Graded Homework #2

Due Oct 2 at 8:59pm **Points** 100 **Questions** 20 **Available** Sep 12 at 2pm - Oct 2 at 8:59pm 20 days **Time Limit** None

Instructions

Graded Homework #2 covers the topics in **Weeks 1, 2, 3, 4 and 5** 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!

This quiz was locked Oct 2 at 8:59pm.

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	25,309 minutes	95 out of 100

Correct answers are hidden.

Score for this quiz: **95** out of 100
Submitted Oct 2 at 8:59pm
This attempt took 25,309 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

Incorrect

Question 3

0 / 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://gatech.box.com/s/qdkpwlxxo0tyxs4kw0m8wyxec5fbhvc7 (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?

● 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

● When height increases by 1 unit, salary increases by 68.89%

Question 6

5 / 5 pts

One of the power companies in Atlanta is trying to decide whether the residents will pay the defaulted bills or not. They collected the data on their credit history, marital status and household income. The analyst in this company decides to use a linear model to solve it. This analyst chooses the right model.

○ True

● False

Question 7

5 / 5 pts

Some betting site makes the prediction that for the next World Cup, France has a 16.67% probability to win. What are the odds for France winning the next World Cup?

○ 5:1

● 1:5

○ 1:6

○ 6:1

Question 8

5 / 5 pts

To detect whether a patient has a certain disease, we can also use confusion matrix. For disease A, we use the cutoff value p = 0.2, and the confusion matrix is as follows. Here we use 0 to denote not having disease A and 1 to denote having disease A. Please use the information provided below to answer the following two questions.

	Predicted Value = 0	**Predicted Value = 1**
True Value = 0	230	54
True Value = 1	9	20

What is the specificity?

● 230/(230+54)

○ 230/(230+20)

○ 230/(230+9)

○ 230/(230+9+54+20)

Question 9 5 / 5 pts

We know that for disease detection, Type II error is much more costly than Type I error. In other words, we try to best detect every potential patient. What can we do to reduce Type II error?

○ Get larger samples to test the accuracy of the confusion matrix

○ Increase the cutoff value p

● Lower the cutoff value p

○ There is no effective way to do it

For the following five questions, we want to study whether smoking will influence the probability of getting heart disease. To make it easier, we choose age (variable name: age0) and number of cigarettes smoked per day (variable name: ncigs0) as two independent variables to predict the event (event: 0 = no; 1 = yes) of coronary heart disease (variable name: chd69)

To load data, install the package “epitools” in your R console and load the data.

Install.packages(“epitools”)

library(epitools)

data(wcgs)

Perform logistic regression and answer the questions below.

Question 10 5 / 5 pts

Which of the following is the correct form of the logistic regression?

● p(chd69) = exp(-6.36 + 0.02*ncigs0 + 0.08*age0)/[1 + exp(-6.36 + 0.02*ncigs0 + 0.08*age0)]

○ Log(chd69) = exp(-6.36 + 0.02*ncigs0 + 0.08*age0)/[1 + exp(-6.36 + 0.02*ncigs0 + 0.08*age0)]

○ Log(p(chd69)/1-p(chd69)) = exp(-6.36 + 0.02*ncigs0 + 0.08*age0)/[1 + exp(-6.36 + 0.02*ncigs0 + 0.08*age0)]

○ Logit(p(chd69)) = exp(-6.36 + 0.02*ncigs0 + 0.08*age0)/[1 + exp(-6.36 + 0.02*ncigs0 + 0.08*age0)]

Question 11 5 / 5 pts

How to interpret the coefficient of age?

○ If age increases by 1 unit, the natural log of the odds of getting heart disease increases by 0.08.

○ If age increases by 1 unit, the odds of getting heart disease increase by a factor of exp(0.08).

○ If age increases by 1 unit, the odds of getting heart disease increase by roughly 100*0.08 percent.

● All of these statements.

Question 12 5 / 5 pts

How to interpret the coefficient of ncigs0?

● If ncigs0 increase by 1 unit, the natural log of the odds of getting heart disease increases by 0.02.

○ If ncigs0 increase by 1 unit, the odds of getting heart disease increases by 0.02.

○ If ncigs0 increases by 1 unit, the odds of getting heart disease increase by exp(0.02).

○ All of these statements.

Question 13 5 / 5 pts

For a 35-yr old man who smokes 10 cigarettes a day, what is the predicted probability of getting coronary heart disease?

○ exp(1-6.36 + 0.02*35 + 0.08*10)/[1 + exp(-6.36 + 0.02*35 + 0.08*10)]

● exp(-6.36 + 0.02*10 + 0.08*35)/[1 + exp(-6.36 + 0.02*10 + 0.08*35)]

○ exp(-6.36 + 0.02*35 + 0.08*10)/[1 + exp(-6.36 + 0.02*35 + 0.08*10)]

○ exp(1-6.36 + 0.08*35 + 0.02*10)/[1 + exp(-6.36 + 0.08*35 + 0.02*10)]

Question 14

5 / 5 pts

A 30-yr old man reduces the number of cigarettes he smokes from 10 cigarettes per day to 9 per day. What is the absolute change in the natural log of predicted odds of getting coronary heart disease?

○ exp(-6.36 + 0.02*1 + 0.08*30)

○ exp(-6.36 + 0.02*1 + 0.08*30)/[1 + exp(-6.36 + 0.02*1 + 0.08*30)]

● 0.02

○ None of the above

Question 15

5 / 5 pts

Which 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 16

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

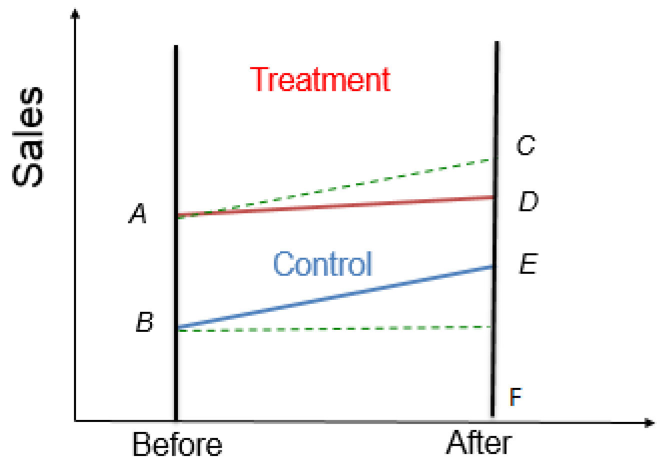
○ None of the above

Consider the following regression model for the sale of cigarettes. County A requires that all cigarette packets contain pictorial images of adverse effects of cigarette smoking. We want to see the impact of this law on sales of cigarette packs. We compare sales before and after passing this law in County A and with County B that did not have any such law.

Y = b0 + b1*(T) + b2*(t) + b3*(T*t)+e

Where T = 1 for County A and 0 for County B

t = 1 indicates after passing the law; t = 0 before passing the law



Question 17

5 / 5 pts

Which of the following terms gives an estimate of the total number of cigarettes sold in County A after passing the law?

○ C-F

● D-F

○ E-F

○ D-E

Question 18

5 / 5 pts

Which of the following terms gives an estimate of the total number of cigarettes sold in County B after passing the law?

○ D-E

○ C-E

○ D-F

● E-F

Question 19

5 / 5 pts

Which of the following terms gives an estimate of comparing the effect of passing the law in County A?

○ b1 + b0

● b2 + b3

○ b0 + b2

○ b3

Question 20

5 / 5 pts

Which of the following terms represents the treatment effect of passing the law?

○ C-F

● C-D

○ C-E

○ E-F

Quiz Score: **95** out of 100