Graded Homework #2

Due Feb 26 at 11:59pm **Points** 100 **Questions** 20
Available Feb 7 at 8am - Feb 26 at 11:59pm 20 days **Time Limit** None

Instructions

Graded Homework #2 covers the topics in **Weeks 1, 2, 3, 4 ,5 and 6** and is worth **10% of your overall grade**. You may work on the homework for as long as you like within the given window. You are allowed only **ONE** attempt for submission. Please note that your answers will automatically save as you key them. As long as you do not click submit, you can enter and exit the assignment as many times as necessary during the time period that it is available. Again, please note, **you should only click "submit" when you are completely finished with the assignment and ready to submit it for grading**.

Good luck!

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	4,374 minutes	100 out of 100

Correct answers are hidden.

Score for this quiz: **100** out of 100
Submitted Feb 21 at 11:28pm
This attempt took 4,374 minutes.

Question 1 — 5 / 5 pts

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?

If R Squared increases, this variable is significant.

If R Squared decreases, this variable is not significant.

Individually R squared cannot tell about variable importance. We can't say anything about it right now.

None of these.

Question 2

5 / 5 pts

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:

Age is a good predictor of health

Health is a good predictor of age

Age is a poor predictor of health

None of these

Question 3

5 / 5 pts

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?

There exists homoskedasticity in the model.

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

We try to build a model for NBA players' salary.

Download the dataset ***nba2017.csv*** from here:
https://www.dropbox.com/s/pe3urv1mv9s8mwb/nba2017%20%281%29.csv

(https://www.dropbox.com/s/pe3urv1mv9s8mwb/nba2017%20%281%29.csv)
(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.

Question 4

5 / 5 pts

For Model_1, what is the interpretation for the coefficient of height?

- [x] One unit increase in height increases salary by 2253985 units
- [ ] One unit increase in height increases salary by 874758 units
- [ ] One unit increase in height decreases salary by 677390 units
- [ ] One unit increase in height increases salary by 677390 units

Question 5

5 / 5 pts

For Model_2, what is the interpretation for the coefficient of height?

- [ ] When height increases by 1%, salary increases by 68.89%
- [ ] When height increases by 1 unit, salary increases by 0.6889%
- [ ] When height increases by 1% unit, salary increases by 0.6889 units
- [x] When height increases by 1 unit, salary increases by 68.89%

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. 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)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.98998    1.13995   -3.50  0.00047 ***
gre          0.00226    0.00109    2.07  0.03847 *
gpa          0.80404    0.33182    2.42  0.01539 *
rank2       -0.67544    0.31649   -2.13  0.03283 *
rank3       -1.34020    0.34531   -3.88  0.00010 ***
rank4       -1.55146    0.41783   -3.71  0.00020 ***
```

Question 6

5 / 5 pts

By how much do the 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.

- 0.6709
- 0.8040
- 0.6754
- 0.5089

Question 7

5 / 5 pts

Choose the wrong statement from the following:

- Change in odds from Rank 2 to Rank 3 is higher than change in odds from Rank 3 to Rank 4.
- Change in a unit of GPA corresponds to higher odds of getting an admit than a change in a unit of GRE.
- 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.
- All of the above are wrong.

Question 8

5 / 5 pts

Given the following Confusion Matrix:

		Predicted Values	
		No	Yes
Actual Values	No	55	11
	Yes	7	117

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

- 0.94

- 0.91
- 0.90
- (•) 0.83

Question 9

5 / 5 pts

Choose the correct option from the following about the effect of an increase in the cut-off value?

- (•) The True Positives will decrease, and the True Negatives will increase.
- 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
- Both False Negatives and True Negatives will decrease.
- The False Positives will increase, and the False Negatives will decrease

Question 10

5 / 5 pts

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%

(Give the answer up to 2 decimal places)

45.56

Question 11

5 / 5 pts

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

Data for this Date Range	
Aug. 31, 2016	-0.12%
July 31, 2016	3.56%
June 30, 2016	0.09%
May 31, 2016	1.53%
April 30, 2016	0.27%
March 31, 2016	6.60%
Feb. 29, 2016	-0.41%
Jan. 31, 2016	-5.07%

Source: Ycharts.com

(Give the answer up to 2 decimal places, if your answer is 0.51% enter 0.51)

0.81

Question 12

5 / 5 pts

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

(Give the answer up to 2 decimal places, if your answer is 0.51% enter 0.51)

0.76

Question 13

5 / 5 pts

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

(Give the answer up to 2 decimal places)

0.3

Question 14

5 / 5 pts

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

- A law that changed the tax rate for some subjects, but not others.
- A hurricane that hits a few stores among a large sample of stores.
- Minimum wage is changed in one state but not another.
- All of the above.

Question 15

5 / 5 pts

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

Checking for correlations between independent variables

Regressing on other independent variables and checking for significant coefficients

Checking for causality between independent variables

None of the above

Question 16

5 / 5 pts

Which of the following is not an example of selection bias?

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

Mailing all houses in different neighbourhood in the state and using the few responses you have received back.

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

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

Question 17

5 / 5 pts

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

- Reducing fat could reduce the risk of heart diseases.
- Increasing exercise can reduce the risk of heart diseases.
- Reducing fat and increasing exercise can together decrease the risk of heart diseases.
- None of the above.

The Earned Income Tax Credit (EITC) is a refundable tax credit for low and moderate-income workers. The amount depends on the 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.

Dataset:

Link (https://www.dropbox.com/s/9hb7obyt95gew2m/eitc%20%281%29.csv)

Question 18

5 / 5 pts

In the above problem, what is the control group and the treatment group respectively?

- Women with children, women without children.
- Women who are employed, women who are unemployed.
- Women without children, Women with children.
- Women who are unemployed, women who are employed.

Question 19

5 / 5 pts

What variables should need to be constructed for the above problem?

- eitc$postbill= as.numeric(eitc$year >=1993) and eitc$kids = as.numeric(eitc$children >1)
- eitc$postbill= as.numeric(eitc$year >1993) and eitc$kids = as.numeric(eitc$children >1)
- eitc$postbill= as.numeric(eitc$year >=1993) and eitc$kids = as.numeric(eitc$children >=1)
- eitc$postbill= as.numeric(eitc$year >1993) and eitc$kids = as.numeric(eitc$children >=1)

Question 20

5 / 5 pts

What is the value of the difference in differences estimate?

(Hint: 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.

Subset the data on the basis of each group you need and then calculate the mean of the work variable.)

- 0.054
- 0.047
- 0.065
- 0.031

Quiz Score: **100** out of 100