stk310 Practical Assignment A5 – Suggested Solution

Note that, where applicable, the given answers are from the SAS output. The answers from the R output will be equivalent, but might differ slightly with respect to the number of decimal places given.

Question 1

SAS Code & Output

```
goptions reset=all;
proc import out=sasuser.cereal datafile='c:\cereal.csv'
                      dbms=csv replace;
       getnames=yes;
       datarow=2;
run;
data pap;
set sasuser.cereal;
y=Sales_of_paul_se_pap;
x2=Price_of_paul_se_pap;
x3=Price_of_Three_Bears_Bran;
x4=Sales_of_Three_Bears_Bran;
keep y x2 x3 x4;
run;
goptions reset=all;
title1 'Multiple regression using proc reg: select model using adjusted R-square';
proc reg data=pap plot=none;
       model y=x2 x3 x4 / selection=adjrsq;
run;
```

Multiple regression using proc reg: select model using adjusted R-square

The REG Procedure
Model: MODEL1
Dependent Variable: y

Adjusted R-Square Selection Method

Number of Observations Read 16

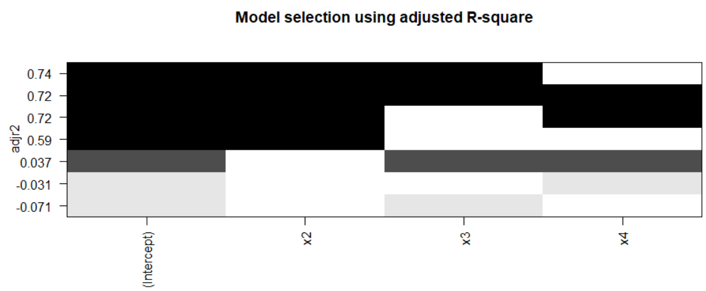
Number of Observations Used 16

Number in Model	Adjusted R-Square	R-Square	Variables in Model
2	0.7408	0.7754	x2 x3
3	0.7229	0.7783	x2 x3 x4
2	0.7196	0.7570	x2 x4
1	0.5873	0.6148	x2
2	0.0371	0.1655	x3 x4
1	-.0310	0.0378	x4
1	-.0711	0.0003	x3

Copyright reserved: Dr Paul J van Staden, University of Pretoria, Department of Statistics

R Code & Output

```
> pap <- read.csv("c:\\cereal.csv", header = T)
> y <- pap$Sales.of.paul.se.pap
> x2 <- pap$Price.of.paul.se.pap
> x3 <- pap$Price.of.Three.Bears.Bran
> x4 <- pap$Sales.of.Three.Bears.Bran
> # Multiple regression: select model using adjusted R-square
> library(leaps)
> ms <- regsubsets(y ~ x2 + x3 + x4, data = pap, nbest = 10)
> plot(ms, main = "Model selection using adjusted R-square", scale = "adjr2")
```



The best possible linear regression model based upon the adjusted coefficient of determination is $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$ with $\bar{R}^2 = 0.7408$.

Question 2

SAS Code & Output

```
goptions reset=all;
title1 'Partial correlation between Y & X2';
proc corr data=pap;
       var y x2;
       partial x3;
run;
goptions reset=all;
title1 'Partial correlation between Y & X3';
proc corr data=pap;
       var y x3;
       partial x2;
run;
```

Partial correlation between Y & X2

The CORR Procedure

1 Partial Variables: x3

2 Variables: y x2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Partial Variance	Partial Std Dev
x3	16	34.31875	3.24925	549.10000	28.50000	40.60000		
y	16	76.50000	20.37318	1224	32.00000	115.00000	444.58623	21.08521
x2	16	31.04375	5.35250	496.70000	22.60000	42.40000	23.85226	4.88388

```
Pearson Partial Correlation Coefficients, N = 16
        Prob > |r| under H0: Partial Rho=0

                          y                 x2

y                   1.00000           -0.88053
                                        <.0001

x2                 -0.88053            1.00000
                     <.0001

              Partial correlation between Y & X3

                      The CORR Procedure
                 1 Partial Variables: x2
                 2 Variables:         y x3

                      Simple Statistics
```

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Partial Variance	Partial Std Dev
x2	16	31.04375	5.35250	496.70000	22.60000	42.40000		
y	16	76.50000	20.37318	1224	32.00000	115.00000	171.30364	13.08830
x3	16	34.31875	3.24925	549.10000	28.50000	40.60000	8.78986	2.96477

```
Pearson Partial Correlation Coefficients, N = 16
        Prob > |r| under H0: Partial Rho=0

                          y                 x3

y                   1.00000            0.64568
                                        0.0093

x3                  0.64568            1.00000
                     0.0093
```

R Code & Output

```
> # Partial correlation coefficients
> yx2x3 <- cbind(y, x2, x3)
> library(ppcor)
> pcor(yx2x3)
$estimate
            y         x2        x3
y   1.0000000 -0.8805272 0.6456831
x2 -0.8805272  1.0000000 0.7394412
x3  0.6456831  0.7394412 1.0000000
```

(a) $r_{12.3} = -0.88053$ Holding the price of **Three Bears Bran** constant, there is a negative linear relation between the price of paul-se-pap and the demand for paul-se-pap. In effect, if the price of paul-se-pap increases and the price of **Three Bears Bran** stays constant, the demand for paul-se-pap will decrease.

(b) $r_{13.2} = 0.64568$ Holding the price of paul-se-pap constant, there is a positive linear relation between the price of **Three Bears Bran** and the demand for paul-se-pap. That is, if the price of **Three Bears Bran** increases and the price of paul-se-pap stays constant, the demand for paul-se-pap will increase.

Copyright reserved: Dr Paul J van Staden, University of Pretoria, Department of Statistics

Question 3

SAS Code & Output

```
data logs;
set pap;
lny=log(y);
lnx2=log(x2);
lnx3=log(x3);
run;
goptions reset=all;
title1 'Multiple regression using proc reg: inference & prediction';
proc reg data=logs plot=none;
       model lny=lnx2 lnx3 / alpha=0.01 clb clm;
       id x2 x3;
       test lnx2=-3;
       output out=reg_out predicted=lnyhat lclm=lnmeanlo uclm=lnmeanup
       r=residual;
run;
goptions reset=all;
title1 'Mean prediction';
proc iml;
use reg_out;
read all var{lnyhat lnmeanlo lnmeanup};
i = 12;
yhat36 = exp(lnyhat[i]);
meanlo36 = exp(lnmeanlo[i]);
meanup36 = exp(lnmeanup[i]);
print 'Predicted mean number of boxes of paul-se-pap sold:';
print yhat36[label=none];
print '95% confidence interval for mean number of boxes of paul-se-pap sold:';
print meanlo36[label=none] meanup36[label=none];
quit;
goptions reset=all;
title1 'Verifying normality assumption';
proc univariate data=reg_out normal;
       var residual;
run;
```

Multiple regression using proc reg : inference & prediction

The REG Procedure
Model: MODEL1
Dependent Variable: lny

Number of Observations Read 16

Number of Observations Used 16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.02420	0.51210	18.14	0.0002
Error	13	0.36701	0.02823		
Corrected Total	15	1.39120			

Root MSE	0.16802	**R-Square**	0.7362
Dependent Mean	4.29816	**Adj R-Sq**	0.6956
Coeff Var	3.90915		

Copyright reserved: Dr Paul J van Staden, University of Pretoria, Department of Statistics

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	5.55680	1.59934	3.47	0.0041	0.73914	10.37445
lnx2	1	-1.76360	0.29328	-6.01	<.0001	-2.64703	-0.88017
lnx3	1	1.35247	0.51640	2.62	0.0212	-0.20306	2.90800

Output Statistics

Obs	x2	x3	Dependent Variable	Predicted Value	Std Error Mean Predict	99% CL Mean		Residual
1	**22.6**	**34.9**	4.7449	4.8626	0.1038	4.5500	5.1752	-0.1177
2	**25.4**	**28.5**	4.5326	4.3826	0.0926	4.1038	4.6615	0.1500
3	**30.7**	**40.6**	4.4308	4.5270	0.0981	4.2315	4.8225	-0.0962
4	**29.1**	**36.4**	4.6151	4.4737	0.0594	4.2948	4.6526	0.1414
5	**27.3**	**32.1**	4.5218	4.4163	0.0537	4.2544	4.5782	0.1055
6	**27.7**	**36.6**	4.4886	4.5681	0.0700	4.3572	4.7790	-0.0794
7	**35.9**	**37.6**	4.1271	4.1472	0.0643	3.9534	4.3410	-0.0201
8	**32.3**	**34.9**	4.4188	4.2328	0.0443	4.0994	4.3662	0.1860
9	**26.0**	**31.3**	4.3820	4.4682	0.0633	4.2774	4.6590	-0.0862
10	**28.9**	**32.0**	4.3175	4.3116	0.0514	4.1567	4.4665	0.005864
11	**37.7**	**36.5**	4.0775	4.0208	0.0680	3.8159	4.2256	0.0568
12	**36.0**	**36.0**	4.3820	4.0835	0.0591	3.9053	4.2617	0.2985
13	**28.2**	**29.4**	4.1109	4.2403	0.0810	3.9963	4.4842	-0.1294
14	**29.6**	**31.2**	4.0775	4.2352	0.0601	4.0540	4.4163	-0.1576
15	**42.4**	**35.8**	3.4657	3.7874	0.0960	3.4982	4.0766	-0.3217
16	**36.9**	**35.3**	4.0775	4.0134	0.0644	3.8195	4.2073	0.0641

Sum of Residuals	0
Sum of Squared Residuals	0.36701
Predicted Residual SS (PRESS)	0.61925

Test 1 Results for Dependent Variable lny

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.50175	17.77	0.0010
Denominator	13	0.02823		

Mean prediction

Predicted mean number of boxes of paul-se-pap sold:

59.352787

95% confidence interval for mean number of boxes of paul-se-pap sold:

49.666577 70.928048

```
Verifying normality assumption

The UNIVARIATE Procedure
Variable: residual (Residual)

Tests for Normality

Test                  Statistic         p Value
Shapiro-Wilk        W     0.981562 Pr < W      0.9747
Kolmogorov-Smirnov  D     0.131712 Pr > D     >0.1500
Cramer-von Mises    W-Sq  0.030616 Pr > W-Sq  >0.2500
Anderson-Darling    A-Sq  0.199871 Pr > A-Sq  >0.2500
```

R Code & Output

```
> # Multiple regression: inference & prediction
> lny <- log(y)
> lnx2 <- log(x2)
> lnx3 <- log(x3)
> (lrm <- lm(lny ~ lnx2 + lnx3, data = pap))

Call:
lm(formula = lny ~ lnx2 + lnx3, data = pap)

Coefficients:
(Intercept)         lnx2         lnx3
      5.557       -1.764        1.352

> summary(lrm)

Call:
lm(formula = lny ~ lnx2 + lnx3, data = pap)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32165 -0.10155 -0.00711  0.11448  0.29853

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   5.5568     1.5993   3.474   0.00411 **
lnx2         -1.7636     0.2933  -6.013 0.0000435 ***
lnx3          1.3525     0.5164   2.619   0.02122 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.168 on 13 degrees of freedom
Multiple R-squared:  0.7362,  Adjusted R-squared:  0.6956
F-statistic: 18.14 on 2 and 13 DF,  p-value: 0.0001731

> confint(lrm, level = 0.99)
                 0.5 %    99.5 %
(Intercept)  0.7391426 10.374455
lnx2        -2.6470336 -0.880168
lnx3        -0.2030595  2.907997
> # Mean prediction
> i <- 12
> exp(predict(lrm, interval="confidence"))[i,]
     fit      lwr      upr
59.35279 52.23328 67.44269
> # Verifying normality assumption
> shapiro.test(lrm$residuals)

        Shapiro-Wilk normality test

data:  lrm$residuals
W = 0.98156, p-value = 0.9747
```

(a) $\ln Y_i = \underset{(1.59934)}{5.55680} - \underset{(0.29328)}{1.76360}\ln X_{2i} + \underset{(0.51640)}{1.35247}\ln X_{3i} + \hat{u}_i$

$\hat{\beta}_2 = -1.76360$ If the price of **Three Bears Bran** is unchanged and the price of **paul-se-pap** increases by 1% per box, the mean demand for **paul-se-pap** will decrease by 1.76%.

$\hat{\beta}_3 = 1.35247$ If the price of **paul-se-pap** remains constant and the price of **Three Bears Bran** increases by 1% per box, the mean demand for **paul-se-pap** will increase by 1.35%.

(b)

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

Test statistic value $t = 3.47$

Since p-value $= 0.0041 < 0.01$, H_0 is rejected at a 1% significance level.

$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

Test statistic value $t = -6.01$

Since p-value $< .0001 < 0.01$, H_0 is rejected at a 1% significance level.

$H_0: \beta_3 = 0$
$H_1: \beta_3 \neq 0$

Test statistic value $t = 2.62$

Since p-value $0.01 < 0.0212 < 0.05$, H_0 cannot be rejected at a 1% significance level, but is rejected at a 5% significance level.

We conclude that all the individual regression coefficients are statistically significant.

(c) $H_0: \beta_2 = \beta_3 = 0$
H_1: At least one of β_2 and β_3 is not zero

Test statistic value $F = 18.14$

Since p-value $= 0.0002 < 0.01$, H_0 is rejected at a 1% significance level.

Thus the overall regression model is statistically significant in that at least one of β_2 and β_3 is a significant parameter.

Copyright reserved: Dr Paul J van Staden, University of Pretoria, Department of Statistics

(d) $H_0\colon \beta_2 = -3$
$H_1\colon \beta_2 \neq -3$

Test statistic value $\Rightarrow$ $F = 17.77$

99% confidence interval for β_2 $\Rightarrow$ $(-2.64703\,,-0.88017)$

Because p-value $= 0.0010 < 0.01$ and because $\beta_2^* = -3$ does not fall within the 99% confidence interval for β_2, H_0 is rejected at a 1% significance level.

So, if the price of **Three Bears Bran** remains unchanged and the price of **paul-se-pap** increases by 1%, the mean demand for **paul-se-pap** will not decrease by 3%.

(e) Estimated mean demand for **paul-se-pap** $\Rightarrow$ $e^{4.0835} = 59.352787$

99% confidence interval for the mean demand for **paul-se-pap**:

$$(e^{3.9053}\,,e^{4.2617}) = (49.666577\,,70.928048)$$

(f) $H_0\colon u_i$ normal
$H_0\colon u_i$ not normal

Test statistic value $\Rightarrow$ $W = 0.981562$

p-value $= 0.9747 > 0.1$, so the null hypothesis cannot even be rejected at a 10% significance level.

Therefore u_i follows a normal distribution.

Question 4

SAS Code & Output

```
data onlylogs;
set logs;
keep lny lnx2 lnx3;
run;
goptions reset=all;
title1 'Multiple regression using proc iml';
proc iml;
use onlylogs;
read all into matrix;
n=nrow(matrix);
y=matrix[,1];
x=j(n,1,1)||matrix[,2:3];
print 'y:' y[label=none] 'X:' x[label=none];
bhat=inv(x`*x)*x`*y;
print 'Vector of parameter estimates:' bhat[label=none];
k=ncol(x);
uhat=y-x*bhat;
mse=uhat`*uhat/(n-k);
print 'Mean square error:' mse[label=none];
varcovb=mse#inv(x`*x);
print 'Covariance matrix for estimates:' varcovb[label=none];
```

Copyright reserved: Dr Paul J van Staden, University of Pretoria, Department of Statistics

```
stderr=sqrt(vecdiag(varcovb));
t=bhat/stderr;
t_pvalue=2#(1-probt(abs(t),n-k));
print 'Standard errors of estimates:' stderr[label=none];
print 'Test statistic values for t-tests:' t[label=none];
print 'p-values for t-tests:' t_pvalue[label=none];
ybar=sum(y)/n;
ess=bhat`*x`*y-n#ybar##2;
tss=y`*y-n#ybar##2;
rsquare=ess/tss;
print 'Coefficient of determination:' rsquare[label=none];
adjrsq=1-(1-rsquare)#(n-1)/(n-k);
print 'Adjusted coefficient of determination:' adjrsq[label=none];
f=(rsquare/(k-1))/((1-rsquare)/(n-k));
f_pvalue=1-probf(f,k-1,n-k);
print 'Test statistic value for F-test:' f[label=none];
print 'p-value for F-test:' f_pvalue[label=none];
quit;
```

```
Multiple regression using proc iml

y: 4.7449321 X: 1 3.1179499 3.5524868
   4.5325995    1 3.2347492 3.3499041
   4.4308168    1 3.4242627 3.7037681
   4.6151205    1 3.3707382 3.5945688
   4.5217886    1 3.3068867  3.468856
   4.4886364    1 3.3214324 3.6000482
   4.1271344    1 3.5807373 3.6270041
   4.4188406    1 3.4750672 3.5524868
   4.3820266    1 3.2580965 3.4436181
   4.3174881    1 3.3638416 3.4657359
   4.0775374    1 3.6296601 3.5973123
   4.3820266    1 3.5835189 3.5835189
   4.1108739    1  3.339322 3.3809947
   4.0775374    1 3.3877744 3.4404181
   3.4657359    1 3.7471484 3.5779479
   4.0775374    1 3.6082116  3.563883

Vector of parameter estimates: 5.5567987
                               -1.763601
                               1.3524687

Mean square error: 0.0282313

Covariance matrix for estimates: 2.5578915 -0.034732 -0.690172
                                 -0.034732 0.0860117 -0.073508
                                 -0.690172 -0.073508 0.2666652

Standard errors of estimates:  1.599341
                              0.2932775
                              0.5163963
```

Copyright reserved: Dr Paul J van Staden, University of Pretoria, Department of Statistics

```
Test statistic values for t-tests: 3.4744303
                                   -6.01342
                                   2.6190517

p-values for t-tests: 0.0041108
                      0.0000435
                      0.0212243

Coefficient of determination: 0.7361943

Adjusted coefficient of determination: 0.6956088

Test statistic value for F-test: 18.139347

p-value for F-test: 0.0001731
```

R Code & Output

```
> # Matrix approach to multiple regression
> n <- nrow(pap)
> (y <- matrix(lny, nrow = n, ncol = 1))
          [,1]
 [1,] 4.744932
 [2,] 4.532599
 [3,] 4.430817
 [4,] 4.615121
 [5,] 4.521789
 [6,] 4.488636
 [7,] 4.127134
 [8,] 4.418841
 [9,] 4.382027
[10,] 4.317488
[11,] 4.077537
[12,] 4.382027
[13,] 4.110874
[14,] 4.077537
[15,] 3.465736
[16,] 4.077537
> (X <- cbind(matrix(1, nrow = n, ncol = 1), matrix(c(lnx2, lnx3), nrow = n, ncol = 2)))
      [,1]     [,2]     [,3]
 [1,]    1 3.117950 3.552487
 [2,]    1 3.234749 3.349904
 [3,]    1 3.424263 3.703768
 [4,]    1 3.370738 3.594569
 [5,]    1 3.306887 3.468856
 [6,]    1 3.321432 3.600048
 [7,]    1 3.580737 3.627004
 [8,]    1 3.475067 3.552487
 [9,]    1 3.258097 3.443618
[10,]    1 3.363842 3.465736
[11,]    1 3.629660 3.597312
[12,]    1 3.583519 3.583519
[13,]    1 3.339322 3.380995
[14,]    1 3.387774 3.440418
[15,]    1 3.747148 3.577948
[16,]    1 3.608212 3.563883
> # Vector of parameter estimates
> (bhat <- solve(t(X) %*% X) %*% t(X) %*% y)
          [,1]
[1,]  5.556799
[2,] -1.763601
[3,]  1.352469
> # Mean square error
> k <- ncol(X)
> uhat <- y - X %*% bhat
> (mse <- as.numeric(t(uhat) %*% uhat / (n - k)))
[1] 0.02823133
```

```
> # Covariance matrix for estimates
> (varcovb <- mse * solve(t(X) %*% X))
            [,1]        [,2]        [,3]
[1,]  2.55789150 -0.03473185 -0.69017210
[2,] -0.03473185  0.08601171 -0.07350782
[3,] -0.69017210 -0.07350782  0.26666517
> # Standard errors of estimates
> (stderr <- as.matrix(sqrt(diag(varcovb))))
          [,1]
[1,] 1.5993410
[2,] 0.2932775
[3,] 0.5163963
> # Test statistic values for t-tests
> (t <- bhat / stderr)
          [,1]
[1,]  3.474430
[2,] -6.013420
[3,]  2.619052
> # p-values for t-tests
> (t_pvalue <- 2 * pt(abs(t), n - k, lower.tail = FALSE))
              [,1]
[1,] 0.00411079586
[2,] 0.00004349936
[3,] 0.02122431800
> # Coefficient of determination
> ybar <- mean(y)
> ESS <- t(bhat) %*% t(X) %*% y - n * ybar ^ 2
> TSS <- t(y) %*% y - n * ybar ^ 2
> (Rsquare <- as.numeric(ESS / TSS))
[1] 0.7361943
> # Adjusted coefficient of determination
> (AdjRsq <- 1 - (1 - Rsquare) * (n - 1) / (n - k))
[1] 0.6956088
> # Test statistic value for F-test
> (F <- (Rsquare / (k - 1)) / ((1 - Rsquare) / (n - k)))
[1] 18.13935
> # p-value for F-test
> (F_pvalue <- pf(F, k - 1, n - k, lower.tail = FALSE))
[1] 0.0001731199
```

(a) $\ln Y_i = \underset{(1.599341)}{5.55667987} - \underset{(0.2932775)}{1.763601}\ln X_{2i} + \underset{(0.5163963)}{1.3524687}\ln X_{3i} + \hat{u}_i$

(b) $\hat{\sigma}^2 = 0.0282313$ $\quad R^2 = 0.7361943$ $\quad \bar{R}^2 = 0.6956088$

(c)

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

Test statistic value $t = 3.4744303$

Because p-value $= 0.0041108 < 0.01$, H_0 is rejected at a 1% significance level.

$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

Test statistic value $t = -6.01342$

Because p-value $= 0.0000435 < 0.01$, H_0 is rejected at a 1% significance level.

Copyright reserved: Dr Paul J van Staden, University of Pretoria, Department of Statistics

$H_0: \beta_3 = 0$
$H_1: \beta_3 \neq 0$

Test statistic value $t = 2.6190517$

Because p-value $0.01 < 0.0212243 < 0.05$, H_0 cannot be rejected at a 1% significance level, but is rejected at a 5% significance level.

We conclude that all the individual regression coefficients are statistically significant.

(d) $H_0: \beta_2 = \beta_3 = 0$
H_1: At least one of β_2 and β_3 is not zero

Test statistic value $F = 18.139347$

Since p-value $= 0.0001731 < 0.01$, H_0 is rejected at a 1% significance level.

So we conclude that the overall regression model is statistically significant.

Copyright reserved: Dr Paul J van Staden, University of Pretoria, Department of Statistics