Homework 1 - Solutions

Question 1

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.
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.

Sol: B Explanation: The intercept is the catchall term that takes within itself anything that is not being predicted/accounted for by the independent variables.

Question 2

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
D. 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 normally distributed with mean equals to zero and constant variance (property of homoscedasticity of variances)

Questions 3 - 6

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 3

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

C. 0.81

D. 1

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

Question 4

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), but not exactly.

Question 5

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?

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

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 6

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

C. 0.65, -17.58 and 3.93

D. 0.65, 8.28 and 0.16

Ans: B Explanation: First, change the dataset into proper units. Convert speed from miles per hour 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 3 to 6

```
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')
```

3. C Explanation: cor(cars$speed, cars$dist) = 0.806

4. C Explanation: Well correlated because the value is positive and close to 1 but not exactly 1 – so not perfectly correlated but well correlated.

5. 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

6. B Explanation: First, change the dataset into proper units. Convert speed from miles per hour 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 7

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.

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.

Answer: C Explanation The null hypothesis of every parameter in the linear regression model is that the coefficient of the parameter is not different from zero (there is no relationship between the X variables and the Y variable). The p-value lies between 0 and 1 and if the p-value is closer to zero, then you can reject the null-hypothesis. The p-value in this question 1.67e-14 which is pretty close to zero. Hence, we can easily reject the null hypothesis which means the coefficient of that parameter is significantly different than zero which means there is a relationship between the parameter and the dependent variable.

Question 8

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

C. 2

D. 3

E. 4

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

Question 9

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

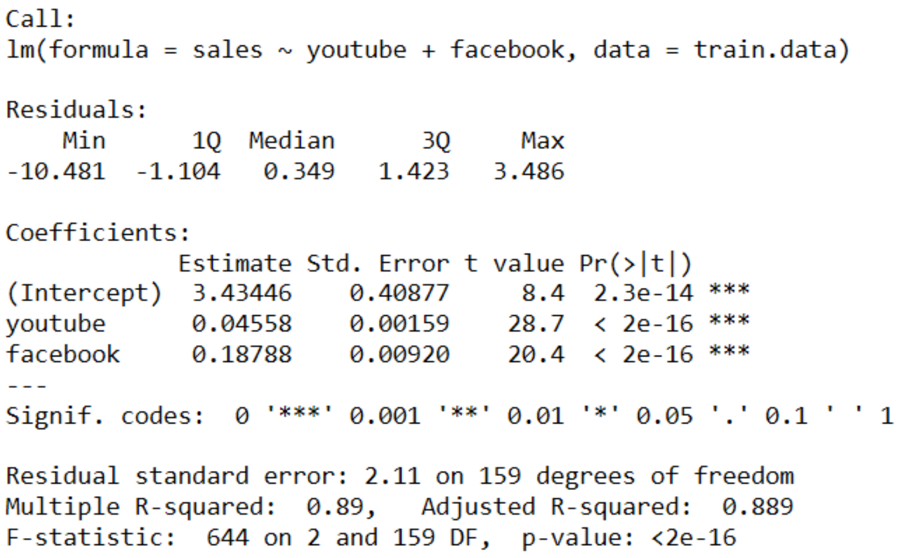
D. At the same age level, people living in urban areas consume more calories than people living in rural areas

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

Ans. D Explanation: D is the correct interpretation of regression coefficient. To be specific, people in rural area consume 40 calories less than those in urban areas, while holding age constant.

Question 10

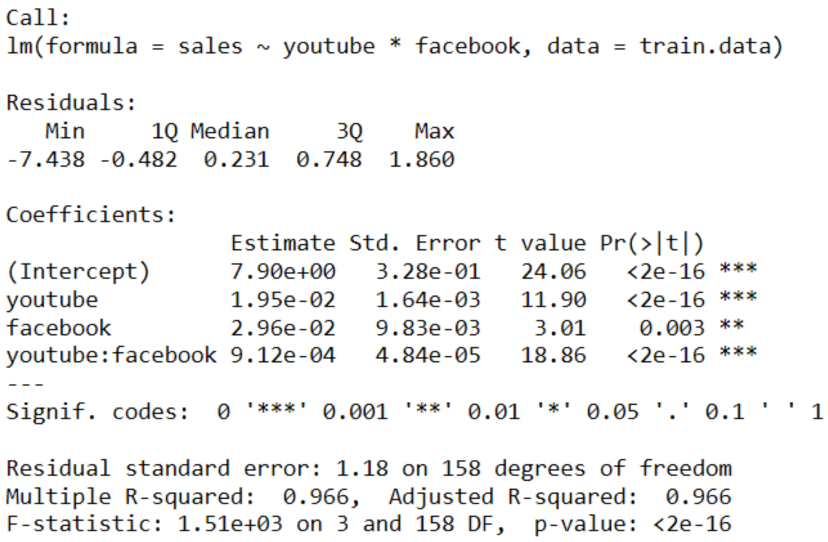
We look to understand and predict sales, by regressing it on advertising budgets spent on YouTube and Facebook. First run without the interaction term:
*sales = b0 + b1*youtube + b2*facebook*



```
Call:
lm(formula = sales ~ youtube + facebook, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.481  -1.104   0.349   1.423   3.486 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.43446    0.40877     8.4  2.3e-14 ***
youtube      0.04558    0.00159    28.7  < 2e-16 ***
facebook     0.18788    0.00920    20.4  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.11 on 159 degrees of freedom
Multiple R-squared:  0.89,    Adjusted R-squared:  0.889 
F-statistic:  644 on 2 and 159 DF,  p-value: <2e-16
```

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

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



```
Call:
lm(formula = sales ~ youtube * facebook, data = train.data)

Residuals:
   Min     1Q Median     3Q    Max 
-7.438 -0.482  0.231  0.748  1.860 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      7.90e+00   3.28e-01   24.06   <2e-16 ***
youtube          1.95e-02   1.64e-03   11.90   <2e-16 ***
facebook         2.96e-02   9.83e-03    3.01    0.003 ** 
youtube:facebook 9.12e-04   4.84e-05   18.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.18 on 158 degrees of freedom
Multiple R-squared:  0.966,	Adjusted R-squared:  0.966 
F-statistic: 1.51e+03 on 3 and 158 DF,  p-value: <2e-16
```

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

- **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

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

Question 11

From Previous Info,
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.

- **A.** 65.5 unit
- **B.** 20.305 units
- **C.** 59.5 units
- **D.** 68.5 units
- **E.** 19 units

Ans. C Explanation: For
sales = b0 + b1*youtube + b2*facebook + b3*(youtube*facebook)
Change in sales = b1*(change in YouTube) + b3*Facebook*(change in YouTube)

Question 12

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.

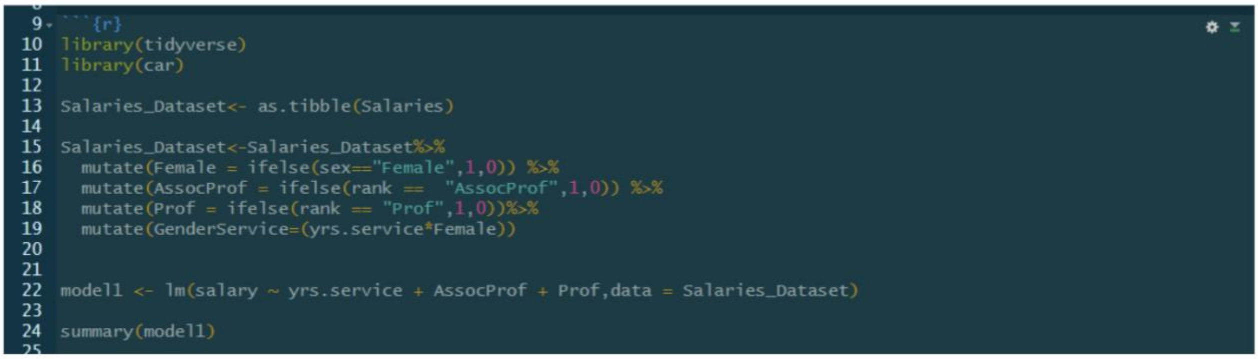
B. The Years.service coefficient is approximately -160 and with respect to the threshold can be used to reject the null hypothesis.

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.

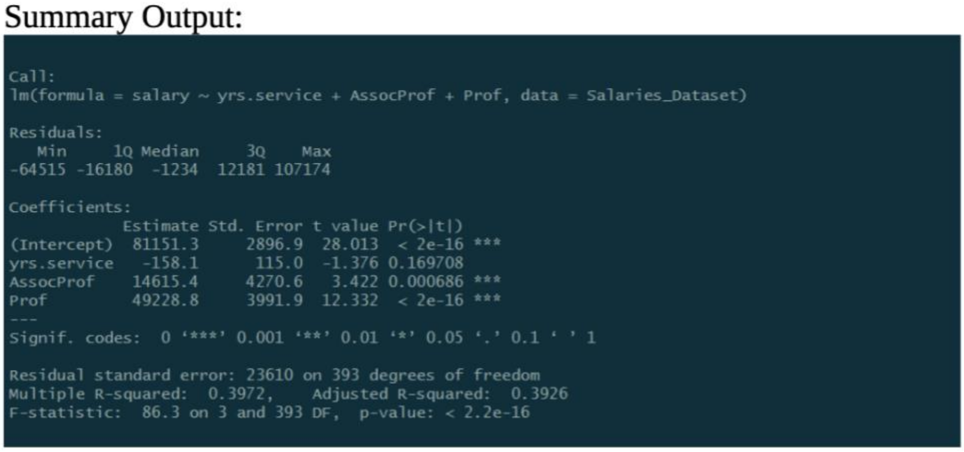
E. The Years.service coefficient is approximately 150 and with respect to the threshold cannot be used to reject the null hypothesis.

Answer: C



~~~
```{r}
library(tidyverse)
library(car)

Salaries_Dataset<- as.tibble(Salaries)

Salaries_Dataset<-Salaries_Dataset%>%
  mutate(Female = ifelse(sex=="Female",1,0)) %>%
  mutate(AssocProf = ifelse(rank ==  "AssocProf",1,0)) %>%
  mutate(Prof = ifelse(rank == "Prof",1,0))%>%
  mutate(GenderService=(yrs.service*Female))


model1 <- lm(salary ~ yrs.service + AssocProf + Prof,data = Salaries_Dataset)

summary(model1)
~~~

Summary Output:



```
Call:
lm(formula = salary ~ yrs.service + AssocProf + Prof, data = Salaries_Dataset)

Residuals:
   Min     1Q Median     3Q    Max 
-64515 -16180  -1234  12181 107174 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  81151.3     2896.9  28.013  < 2e-16 ***
yrs.service   -158.1      115.0  -1.376 0.169708    
AssocProf    14615.4     4270.6   3.422 0.000686 ***
Prof         49228.8     3991.9  12.332  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23610 on 393 degrees of freedom
Multiple R-squared:  0.3972,	Adjusted R-squared:  0.3926 
F-statistic:  86.3 on 3 and 393 DF,  p-value: < 2.2e-16
```

So the yrs.service coefficient is -158.1 with a p-value of 0.1697 which is significantly above the threshold set of 0.05. Hence cannot reject Null hypothesis.

Question 13

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

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

D. Predictor variables are linearly independent of each other

Solution: C

We assume constant variance of error (homoscedasticity) and C directly contradicts this.

Question 14-16:

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/

Code Solution:

```
> library(lmtest)
Loading required package: zoo

Attaching package: ‘zoo’

The following objects are masked from ‘package:base’:

    as.Date, as.Date.numeric

Warning message:
package ‘lmtest’ was built under R version 3.5.3
> library(MASS)
> dat <- Boston
> head(dat)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
> ols_reg4 = lm(medv ~ ., dat)
> summary(ols_reg4)

Call:
lm(formula = medv ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.595  -2.730  -0.518   1.777  26.199 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
```

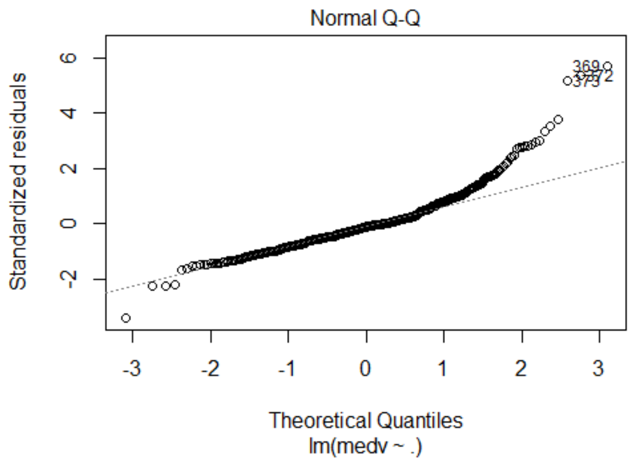
```
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,  Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16

> plot(ols_reg4)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
> shapiro.test(residuals(ols_reg4))

Shapiro-Wilk normality test

data:  residuals(ols_reg4)
W = 0.90138, p-value < 2.2e-16

> dwtest(ols_reg4)

Durbin-Watson test

data:  ols_reg4
DW = 1.0784, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

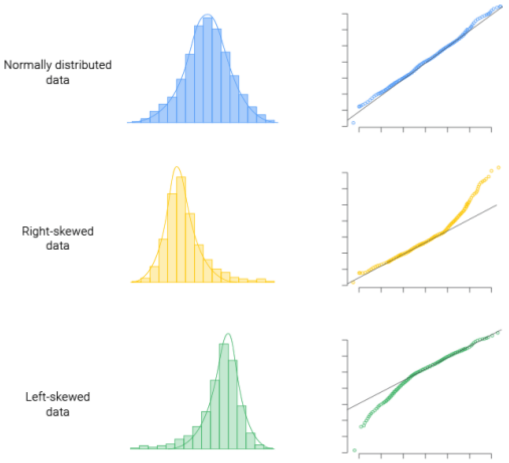
Question 14: Let’s check if our residuals are normal by doing visual inspection of a diagnostic plot. (Normal Q-Q plot). What do you observe?

A. There seems to be non-normality with distribution being left-skewed.
B. There seems to be perfect normality.
C. There seems to be non-normality with distribution being right-skewed.
D. None of the above.

Correct Answer: C

Explanation: As can be seen in the plot, towards the right end of Theoretical Quantiles, the Standardised residuals of the data at hand shift left, and in terms of statistics, it is called being Skewed towards the right.
One can think of a QQ plot as plotting 2 curves (bell shaped curves) on the two axis, Theoretical ones on the X axis and Standardised ones on the Y axis.





Question 15: 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.2?

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.
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.

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 above.

Question 16: 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.
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. The Durbin-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 17

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?
A. Dependent and independent variables have linear relationship
B. Dependent and independent variables follow normal distribution
C. Heteroscedasticity (non-constant variance) is observed in original model
D. 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

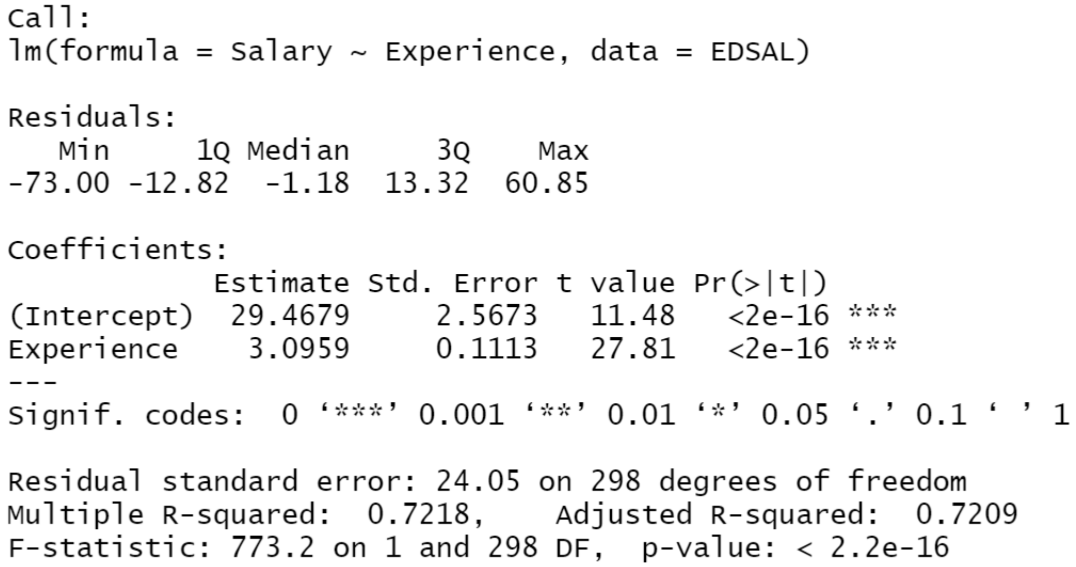
Question 18

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

A. Lin-Log
B. Log-Log
C. Log-Lin
D. Lin-Lin

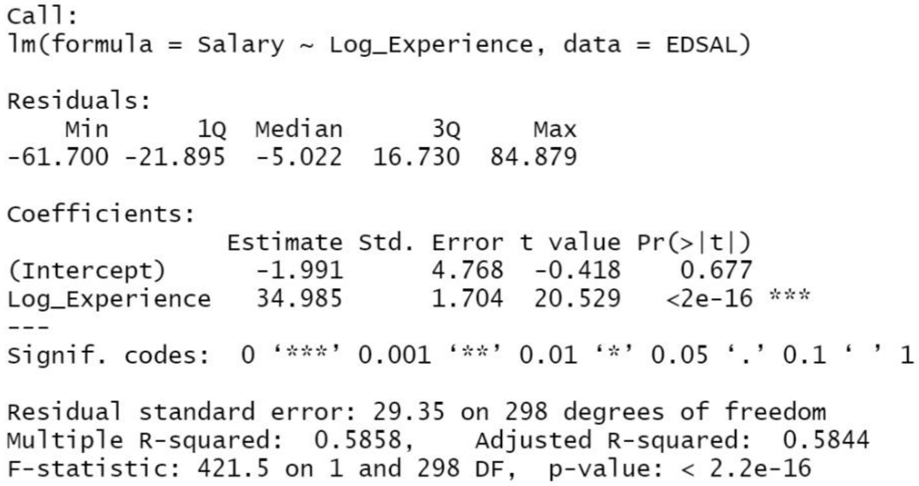
Answer: C
Explanation:

Linear-Linear Model:



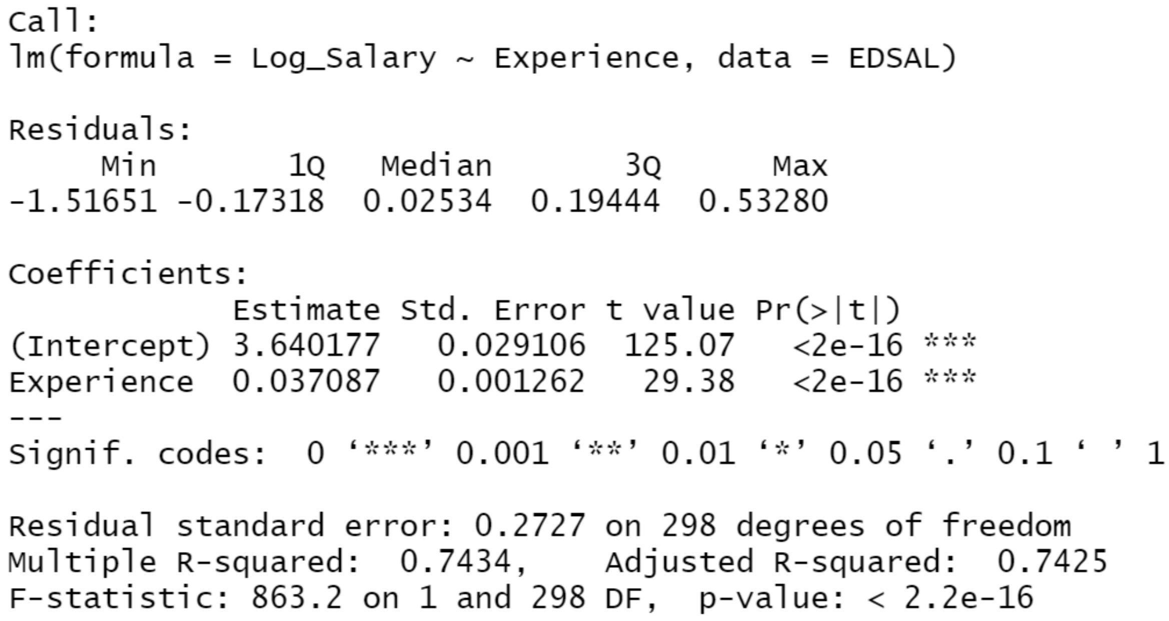
```
Call:
lm(formula = Salary ~ Experience, data = EDSAL)

Residuals:
   Min     1Q Median     3Q    Max 
-73.00 -12.82  -1.18  13.32  60.85 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  29.4679     2.5673   11.48   <2e-16 ***
Experience    3.0959     0.1113   27.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.05 on 298 degrees of freedom
Multiple R-squared:  0.7218,	Adjusted R-squared:  0.7209 
F-statistic: 773.2 on 1 and 298 DF,  p-value: < 2.2e-16
```

Linear-Log Model:



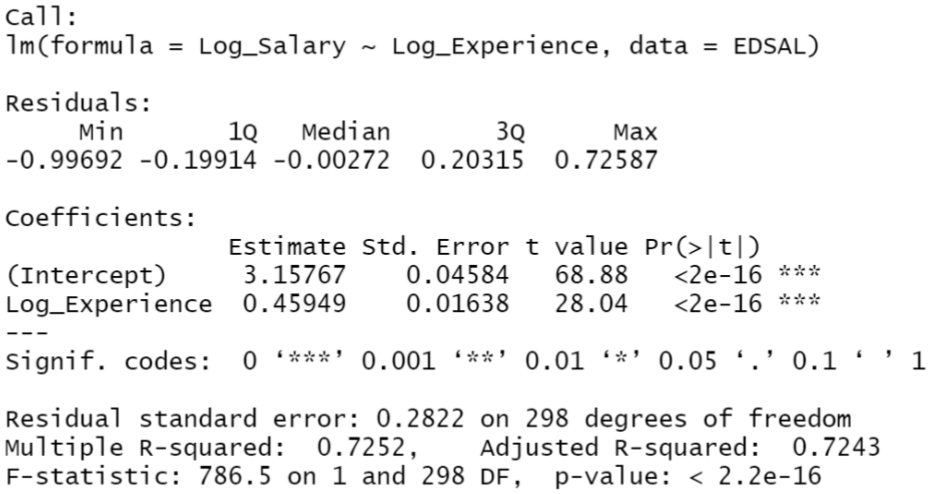
```
Call:
lm(formula = Salary ~ Log_Experience, data = EDSAL)

Residuals:
    Min      1Q  Median      3Q     Max 
-61.700 -21.895  -5.022  16.730  84.879 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -1.991      4.768  -0.418    0.677    
Log_Experience   34.985      1.704  20.529   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.35 on 298 degrees of freedom
Multiple R-squared:  0.5858,	Adjusted R-squared:  0.5844 
F-statistic: 421.5 on 1 and 298 DF,  p-value: < 2.2e-16
```

Log-Linear Model:



```
Call:
lm(formula = Log_Salary ~ Experience, data = EDSAL)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.51651 -0.17318  0.02534  0.19444  0.53280 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.640177   0.029106  125.07   <2e-16 ***
Experience  0.037087   0.001262   29.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2727 on 298 degrees of freedom
Multiple R-squared:  0.7434,    Adjusted R-squared:  0.7425 
F-statistic: 863.2 on 1 and 298 DF,  p-value: < 2.2e-16
```

Log-Log Model:



```
Call:
lm(formula = Log_Salary ~ Log_Experience, data = EDSAL)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.99692 -0.19914 -0.00272  0.20315  0.72587 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.15767    0.04584   68.88   <2e-16 ***
Log_Experience  0.45949    0.01638   28.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2822 on 298 degrees of freedom
Multiple R-squared:  0.7252,    Adjusted R-squared:  0.7243 
F-statistic: 786.5 on 1 and 298 DF,  p-value: < 2.2e-16
```

<u>Question 19</u>

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
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

Answer: B
Explanation:

$InY = b_0 + b_1 * X \Rightarrow \frac{\mathrm{dY}}{\mathrm{Y}} = b_1 * dX$ where $\frac{dY}{Y}$ is the percentage change in Y and dX is the change in X

$InY = b_0 + b_1 * X \Rightarrow Y = e^{b_0+b_1*X} \Rightarrow$ Increase X by 1 unit $\Rightarrow Y(X+1) = e^{b_0+b_1*(X+1)} \Rightarrow$ Percentage change of Y $= \frac{Y(X+1)}{Y(X)} - 1 = \frac{e^{b_0+b_1*(X+1)}}{e^{b_0+b_1*X}} - 1 = e^{b_1} - 1 = (e^{b_1} - 1) * 100\%$

Question 20

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

Answer: D
Explanation:

$InY = b_0 + b_1 * InX \Rightarrow \frac{dY}{Y} = b_1 * \frac{dX}{X}$ where $\frac{dY}{Y}$ is the percentage change in Y and $\frac{dX}{X}$ is the percentage change in X

$InY = b_0 + b_1 * InX \Rightarrow \mathrm{Y} = \mathrm{e}^{\mathrm{b_0+b_1*InX}} \Rightarrow$ Increase X by 1% $\Rightarrow \mathrm{Y(InX + 0.01)} = e^{b_0+b_1*[In(X)+0.01]} \Rightarrow$ percentage change of Y $= \frac{\mathrm{Y(InX+0.01)}}{\mathrm{Y(InX)}} - 1 = \frac{\mathrm{e^{b_0+b_1*(InX+0.01)}}}{\mathrm{e^{b_0+b_1*InX}}} - 1 = e^{0.01b_1} - 1 = (e^{0.01b_1} - 1) * 100\%$