}{\sqrt{S_{11}S_{22}}} = 0$$

we have that $S_{12} = 0$ and so

$$\hat{\beta}_{1b} = 0$$

[3 marks]

Since $\hat{\beta}_{1b} = 0$, we have that $\hat{\beta}_{0b} = \overline{x}_2$ and $\hat{x}_{2i} = \overline{x}_2$, $(i = 1, \ldots, n)$. Hence the residuals from this model are

$$e_{b;i} = x_{2i} - \overline{x}_2.$$

[2 marks]

c) Now we have the model $e_{a;i} = \beta_{0c} + \beta_{1c} e_{b;i} + \varepsilon_{c;i}$.
First we note that

$$\sum_{i=1}^{n} e_{a;i} = \sum_{i=1}^{n} (y_i - \overline{y}) + \frac{S_{1y}}{S_{11}} \sum_{i=1}^{n} (x_{1i} - \overline{x}_1) = 0$$

$$\sum_{i=1}^{n} e_{b;i} = \sum_{i=1}^{n} (x_{2i} - \overline{x}_2) = 0$$

Hence $\hat{\beta}_{0c} = 0$ and

$$\hat{\beta}_{1c} = \frac{\sum e_{a;i} e_{b;i}}{\sum e_{b;i}^2}$$

[3 marks]

Next we have

$$\begin{aligned}
\sum_{i=1}^{n} e_{b;i}^2 &= \sum_{i=1}^{n} (x_{2i} - \overline{x}_2)^2 = S_{22} \\
\sum_{i=1}^{n} e_{a;i} e_{b;i} &= \sum_{i=1}^{n} \left[(y_i - \overline{y}) + \frac{S_{1y}}{S_{11}} (x_{1i} - \overline{x}_1) \right] (x_{2i} - \overline{x}_2) \\
&= \sum_{i=1}^{n} (x_{2i} - \overline{x}_2)(y_i - \overline{y}) - \frac{S_{1y}}{S_{11}} \sum_{i=1}^{n} (x_{1i} - \overline{x}_1)(x_{2i} - \overline{x}_2) \\
&= S_{2y} - \frac{S_{1y} S_{12}}{S_{11}} \\
&= S_{2y} - r_{1,2} S_{1y} \sqrt{\frac{S_{22}}{S_{11}}} \\
&= S_{2y}
\end{aligned}$$

Hence the slope of the least squares line is

$$\hat{\beta}_{1c} = \frac{S_{2y}}{S_{22}}$$

which we note is exactly the slope that would be found from fitting the simple regression model with response Y and covariate X_2. **[5 marks]**

d) From (b) we have that the slope of the line is

$$\hat{\beta}_{1d} = r_{1,2} \sqrt{\frac{S_{22}}{S_{11}}}$$

and so the intercept is

$$\hat{\beta}_{0d} = \overline{x}_2 - r_{1,2} \sqrt{\frac{S_{22}}{S_{11}}} \overline{x}_1.$$

Hence the residuals become

$$e_{d;i} = (x_{2i} - \overline{x}_2) - r_{1,2} \sqrt{\frac{S_{22}}{S_{11}}} (x_{1i} - \overline{x}_1)$$

To prove that all of the $e_{d;i} = 0$ we will show that $\sum e_{d;i}^2 = 0$. Since every $e_{d;i}^2 \geqslant 0$ then the only way that $\sum e_{d;i}^2$ can be 0 is if every $e_{d;i}^2 = 0$ and hence $e_{d;i} = 0$, $i = 1, \ldots, n$.

$$\begin{aligned}
\sum_{i=1}^{n} e_{d;i}^2 &= \sum_{i=1}^{n}\left[(x_{2i}-\overline{x}_2) - r_{1,2}\sqrt{\frac{S_{22}}{S_{11}}}(x_{1i}-\overline{x}_1)\right]^2 \\
&= \sum_{i=1}^{n}(x_{2i}-\overline{x}_2)^2 - 2r_{1,2}\sqrt{\frac{S_{22}}{S_{11}}}\sum_{i=1}^{n}(x_{1i}-\overline{x}_1)(x_{2i}-\overline{x}_2) + r_{1,2}^2\frac{S_{22}}{S_{11}}\sum_{i=1}^{n}(x_{1i}-\overline{x}_1)^2 \\
&= S_{22} - 2r_{1,2}\sqrt{\frac{S_{22}}{S_{11}}}S_{12} + r_{1,2}^2 S_{22} \\
&= S_{22} - 2r_{1,2}^2 S_2 2\frac{S_{12}}{\sqrt{S_{11}S_{22}}} + r_{1,2}^2 S_{22} \\
&= S_{22} - 2S_{22} + S_{22} \qquad (\text{since } r_{1,2}^2 = 1) \\
&= 0
\end{aligned}$$

Hence $e_{d;i} = 0$, $i = 1, \ldots, n$. **[6 marks]**

Now we recall that the slope of the regression from part (c) requires division by the sum of squares of the residuals from part (b) but we just showed that, when $|r_{1,2}| = 0$, those residuals satisfy

$$\sum_{i=1}^{n} e_{d;i}^2 = 0$$

and so the slope of the regression line is indeterminate since it requires d division by 0.
[2 marks]

Q. 2 Textbook 3.5

Here is the SAS code that I used. You can also find this code in the SAS file posted on the website. Note that, for simplicity of code, I am using a single `PROC REG` to fit all 5 models but it is perfectly acceptable to fit each separately also. The output will be identical other than the naming of the models.

```
Libname S3A3 "D:\STAT 3A03\Fall 2017\Data";

PROC IMPORT out=S3A3.exam
   datafile="D:\STAT 3A03\Fall 2017\Data\Examination.txt"
       DBMS=DLM REPLACE;
   Getnames=yes;
   Datarow=2;
run;

PROC REG Data=S3A3.exam plots=none;
    Model F=P1 P2;
    Model F=P1;
    Model F=P2;
    Model P1=P2;
    Model P2=P1;
run;
```

The SAS output for each of the five models is:

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: F

Number of Observations Read	22
Number of Observations Used	22

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2314.26087	1157.13043	74.07	<.0001
Error	19	296.83004	15.62263		
Corrected Total	21	2611.09091			

Root MSE	3.95255	R-Square	0.8863
Dependent Mean	81.36364	Adj R-Sq	0.8744
Coeff Var	4.85788		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-14.50054	9.23565	-1.57	0.1329
P1	1	0.48834	0.23299	2.10	0.0497
P2	1	0.67204	0.17928	3.75	0.0014

[3 marks]

The SAS System

The REG Procedure
Model: MODEL2
Dependent Variable: F

Number of Observations Read	22
Number of Observations Used	22

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2094.74806	2094.74806	81.14	<.0001
Error	20	516.34285	25.81714		
Corrected Total	21	2611.09091			

Root MSE	5.08106	R-Square	0.8023
Dependent Mean	81.36364	Adj R-Sq	0.7924
Coeff Var	6.24487		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-22.34244	11.56395	-1.93	0.0676
P1	1	1.26052	0.13994	9.01	<.0001

[3 marks]

The SAS System

The REG Procedure
Model: MODEL3
Dependent Variable: F

Number of Observations Read	22
Number of Observations Used	22

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2245.63144	2245.63144	122.89	<.0001
Error	20	365.45947	18.27297		
Corrected Total	21	2611.09091			

Root MSE	4.27469	R-Square	0.8600
Dependent Mean	81.36364	Adj R-Sq	0.8530
Coeff Var	5.25381		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.85355	7.56181	-0.25	0.8089
P2	1	1.00427	0.09059	11.09	<.0001

[3 marks]

The SAS System

The REG Procedure
Model: MODEL4
Dependent Variable: P1

Number of Observations Read	22
Number of Observations Used	22

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1030.57733	1030.57733	71.62	<.0001
Error	20	287.78630	14.38932		
Corrected Total	21	1318.36364			

Root MSE	3.79333	R-Square	0.7817
Dependent Mean	82.27273	Adj R-Sq	0.7708
Coeff Var	4.61067		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	25.89805	6.71029	3.86	0.0010
P2	1	0.68033	0.08039	8.46	<.0001

[3 marks]

The SAS System

The REG Procedure
Model: MODEL5
Dependent Variable: P2

Number of Observations Read	22
Number of Observations Used	22

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1740.54719	1740.54719	71.62	<.0001
Error	20	486.04372	24.30219		
Corrected Total	21	2226.59091			

Root MSE	4.92972	R-Square	0.7817
Dependent Mean	82.86364	Adj R-Sq	0.7708
Coeff Var	5.94920		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-11.66887	11.21954	-1.04	0.3107
P1	1	1.14901	0.13577	8.46	<.0001

[3 marks]

Based on these outputs we have the following estimates using the notation of the question

$$\begin{array}{rcl} \hat{\beta}_0 &=& -14.50054 \\ \hat{\beta}_1 &=& 0.48834 \\ \hat{\beta}_2 &=& 0.67204 \end{array}$$

$$\begin{array}{rcl} \hat{\beta}'_0 &=& -22.34244 \\ \hat{\beta}'_1 &=& 1.26052 \end{array} \qquad \begin{array}{rcl} \hat{\alpha}_0 &=& 25.89805 \\ \hat{\alpha}_2 &=& 0.68033 \end{array}$$

$$\begin{array}{rcl} \hat{\beta}''_0 &=& -1.85355 \\ \hat{\beta}'_2 &=& 1.00427 \end{array} \qquad \begin{array}{rcl} \hat{\alpha}'_0 &=& -11.66887 \\ \hat{\alpha}_1 &=& 1.14901 \end{array}$$

a) For this part we have

$$\hat{\beta}_1 + \hat{\beta}_2\hat{\alpha}_1 \;=\; 0.48824 + 0.67204 \times 1.14901 \;=\; 1.26052 \;=\; \hat{\beta}'_1$$

[5 marks]

b) For this part we have

$$\hat{\beta}_2 + \hat{\beta}_1\hat{\alpha}_2 \;=\; 0.67201 + 0.48824 \times 0.68033 \;=\; 1.00427 \;=\; \hat{\beta}'_2$$

[5 marks]

Q. 3 a)

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} \\ &= \begin{pmatrix} 0.65 & -0.20 & -0.15 \\ -0.20 & 0.40 & 0.00 \\ -0.15 & 0.00 & 0.05 \end{pmatrix} \begin{pmatrix} 51 \\ 24 \\ 182 \end{pmatrix} \\ &= \begin{pmatrix} 1.05 \\ -0.60 \\ 1.45 \end{pmatrix}\end{aligned}$$

[3 marks]

b) The fitted values are

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 2.50 \\ 3.95 \\ 5.40 \\ 6.85 \\ 8.30 \\ 1.90 \\ 3.35 \\ 4.80 \\ 6.25 \\ 7.70 \end{pmatrix}$$

[3 marks]

The error sum of squares is then

$$\begin{aligned}\text{SSE} &= (5-2.5)^2 + (2-3.95)^2 + (6-5.4)^2 + (5-6.85)^2 + (9-8.30)^2 \\ &\quad +(0-1.9)^2 + (4-3.35)^2 + (5-4.80)^2 + (8-6.25)^2 + (7-7.7)^2 \\ &= 21.95\end{aligned}$$

Hence an unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p-1} = \frac{21.95}{7} = 3.1357.$$

[3 marks]

c) The standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ are

$$\begin{aligned}\text{s.e.}(\hat{\beta}_1) &= \sqrt{0.40 \times 3.1357} = 1.1199 \\ \text{s.e.}(\hat{\beta}_1) &= \sqrt{0.05 \times 3.1357} = 0.3960\end{aligned}$$

[2 marks]

The appropriate critical value is $t_{7;0.025} = 2.36$ Hence the confidence interval for β_1 is

$$\begin{aligned}\hat{\beta}_1 \pm 2.36\text{s.e.}(\hat{\beta}_1) &= -0.60 \pm 2.36 \times 1.1199 \\ &= (-3.243,\ 2.043)\end{aligned}$$

and that for β_2 is

$$\begin{aligned}\hat{\beta}_2 \pm 2.36\text{s.e.}(\hat{\beta}_2) &= 1.45 \pm 2.36 \times 0.3960 \\ &= (0.516,\ 2.384)\end{aligned}$$

[2 marks]

d) The test statistic we must use to test this hypothesis is

$$t \;=\; \frac{\hat{\beta}_2 - 1}{\text{s.e.}(\hat{\beta}_2)}.$$

Hence the observed test statistic value is

$$t_{\text{obs}} \;=\; \frac{1.45 - 1}{0.396} \;=\; 1.136.$$

[2 marks]

If $\beta_2 = 1$ then this would come from a t_7 distribution. The 5% critical value for a two-sided test is then $t_{7;0.025} = 2.36$ and since $|t_{\text{obs}}| < t_{7;0.025}$ we fail to reject H_0 and conclude that there is insufficient evidence against the null hypothesis for us to conclude it is false. Hence it is perfectly possible that the true value of β_2 is equal to 1, given the observed data. [2 marks]

e) This is estimation of the mean response when $\boldsymbol{x}_0 \;=\; (1, 1, 3)^t$. Hence the point estimate is

$$\hat{\mu} \;=\; \hat{\beta}_0 + \hat{\beta}_1 + 3\hat{\beta}_2 \;=\; 4.8$$

[1 mark]

The standard error of $\hat{\mu}$ is

$$\begin{aligned}
\text{s.e.}(\hat{\mu}) \;&=\; \hat{\sigma}\sqrt{\boldsymbol{x}_0^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_0} \\
&=\; \sqrt{3.1357} \times \sqrt{1 \times (0.65 - 0.20 - 0.45) + 1 \times (-0.20 + 0.4 + 0) + 3 \times (-0.15 + 0 + 0.15)} \\
&=\; \sqrt{3.1357 \times 0.2} \\
&=\; 0.7919
\end{aligned}$$

[2 marks]

Hence a 95% confidence interval is

$$\begin{aligned}
\hat{\mu} \;\pm\; 2.36\text{s.e.}(\hat{\mu}) \;&=\; 4.8 \;\pm\; 2.36 \times 0.7919 \\
&=\; (2.931,\; 6.669)
\end{aligned}$$

[1 mark]

f) In this case we are predicting a single value with $\boldsymbol{x}_0 \;=\; (1, 0, 5)^t$. The point predictor is

$$\hat{y}_0 \;=\; \hat{\beta}_0 + 5\hat{\beta}_2 \;=\; 8.3$$

[1 mark]

The standard error of prediction is

$$\begin{aligned}
\text{s.e.}(\hat{y}_0) \;&=\; \hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_0} \\
&=\; \sqrt{3.1357} \times \sqrt{1 + 1 \times (0.65 - 0 - 0.75) + 0 \times (-0.20 + 0 + 0) + 5 \times (-0.15 + 0 + 0.25)} \\
&=\; \sqrt{3.1357 \times 1.4} \\
&=\; 2.0952
\end{aligned}$$

[2 marks]

Hence a 95% prediction interval is

$$\begin{aligned}
\hat{y}_0 \;\pm\; 2.36\text{s.e.}(\hat{y}_0) \;&=\; 8.3 \;\pm\; 2.36 \times 2.0952 \\
&=\; (3.355,\; 13.245)
\end{aligned}$$

[1 mark]

Q. 4 **a)**

$$\begin{aligned} a &= p = 6 \\ b &= n - p - 1 = 24 - 6 - 1 = 17 \\ c &= n - 1 = 23 \\ d &= 112.64325 * a = 675.85947 \\ e &= 831.50958 - 675.85947 = 155.65011 \\ f &= e/b = 155.65011/17 = 9.155889 \\ g &= 112.643245/f = 112.643245/9.155889 = 12.3028 \end{aligned}$$

[7 marks]

b) The null and alternative hypotheses are

$$\begin{aligned} H_0 &: \quad \beta_{\text{tax}} = \beta_{\text{age}} = \beta_{\text{bed}} = \beta_{\text{bath}} = \beta_{\text{space}} = \beta_{\text{lot}} = 0 \\ H_1 &: \quad \text{At least one of these parameters is non-zero} \end{aligned}$$

[2 marks]

The F statistic found above has $F_{\text{obs}} = 12.3028$. If H_0 is true then this should be an observation from an $F_{6,17}$ distribution. From the table we know $F_{0.05;6,17} < F_{0.05,6,15} = 2.79$.
Since $F_{\text{obs}} > 2.79$ we reject H_0 and conclude that at least one of the 6 variables is related to sale price. **[3 marks]**

c)

$$\hat{\sigma} = \sqrt{\text{MSE}} = \sqrt{9.155889} = 3.0259$$

[2 marks]

d) The required correlation is the square root of the coefficient of determination, R^2, and **must** be positive.

$$r = \sqrt{R^2} = \sqrt{\frac{675.85947}{831.50958}} = 0.9016$$

[3 marks]

e) The null and alternative hypotheses being tested here are

$$\begin{aligned} H_0 &: \quad \beta_{\text{age}} = \beta_{\text{bath}} = \beta_{\text{space}} = \beta_{\text{lot}} = 0 \\ H_1 &: \quad \text{At least one of these parameters is non-zero} \end{aligned}$$

[2 marks]

The test statistic is

$$\begin{aligned} F &= \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{full})/(\text{df}_{\text{red}} - \text{df}_{\text{full}})}{\text{MSE}_{\text{full}}} \\ &= \frac{(194.98312 - 155.65011)/4}{9.155889} \\ &= 1.0740 \end{aligned}$$

[3 marks]

If H_0 is true then $F \sim F_{4,17}$. From the table we see that the critical value is $F_{0.05;4,17} > F_{0.05;4,15} = 3.06$.
Since $1.0740 < 3.06$ we cannot reject the null hypothesis and so we conclude that taxes and number of bedrooms alone are sufficient in the model for sale price of a house. **[3 marks]**