⚠ This quiz has been regraded; your score was not affected.

Graded Homework #1

Due Feb 5 at 11:59pm **Points** 100 **Questions** 20
Available Jan 17 at 6pm - Feb 5 at 11:59pm 19 days **Time Limit** None

Instructions

Graded Homework #1 covers the topics in **Weeks 1, 2, 3** and is worth **10% of your overall grade**. You may work on the homework for as long as you like within the given window. 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 as you only have ONE ATTEMPT**.

Also, please remember that you are to complete this assignment on your own. Any help given or received constitutes cheating. If you have any general questions about the assignment, please post to the Piazza board. **If your question involves specific references to the answer to a question or questions, please be sure to mark your post as private**. **Please choose the closest answers.**

Good luck!

This quiz was locked Feb 5 at 11:59pm.

Attempt History

	Attempt	Time	Score	Regraded
LATEST	Attempt 1	7 minutes	100 out of 100	100 out of 100

Score for this quiz: **100** out of 100
Submitted Feb 3 at 5:53pm
This attempt took 7 minutes.

Question 1 — 5 / 5 pts

Choose the correct statement regarding the sum of residuals calculated using Ordinary Least Squares (OLS).

A. The sum of residuals will always be nonzero whatever the form of the linear regression as long as you are using OLS to estimate the coefficients.

Correct!

B. The sum of residuals will always be equal to zero if you include intercept term in your model and you are using OLS to estimate the coefficients

C. The sum of residuals may or may not be zero when using OLS and the R software makes mathematical adjustments to make it zero.

D. The sum of residuals may or may not be zero when using OLS and the R software makes no mathematical adjustments because it is not needed.

Question 2

5 / 5 pts

Choose the correct statement regarding the error terms in the assumption of Ordinary Least Squares (OLS).

A. The error terms are normally distributed with a constant non-zero mean and constant Variance.

B. The variance of error terms may or may not be constant as long as the terms are normally distributed with mean equal to zero.

C. The error terms follow lognormal distribution with mean equal to zero and constant variance.

Correct!

D. The error terms are normally distributed with mean equal to zero and a constant variance.

Questions 3-6 - Details

The National Traffic Study Institute is conducting a study to find out the relationship between the speed at which the car is moving and the distance it takes to stop after applying the brakes. You were hired as a statistician to work on this problem. The data can be accessed as follows:

install.packages("Ecdat")

library(Ecdat)

data(cars)

You can easily see that these are the variables present in the dataset and the corresponding units using help command on R console – speed (in mph) and dist (in ft).

Use this dataset for the following 4 questions.

Question 3

5 / 5 pts

Let's try to find out if there is a correlation between the distance needed to stop and the speed at which the car is moving. What correlation value do you find when doing this in R?

A. 0

B. 0.72

Correct!

C. 0.81

D. 1

Question 4

5 / 5 pts

Would you say that distance to stop and speed of the car are?

A. Not correlated

B. Inversely correlated

Correct!

C. Well correlated

D. Perfectly correlated

Question 5

5 / 5 pts

Now, let's fit a linear model with distance needed to stop as the response and speed as the predictor. What is the percent variation explained by speed, intercept, and coefficient of speed?

Correct!

A. 0.65, -17.58 and 3.93

B. 0.65, 17.58 and 3.93

C. 0.65, 8.28 and 0.16

D. 0.89, 0 and 0.31

Question 6

5 / 5 pts

Now suppose we need to change the units of distance needed to stop from feet to meters and speed from mph to meters per second because we need the results to be standard units. What would be the results for percent variation explained by speed, intercept, and coefficient of speed?

- A. 0.65, -5.36, and 1.19
- B. 0.65, -5.36, and 2.68

Correct!

- C. 0.65, -17.58, and 3.93
- D. 0.65, 8.28, and 0.16

Question 7

5 / 5 pts

If p-value of a particular parameter in your linear regression model is equal to 1.67e-14, what does it say about the coefficient of that parameter?

- A. The null hypothesis corresponding to this parameter can be rejected and hence coefficient of the parameter is equal to zero.
- B. The nature of the coefficient of the parameter is ambiguous and hence we need to change the model.

Correct!

- C. The null hypothesis corresponding to this parameter can be rejected and hence coefficient of the parameter is significant and hence not equal to zero.

D. The null hypothesis corresponding to this parameter can be accepted and hence coefficient of the parameter is equal to zero.

Question 8

5 / 5 pts

For the following dataset: we regress calorie consumption of individual based on their region, the level of urban development of the place they live in and age:

Response Variable: Calorie Consumption (continuous, numeric)

Independent Variables: Region (3 categories: Midwest, East, West)

Level of Urban Development (2 categories: Urban, Rural)

Age (integer)

To transform region variable into categorical variables for regression, how many dummy variables do we need to insert into the regression?

A. 0

B. 1

Correct!

C. 2

D. 3

E. 4

Question 9

5 / 5 pts

For the following dataset: we regress calorie consumption of individual based on their region, the level of urban development of the place they live in and age:

Response Variable: Calorie Consumption (continuous, numeric)

Independent Variables: Region (3 categories: Midwest, East, West)

Level of Urban Development (2 categories: Urban, Rural)

Age (integer)

Suppose we transform the urban development variable into Rural_True dummy variable (Urban =0, Rural = 1) and run the regression of Calorie Consumption against Age and Rural_True:

	Estimate	S.E.	t-value	Pr >|t|
Intercept	200.89	5.68	5.82	0.008
Age	114.82	0.056	6.89	0.001
Rural_True	-40.22	1.23	-4.89	0.012

Which statement about Level of Urban Development is TRUE (confidence Interval = 95%)?

A. Rural_True does not have a significant effect on the response variable

B. All people living in urban areas consume less calories than all people living in rural areas

C. All people living in urban areas consume more calories than all people living in rural areas

Correct!

D. At the same age level, people living in urban areas consume more calories than people living in rural areas

E. At the same age level, people living in urban areas consume less calories than people living in rural areas

Question 10

5 / 5 pts

We look to understand and predict sales, by regressing it on advertising budgets spent on YouTube and Facebook. First run without the interaction term:

*sales = b0 + b1*youtube + b2*facebook*

The interaction between the two forms of advertising is also observed as follows:

*sales = b0 + b1*youtube + b2*facebook + b3*(youtube*facebook)*

What can we say about the interaction effects? (confidence interval = 95%)

Correct!

- (selected) A. Interaction effects exist between YouTube and Facebook since p-value is much smaller than the error rate. We should include the interaction term in the regression model to explain the variability in the data.
- B. Interaction effects exist between YouTube and Facebook since p-value is much larger than error rate. We should include the interaction term in the regression model to explain the variability in the data.
- C. Interaction effects do not exist between YouTube and Facebook since p-value is close to zero.
- D. Interaction effects do not make sense because we already know that the two are independent of each other.
- E. Interaction effects are irrelevant when conducting regression analysis.

Question 11 Original Score: 5 / 5 pts Regraded Score: 5 / 5 pts

⓵ This question has been regraded.

From previous information, what is the expected increase in the number of units of sales, if an extra $1000 is invested in YouTube advertising? (Assume $45 invested in Facebook advertising)

orrect Answer

A. 65.5 unit

B. 20.305 units

ou Answered

C. 60.54 units

D. 68.5 units

E. 19 units

Question 12 5 / 5 pts

This question requires you to build the linear regression model below and answer the questions on the basis of your results.

You will/may require the following dependencies:

tidyverse

Load the Salaries dataset from the car package as below:

install.packages("car")

library(car)

Load the dataset into a tibble as below:

Salaries_Dataset<- as.tibble(Salaries)

Now create indicator variables for the ‘rank’ column, specifically with the base case of AsstProf (i.e create AssocProf and Prof variables with 1 denoting the positive case and 0 the negative).

Create a linear regression model for the following

Salary = b0 + b1* Years.service + b2*AssocProf (dummy variable) + b3*Prof (dummy variable)

Select the correct result (assume a p-value threshold of 5%):

A. The Years.service coefficient is approximately 450 and with respect to the threshold can be used to reject the null hypothesis.

B. The Years.service coefficient is approximately -160 and with respect to the threshold can be used to reject the null hypothesis.

Correct!

C. The Years.service coefficient is approximately -160 and with respect to the threshold cannot be used to reject the null hypothesis.

D. The Years.service coefficient is approximately 450 and with respect to the threshold cannot be used to reject the null hypothesis.

E. The Years.service coefficient is approximately 150 and with respect to the threshold cannot be used to reject the null hypothesis.

Question 13

5 / 5 pts

Which of the following is not an assumption of Linear Regression?

A. The expected value of response variable Y at each predictor value X corresponds to a straight line.

B. Error terms are independent and identically distributed normal Random variables, with mean 0 and variance σ^2.

Correct!

C. The error terms vary with the predictor variable(s) X.

D. Predictor variables are linearly independent of each other.

Questions 14 -16 - Details

The Boston city government wants to hire you as a consultant to help determine the factors that influence housing values in the city. Armed with your knowledge of linear regression in R, you decide to earn your pay check and accept the contract.

Load the Boston library dataset from the MASS package.

```
> library(MASS)

> dat <- Boston

> head(dat)
```

Fit a linear model with median value of owner-occupied homes as response and all other variables except age and indus as predictors. We shall now try and study visually the various aspects of the regression model we have just built. Use the following command to view the various diagnostic plots of your model.

```
>  plot( <regression_model_object> )
```

Hint : Use the following link to help yourself better understand the various diagnostic plots **https://www.theanalysisfactor.com/linear-models-r-diagnosing-regression-model/ (https://www.theanalysisfactor.com/linear-models-r-diagnosing-regression-model/)**

Question 14

5 / 5 pts

Let's check if our residuals are normal by doing visual inspection of a diagnostic plot. (Normal Q-Q plot). What do you observe?

A. There seems to be non-normality with distribution being left-skewed.

B. There seems to be perfect normality.

Correct!

C. There seems to be non-normality with distribution being right-skewed.

D. None of the above.

Question 15

5 / 5 pts

Let's run a formal test to confirm if there is indeed a non-normality. This test is called Shapiro-Wilk normality test and the run command for the same is shapiro.test(residuals(your lm object)). The null hypothesis is that the residuals are normal. Now does the result from this test match your results from Q 14 ?

A. No, we get a low p-value in Shapiro-Wilk test which means the residuals are normally distributed whereas visual inspection in Q.2 led us to believe that there is a non-normal distribution of residuals.

Correct!

B. Yes, we get a low p-value in Shapiro-Wilk test which means the residuals are not normally distributed and visual inspection in Q.2 also led to the conclusion that there is a non-normal distribution of residuals.

C. No, we get a low p-value in Shapiro-Wilk test which means the residuals are non-normally distributed whereas visual inspection in Q.2 led us to believe that there is a normal distribution of residuals.

D. Yes, we get a low p-value in Shapiro-Wilk test which means the residuals are normally distributed whereas visual inspection in Q.2 led us to believe that there is a normal distribution of residuals.

Question 16

5 / 5 pts

Let's check for any autocorrelation in the data. Durbin-Watson statistic is used for that. The function "dwtest" in the package "lmtest" can be used for this. dwtest takes your linear model as input. The NULL hypothesis for this test is that the errors are uncorrelated. Let's use that. Type the following code to get ready Install.packages('lmtest') require(lmtest) What does the test tell you?

A. The small p-value indicates that there is no autocorrelation.

Correct!

B. The small p-value indicates that there might be autocorrelation.

C. The large p-value indicates that there is no autocorrelation.

D. The large p-value indicates that there might be autocorrelation.

Questions 17-20 - Details

For below questions, use the file EDSAL.csv.

Download the EDSAL.csv file (link: **https://gatech.app.box.com/s/kdwefbb35qkluzp2yt5vl32gk7hlrp7o (https://www.dropbox.com/s/gdwpbkdw72oq4mm/EDSAL.csv?dl=0)**) and upload it to a dataframe (in R). The three variables are Education, Experience and Salary. Code to load the data set is as follows:

> EDSAL = read.csv("EDSAL.csv", header = TRUE)

Run 4 linear regressions using the lm function in R. (note – you have to use the natural log)

- Lin-Lin: Use Salary as the dependent variable and Experience as the independent variable.
- Lin-Log: Use Salary as the dependent variable and log(Experience) as the independent variable.
- Log-Lin: Use log(Salary) as the dependent variable and Experience as the independent variable.
- Log-Log: Use log(Salary) as the dependent variable and log(Experience) as the independent variable.

Question 17

5 / 5 pts

For which of following situations are we most likely to consider log transformation?

A. Dependent and independent variables have linear relationship

B. Dependent and independent variables follow normal distribution

Correct!

C. Heteroscedasticity (non-constant variance) is observed in original model

D. High R-squared score is observed in original model

Question 18

5 / 5 pts

Which of the 4 fitted models has the highest R-square value?

A. Lin-Log

B. Log-Log

Correct!

C. Log-Lin

D. Lin-Lin

Question 19

5 / 5 pts

Which is the interpretation of the slope coefficient for the Log-Lin model?

A. Increasing Experience by 1 unit leads to 0.037087 units increase in Salary

Correct!

B. Increasing Experience by 1 unit leads to ((e^0.037087)-1) *100% increase in Salary

C. Increasing Experience by 1% leads to 0.01*0.037087 units increase in Salary

D. Increasing Experience by 1% leads to (e^ (0.037087*0.01)-1) *100% increase in Salary

Question 20

5 / 5 pts

Which is the interpretation of the slope coefficient for the Log-Log model?

A. Increasing Experience by 1 unit leads to 0.45949 units increase in Salary

B. Increasing Experience by 1 unit leads to ((e^0.45949)-1) *100% increase in Salary

C. Increasing Experience by 1% leads to 0.01*0.45949 units increase in Salary

Correct!

D. Increasing Experience by 1% leads to (e^ (0.45949*0.01)-1) *100% increase in Salary

Quiz Score: **100** out of 100