**Homework 1 - Solutions**

**Question 1**

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

1. The error terms are normally distributed with a constant non-zero mean and constant Variance
2. The variance of error terms may or may not be constant as long as the terms are normally distributed with mean equal to zero
3. The error terms follow lognormal distribution with mean equal to zero and constant variance
4. The error terms are normally distributed with mean equal to zero and a constant variance

**Sol: D Explanation:** The OLS model assumes that the error terms (residuals) are normallydistributed with mean equals to zero and constant variance (property of homoscedasticity of variances)

**Questions 2-5**

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 5 questions.

**Question 2**

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?

1. 0
2. 0.72
3. 0.81
4. 1

**Ans: C Explanation**: cor(cars$speed, cars$dist) = 0.806

**Question 3**

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

A. Not correlated

B. Inversely correlated

C. Well correlated

D. Perfectly correlated

**Ans: C Explanation:** Well-correlated because the value is close to 1 (perfect correlation), butnot exactly.

**Question 4**

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?

1. 0.65, -17.58 and 3.93
2. 0.65, 17.58 and 3.93
3. 0.65, 8.28 and 0.16
4. 0.89, 0 and 0.31

**Ans: A Explanation:** Percent variation explained by speed is the R-squared value = 0.65;intercept of speed (from the regression summary table) = -17.58; coefficient of speed (from the table again) = 3.93

**Question 5**

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 coeffiient of speed?

1. 0.65, -5.36 and 1.19
2. 0.65, -5.36 and 2.68
3. 0.65, -17.58 and 3.93
4. 0.65, 8.28 and 0.16

**Ans: B Explanation:** First, change the dataset into proper units. Convert speed from miles perhour to meters per second (multiply by 0.44704); convert feet into meters again by multiplying by a conversion factor (0.3048). Then, reuse the same steps as in Qn4 to get regression summary and look for the same variable outputs.

**CODE FOR QN 2-5**

library(Ecdat)

load(cars)

cor(cars$speed, cars$dist)

lm <- lm(dist ~ speed, data=cars)

summary(lm)

## code to change units

new.dat <- data.frame(speed=7.5)

predict(lm, newdata = new.dat, interval = 'confidence')

1. **C Explanation:** cor(cars$speed, cars$dist) = 0.806
2. **C Explanation:** Well correlated because the value is positive and close to 1 but not exactly
3. – so not perfectly correlated but well correlated.
4. **A Explanation:** percent variation explained by speed is the R-squared value = 0.65; interceptof speed (from the regression summary table) = -17.58; coefficient of speed (from the table again) = 3.93
5. **B Explanation:** First, change the dataset into proper units. Convert speed from miles perhour to meters per second (multiply by conversion factor); convert feet into meters again by multiplying by a conversion factor. Then, reuse the same steps as in Qn4 to get regression summary and look for the same variable outputs.

**Question 6**

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?

1. 0
2. 1
3. 2
4. 3
5. 4

**Ans. C Explanation:** Only two variables are needed because one factor levelis taken as base line.

**Question 7**

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*

*A screenshot of text

Description automatically generated*

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

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

*A close up of text on a white background

Description automatically generated*

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

1. 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.
2. 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.
3. Interaction effects do not exist between YouTube and Facebook since p-value is close to zero
4. Interaction effects do not make sense because we already know that the two are independent of each other
5. Interaction effects are irrelevant when conducting regression analysis

**Ans. A Explanation**: p-value of interaction term is close to zero, meaning that the interactionbetween Facebook and YouTube is significant.

**Question 8**

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 thresholdcan be used to reject the null hypothesis.

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

**C.** The Years.service coefficient is approximately -160 and with respect to the thresholdcannot be used to reject the null hypothesis.

**D.** The Years.service coefficient is approximately 450 and with respect to the thresholdcannot be used to reject the null hypothesis.

**E.** The Years.service coefficient is approximately 150 and with respect to the thresholdcannot be used to reject the null hypothesis.

**Answer: C**

**A screenshot of a cell phone

Description automatically generated**

**Instructions for Q9 and 10**

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/>

**Question 9:** 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 compare the results of this test to a diagnostic plot (by checking if our residuals are normal or not by performing a visual inspection). Do they match?

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

B. Yes, we get a low p-value in Shapiro-Wilk test which means the residuals are not normally distributed and visual inspection of diagnostic plot 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 inspectionof diagnostic plot 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 of diagnostic plot led us to believe that there is a normal distribution of residuals.

**Correct Answer: B**

Explanation: The Shapiro Wilk test produces a value of 0.9 with a p-value < 0.05. The null-hypothesis of this test is that the population is normally distributed. Thus, on the one hand, if the p value is less than the chosen alpha level (typically 95%, and hence we test against 0.05), then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. Therefore, for this case, the population of residuals are not normally distributed, which can be seen in the residual graphs.

**Question 10:** Let’s check for any autocorrelation in the data. Durbin-Watson statistic is usedfor 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.

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.

**Correct Answer: B**

**Explanation:** The Durbin-Watson test produced a value of 1.08 with p-value < 0.05. TheDurbin-Watson test tests the null hypothesis that linear regression residuals are uncorrelated, against the alternative hypothesis that autocorrelation exists. The small p value is thus indicative of the fact there indeed is a correlation amongst the residuals and are not independently distributed.

**Question 11**

For below questions, use the file EDSAL.csv. Download the EDSAL.csv file (link: https://gatech.app.box.com/s/kdwefbb35qkluzp2yt5vl32gk7hlrp7o) 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.

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

1. Dependent and independent variables have linear relationship
2. Dependent and independent variables follow normal distribution
3. Heteroscedasticity (non-constant variance) is observed in original model
4. High R-squared score is observed in original model

**Answer: C Explanation:**

The main purpose of log-transformation is to:

1. Achieve a more linear relationship
2. Make a distribution more normal
3. Make the variance more constant
4. Get a better fit such as higher R-squared

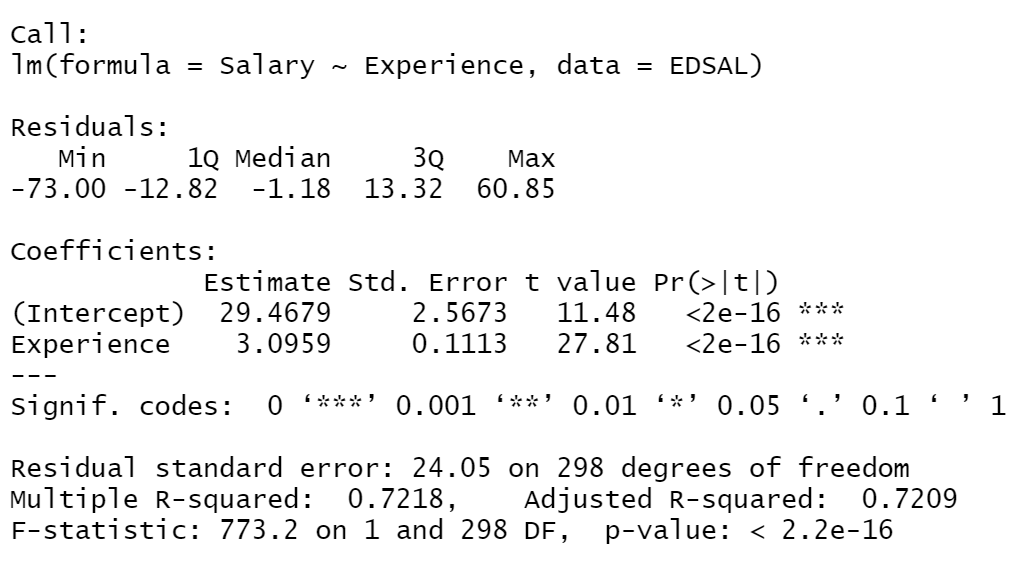
**Question 12**

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

1. Lin-Log
2. Log-Log
3. Log-Lin
4. Lin-Lin

**Answer: C Explanation:**

Linear-Linear Model:



Linear-Log Model:

