UNIVERSITY OF CALGARY

Faculty of Science
Data603: Statistical Modeling with Data
Fall 2019 Quiz 1

Student ID Number: ………………**SOLUTIONS…**……………………………..……………………..

Last Name: ………………………………………… **First Name:** ……………………………………….…..

Time: 60 min. Date: Nov 7th 2019

Examination Rules

1. This is an open book quiz. You are allowed to use the material or notes you have been studying.
2. Use of a personal laptop is permitted.
3. No additional time will be granted to fill in the forms.

For an instructor

Question	1 (a-g)	2 (a-c)	3(a-d)	4	5	6	7 (a-b)	Total
Total Marks	20	9	10	5	2	8	6	50
Actual Marks								

Use R output 1 to answer question 1 a)- 1 g)

Question 1

A charge for shipping a package in a regional express delivery company is based on the package weight and distance shipped. The company's profit per package depends on the package size (volume of space that it occupies) and the size and nature of the load on the delivery truck. The company recently conducted a study to investigate the relationship between the cost of shipment, variable Cost (in dollars) and the variables that control the shipping charge—package weight, variable Weight (in pounds), and distance shipped, variable Distance (in miles). Twenty packages were randomly selected from large number received for shipment and a detailed analysis of the cost of shipment was made for each package.

1a) (3 mark) Construct the ANOVA table for the model.

```
## Model 2: Cost ~ Distance + Weight
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     19 452.09
## 2     17  37.90  2    414.18 92.888 7.066e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sources	Df	S S	MS	F	p-value
Regression	2	414.18	207.09	92.88	7.066e-10
Error	17	37.9	2.229		
Total	19	452.09			

1b) (3 mark) Test the hypotheses for **the full model**. Use significance level 0.05.

$$H_0: \quad \beta_1 = \beta_2 = 0$$
$$H_a: \quad \text{at least one} \beta_i \text{not equal to zero i} = \text{Distance, Weight}$$

Fcal= 92.88 with p-value = p-value: 7.066e-10 <0.05 so we reject Ho at $\alpha = 0.05$

Therefore, at least one of the predictors must be related with Cost.

1c) (3 mark) Test the hypothesis that the average cost of shipment increases as the weight increases when distance is held constant. Use significance level 0.05.

```
Coefficients:
##               Estimate Std. Error t value Pr(>|t|)

## Weight       1.292414   0.137842   9.376 3.95e-08 ***
```

$$H_0:\quad \beta_{Weight} = 0$$
$$H_a:\quad \beta_{Weight} \neq 0$$

The individual coefficient t test shows that tcal=9.376 with the p-value for Weight =3.95e-08< 0.05, so we should clearly reject the null hypothesis. Therefore, the predictor Weight is significantly related with Cost.

1d) (3 mark) Check Individual t test for predictors. What is the first order model?

```
Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.672757   0.891147  -5.244 6.60e-05 ***
## Distance     0.036936   0.004602   8.026 3.49e-07 ***
## Weight       1.292414   0.137842   9.376 3.95e-08 ***
## ---
```

$\widehat{Cost} = -4.672757 + 0.036936 Distance + 1.292414 Weight.$

1e) (3 mark) Obtain 99% confidence interval of the regression coefficient for Distance variable.

```
confint(model,level=0.99)

##                   0.5 %      99.5 %
## (Intercept) -7.25550628 -2.09000697
## Distance     0.02359865  0.05027238
## Weight       0.89291631  1.69191154
```

From the output, a 99% confidence Interval for Distance = (0.02359865,0.05027238) which means that the average cost increases by \$0.02359865 to \$0.05027238 for every 1 mile increase in Distance.

1f) (3 mark) Find the R^2_{adj} value from the model in part d) and interpret it.

```
## Residual standard error: 1.493 on 17 degrees of freedom
## Multiple R-squared:  0.9162, Adjusted R-squared:  0.9063
## F-statistic: 92.89 on 2 and 17 DF,  p-value: 7.066e-10
```

$R^2 Adj$=0.9063 implies that the variation in Cost can be explained by this model (with independent variable Distance and Weight) 90.63%.

1g) (2 mark) Obtain 95% prediction interval for the cost of shipment y when Weight = 6.5 pounds and Distance=150 miles.

```
newdata2 = data.frame(Weight=6.5, Distance=150)
predict(model,newdata2,interval="predict")

##        fit      lwr      upr
## 1 9.268261 5.960167 12.57636
```

95% prediction interval for the cost of shipment y when Weight = 6.5 pounds and Distance=150 miles is between $5.960167 to $12.57636.

Use R output 2 to answer question 2 a)- 2 c)

Question 2:

2a) (3 mark) Which model would you suggest to predict the cost of shipment?

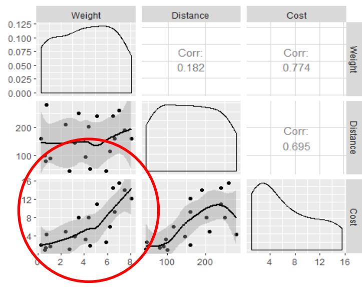
(a) $\widehat{Cost} = \hat{\beta}_0 + \hat{\beta}_1 Weight + \hat{\beta}_2 Distance$

(b) $\widehat{Cost} = \hat{\beta}_0 + \hat{\beta}_1 Weight + \hat{\beta}_2 Distance + \hat{\beta}_3 Weight * Distance$

(c) $\widehat{Cost} = \hat{\beta}_0 + \hat{\beta}_1 Weight + \hat{\beta}_2 Weight^2 + \hat{\beta}_3 Distance + \hat{\beta}_4 Weight * Distance$

(d) $\widehat{Cost} = \hat{\beta}_0 + \hat{\beta}_1 Weight + \hat{\beta}_2 Weight^2 + \hat{\beta}_3 Distance + \hat{\beta}_4 Distance^2 + \hat{\beta}_4 Weight * Distance$

2b) (3 mark) Provide any supporting details why the model selected in part 2a) is considered to be the best fit model by using the figure from R output 2.



The scatterplot between Cost and Weight seems to have a nonlinear curve.

2c) (3 mark) Give the value of R^2_{adj} and RMSE from the model chosen in part 2a).

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.4746969  0.4584500   1.035 0.316870
## Distance          0.0090777  0.0026535   3.421 0.003791 **
## Weight           -0.5781705  0.1706879  -3.387 0.004062 **
## I(Weight^2)       0.0867388  0.0193380   4.485 0.000436 ***
## Distance:Weight   0.0072587  0.0006176  11.753 5.74e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4346 on 15 degrees of freedom
## Multiple R-squared:  0.9937, Adjusted R-squared:  0.9921
## F-statistic: 594.6 on 4 and 15 DF,  p-value: 2.541e-16
```

R^2_{adj}= 0.9921   RMSE= 0.4346

Use R output 3 to answer question 3

Question 3: An investor investigates the factors that affect the sale price of ocean side condominium units. The following variables shown below were measured.

x1 =Floor height (x1 =1,2,...,8)

x2 =Distance from elevator (x2 =1,2,...,15)

$$x3 = \begin{cases} 1 \text{ if an ocean view} \\ 0 \text{ if not} \end{cases} \qquad x4 = \begin{cases} 1 \text{ if an end unit} \\ 0 \text{ if not} \end{cases} \qquad x5 = \begin{cases} 1 \text{ if furnished} \\ 0 \text{ if not} \end{cases}$$

3a) (2 mark) What model selection technique is used for building this model?

```
goodmodel=ols_step_both_p(model, pent = 0.1, prem = 0.3, details=TRUE)

## Stepwise Selection Method
## ---------------------------
## Candidate Terms:
## 1. FLOOR
## 2. DIST
## 3. factor(VIEW)
## 4. factor(END)
## 5. factor(FURNISH)
##
## We are selecting variables based on p value...
## Stepwise Selection: Step 1
```

Stepwise technique as the function ols-step-both-p was used.

3b) (2 mark) From the output, which predictor(s) is (are) dropped from the full model?

The predictor END was dropped

3c) (3 mark) From the output, which predictor is declared as the best predictor of the sale price in all possible simple linear regression models? Provide t-value to support your answer.

```
## We are selecting variables based on p value...
## Stepwise Selection: Step 1
##
## - factor(VIEW) added
##                          Model Summary
## -----------------------------------------------------------------
## R                        0.579       RMSE                 27.701
## R-Squared                0.335       Coef. Var            13.762
## Adj. R-Squared           0.332       MSE                 767.357
## Pred R-Squared           0.322       MAE                  22.134
## -----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##                                  ANOVA
## --------------------------------------------------------------------------
##                   Sum of
##                  Squares         DF    Mean Square       F          Sig.
## --------------------------------------------------------------------------
## Regression     80049.921          1      80049.921    104.319     0.0000
## Residual      158842.854        207        767.357
## Total         238892.775        208
## --------------------------------------------------------------------------
##                                      Parameter Estimates
## -----------------------------------------------------------------------------------------------
##         model         Beta    Std. Error    Std. Beta        t         Sig       lower       upper
## -----------------------------------------------------------------------------------------------
##   (Intercept)      181.050         2.756                  65.684     0.000     175.615     186.484
## factor(VIEW)1       39.163         3.834        0.579     10.214     0.000      31.604      46.723
## -----------------------------------------------------------------------------------------------
##
```

The best predictor= View t-value= 10.214

3d) (3 mark) After using the model selection procedure, write the first order model (substitute all regression coefficients) .

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        185.1932     5.0438  36.717  < 2e-16 ***
## factor(VIEW)1       40.3347     3.4599  11.658  < 2e-16 ***
## FLOOR               -3.7359     0.7465  -5.004 1.21e-06 ***
## DIST                 1.6793     0.3717   4.518 1.06e-05 ***
## factor(FURNISH)1 -32.4497     9.5885  -3.384 0.000856 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.42 on 204 degrees of freedom
## Multiple R-squared:  0.4908, Adjusted R-squared:  0.4808
## F-statistic: 49.16 on 4 and 204 DF,  p-value: < 2.2e-16
```

$$Sale\ \hat{P}raice = 185.1932 + 40.3347View - 3.7359Floor + 1.6793Dist - 32.4497Furnish$$

Use R output 4 to answer question 4

Question 4: From the output, the model contains one categorical independent variable, rank, with 4 levels (Level 1, 2, 3, and 4). Write all (sub)-regression models for each rank (substitute all regression coefficients). (5 mark)

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      42.000      2.259  18.593  < 2e-16 ***
## factor(rank)2    10.571      4.005   2.640 0.013613 *
## factor(rank)3    14.875      3.830   3.884 0.000602 ***
## factor(rank)4    25.102      1.275   6.471  < 2e-16 ***

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.795 on 36 degrees of freedom
## Multiple R-squared:  0.889 , Adjusted R-squared:  0.856
## F-statistic: 14.515 on 3 and 36 DF,  p-value: 0.001319
```

$$\hat{Response} = \begin{cases} \hat{\beta}_0 = 42 & \text{, if Rank} = 1 \\ \hat{\beta}_0 + \hat{\beta}_1 = 42 + 10.571 = 62.571 & \text{, if Rank} = 2 \end{cases}$$

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_2 = 42 + 14.875 = 56.875 & \text{, if Rank} = 3 \\ \hat{\beta}_0 + \hat{\beta}_3 = 42 + 25.102 = 67.102 & \text{, if Rank} = 4 \end{cases}$$

Question 5: Multiple regression analysis is used when

(a) there is not enough data to carry out simple linear regression analysis.

(b) the dependent variable depends on more than one independent variable.

(c) the independent variable cannot carry categorical data.

(d) the response variable carries categorical data with 4 levels (0,1,2,3).

Use R output 5 to answer question 6

Question 6: Mike and Bill are asked by Snow Kingdom Resort to analyse some data that might help in predicting how many customers to expect on a given day. The resort manager supplies data for a random sample of 32 days. The information includes

Skier: the number of customers

Snow: the number of inches of snow on the ground at noon

Weekend: whether the day fell on a weekend (0= weekday, 1= weekend)

Temperature: the highest temperature (degree Fahrenheit)

Give the interpretation of the effect of weekend, snow and temperature on the number of customers. (8 mark)

```
intmodel<-lm(skiers ~factor(weekend)+snow+temperature+factor(weekend)*temperature,dat
a=ski )
summary(intmodel)
## Call:
## lm(formula = skiers ~ factor(weekend) + snow + temperature +
##     factor(weekend) * temperature, data = ski)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -153.60  -39.89   11.32   36.98  111.18
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   552.949     65.946   8.385 5.38e-09 ***
## factor(weekend)1              208.279     51.697   4.029  0.00041 ***
## snow                            3.913      1.665   2.350  0.02635 *
## temperature                    -6.037      1.975  -3.056  0.00500 **
## factor(weekend)1:temperature  -13.021      4.984  -2.613  0.01450 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.28 on 27 degrees of freedom
## Multiple R-squared:  0.859,  Adjusted R-squared:  0.8382
## F-statistic: 41.14 on 4 and 27 DF,  p-value: 4.1e-11
```

Estimated Model:

$$\widehat{Number\ of\ customers} = \hat{\beta}_0 + \hat{\beta}_1 weekend + \hat{\beta}_2 snow + \hat{\beta}_3 temperature + \hat{\beta}_4 weekend * temperture$$

Weekend:

$$\widehat{Number\ of\ customers} = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 snow + \hat{\beta}_3 temperature + \hat{\beta}_4 temperture & \text{, if it is on weekend} --(1) \\ \hat{\beta}_0 + \hat{\beta}_2 snow + \hat{\beta}_3 Temperature & \text{, if it is on weekday} --(2) \end{cases}$$

$(1) - (2) = weekend - weekday: \hat{\beta}_1 + \hat{\beta}_4\ temperture$=208.279-13.021temperature which means the average difference of customers between Weekend and Weekday is 208.279-13.021temperature. Hence when the temperature is higher (every 1 degree increase in temperature), on average, customers would prefer to ski on weekdays (decrease by 208.279-13.021temperature)

Snow: $\hat{\beta}_2 = 3.913$

For every one inch increase in snow on the ground at noon, the average number of customers will increase by 3.913 customers for a given amount of other predictors(are held constant)

Temperature: $(\hat{\beta}_3 + \hat{\beta}_4\, weekend)temperture$= $(-6.037 - 13.021weekend)temperture$

On weekend, for every one degree increase in temperature, the number of customer will decrease by 19.058 customers (6.037+13.021)

On weekdays, for every one degree increase in temperature, the number of customer will decrease by 6.037 customers

Use R output 6 to answer question 7

Question 7: In the oil industry, water that mixes with crude oil during production and transportation must be removed. Chemists have found that the oil can be extracted from the water/oil mix electrically. Researchers at the University of Bergen (Norway) conducted a series of experiments to study the factors that influence the voltage (y) required to separate the water from the oil. The seven independent variables investigated in the experimental study and the sample data for 5 experiments are given in the table below.

EXPERIMENT NUMBER	VOLTAGE y (kw/cm)	DISPERSE PHASE VOLUME x_1 (%)	SALINITY x_2 (%)	TEMPERATURE x_3 (°C)	TIME DELAY x_4 (hours)	SURFACTANT CONCENTRATION x_5 (%)	SPAN:TRITON x_6	SOLID PARTICLES x_7 (%)
1	.64	40	1	4	.25	2	.25	.5
2	.80	80	1	4	.25	4	.25	2
3	3.20	40	4	4	.25	4	.75	.5
4	.48	80	4	4	.25	2	.75	2
5	1.72	40	1	23	.25	4	.75	2

7a) (3 mark) From the output 6, what is the best fitted model?

```
model8<-lm(Voltage~Volume+I(Volume^2)+Salinity+Surfactant+Volume*Salinity+Volume*Surf
actant,data=wateroil)
summary(model8)
## Call:
## lm(formula = Voltage ~ Volume + I(Volume^2) + Salinity + Surfactant +
##     Volume * Salinity + Volume * Surfactant, data = wateroil)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.54000 -0.09000  0.01333  0.12500  0.64000
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.0666667  0.1951827   5.465 0.000144 ***
## Volume            -0.1155417  0.0232047  -4.979 0.000320 ***
## I(Volume^2)        0.0012552  0.0003047   4.119 0.001423 **
## Salinity           0.6400000  0.1781766   3.592 0.103700
## Surfactant         1.1800000  0.2672650   4.415 0.000843 ***
## Volume:Salinity   -0.0078333  0.0028172  -2.781 0.016634 *   √
## Volume:Surfactant -0.0120000  0.0042258  -2.840 0.014906 *   √
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3381 on 12 degrees of freedom
## Multiple R-squared:  0.8671, Adjusted R-squared:  0.8107
## F-statistic: 13.05 on 6 and 12 DF,  p-value: 0.0001211
```

7b) (3 mark) Provide good supporting details why did you choose the model in part 7a).

In fact, all interaction terms are significant to be kept in the model and it has the highest R^2 adj =0.8107 with the lowest RMSE =0.3381 among valid models. Although Radj in model 6 is higher than the model 8, the interaction term Salinity:Surfactant is insignificant as the p-value =0.53882 , so we clearly would not select this model.