Fall 2021 MGT 6203 FINAL EXAM

**PART1 – THEORY**

Week 1

**Q1)**What would be the null hypothesis for the regression *F*-test for the following equation

  Y= β0 +β1x1 + β2x2  + u

**A.     β1 = 0 and β2 = 0**

B.    β1 = 0 or β2 = 0

C.      β0 = 0 andβ1 = 0 and β2= 0

D.   β0 = 0 orβ1= 0 or  β2 = 0

**Ans: A**, the regression F-statistics test the joint null hypothesis that the values of all the coefficients of explanatory variables except the constant are 0. The intercept is not included in the hypothesis.  (Week 1, slide 24 and 32)

Week 2

**Q2)**We have the following regression equation:

life\_span = B0 + B1\*HoursExercisePerWeek +  B2\*Smoker + B3\*Exer\_Smoker

**HoursExercisePerWeek** is a continuous variable

**Smoker**is a dummy variable (1 is smoker, 0 is non-smoker)

**Exer\_Smoker** is an interaction term HoursExercisePerWeek\* Smoker

Which variable impacts the intercept on the y axis?

1. HoursExercisePerWeek
2. Smoker
3. Exer\_Smoker
4. All of the above

Answer: B

HoursExercisePerWeek and Exer\_Smoker impact the slope.  Wk 2 page 26 slide 1

Week 3

**Q3)**As  X increases by 1%, y changes by b1% holding all other factors constant. Which model can be interpreted like this?

1. Log-Linear Model
2. **Log-Log Model**
3. Linear-Log Model
4. Polynomial

**Answer: B**

**Explanation: Lesson 3, slide 17**

Week 4

**Q4)**  A group of 20 students spend between 0 and 6 hours studying for an exam (number of hours being the predictor). The response variable has 2 outcomes: **Pass=1, Fail=0.** The logistic regression analysis gives the following output:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Coefficient** | **Std. Error** | **z-value** | **P-value** |
| **Intercept** | −4.0777 | 1.7610 | −2.316 | 0.0206 |
| **Hours** | 1.5046 | 0.6287 | 2.393 | 0.0167 |

What is the probability of passing the exam for a student who studies 5 hours?

A. 0.87

**B. 0.96**

C. 0.34

D. 0.61

*p= e(b0+b1\*x)/(1+ e(b0+b1\*x))*

*= e(-4.078+5\*1.505)/(1+ e(-4.078+5\*1.505))*

*= 0.96*

Week 4, slide 19

Week 5

**Q5).**Which of the following is **not** needed to establish causation?

a)     Hypothesized cause must precede its anticipated effect

b)    Other possible explanations that can cause the effect must be ruled out

c)     Change in cause must lead to a change in effect

d)    The effect must always have a reverse impact on the cause

Answer: D. (Module 5, slide 3)

Week 6

**Q6)**Using the stock monthly return in the chart below – calculate the Compounded return. You may use the standalone built-in calculator app on your computer, but advanced statistical tools like excel, python, r are prohibited and will trigger a violation.

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Price | Dividend | Stock Split |
| Jan | 45 |  |  |
| Feb | 26 |  | 2 for 1 |
| March | 27.2 | 0.6 |  |
| April | 17 |  | 3 for 2 |

1. 24.25%
2. 15.75%
3. -64.44%
4. -35.25%

Answer: B 15.75%. Lesson 6, slides 6-7.

Calculate the returns each month:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Date | Price | Dividend | Stock Split | Return | Calculation |
| Jan | 45 |  |  |  |  |
| Feb | 26 |  | 2 for 1 | 0.155 | *(26\*2)/45 -1 =  0.155* |
| March | 27.2 | 0.6 |  | 0.069 | *((27.2 + 0.6)/26)-1 =  0.069* |
| April | 17 |  | 3 for 2 | -0.1727 | *(17\*1.5/27.2)-1 = -0.0625* |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Compound return = 𝑟1 + 1 × 𝑟2 + 1 × ⋯ × 𝑟𝑛 + 1 - 1

**Compound return  =  (1+0.155)\*(1+0.069)\*(1-0.0625)-1  = 0. 1575**

Week 7

**Q7)**

|  |
| --- |
| Suppose you invested in a fund for 1 year. The fund return was 20% and risk-free rate was 7%. The standard deviation of the excess return was 5% and beta was 0.7. What was the fund’s Sharpe ratio? |
| A. 0.2857 |
| B. 4 |
| C. 2.6 |
| D. 0.1857 |
|  |
| **SOLUTION:** |
| **C. 2.6**  **Sharpe Ratio = (Return – Risk free rate)/Standard Deviation of excess return**  **Sharpe Ratio = 0.20-0.07/0.05 = 2.6** |

Week 8

**Q8)**

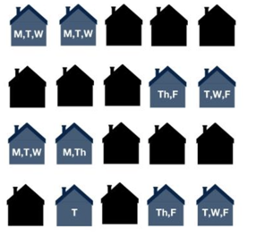
Which of the following is most likely to happen if in the long-term excessive investments are made into systematic factor funds?

1. The average annual return premiums of the factors will go down
2. The average annual return premium of the factors will go up
3. The average annual return premiums of the factors will not change
4. None of the above

**Answer: (A) Explanation: “Factor investing is becoming popular”; “As investors invest more money into these factors, the performance might dissipate”**

**(Week 8 Lesson 2 Slide 13)**

Week 9



**Q9)**The picture above shows the houses that watched the evening news, and what day they watched it for the past week. Calculate Total GRP. Note: Assume each house has a television and each house is worth 5 rating points.

     A. 120

**B**.**110**

     C. 130

     D. 115

**Explanation: week 9, lesson 1, slide 4**

**There are 20 houses,**

**Ratings per house = 100/20**

**GRP =   Monday = 4 \*5 = 20**

**Tuesday = 6 \* 5 = 30**

**Wednesday = 5 \* 5 = 25**

**Thursday = 3 \* 5 = 15**

**Friday = 4 \* 5 = 20**

**-------------------------**

**Total = 110**

**Q10)**  A $50,000 advertising budget at a CPM of 10 can get how many impressions?

A. 5,000

B. 500,000

C. 50,000,000

**D. 5,000,000**

Week 10

**Q11)**Which of the following is TRUE regarding Google’s SEM Advertising

1. The advertisement that places the highest CPC Bid will **ALWAYS**have a higher ad rank and be displayed first
2. The ad rank or position of an ad is a determined by the CPC bid **AND**Quality Score
3. The Quality score for an ad considers only ad relevance
4. Both B and C

Answer: B; CPC Bid x Quality Score = Ad Rank (position). C is incorrect because Quality score is determined by relevance, CTR and landing page

Week 10 slide 15

**Q12)**Which metric is the riskiest for a company that publishes ads?

1. CPS – Cost per sale
2. CPM – Cost per Mille (Thousand)
3. CPC – Cost per Click
4. CPV – Cost per View

Answer: A – Based on the advertising risk principle, CPS is the riskiest for the publisher

Week 10, slide 33

Week 11

**Q13)**  A small organic pet food business would like to improve their website’s conversion rates. Which of the following is not a method for improving conversion rates?

A. A/B testing

B. Customer Journey Analysis

C. Multivariate testing

D. All of the above are methods for improving conversion rates.

Answer: D, Page 10 S2

Week 12

**Q14)**What are the assumptions of M/M/1 model?

A. Infinite Queue Length, Infinite Customer Population, Poisson Random Arrivals, Single Channel

B. Finite customer population, Exponential random arrivals, FIFO, Multiple Line

C. Normal Random Arrival Times, Double Phase, FIFO, Multiple Channels

D. Finite Customer Population, Finite Queue Length, FIFO, Single Channel

Answer: A, Intro and Managing Queues, Lesson 5 Page 22 Slide 2

**Q15)**In response to the ongoing pandemic, a local restaurant has implemented social distancing measures, which include limited number of people waiting in the system. The owner would like to ensure with a 95% confidence interval (or service level) that no more than 3 customers will be in the system at any time. What would be the utilization rate and arrival rate assuming that the staff can service customer at the rate of one every three minutes?

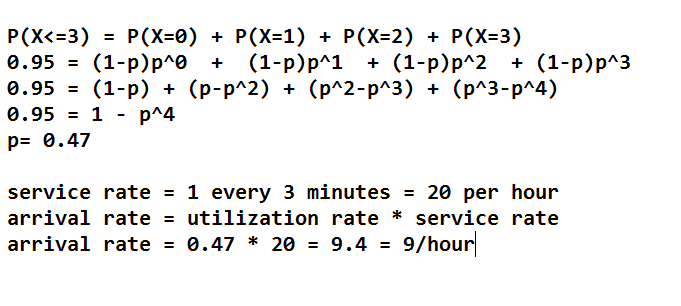
A. Utilization rate = 0.52, Arrival rate = 9/HR

B. Utilization rate = 0.52, Arrival rate = 12/HR

C. Utilization rate = 0.47, Arrival rate = 9/HR

D. Utilization rate = 0.47, Arrival rate = 12/HR

Answer: C, Intro and Managing Queues, Lesson Queueing Theory M/M/1 Model, Slide 24



Week 13

**Q16)**Which of the following would be considered an internal failure cost of quality?

A. Re-testing

B. Processing warranty claims and replace broken parts

C. Inspecting and testing all raw materials

D. Sending employees to train on the latest technology

Answer: A, What are the Costs of Quality? Module 13 Lesson 2; Page 6

**Q17)** After a meeting with your customer they give you the USL (upper specification limit) of 16 and the LSL (lower specification limit) of 2.   You take a sample and find the mean of the sample is 12 and a standard deviation of 2.

What is the Cp (round to two decimal places)?

1. 1.17
2. 0.67
3. 2.34
4. 4.67

Answer: A, (USL-LSL)/(6\*Stdv)= (16-2)/(6\*2) = 1.17

Slide 22

Week 14

**Q18)**Quickest Trippy is a local gas station. They want to predict demand for gasoline and have the following historical data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Month** | **Demand (in gallons)** | **Forecast (in gallons)** | **RSFE** | **MAD** |
| **April** | 13000 | 11000 | 2000 | 2000 |
| **May** | 18700 | 12200 | 6500 | 6500 |
| **June** | 10500 | 16100 | -5600 | 5600 |
| **July** | 18700 | 12740 | 5960 | 5960 |
| **August** | 22000 |  |  |  |

Using Exponential Smoothing with α = 0.6, calculate TS. According to TS value that you calculated, do you need to investigate the model?

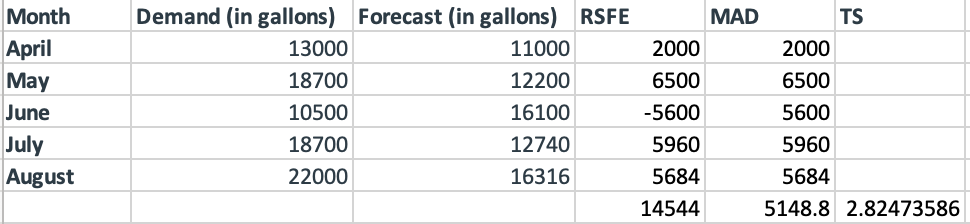
A. TS = 0.56, Do not investigate

B. TS = 0.56, Investigate

C. TS = 0.11, Investigate

D. TS = 2.82, Do not investigate

Answer: D, Forecasting Demand, Lesson Error Method, Slide 28



RSFE = 14544

MAD = 5148.8

TS = RSFE / MAD = 2.82

Since |2.84| < 4, we do not investigate the model

Week 15

**Q19)** A local drugstore is selling a popular brand of air purifier. The demand is averaged at 500 units **annually.** The ordering price each unit is $60 if they order 99 or less units, and they can get a 5% discount if they order 100 or more units. Each time when the drugstore orders the product from the supplier, it costs $50, holding cost for the inventory in the store is estimated to be $10 per unit annually. To keep the total cost low, how many units of the air purifier should the drugstore order every time.

1. 51 units
2. 71 units
3. 100 units
4. 121 units

**Answer:** C

EOQ= (2\*D\*S/H)^0.5=71

Option1: not take the discount    
Total cost=(500\*50/71) + 71/2\*10 +60\*500= 30707

Option2: take the discount, EOQ=100

Total cost=(500\*50/100)+100/2\*10+60\*.95\*500=29250

Option 2 costs less, they should order 100 units.

**Q20)** A local drugstore is selling a popular brand of air purifier. The demand is averaged at 500 units **annually**, and the standard deviation of **daily** demand is estimated to be 2 units.  Each unit costs $60 with no other discounting program and the annual inventory cost is 8% of unit cost. Ordering cost is $50, and it usually takes about 5 days to arrive to the store.  The drugstore store manager would like to keep the in-stock rate above 90% (Z value = 1.28). How many units of the air purifier should be ordered each time? At what stock level (units left in store) should they start a new order (round the answer to the nearest integer)?

1. 72 units, 13units
2. 102 units, 13 units
3. 72 units, 20 units
4. 102 units, 20 units

**Answer:** B.

EOQ=(2\*D\*S/H)^0.5=(2\*500\*50/(60\*0.08))^0.5=102

Lesson 4. Slide 9, lesson 6,7 additional examples.

ROP=(500/365\*5)+1.28\*2\*(5)^0.5=12.57~13

Lesson 5, slide 23, lesson 6,7 additional examples.

**PART2 – CODING**

Week 1

Instructions: For this problem, load the Auto dataset from the ISLR package

**library**(ISLR)    
**library**(car)

## Loading required package: carData

**Q1)** Simple liner regression: Run a linear regression model with mpg as the dependent variable and horsepower as the predictor

**What is the Adjusted R-squared:**

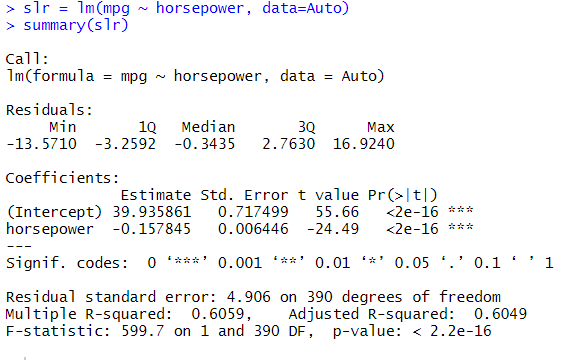
A. 0.6059

B. 0.7174

C. 0.0065

D. 0.6049

#Answer: D Adjusted R-squared:  0.6049



Week2

Using the mtcars data set that is built into R, create 3 linear regression models:

Model 1 is a simple linear regression with mpg as the dependent and hp as the independent

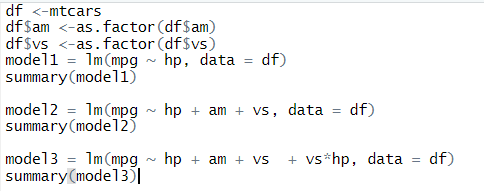
Model 2 is a multiple linear regression with mpg as the dependent and hp as the independent, am as a categorical and vs as a categorical

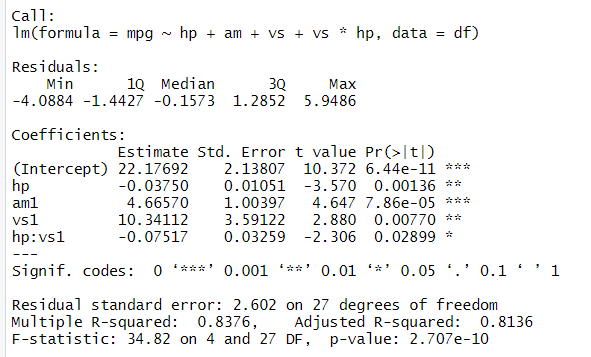
Model 3 is a multiple linear regression with mpg as the dependent and hp as the independent, am as a categorical, vs as a categorical, and vs\*mpg as an interaction term

**Q2)** Which of the 3 models has the highest Adj\_R Squared?

1. Model 1
2. Model 2
3. Model 3

Ans. Model 3





Week4

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. Then create a logistic regression model that uses all the variables in the **Boston** dataset as features, and **Result** as the response. 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.

**Q3)** How should one interpret the coefficient of *rm*?

1. *If rm* increases by 1 unit, the natural log of the odds of the house median value being greater than $30,000 will increase by 2.3549, holding all other variables fixed.
2. If *rm* increases by 1 unit, the odds of the house median value being greater than $30,000 will increase by 2.3549, holding all other variables fixed.
3. If *rm* increases by 1 unit, the probability of the house median value being greater than $30,000 will increase by 2.3549, holding all other variables fixed.
4. If *rm* increases by 1 unit, the number of houses with median value greater than $30,000 will increase by 2.3549, holding all other variables fixed.

**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.

**Q4)** What is the Sensitivity of the model? (Use threshold 0.5 and consider “1” as the positive reference case in the confusion matrix)

1. 0.957
2. 0.143
3. 0.857
4. 0.878

**Answer:**C

**Explanation:**Based on the Confusion Matrix we can compute sensitivity as:

Sensitivity = True Positives / (True Positives + False Negatives) = 72 / (72 + 12) = 0.857

library(MASS)

library(dplyr)

?Boston

library(ROCR)

Boston$Result <- ifelse(Boston$medv > 30,1,0)

names(Boston)

Boston$medv <- NULL

# Apply logistic regression algorithm on Boston data set train

logis <- glm(Result ~ ., data = Boston, family = binomial)

# Answer to QUESTION 1, 2 and 3

summary(logis)

# Predict using the model built

Boston <- Boston %>% mutate(pred = predict(logis, data=Boston, type = "response")) %>% mutate(pred = ifelse(pred >= 0.5,1,0))

# Calculate confusion matrix

xtabs(~Result + pred, data = Boston)

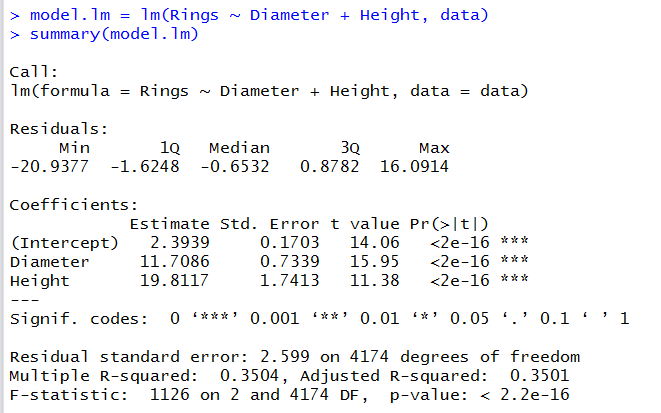
Week 5

Use the abalone.csv dataset to answer questions from 1 to 3:

**Q5)** Using the function lm, create a linear regression that regresses “Rings” onto “Diameter” and “Height” (i.e., “Rings” is the response variable and “Diameter” ,“Height” are the independent variables).  Which one of the following statements is FALSE?

1. Diameter is significant at a 5% Confidence Interval
2. R-Squared is 0.35
3. **The intercept is 2.3939**
4. One unit of increase in height causes the number of rings to increase by 19.81 on average keeping "diameter" constant

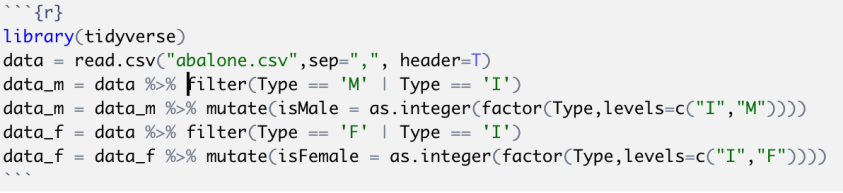
**Explanation:  The intercept is 2.3939 (Lesson 1, Video 7, Slide 4)**

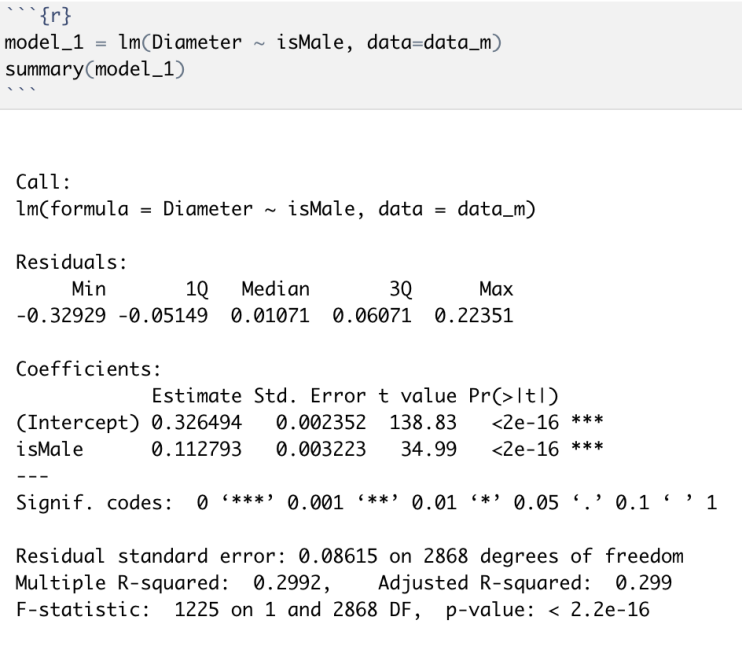


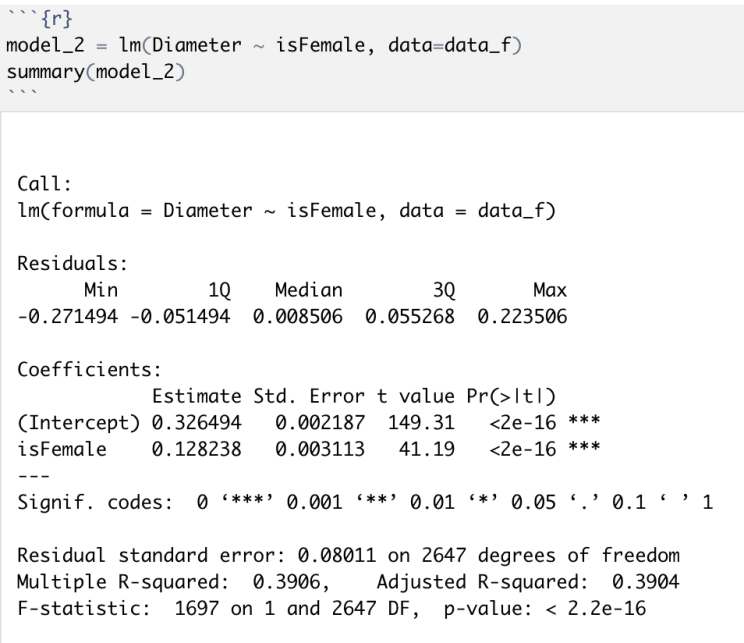
**Q6)** Create two separate datasets from abalone.csv. The first data set will contain Male (M) abalones and infant (I) abalones.  The second dataset will contain Female (F) abalones and infant (I) abalones.  Now use a linear regression model to compute the diameter difference estimator (average difference in diameter) for each dataset (Note: take infant (I) as the reference in each case). What is the diameter difference estimator for Male/Infant and Female/Infant respectively?

1. 0.125, 0.118
2. **0.113, 0.128**
3. 0.128, 0.113
4. 0.118, 0.125

Explanation: The b1 coefficient for each model is 0.113 and 0.128 respectively. See summaries below (Lesson 5/ Video 3 / Slide 4  and Lesson 5/ Video 4/1-10)







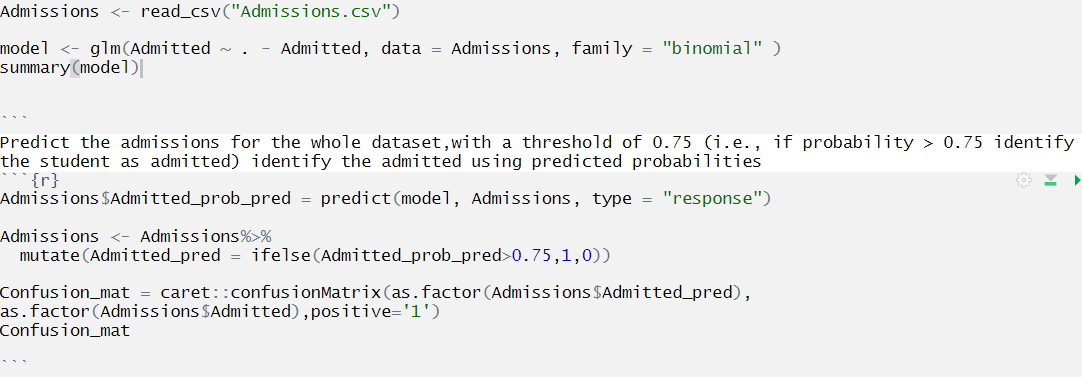
Use the Admissions.csv dataset to answer following questions

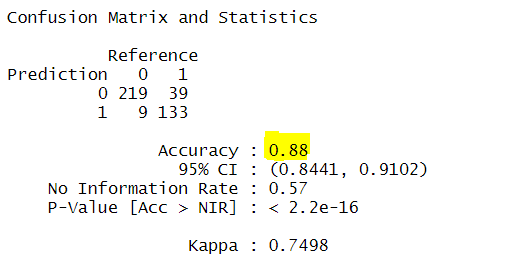
Fit a logistic regression model using Admitted as the response variable and all the other variables as independent variables. Once the model is done, predict the probabilities of getting admitted for all the datapoints. Using a threshold of 0.75 identify the students as admitted using predicted probabilities (i.e., if probability > 0.75 identify the student as admitted).

**Q7)** What is the Accuracy of the logistic regression model?

1. 0.84
2. **0.88**
3. 0.76
4. None of the above

Page Break





**Explanation: Lesson 3 / Video 6 / Slides 3 – 8**

Week 6-7

**For Questions 8-11 you will use the dataset “managers” from the PerformanceAnalytics package in R. This dataset is already in XTS format and no further data cleaning is necessary. Please use this dataset as is and DO NOT impute any null values with 0. The dataset begins on 1996-01-31 and end on 2006-12-31. It contains the following variables:**

•      **HAM1-HAM6:** Columns of monthly returns for six hypothetical asset managers

•      **EDHEC LS EQ:** EDHEC Long-Short Equity hedge fund index

•      **SP500 TR:** S&P 500 total returns

•      **US 10YR TR:** Total return series for US Treasury 10-year bond

•      **US 3m TR:**Total return series for US Treasury 3-month bill

**The dataset "managers” can be found by using the following code:**

if (!require(PerformanceAnalytics)) install.packages("PerformanceAnalytics")

library(PerformanceAnalytics)

data(managers)

**Note: You may/may not need the following dependencies:**

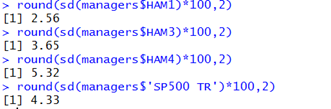
**lubridate package**

**Q8)**

a) Which of the following investments (Hypothetical Asset Manager 1, Hypothetical Asset Manager 3, Hypothetical Manager 4, S&P 500) has the highest standard deviation over the entire period given by the dataset.  (Dataset period: 1996-01-31 to 2006-12-31)

1. Hypothetical Asset Manager 1
2. Hypothetical Asset Manager 3
3. Hypothetical Asset Manager 4
4. S&P 500

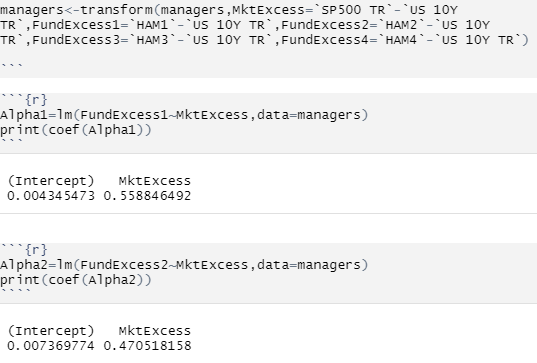
**Answer: C. Hypothetical Asset Manager 4**

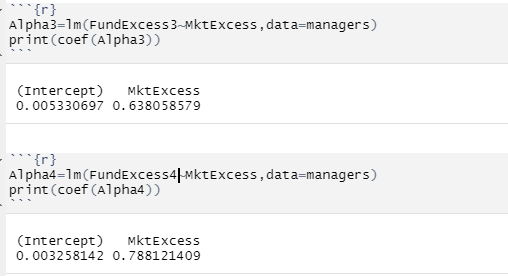


**Q9).** Which manager had the highest Jensen’s Alpha? (Hint: You can use the US 10 Year Treasury Bond as the Risk Free Rate and the S&P 500 Data as the Market Return)

1. Hypothetical Manager 1
2. Hypothetical Manager 2
3. Hypothetical Manager 3
4. Hypothetical Manager 4

**Answer: Hypothetical Manager 2**





**Q10)**Find the Treynor ratio for HAM4 and HAM3 (using the Performance Analytics library). When finding both ratios use “SP500 TR” as your benchmark market return. Also, be careful to select the "US 10Y TR” as your Risk-Free rate. Comparing the Treynor ratio values, which had a stronger performance?

1. HAM3’s Treynor ratio is higher than HAM4, thus HAM3 had more reward per unit of systematic risk than HAM4.
2. HAM3’s Treynor ratio is lower than HAM4, thus HAM3 had less reward per unit of systematic risk than HAM4.
3. HAM4’s Treynor ratio is higher than HAM 3, thus HAM4 had more reward per unit of systematic risk than HAM3.
4. HAM3 and HAM 4 have the same Treynor Ratio and thus have equivalent reward to systematic risk profiles.

Answer:

A. HAM3 is higher, HAM3 fund had a stronger performance than HAM 4 according to the Treynor ratio.

Explanation: Module 7, slide 8

TreynorRatio(managers$HAM3,managers$'SP500 TR',managers$'US 10Y TR')

TreynorRatio(managers$HAM4,managers$'SP500 TR',managers$'US 10Y TR')

#Treynor HAM3 = 0.140099185889986

#Treynor HAM4 = 0.0751525920385063

**Q11)**Find the Treynor Ratio of the following equally weighted Finance Portfolio. Use the Contrafund data to get the data for the risk free rate.

Stocks: Bank of America (BAC), Visa (V), Wells Fargo (WFC), Goldman Sachs (GS)

Date Range: 1st Jan 2017- 31st December 2017

Also, in which month did the portfolio perform the best?

1. 0.004, February
2. 0.004, September
3. 0.002, February
4. 0.002, September

Code: (To get best portfolio return month just look at the dataframe of the monthly portfolio and select the one with highest returns)

library(tidyverse)

library(tidyquant)

library(PerformanceAnalytics)

library(xts)

library(lubridate)

stock\_prices <- c("BAC","V","WFC","GS") %>%

    tq\_get(get  = "stock.prices",

           from = "2017-01-01",

           to   = "2017-12-31")

#stock\_prices

stock\_returns\_monthly <- stock\_prices %>%

    group\_by(symbol) %>%

    tq\_transmute(select     = adjusted,

                 mutate\_fun = periodReturn,

                 period     = "monthly",

                 col\_rename = "Ra")

#stock\_returns\_monthly

weights\_ans <- c(0.25,0.25,0.25,0.25)

portfolio\_returns\_monthly <- stock\_returns\_monthly %>%

    tq\_portfolio(assets\_col  = symbol,

                 returns\_col = Ra,

                 weights = weights\_ans,

                 col\_rename  = "Ra") %>% mutate(date=substring(date,1,7)) %>%

  rename(Date=date)

#portfolio\_returns\_monthly

contra\_fund<-read.csv("contrafund.csv")

contra\_fund$Date<-mdy(contra\_fund$Date)

contra\_fund <- contra\_fund %>% mutate(Date=substring(Date,1,7))

final\_portfolio <- left\_join(portfolio\_returns\_monthly,

                                   contra\_fund,

                                   by = "Date") %>% mutate(Date=paste(Date,"01",sep="-"))

final\_portfolio$Date <- as.Date(as.character(final\_portfolio$Date))

final\_data <-xts(final\_portfolio[,-1],final\_portfolio$Date)

treynor\_portfolio <- TreynorRatio(final\_data$Ra,final\_data$Risk.Free)

treynor\_portfolio

Week 8 (1 dataset 2 questions)

**For Questions 12 & 13, use the provided dataset Final\_Exam\_Factors.csv to construct factor regression models for NVDA and INTC. (Use all factors provided to construct the factor regression model.  Please note, date is not a factor and will not be used as a feature in your model.  Also, the monthly returns for both stocks are the total return and not excess return)**

Date: Monthly dates from 1/31/2007 - 6/30/2018

SMB: The monthly return factor attributed to the size factor

HML: The monthly return factor attributed to the value factor

QMJ: The monthly return factor attributed to the profitability factor

BAB: The monthly return factor attributed to the betting against beta factor

MOM: The monthly return factor attributed to the momentum factor

RF: The monthly risk-free rate

MKT: The monthly return of the broad market (S&P 500 index)

Mkt\_RF: The monthly return factor of the market minus the monthly risk-free rate

NVDA: The monthly return for Nvidia Stock

INTC: The monthly return for Intel Stock

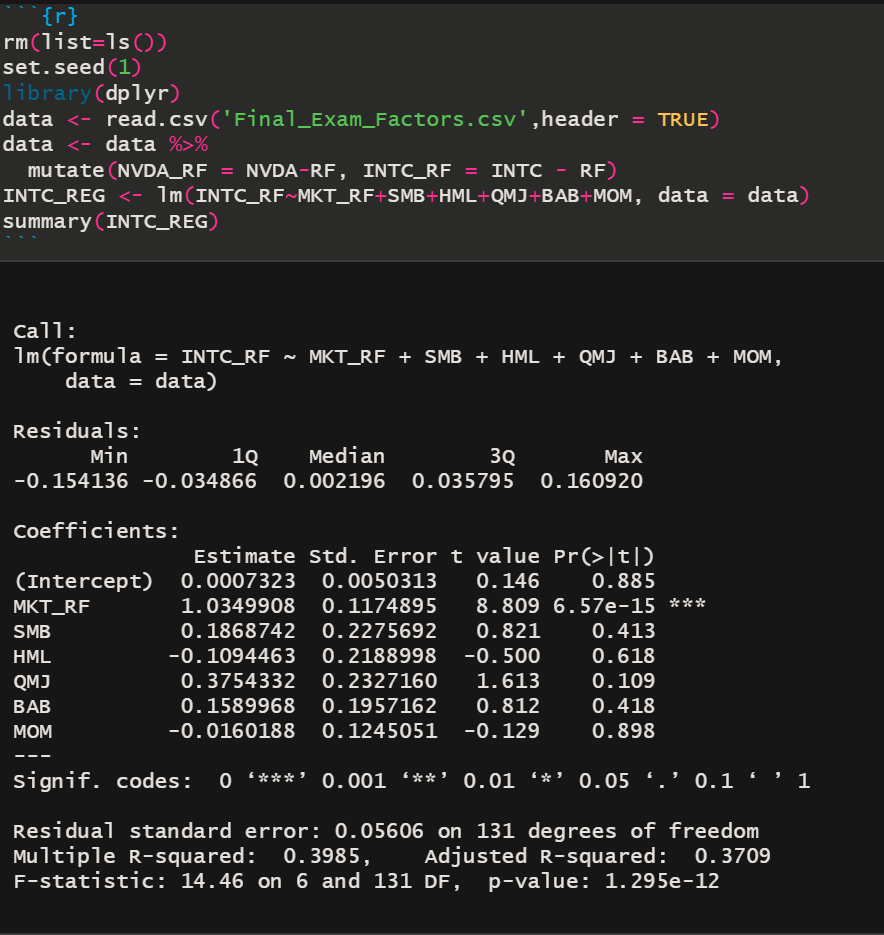
**Note:**Assume a coefficient significance threshold of a = .1

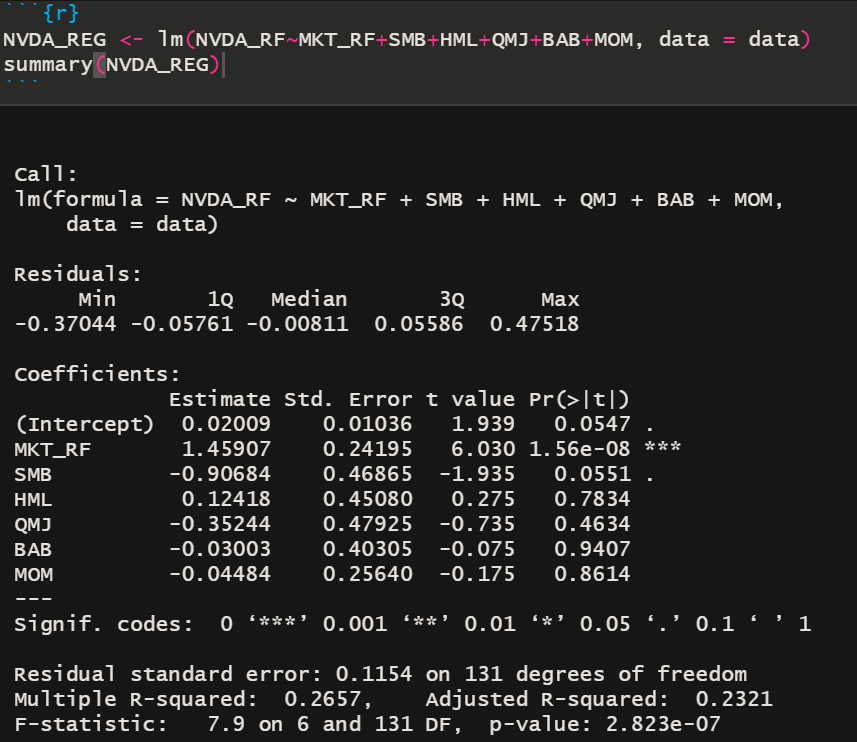
**Use the output of these factor models to answer the following:**

**Q12)**Which of the following statements is TRUE? (Select one)

1. Intel has a statistically significant BAB coefficient at the a=.1 significance threshold with a value of 0.1589968.
2. NVDA has a statistically significant MOM coefficient at the a=.1 significance threshold with a value of –0.04484.
3. Intel has a statistically insignificant HML coefficient at the a=.1 significance threshold with a value of -0.114862.
4. NVDA has a statistically significant QMJ coefficient at the a=.1 significance threshold with a value of –0.35244.
5. **NVDA has a statistically significant SMB coefficient at the a=.1 significance threshold with a value of –0.90684.**

Answer: E; The only significant coefficient for INTC was “MKT\_RF”. The only significant coefficient for NVDA was “SMB” and “MKT\_RF”.





**Q13)**Which of the following statements is TRUE? (Select one)

1. Nvidia has a positive and significant MKT\_RF coefficient at the a=.1 significance threshold.
2. Nvidia has a negative and significant SMB coefficient at the a=.1 significance threshold. Meaning that it is tilted towards large cap.
3. Intel has a positive and significant MKT\_RF coefficient at the a=.1 significance threshold.
4. Intel and Nvidia both exhibit a beta greater than the beta of the market.
5. **All of the above.**

Answer: E. For answers A-C see question 9 solution above. Answer D is correct because both stocks demonstrate MKT\_RF coefficients greater than one which is always the value of the market beta.

Week 9

Please use the Facebook Ad dataset *KAG\_wrangled\_dataset****.csv***for the next question. We advise to solve the questions using R (preferably using *dplyr*library wherever applicable) after reviewing Week 7 and other resources provided for learning *dplyr*in R Learning Guide

NOTE: For no clicks and no amount spent, please consider CPC as 0 for all the questions.

Load the dataset as:

data <- read.csv("KAG\_wrangled\_dataset.csv ",stringsAsFactors = FALSE)

**Q14)**.

Of all the ads with more than 10,000 impressions, what percentage of these ads were targeted at females?  Note:  Please round the answer to 2 decimal places.

A.  51.46%

B.  68.58%

C.  69.51%

**D.  48.54%**

Solution: From HW3 and dplyr library,



\*KAG\_wrangled\_dataset.csv is named as KAG\_conversion\_data in the solution

**Q15 – Q16**

A local travel agency has just hired you to help them understand their customer churn. They have provided you with a dataset that includes the following information:

* Age - Customer age
* Frequent Flyer - Is the customer a frequent flyer
* Annual Income Class - The customers annual income class
* Services Opted - The number of traveling services used by the customer
* Account Synced To Social Media - Is the customer’s social media account linked to their travel account
* Booked Hotel - Does the customer book hotels through the travel agency
* Churn - Did the customer churn (0 = No, 1 = Yes)

**Q15)**

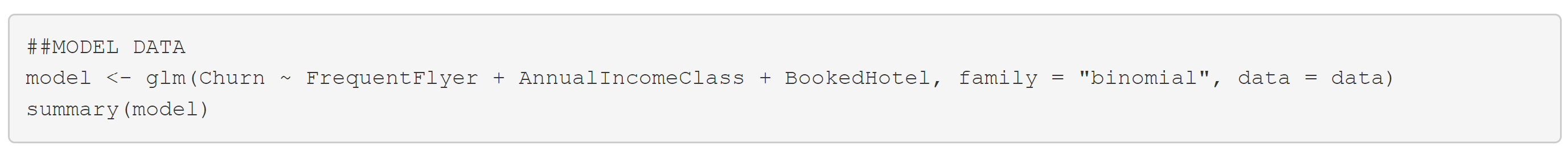
For your initial analysis the travel agency wants you to focus on three features - Frequent Flyer, Annual Income Class, and Booked Hotel. Using these three features, they want you to identify which of the following customer segments has the highest probability of churning. Build your churn model using logistic regression.

1. Frequent Flyer = Yes, Annual Income Class = High , Book Hotel = Yes
2. **Frequent Flyer = No, Annual Income Class = Middle , Book Hotel = No**
3. Frequent Flyer = Yes , Annual Income Class = Middle , Book Hotel = Yes
4. Frequent Flyer = No, Annual Income Class = Low, Book Hotel = Yes

**Answer: B) Frequent Flyer = No, Annual Income Class = Middle, Booked Hotel = No**  
With a linear regression model, each feature is a linear combination of the log odds of churn. Therefore, for each feature we can select the category that increases the log odds of churn the most. Based on the model summary the following categories increase the log odds of churn the most - Frequent Flyer = No, Annual Income Class = Middle, Booked Hotel = No. Alternatively, you could create a data set with each customer segment, predict the churn for each customer segement, and select the customer segement with the highest churn.

**Scatter chart

Description automatically generated with medium confidence**

****

**Text

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Q16)**

The local travel agency would like to update your logistic regression churn model to include: Frequent Flyer, Annual Income Class, Booked Hotel, Age, and Services Opted. They have a list of 5 customers and their attributes, which is shown below. They are interested in identifying all customers with more than a 90% probability of churning. Update your logistic regression churn model, and identify customers with more than a 90% probability of churning.

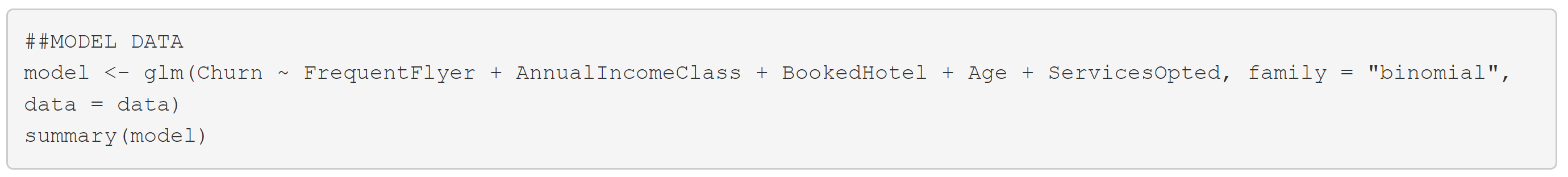
| **Customer Name** | **FrequentFlyer** | **AnnualIncomeClass** | **BookedHotel** | **Age** | **Services Opted** |
| --- | --- | --- | --- | --- | --- |
| James | Yes | Middle | Yes | 21 | 0 |
| Sally | No | High | Yes | 32 | 0 |
| David | Yes | Low | Yes | 43 | 6 |
| Ben | No | Middle | No | 54 | 5 |
| Karen | Yes | High | No | 65 | 1 |

1. **Ben, Karen**
2. Sally, David, Ben
3. James, Karen
4. David, James

**Answer: A) Ben, Karen**

**A picture containing application

Description automatically generated**

****

**Text

Description automatically generated**

**A picture containing text

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

Week 12

**Q17)** You work for a logistics company and your manager has asked you to evaluate their shipping and outbound system queue. Specifically, the manager wants to better understand the relationship between packing rates and the resulting average amount of time a package spends in the system. The warehouse's average service rate is a function of the number of servers working such that (number\_of\_workers\*72 = average\_packages\_serviced\_per\_hour). At any given time, the manager can have 16 to 24 workers in the warehouse. Assuming the warehouse operation meets M/M/1 model assumptions and has an average package arrival rate of 1000 packages/hour, what is the minimum number of workers required to meet the threshold of 9 seconds for the average time a package spends in the system?

1. 16 workers
2. 18 workers
3. **20 workers**
4. 24 workers

**Explanation:** With 18 workers the average time a package spends in the system is 12.16 seconds, and with 20 workers the average time a package spends in the system is 8.2 seconds.

Chart

Description automatically generated

**Q18)**Using the same model assumptions as the previous question and assuming your answer for the previous question being 21 servers, what is the probability that exactly 7 packages are in the system? Round your answer to the nearest hundredth of a percent.

1. **1.87%**
2. 4.12%
3. 10.42%
4. 12.36%

**Explanation:** With 21 servers we can calculate the probability that exactly 7 packages are in the system by applying the formula on Module 12 slide 23 “Probability of n units in the system”

A picture containing text, clock, watch

Description automatically generated

Which yields the following:

Graphical user interface, text, application, email

Description automatically generated

WEEK 13

After taking 8 samples (each sample contains 5 parts) you want to calculate the control limits of the X-bar and R charts.

Copy the following into R

x <- data.frame("sample" =c("sample1", "sample2", "sample3", "sample4", "sample5", "sample6",

                            "sample7", "sample8" ),

                "part1" = c(48.4, 47.9, 41.1, 49.6,    53.8, 58.5, 56.5, 55.1),

                "part2" = c(48.2, 41.7,    54.9,    46.1,    55.9,    53.9,    54,    49.7),

                "part3" = c(45.2,    46.9,    46.8,    46.3,    43.4,    56.7,    51.8,    49.6),

                "part4" = c(53.9,    45.1,    55.5,    54.9,    55.4,    51.9,    54.3,    45.7),

                "Part5" = c(51.4,    50.3,    50.4,    44.1,    56.3,    52,    50.2,    52.5))

D3 <-0

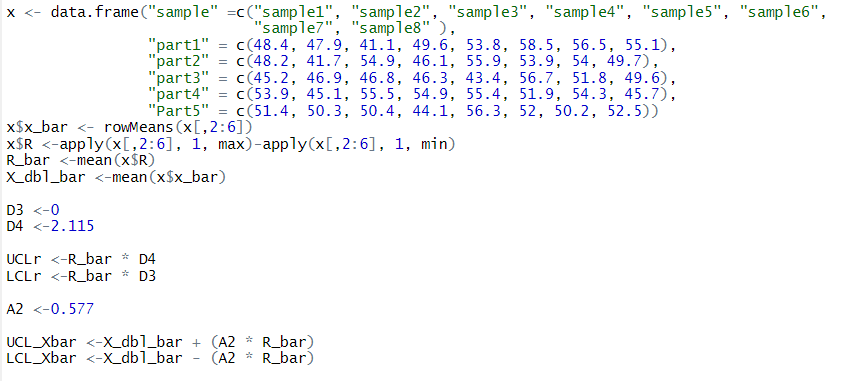
D4 <-2.115

A2 <-0.577

**Q19** What is the UCLr, LCLr, UCL\_Xbar and LCL\_Xbar respectively?  (Hint: you may want to use the rowMeans function and the apply function)

1. UCLr= 45.04, LCLr = 0, UCL\_Xbar = 56.25 and LCL\_Xbar = 20.5419
2. UCLr= 20.5419, LCLr = 0, UCL\_Xbar = 56.25 and LCL\_Xbar = 45.0433
3. UCLr= 56.25, LCLr = 20.5419, UCL\_Xbar = 45.04 and LCL\_Xbar = 0
4. None of the above

Ans: B



Wk14, slides 25-32

Week 14  - Dataset uploaded to files

Use the***Store\_Demand\_Final.csv*** file for Q18-20. You will be using simple exponential smoothing to forecast demand. The dataset has two columns Date and total\_demand with 100 entries.

Data cleaning:

* There are no NA values and no need to remove NA values
* Convert Date column to date format (month, day, year). Hint: when you run class(data$Date), it should return “Date” and not “factor” or “character”
* Convert the csv to an xts object

**Q20)**Model the data with alpha = 0.25 and h =5. What is the RMSE? (Hint: Use accuracy())

1. -22.18246
2. 926.2459
3. 786.8635
4. 909.0658

Ans: (d) 909.0658

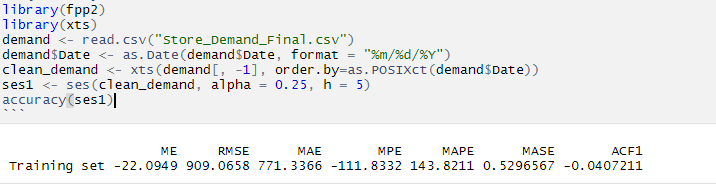
Explanation: SA9 Q3 – Q10, Sample Exam Q9 – 10; Wk14 slide 25

Q21)What is the MAE?

1. 771.3366
2. 926.2459
3. 786.8635
4. -110.3856

Ans: (a) 771.3366

Explanation: SA9 Q3 – Q10, Sample Exam Q9 - 10; Wk14 slide 25



**Q22**Plot the model from Q27, observe plot movement of historical part, what do you see?

1. It is continuously decreasing
2. It is continuously increasing
3. It is moving up and down
4. It is a straight line

Ans: (c) It is moving up and down

Explanation: SA9 Q5

