MGT6203 - HW2

Akshay Gangavalli

1/29/2020

## Questions 1 to 3 - Weeks 1 & 2 Content

**Q.1** In a linear regression problem, we are using “R-squared” to measure goodness-of-fit. We add a feature (variable) in linear regression model and retrain the same model.

Which of the following option is true?

A. If R Squared increases, this variable is significant.  
B. If R Squared decreases, this variable is not significant.  
C. Individually R squared cannot tell about variable importance. We can’t say anything about it right now.  
D. None of these.

C Explanation - “R squared” individually can’t tell whether a variable is significant or not because each time when we add a feature, “R squared” can either increase or stay constant. But, it is not true in case of “Adjusted R squared” (increases when features found to be significant).

**Q.2** A correlation between age and health of a person found to be -1.09. On the basis of this you would tell the doctors that:

A. Age is a good predictor of health

B. Health is a good predictor of age

C. Age is a poor predictor of health

D. None of these

Solution: D Explanation - Correlation coefficient range is between [-1 ,1]. So -1.09 is not possible.

**Q.3** You work for a bank where you are trying to predict the probability of default of a customer based on FICO score and annual income. Which of the following problems can arise while using a multiple linear regression model?

A. There exists homoskedasticity in the model.  
B. The model can produce predicted probabilities that are less than zero and greater than one. C. The model leads to the omitted variable bias as only two independent factors can be included in the model.  
D. The model leads to an overestimation of the effect of independent variables on the dependent variable.

Solution : B Since Linear Regression models and predicts continuous variables, the predicted probability may lie outside the prescribed values for a probability (that is, outside the range of 0 and 1). This is where a Logistic Regression helps.

## Questions 4 and 5 - Week 3 Content

We try to build a model for NBA players’ salary. Download the dataset ***nba2017.csv*** from here:

https://gatech.box.com/s/qdkpwlxxo0tyxs4kw0m8wyxec5fbhvc7

Load the dataset using the code *nba = read.csv("nba2017.csv", header = TRUE)*. Now we take a closer look at the data set. There are four variables salary, Ht(Height),

Exp(Experience) and expsq(the square of Experience).

First, build a model using salary as the response and Ht and Exp as variables and denote it as Model\_1. Build a second model using log(salary) as the response and Ht and Exp as variables, we denote it as Model\_2.

**Q.4** For Model\_1, what is the interpretation for the coefficient of height?

A. One unit increase in height increases salary by 2253985 units

B. One unit increase in height increases salary by 874758 units

1. One unit increase in height decreases salary by 677390 units

D. One unit increase in height increases salary by 677390 units

Answer: A – Code Below

**Q. 5** For Model\_2, what is the interpretation for the coefficient of height?

A. When height increases by 1%, salary increases by 68.89%

1. When height increases by 1 unit, salary increases by 0.6889%
2. When height increases by 1%, salary increases by 0.6889 units
3. When height increases by 1 unit, salary increases by 68.89%

Sol- D (code below)

Sol:

nba <- read.csv("nba2017.csv", header=TRUE)  
model\_1 <- lm(Salary ~ Ht+Exp, data = nba)  
summary(model\_1)

##   
## Call:  
## lm(formula = Salary ~ Ht + Exp, data = nba)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11584274 -2907609 -1371039 1880620 19884672   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12353214 5785552 -2.135 0.0332 \*   
## Ht 2253985 874758 2.577 0.0103 \*   
## Exp 677390 61100 11.087 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5728000 on 496 degrees of freedom  
## Multiple R-squared: 0.2042, Adjusted R-squared: 0.201   
## F-statistic: 63.65 on 2 and 496 DF, p-value: < 2.2e-16

model\_2 <- lm(log(Salary) ~ Ht+Exp, data = nba)  
summary(model\_2)

##   
## Call:  
## lm(formula = log(Salary) ~ Ht + Exp, data = nba)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0100 -0.6698 0.2475 1.0030 2.6381   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.39419 1.39946 6.713 5.24e-11 \*\*\*  
## Ht 0.68885 0.21159 3.256 0.00121 \*\*   
## Exp 0.16738 0.01478 11.325 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.385 on 496 degrees of freedom  
## Multiple R-squared: 0.2151, Adjusted R-squared: 0.2119   
## F-statistic: 67.96 on 2 and 496 DF, p-value: < 2.2e-16

1. A Explanation - From the output summaries above, in model\_1 the coefficient of Ht is 2253985. This means that a one unit increase in height increases salary by 2253985 units.
2. D Explanation - In model\_2, the coefficient of Ht is 0.68885. This means that a one unit increase in height, increases the ln(salary) by 0.68. This can be interpreted as one unit increase in height roughly increases the salary by 68.85%.

## Questions 6 to 9 - Week 4 Content

**Use for 6 and 7**

The following logistic regression is conducted to understand how the

odds of an admit to a university change with respect to the applicant’s

scores, and the prestige of the institution.

GRE refers to the Graduate Record Examinations offered by the students which Universities use as criteria for admission, and GPA being the Grade

Point Average of the applicant’s undergraduate studies.

Rank ranges from 1 to 4, which refers to the prestige of the university of

the. Below is shown the logit fit of the data. Interpret the representation of

rank in the below regression, through the knowledge of indicator and

dummy variables. (Revise week 2)

A screenshot of a cell phone

Description automatically generated

**Q6**. By how much does the value of log(odds) of an admit into a university change, if the applicant completed his undergraduate studied in a Rank 2 University as compared to a Rank 1 university.

A) 0.6709

B) 0.8040

C) 0.6754

D) 0.5089

Marks will be Given for C or D as question was interpreted differently by different students.

Answer C) Change in log odds is 0.6754, as visible in the coefficients.

Answer D) Change in odds is e^(-0.67544) as base case is Rank 1, and Rank 2has an intercept of -0.67544. Change in log(odds) is the intercept, hence change in odds is e^(intercept)

**Q7.** Choose the wrong statement from the following

A) Change in odds from Rank 2 to Rank 3 is higher than change in odds from Rank 3 to Rank 4.

B) Change in a unit of GPA corresponds to higher odds of getting an admit

than a change in a unit of GRE

C) An applicant from a University of Rank lower than 1 can never have

higher odds of an admit than an applicant from a University of Rank 1.

D) All of the above are wrong

Answer: C) Applicant can have a higher GRE Score/GPA. For A - Rank 3-4 is 0.2119 and Rank 2-3 0.2617

**Q8**. Given the following Confusion Matrix:

What is the Specificity of the fitted model denoted by the Confusion

Matrix?

A) 0.94

B) 0.91

C) 0.90

D) 0.83

Option D – The specificity is given by (True Negatives)/(True Negatives +

False Positives) = 55/(55+11) = 0.83

**Q9** Choose the correct option from the following about the effect of

increase in the cut-off value?

A) The True Positives will decrease, and the True Negatives will increase.

B) The change in cut-off should have no effect on the number of true

positives and true negatives of the model, as long as we do not change

the variables in the fitted logistic regression model

C) Both False Negatives and True Negatives will decrease.

D) The False Positives will increase, and the False Negatives will decrease

Option A – When the cut-off value is increased, The True Positives decrease,

True Negatives increase, False Positives decrease, and False Negatives increase

## Questions 10-13 Week 6 Content. (Enter the value)

**Q10.** Consider buying Coca-Cola stock. Calculate the fundamental value of Coca-Cola using the following information.

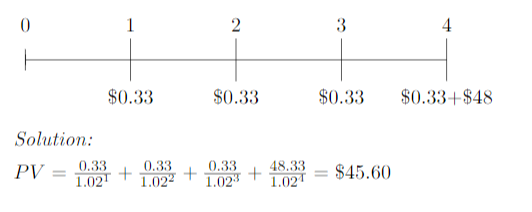
•Quarterly dividend of $0.33/share (assume a dividend was just paid)

•Coca-Cola plans to keep the dividend fixed for the next 4quarters

•Projected price in1 2-months = $48

•Quarterly discount rate of 2%

Solution:



> a1 = rep(1.02, 4)

> a1

[1] 1.02 1.02 1.02 1.02

> a2 = cumprod(a1)

> a2

[1] 1.020000 1.040400 1.061208 1.082432

> payoff = c(0.33, 0.33, 0.33, 48.33)

> payoff

[1] 0.33 0.33 0.33 48.33

> res = sum(payoff / a2)

> res

[1] 45.60113

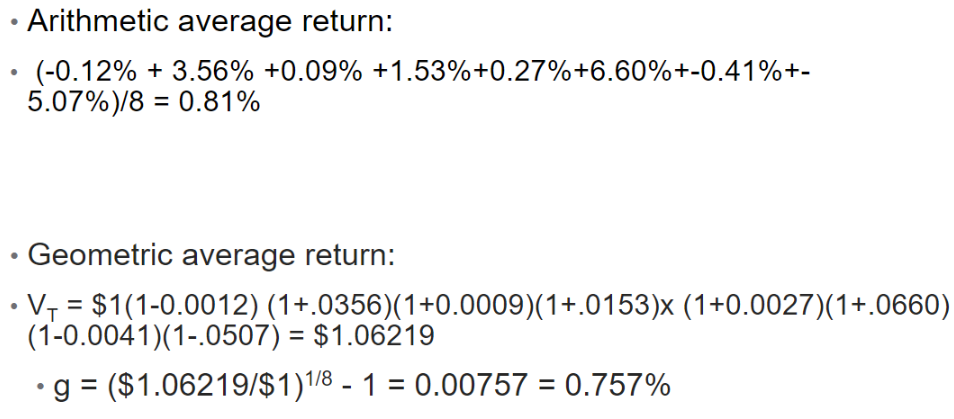
**Q11.** Find the arithmetic average return for the following data of SP 500.

A screenshot of a cell phone

Description automatically generated

**Q12.** Find the geometric average return for the above data of SP 500.

Solution:



**Q13**. Given the annual average return of a portfolio is 8.3% and the standard deviation is 17.57%. With a 3% risk free rate, please calculate the Sharpe ratio of this portfolio.

Solution:



## Questions 14 and 15 - Week 5 Content

**Q14** What of the following is an example of a natural experiment?

a)A law that changed the tax rate for some subjects, but not others.

b)A hurricane that hits a few stores among a large sample of stores.

c)Minimum wage is changed in one state but not another.

d)All of the above.

D (All of the above) Explanation - these examples are directly taken from week 5 notes, slide 15. In a natural experiment, the subjects who might be undergoing treatment are not able to choose if they are in the treatment or control group. This choice is made by an external agent, or a factor like weather, a policy change, etc.

**Q15**. Random assignment (in a randomized controlled experiment) can be assessed by:

1. Checking for correlations between independent variables
2. Regressing on other independent variables and checking for significant coefficients
3. Checking for causality between independent variables
4. None of the above

B (Regressing on other independent variables and checking for significant coefficients) Explanation - If there is random assignment, there should not be any significant coefficients, i.e., the p-values for all other independent variables should be larger than 0.05.

## Questions 16 to 20 - Week 5 Content

**Q16**. Which of the following is not an example of selection bias?

a. Taking a sample of people in the neighbourhood around your house for a statewide survey.

b. Mailing all houses in different neighbourhood in the state and using the few responses you havereceived back.

c. Dividing states into subgroups based on important characteristics and randomly selecting houses to be surveyed.

d. Taking surveys of people who register to participate in the study.

Ans. (c) is the only part where there is no selection bias for a statewide survey.

**Q. 17** Health researchers have looked at a large dataset of disease rates, diet and other health behaviours. The experiments show that there is a strong correlation between increased heart disease is correlated with higher fat diets (a positive correlation), and increased exercise is correlated with less heart disease (a negative correlation). What does this imply? (select all that

apply)

a. Reducing fat could reduce the risk of heart diseases

b. Increasing exercise can reduce the risk of heart diseases

c. Increasing exercise can decreases the risk of heart diseases

d. None of the above

Ans. (d) Correlation is not equal to causation

**Problem statement for next 3 questions: 18-20**

The Earned Income Tax Credit (EITC) is a refundable tax credit for low and moderate-income workers. The amount depends on income and number of children. In 1993, an expansion of EITC was passed, which aimed at providing a tax break for low income individuals with children. The

bill went into effect in 1994. We want to use the data to observe the difference in employment for women with children.

**Q.18** In the above problem, what is the control group and the treatment group respectively.

a. Women with children, women without children

b. Women who are employed, women who are unemployed

c. Women without children, Women with children

d. Women who are unemployed, women who are employed.

Ans: C. The act effects individuals with children, so the treatment group will be women with children and the control group in women without children. We shall observe them both pre and post 1993.

Load the EITC data set. It contains a lot of information about the EITC of women, and factors which come into determining it (the amount they earn, children, employment status etc.). To observe the effect of the bill on women with children, we will have to construct dummy variables to indicate these changes.

**Q. 19** What variables should need to be constructed for the above problem.

a. eitc$postbill= as.numeric(eitc$year >=1993)

eitc$kids = as.numeric(eitc$children >1)

b. eitc$postbill= as.numeric(eitc$year >1993)

eitc$kids = as.numeric(eitc$children >1)

c. eitc$postbill= as.numeric(eitc$year >=1993)

eitc$kids = as.numeric(eitc$children >=1)

d. eitc$postbill= as.numeric(eitc$year >1993)

eitc$kids = as.numeric(eitc$children >=1)

Ans. (d) We want to test effect of women who have kids and don’t, the variable for kids should be greater than or equal to one, and the law went into effect in 1994, hence the year should be greater than 1993.

**Q. 20** To calculate difference in difference estimate we need four values, control group before treatment, control group after treatment, treatment group before treatment and treatment group after treatment.

To get these values we will have to calculate the mean of the work variable for each of the four groups. Calculate these values in R. What is the value of the difference in difference in estimate?(Hint: Subset the data on the basis of each group you need and then calculate the mean of the work variable. )

a.0.054

b.0.047

c.0.065

d.0.031

Ans (b) R Code:

eitc = read.csv("C:/Users/kamya/Downloads/eitc.csv") #change path

# Create two additional dummy variables

# the EITC went into effect in 1994

eitc$post93 = as.numeric(eitc$year >= 1994)

# Treatment group will be all women with children.

eitc$anykids = as.numeric(eitc$children >= 1)

# Compute the four data points needed in the DID calculation:

a = sapply(subset(eitc, post93 == 0 & anykids == 0, select=work), mean)

b = sapply(subset(eitc, post93 == 0 & anykids == 1, select=work), mean)

c = sapply(subset(eitc, post93 == 1 & anykids == 0, select=work), mean)

d = sapply(subset(eitc, post93 == 1 & anykids == 1, select=work), mean)

# Difference in difference parameter

(d-c)-(b-a)

Alternate Solution: Q20)

CODE:

library(readr)

eitc <- read\_csv("eitc.csv")

eitc$time = ifelse(eitc$year >= 1994, 1, 0)

eitc$child = ifelse(eitc$children>0,1,0)

eitc$int = eitc$time \* eitc$child

model<- lm(work~time+child+int, data=eitc)

summary(model)

A screenshot of text

Description automatically generated

Difference in difference estimator is the coefficient of the interaction term which is **0.047**