SAMPLE FINAL EXAM QUESTIONS 2020

Theory - 20 questions

Q1. If the explained sum of squares is 35 and the total sum of squares is 49, what is the residual sum of squares?

A. 10
B. 14
C. 12
D. 18

Solution: B

Explanation: Total Sum of Squares = Explained Sum of Squares + Residual Sum of Squares
SST = SSE + SSR
SSR = SST – SSE = 49 – 35 = 14

Week 2 Questions

Q2. Consider a regression model predicting the amount someone has spent (on some arbitrary product) based on their salary and age. To improve the model we have created an interaction variable between Salary and Age. Assume salary given in dollars, and age in years.
b0 = 500
b1 = 0.35
b2 = 0.12
b3=0.21

AmountSpent = b0 + b1Salary + b2Age + b3SalaryAge

How much more will a 30-year old person with a salary of $20,000 spend than a person who is 2 years older and earns $5000 less than him/her?

a)17500
b)33249.54
c)26949.76
d)26950.24

Solution: (c) Week2 Lesson 4
DifInAmountSpent = b1 * (Dif in Salary) + b2*(Dif in Age) + b3* (Dif in (Salary*Age))
DifInAmountSpent = 0.35*(5000) +0.12*(-2) + 0.21 * (120,000) = 26949.76

Week 3 Questions

Q3 Given the independent variable X and dependent variable Y, we regress Y on log(X) and get the following formula: Y = b_1 * log(X) + b_0 where b_1 and b_0 are the estimated coefficient and intercept respectively from OLS regression. How should we interpret it?

A. As X increases by 1 unit, Y increases by b_1 units
B. As X increases by 1%, Y increases by 0.01b_1 units
C. As X increases by 1 unit, Y increases by 100($e^{b_1} - 1$)%
D. As X increases by 1%, Y increases by 100($e^{0.01b_1} - 1$)%

Answer: B
As X increases by 1%, log(X) becomes log(X) + 0.01 and y_new = b_1*(log(X) + 0.01) + b_0 = y_old + 0.01b_1
Refer to Week 3 TA session notes

Q4 A speculative fund manager wants to take advantage of a mispricing in the market, where he sees a **XYZ** stock trading at $98.00 five minutes before the closing bell and decides to buy the stock. However, his order cannot be placed in time and the market closes with the **XYZ** stock price at $98.00. The next day once the market opens, he sees the best bid and offer prices of the stock **XYZ** (in $) as follows:

Bid: $99.00	**Ask: $99.50**

What is the delay cost per share (in basis points) he will incur if he places a market order immediately? (rounded to the closest integer)?

a) 102 bps
b) 153 bps
c) 51 bps
d) There is no delay cost and the order will get executed at $98

Solution: b)
If the speculator wants to buy a stock on a market order, it'll be executed in the ask price, which is $99.50. Hence, delay cost of (99.50-98) = $1.50. Basis points = (1.5/98) *10^4 = 153.06 bps

Q5 Which of the following statement is correct regarding the "factor regression"?

$$r_t^{\text{fund}} - r_t^f = \alpha + \beta_1\left(r_t^m - r_t^f\right) + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 MOM_t + \beta_5 BAB_t + \beta_6 QMJ_t + \varepsilon_t$$

A) The coefficients on Mkt-rf; SMB; HML; MOM; BAB; and QMJ tell us about exposure to the different factors
B) A positive coefficient on MOM indicates that the fund is tilted toward low momentum stocks
C) A positive and significant intercept term indicates that the fund manager has underperformed
D) All of the above

Solution A: (Week 8 Lesson 2)

Q6 Let us consider that in a small town, there are 100 households that possess a TV. Out of the 70 households that use the TV, 50 are viewing a particular channel. The HUT, Rating and Share (for the particular channel) are respectively?

a. 50/100, 50/70, 70/100
b. 20/70, 20/50, 20/100
c. 70/100, 50/70, 50/100
d. 70/100, 50/100, 50/70

Answer: D. (Week 10 Lesson 3)
HUT = (Households using TV) / (Total TV Households)
Rating = (Households viewing the channel) / (Total TV Households)
Share = (Households viewing the channel) / (Households using TV)

Q7. There are various types of reports in Google Analytics, one of them shows us how people engaged with our website, including which pages they viewed, their landing and exit pages. What is this report?

A. Acquisition Report
B. Conversion Report
C. Behaviour Report
D. Audience Report

Answer: C) Refer to Week 11 Lesson 2

Week 13 Questions

Q8 Food served at a restaurant should be between 38°C and 49°C when it is delivered to the customer. The process used to keep the food at the correct temperature has a standard deviation of 2°C and the mean value for these temperature is 40°C. According to process capability index (Cpk), is this process capable of meeting the requirements?

a. Yes, Cpk is 3.333
b. Yes, Cpk is 4
c. No, Cpk is 0.3333
d. No, Cpk is 0.4

Answer: C
Explanation:

Cpk = Minimum of [{(upper specification-$\bar{\bar{x}}$)/3s}, {($\bar{\bar{x}}$-lower specification)/3s}]
Cpk = Minimum of [{(49 – 40)/ (3*2)}, {(40 – 38)/ (3*2)}] = 0.3333
Cpk = Minimum of [{1.5}, {0.3333}] = 0.3333
Process is not capable because Cpk<1

Q9 Consider a steel rod cutting process. A critical dimension is the rod length. David has taken 4 rods per day for the past 10 days and measured them. The data from his samples are given in the table below:

Sample Number	Rod lengths (mm)			
	(i)	(ii)	(iii)	(iv)
1	144	146	154	146
2	151	150	134	153
3	145	139	143	152
4	154	146	152	148
5	157	153	155	157
6	157	150	145	147
7	149	144	137	155
8	141	147	149	155
9	158	150	149	156
10	145	148	152	154

What are the upper and lower control limits for the R chart? (Given D4 = 2.282, D3 = 0)

a. UCLr = 25.4845 mm, LCLr = 0.00 mm
b. UCLr = 24.7687 mm, LCLr = 1.32 mm
c. UCLr = 27.8445 mm, LCLr = 0.67 mm
d. UCLr = 26.4712 mm, LCLr = 0.00 mm

Answer: D
$\bar{R} = 11.6$ (average of all ranges)
UCLr = D4 * $\bar{R}$
UCLr = 2.282 * 11.6
UCLr = 26.4712
LCLr = D3 * $\bar{R}$
LCLr = 0 * 11.6
LCLr = 0

Week 14 Questions

Use the following for Questions 18-20:
Bobby Dodd works at Football Inc., a seller of high-quality footballs. He is interested in forecasting demand for his footballs that are sold weekly to Ga Tech using exponential smoothing. Assume an initial forecast of 175 and the demand data below:

Week	Demand	Forecast
1	180	175

2	168	
3	159	
4	175	

Q10 Using Exponential Smoothing and α = 0.7, forecast how many footballs need to be produced for week 4?

(A) 178.5
(B) 167.51
(C) 162.65
(D) 175.35

Ans: (C) Using the formula $F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1})$ for forecasting, we get the following table

Week	Demand	Forecast
1	180	175
2	168	178.5
3	159	171.15
4	175	**162.645**

(Week 14, Lesson 2)

Coding - 20 questions

Q1 Please use the data set 'direct_marketing.csv' to answer the following question.
We are interested in the effects of categorical variable 'Gender' and numerical variable 'Salary' on 'AmountSpent'. After running the linear regression of 'AmountSpent' on 'Gender' and 'Salary', which of the following is correct?

A. The intercept is significant on 95% confidence level
B. Male customers spend less than female customers
C. The 'Salary' effect is not significant on 95% confidence level
D. For one certain customer, higher salary generally means higher amount spent

Answer: D.
The code:

```{r}
library(readr)
```

```{r}
dirmrt <- read_csv("direct_marketing.csv")
```

```{r}
model = lm(AmountSpent ~ Gender + Salary, data=dirmrt)
```

```{r}
summary(model)
```

```
Call:
lm(formula = AmountSpent ~ Gender + Salary, data = dirmrt)

Residuals:
    Min      1Q  Median      3Q     Max
-218.06  -32.31   -5.37   28.28  374.28

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.515e+00  4.680e+00  -0.537    0.591
GenderMale   3.867e+00  4.503e+00   0.859    0.391
Salary       2.180e-03  7.357e-05  29.626   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68.72 on 997 degrees of freedom
Multiple R-squared:  0.4898,    Adjusted R-squared:  0.4888
F-statistic: 478.6 on 2 and 997 DF,  p-value: < 2.2e-16
```

According to the output, the intercept is not significant on 95% confidence level, answer A is incorrect. GenderMale's coefficient is positive, answer B is incorrect. The 'Salary' effect is significant on 95% confidence level, answer C is incorrect. For one specific customer, since Salary's coefficient is positive, higher salary means higher amount spent.

Q2 We are interested in predicting which houses have median value greater than $30,000 using the variables in the **Boston housing dataset**.
The dataset **Boston** will be loaded into the environment once the following command is run:
library(MASS)
Please run the following code in R to learn more about the variables in the dataset:
?Boston
Create a new binary variable **Result** with a value of 1 if the **medv** (median value of owner-occupied homes in $1000s) variable is greater than $30k and 0 otherwise. Create a logistic regression to model this question using all the variables in the **Boston** dataset. Please do not forget to remove the **medv** variable while building the model. Use the information from the model to answer the following two questions. Select the closest answer.
How should one interpret the coefficient of *rm*?

A. If *rm* increases by 1 unit, the natural log of the odds of the house median value being greater than $30,000 increases by 2.3549.
B. If *rm* increases by 1 unit, the odds of the house median value being greater than $30,000 increase by 2.021.
C. If *rm* increases by 1 unit, the odds of the house median value being greater than $30,000 increase by exp(2.854).
D. All of the above.

Answer: A

Explanation: If *rm* increases by 1 unit, the natural log of the odds of the house median value being greater than $30,000 increases by 2.3549. This is the direct interpretation of the coefficient of *rm* = 2.3549.

Please use the data set UPS_KO.csv to answer the following questions. For each column:
Date: This column represents date from 09/2014 to 08/2019.
Mkt_RF: This column represents market premium (i.e., Market return – risk_free rate).
SMB: This column represents size factor.
HML: This column represents value factor.
RF: This column represents risk free rate.
UPS: This column represents return of UPS.
KO: This column represents return of KO.

Q3. Estimate a three-factor model by regressing return in excess of the risk free rate on Mkt_rf; SMB; and HML for both UPS and KO
The coefficient of SMB for the three factor model for KO suggests that:

A. KO is tilted towards small cap stocks
B. KO is tilted towards large cap stocks
C. KO is tilted towards value stocks
D. KO is tilted towards growth stocks

Solution: B) Week 8 Lesson 2

```
data <- read.csv("UPS_KO.csv")
head(data)
```

Date	Mkt_RF	SMB	HML	RF	UPS	KO
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
201908	-0.0258	-0.0242	-0.0499	0.0016	-0.006779911	0.045791340
201907	0.0119	-0.0207	0.0014	0.0019	0.156870364	0.041650737
201906	0.0693	0.0033	-0.0108	0.0018	0.122018241	0.036433906
201905	-0.0694	-0.0120	-0.0239	0.0021	-0.125211831	0.001426832
201904	0.0396	-0.0168	0.0193	0.0021	-0.049400387	0.056087991
201903	0.0110	-0.0313	-0.0407	0.0019	0.022872350	0.033524475

```
library(stargazer)
factor <- lm(KO ~ Mkt_RF+SMB+HML, data = data)
stargazer(factor, align = TRUE, type = "text", out = "factorModelFull.html")
summary(factor)
```

```
================================================
                        Dependent variable:
                    ----------------------------
                                KO
------------------------------------------------
Mkt_RF                        0.540***
                               (0.104)

SMB                           -0.947***
                               (0.149)

HML                            0.003
                               (0.147)

Constant                       0.003
                               (0.004)

------------------------------------------------
Observations                     60
R2                              0.490
Adjusted R2                     0.462
Residual Std. Error       0.027 (df = 56)
F Statistic           17.899*** (df = 3; 56)
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
Call:
lm(formula = KO ~ Mkt_RF + SMB + HML, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.071984 -0.023333  0.000469  0.020360  0.052208

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.002568   0.003706   0.693    0.491
Mkt_RF       0.540413   0.103973   5.198 2.94e-06 ***
SMB         -0.946926   0.149202  -6.347 4.13e-08 ***
HML          0.002848   0.146981   0.019    0.985
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02742 on 56 degrees of freedom
Multiple R-squared:  0.4895,    Adjusted R-squared:  0.4622
F-statistic:  17.9 on 3 and 56 DF,  p-value: 2.87e-08
```

Q4. Use excess returns of UPS and KO and run three-factor (i.e., Mkt_RF, SMB, HML) models for both UPS and KO. Which of the following factor is statistically significant at the 0.01 significance level for both models?

A. Market premium

B. SMB

C. HML

D. None of above

Answer for question 4: A

````
```{r}
library(readr)
library(dplyr)
```

```{r}
UPS_KO <- read_csv("UPS_KO.csv")
```

```{r}
UPS_KO <-
  UPS_KO %>%
  mutate(UPS_RF = UPS - RF) %>%
  mutate(KO_RF = KO - RF)
```

```{r}
factor_model_UPS = lm(UPS_RF ~ Mkt_RF + SMB + HML, UPS_KO)
summary(factor_model_UPS)
```

```{r}
factor_model_KO = lm(KO_RF ~ Mkt_RF + SMB + HML, UPS_KO)
````
````

Output for KO:

```r
factor_model_KO = lm(KO_RF ~ Mkt_RF + SMB + HML, UPS_KO)
summary(factor_model_KO)
```

```
Call:
lm(formula = KO_RF ~ Mkt_RF + SMB + HML, data = UPS_KO)

Residuals:
      Min        1Q    Median        3Q       Max
-0.072350 -0.023079  0.000607  0.020961  0.051588

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.001875   0.003697   0.507    0.614
Mkt_RF       0.540239   0.103711   5.209 2.82e-06 ***
SMB         -0.944037   0.148826  -6.343 4.18e-08 ***
HML          0.007514   0.146611   0.051    0.959
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02735 on 56 degrees of freedom
Multiple R-squared:  0.4894,    Adjusted R-squared:  0.462
F-statistic: 17.89 on 3 and 56 DF,  p-value: 2.887e-08
```

Output for UPS:

```r
factor_model_UPS = lm(UPS_RF ~ Mkt_RF + SMB + HML, UPS_KO)
summary(factor_model_UPS)
```

```
Call:
lm(formula = UPS_RF ~ Mkt_RF + SMB + HML, data = UPS_KO)

Residuals:
      Min        1Q    Median        3Q       Max
-0.132991 -0.028052 -0.001813  0.034269  0.139366

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.795e-05  6.508e-03  -0.012    0.990
Mkt_RF       1.185e+00  1.825e-01   6.489  2.4e-08 ***
SMB         -5.048e-02  2.620e-01  -0.193    0.848
HML          3.858e-01  2.581e-01   1.495    0.141
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04814 on 56 degrees of freedom
Multiple R-squared:  0.4568,    Adjusted R-squared:  0.4277
F-statistic:  15.7 on 3 and 56 DF,  p-value: 1.584e-07
```

Q5. Please use the Facebook Ad dataset ***KAG_conversion_data_wrangled.csv*** for the next set of questions. We advise to solve these questions using R (preferably using *dplyr* library wherever applicable) after reviewing the code provided for Week 11 and other resources provided for learning *dplyr* in R Learning Guide

Load the dataset as:

data <- read.csv("KAG_conversion_data_wrangled.csv",stringsAsFactors = FALSE)

Which ad (ad_id) has the highest cost per click?

A. 1121223
B. 1121129
C. 1121413
D. 1121229

Answer: A. 1121223

Code: data %>% filter(CPC== max(CPC)) %>% select(ad_id)

Q6. What percentage of ads (ad_id) have a cost per click (CPC) of at least $1? (>=1). NOTE: Please round the answer to 2 decimal places

A. 59.49%
B. 76.29%
C. 78.83%
D. 85.405

Answer: C 78.83%
Code: round((data %>% filter(CPC >=1) %>% nrow())/(data %>% nrow())*100,2)

Use the following for Q7-8

In the dataset "Queue", you are given the data of the first three days of the week for a Bank of America outlet based in Downtown Atlanta. All times are in minutes.

Q7 The arrival rate per hour on Monday is 125.71 customers/hour calculated as – numbers of customers arrived divided by the total number of hours they arrived in (880/7 = 125.71). On the basis of the above, calculate the arrival rate (in hours) on Tuesday and Wednesday. Note: Consider total working hours on Tuesday and Wednesday as 8 hours and 7 hours respectively.
Now calculate the utilization rate on Monday, Tuesday and Wednesday using the above calculated arrival rates and the following service rates. Monday is 140 customer/hour, Tuesday is 120 customers/hour and on Wednesday is 200 customers/hour?

a. utilization Monday = 0.69, utilization Tuesday = 0.65, utilization Wednesday = 0.73
b. utilization Monday = 0.89, utilization Tuesday = 0.75, utilization Wednesday = 0.83
c. utilization Monday = 0.69, utilization Tuesday = 0.65, utilization Wednesday = 0.83
d. utilization Monday = 0.89, utilization Tuesday = 0.75, utilization Wednesday = 0.73

Answer: (d)
utilization Monday = 0.89, utilization Tuesday = 0.75, utilization Wednesday = 0.73

```
Queue <- read_csv("C:/Queue.csv")
head(Queue)

# QUESTION 1
#Monday <- Queue[Queue$Day == 'Monday',]
Monday <- filter(Queue, Day == "Monday")
Tuesday <- filter(Queue, Day == "Tuesday")
Wednesday <- filter(Queue, Day == "Wednesday")

# Number of customers arriving on Monday
number_of_customers_mon = nrow(Monday)
number_of_customers_tue = nrow(Tuesday)
number_of_customers_wed = nrow(Wednesday)

# Average hourly arrival rate
lambda_mon = number_of_customers_mon/7
lambda_tue = number_of_customers_tue/8
lambda_wed = number_of_customers_wed/7

mu_mon = 140
mu_tue = 120
mu_wed = 200

# Utilization rate
utilization_mon = lambda_mon/mu_mon
utilization_tue = lambda_tue/mu_tue
utilization_wed = lambda_wed/mu_wed

print(paste("The utilisation rate on Monday is", utilization_mon))
print(paste("The utilisation rate on Tuesday is", utilization_tue))
print(paste("The utilisation rate on Wednesday is", utilization_wed))
```

Q8 Using the queuing package in R, develop a MM1 model and generate a report of the queue characteristics **for Tuesday** using the arrival rates and service rates mentioned in the previous question. (take n = 20) Hint: Refer to the code in the "practice questions" file for week 12. What is the average number of customers in the queue and the average time a customer waits in the queue?

a. Average time = 0.025, Average number of customers = 2.25
b. Average time = 0.015, Average number of customers = 2.05
c. Average time = 0.035, Average number of customers = 3.15
d. Average time = 0.05, Average number of customers = 3.25

Answer: (a) Average time = 0.025 hours, Average number of customers = 2.25

Queue Package in R

```{r}
library(queueing)
#Inputs to our MM1 model
i_mm1 <- NewInput.MM1(lambda = 90, mu = 120, n = 20)

## Build the model
o_mm1 <- QueueingModel(i_mm1)

Report(o_mm1)
```

```
The inputs of the M/M/1 model are:
lambda: 90, mu: 120, n: 20

The outputs of the M/M/1 model are:

The probability (p0, p1, ..., pn) of the n = 20 clients in the system are:
0.25 0.1875 0.140625 0.1054688 0.07910156 0.05932617 0.04449463 0.03337097 0.02502823 0.01877117 0.01407838
0.01055878 0.007919088 0.005939316 0.004454487 0.003340865 0.002505649 0.001879237 0.001409428 0.001057071
0.000792803
The traffic intensity is: 0.75
The server use is: 0.75
The mean number of clients in the system is: 3
The mean number of clients in the queue is: 2.25
The mean number of clients in the server is: 0.75
The mean time spend in the system is: 0.0333333333333333
The mean time spend in the queue is: 0.025
The mean time spend in the server is: 0.00833333333333333
The mean time spend in the queue when there is queue is: 0.0333333333333333
The throughput is: 90
```

Use the following for Q9-10

You need Daily_Demand.csv for this and using simple exponential smoothing to forecast demand. Convert the csv to an xts object. (Make sure you only have 2 columns, Date and total while converting)

Q9) Model the data with alpha = 0.35 and h =5. What is the RMSE? (Hint: Use accuracy())

(a) 95
(b) 200
(c) 35
(d) 70

Ans: (a) 95

Q10) What is the MAE?

(a) 32
(b) 74
(c) 124
(d) 102

Ans: (b) 74

```
data<- read.csv("Daily_Demand.csv")%>%select(Date,total)

data.xts<-xts(data[, -1], order.by=as.POSIXct(data$Date))

model<-ses(data.xts, alpha = .35, h = 5)
accuracy(model)
```

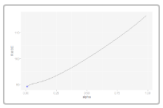
```
                     ME      RMSE       MAE        MPE      MAPE       MASE
Training set -1.305865 95.84131 74.53825 -8.060868 25.58316 0.2477396
                    ACF1
Training set -0.114236
```

```{r}
alpha <- seq(.01, .99, by = .01)
RMSE <- NA
for(i in seq_along(alpha)) {
  fit <- ses(data.xts, alpha = alpha[i], h = 5)
  RMSE[i] <- accuracy(fit)[1,2]
}


# convert to a data frame and idenitify min alpha value
alpha.fit <- data_frame(alpha, RMSE)
alpha.min <- filter(alpha.fit, RMSE == min(RMSE))
print(alpha.min)

ggplot(alpha.fit, aes(alpha, RMSE)) +
  geom_line() +
  geom_point(data = alpha.min, aes(alpha, RMSE), size = 2, color = "blue")

```



alpha <dbl>	RMSE <dbl>
0.01	89.25601