**Graded Homework 1**

**Solution Manual**

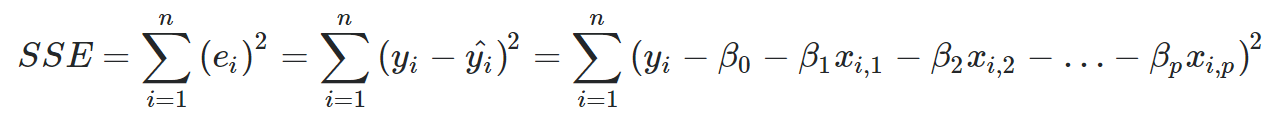
**Question 1. Choose the correct statement regarding the sum of residuals calculated using Ordinary Least Squares (OLS).**

Answer 1. (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.

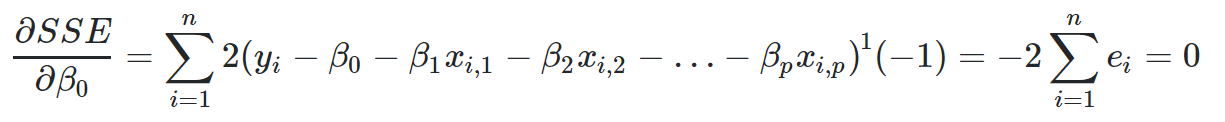
When an intercept is included in multiple linear regression for a data set with p points, the line is of the form:



In OLS (Ordinary Least Square) Regression, the objective is to minimize the SSE (Sum of Squared Errors) to 0, which is obtained as follows:



And to find all the betas, we differentiate all the above equation against all of betas individually including *β0.* We obtain ‘p+1’ equations to solve for all the ‘p+1’ betas. Differentiating against *β0*, we get:



The above equation is responsible for ensuring that the sum of residuals is 0. Therefore, C and D are incorrect options.

We now must decide between A and B. One form of regression under consideration is “Regression passing through Origin” (which is ideally not considered to be True OLS) which is a special case. Absence of intercept still means that the SSE is minimized, but it does not guarantee that SSE = 0. It is only when the intercept is included that the residuals sum up to 0.

Thus, B is the correct option. One can think of the intercept as the catchall term, which provides for shifting the line of regression closer to or away from the X-axis (say the response terms are of the order 1000s, and thus the *β0* value helps move the line of regression the said order, providing for finer tuned remaining betas).

**We understand that some ambiguity was observed regarding the forms of regression. Here, the question was designed keeping in mind the concept of the importance of the intercept term. While the true answer should be (B), we shall consider both (A) and (B) as the right choice.**

**Question 2. Consider a simple linear regression model. Y ~ X. A fit for this model will be a line in the X-Y plane. Now suppose we fit a model for the opposite relationship: X ~ Y (that is, X dependent, and Y independent). This new model yields a new regression line in the X-Y plane. Which of the following statements are true about these two lines?**

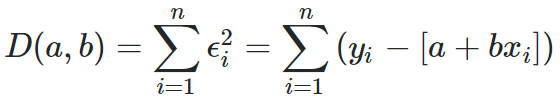
Answer 2. (B) These lines will always intersect at the mean (Xmean, Ymean)

If we have that y is a dependent variable on x, (and for their respective means and ), then the least squares regression Y ~ X line is given as:

Similarly, the line for X ~ Y is given as:

Clearly, these two lines meet at the means and

The above equation is obtained using the least square estimator method for estimating the values a and b while minimizing the residuals:



2

The final line is and we find a and b using equations obtained by placing

which gives us and

We can also write a small piece of code to check the same:

> mydata = read.csv(‘EDSAL.csv')

> YregX <- lm(Salary~Experience, mydata)

> summary(YregX)

Call:

lm(formula = Salary ~ Experience, data = mydata)

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

> XregY <- lm(Experience~Salary, mydata)

> summary(XregY)

Call:

lm(formula = Experience ~ Salary, data = mydata)

Residuals:

Min 1Q Median 3Q Max

-11.430 -4.867 -1.787 3.489 20.805

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.475177 0.841711 -1.753 0.0807 .

Salary 0.233146 0.008385 27.806 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.601 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

> plot(mydata$Experience, mydata$Salary)

> abline(29.4679, 3.0959)

> abline(1.475177/0.233146, 1/0.233146)

>

> mean(mydata$Experience)

[1] 19.39333

> mean(mydata$Salary)

[1] 89.50846

The graph produced is as follows (which meets at the mean of the Xs and Ys)

A screenshot of a cell phone

Description automatically generated

> abline(29.4679, 3.0959)

> abline(1.475177/0.233146, 1/0.233146)

Kindly note that while plotting the two line on the same axis, the line from X ~ Y needs to be transformed to X-Y co-ordinate as shown above in the second line of code. Another point to note is that the difference in the slope for these lines is large as randomness increases in the data. If you consider data that is perfectly describable by a linear equation, then one would obtain the same line in both the cases (i.e. Y ~ X and X ~ Y).

For example, considering Data Points {(1,3), (2,5), (3,7), (4,9), (5,11)}, the lines obtained would be y = 2x + 1 and x = 0.5y + 0.5, which are basically the same overlapping lines.

**Code for Questions 3 to 7**

> # loading data

> library(Ecdat)

> load(cars)

>

> # correlation

> cor(cars$speed, cars$dist)

[1] 0.8068949

>

> # simple lin reg model and summary

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

>

> summary(lm)

Call:

lm(formula = dist ~ speed, data = cars)

Residuals:

Min 1Q Median 3Q Max

-29.069 -9.525 -2.272 9.215 43.201

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -17.5791 6.7584 -2.601 0.0123 \*

speed 3.9324 0.4155 9.464 1.49e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

> # changing units to meters per second and meters

> new.speed <- cars$speed\*0.44704

> new.dist <- cars$dist\*0.3048

>

> # new lin reg model after updating units

> lm.new <- lm(new.dist~new.speed)

>

> summary(lm.new)

Call:

lm(formula = new.dist ~ new.speed)

Residuals:

Min 1Q Median 3Q Max

-8.8603 -2.9033 -0.6925 2.8086 13.1678

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.3581 2.0600 -2.601 0.0123 \*

new.speed 2.6812 0.2833 9.464 1.49e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.688 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

> # predict stopping distance with 7/5m/s speed and confidence interval

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

>

> predict(ols\_reg2, testdata, interval="predict")

fit lwr upr

1 14.7526 5.227095 24.2781

>

> predict(ols\_reg2, testdata, interval="predict", level = 0.9)

fit lwr upr

1 14.7526 6.806648 22.69855

**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?**

Answer 3. (C) 0.81

As can be seen from the code above, the cor() function in R returns a value of 0.8068 as the correlation between Speed and Distance

**Question 4. Would you say that distance to stop and speed of the car are?**

Answer 4. (C) Well Correlated

The 2 covariates are well-correlated with a correlation value which is close to 1 in the acceptable range [-1, 1].

**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?**

Answer 5. (A) 0.65, -17.58 and 3.93

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 regression summary table) = 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?**

Answer 6. (B) 0.65, -5.36, and 2.68

First, change the dataset into proper units. Convert speed from miles per hour to meters per second (multiply by 0.44704) and 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.

Percent variation explained by speed is the R-squared value = 0.65

Intercept of speed (from the regression summary table) = -5.36

Coefficient of speed (from the regression summary table) = 2.68

**Question 7. Now suppose your model is ready and you are asked to make a prediction for distance a car will need to stop in meters if it is moving at a speed of 7.5 m/s. You are required to report the predicted distance and also the lower and upper bound for the 90% confidence interval. What will you report?**

Answer 7. (D) None of the Above

To the calculate the above given metrics, we use the R function predict() as follows:

> predict (lm.new, newdata = new.dat, interval ='confidence', level=0.90)

fit lwr upr

[1] 14.7508 13.60107 15.90053

This does not match any of the given options. Take a note of the argument ‘interval’. While calculating a range for prediction, we may come across two types intervals – ‘confidence’ and ‘predict’ and it is important to understand the difference between the two.

> predict(ols\_reg2, testdata, interval=’predict’, level = 0.90)

fit lwr upr

[1] 14.7526 6.806648 22.69855

Confidence intervals tell you about how well you have determined the mean. The key point is that the confidence interval tells you about the likely location of the true population parameter.

Prediction intervals tell you where you can expect to see the next data point sampled. The key point is that the prediction interval tells you about the distribution of values, not the uncertainty in determining the population mean.

Prediction intervals must account for both the uncertainty in knowing the value of the population mean, plus data scatter. So, a prediction interval is always wider than a confidence interval.

Let us consider another example [Thanks to Serhii Kushchenko on stats stack exchange for this explanation]. Suppose you try to predict the people's weight from their height, gender(Male, Female) and Calories Consumed (Less than 1000, 1000 to 2000, Above 2000). Given the large population of the earth (in billions), one would expect lots of people with the same height, gender and calories consumed having different weights, which may vary for different reasons. But what we may talk about with some certainty is that on average, we would obtain a somewhat constant value when sampled multiple times.

One task might be to forecast the weight of some specific person. And we don't know the living circumstances of that individual. Here the prediction interval must be used. But in a practical environment, the task is to predict the average weight of all the people having the same values of all three explanatory variables at hand. Here we use the confidence interval. Likewise, in the problem, we wish to check, given that the vehicle is moving at 7.5 m/s, what would be on average, the likely value of the stopping distance.

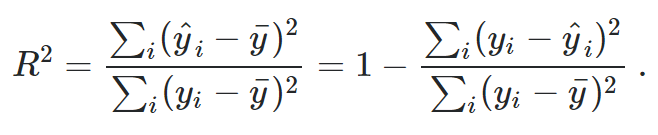
We observe that the two are centered around the same point, but prediction interval is much wider than the confidence interval.

**NOTE: We understand that this concept is highly statistical in nature, and thus, would award points to both (A) and (D).**

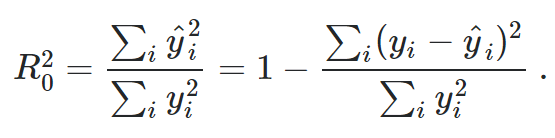
**Question 8. Your boss is not happy with the intercept term and asks you to try dropping it. Let’s drop the intercept as your boss asked. Which of the following finding will you report back to boss concerning the percentage of variation explained by the model?**

Answer 8. (B) In the R output, the percentage of variation explained increases, but this does not mean a better fit as percentage of variation explained is artificially inflated in R output when you remove the intercept term

The regular way calculating R2 in R is as follows:



But when dropping the intercept term, R uses a modified version which is as follows:



R2 becomes higher without intercept, not because the model is better, but because the definition of R2 used is another one! R2 is an expression of a comparison of the estimated model with some standard model, expressed as reduction in sum of squares compared to sum of squares with the standard model. In the model with intercept, the comparison sum of squares is around the mean. Without intercept, it is around zero! The latter is usually much higher, so it easier to get a large reduction in sum of squares. However, setting y=0 at x=0 introduces bias in the data and thus the co-efficient may not be truly significant

**Question 9. We regress calorie consumption of individual based on their region, the level of urban development of the place they live in and age. To transform region variable into categorical variables for regression, how many dummy variables do we need to insert into the regression?**

Answer 9. (C) 2

We have three categorical variables, only two variables are needed because one factor level is taken as base line.

**Question 10. We regress calorie consumption of individual based on their region, the level of urban development of the place they live in and age. Which statement about Level of Urban Development is TRUE (confidence Interval = 95%)?**

Answer 10. (D) At the same age level, people living in urban areas consume more calories than people living in rural areas.

The confidence level is 95%, and the p-value for Rural\_True (0.012) is less than 0.05. Therefore, Rural\_True has a significant effect on the response variable. A is incorrect. B and C fail to consider ageing effect; thus B and C are incorrect. People in rural area consume 40 calories less than those in urban areas, while holding age constant. Hence Choice D is the correct answer

**Question 11. Suppose we regress satisfaction against food type, condiment type and interaction between food type and condiment type:**

**Satisfaction = b0 + b1\* Food + b2\* Condiment + b3\* Food\*Condiment**

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

Answer 11. (A) Interaction effects exist between food and condiment since p-value is much smaller than the significance level/alpha = 5%. We should include the interaction term in the regression model to explain the variability in the data.

Since the p value of the interaction effect is close to zero, and it is less than 5% (error rate), we will reject the null hypothesis that interaction effect does not exist between food and condiment. Thus, A is correct answer. If food and condiment are independent of each other, the interaction effect should not be significant, making D is incorrect. Interaction effects are relevant when conducting regression analysis, E is incorrect.

**Question 12. Suppose we regress satisfaction against food type, condiment type and interaction between food type and condiment type:**

**Satisfaction = b0 + b1\* Food + b2\* Condiment + b3\* Food\*Condiment**

**What is the expected value of satisfaction if food is hot dog and condiment is chocolate sauce? (The coefficient is 1 when the food is hot dog and condiment is chocolate sauce.)**

Answer 12. (E) 65.317

77.320 + 0.141 + 1.863 - 14.007 = 65.317

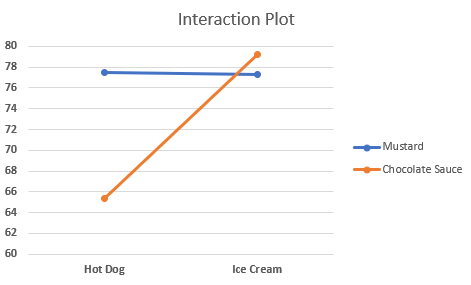
The total satisfaction is obtained by adding Intercept + Effects of Food (when Hot dog = 1) + Effects of Condiment (when Chocolate Sauce = 1) + Interaction Effects of Hot dog and Chocolate Sauce.

**NOTE: This question originally had a typo (64.317 instead of 65.317). We updated the quiz, but it didn’t reflect for all students – so asked them to pick the closest answer.**

**Question 13. When plotting the expected value for food (one line for chocolate sauce and one line for mustard), which plot are you most likely to see?**

Answer 13. (E) Since the interaction effect is significant between food and condiment, we should expect to see the two lines cross each other. Therefore, A and C cannot be correct. In terms of choice B, from the statistic result table in the previous question, hot dog with mustard sauce should be more pleasant than hot dog with chocolate sauce. Choice B is also not correct.

In terms of choice D, if the mustard line is flat, it means that the food condiment has no significant influence on the satisfaction. However, the statistic table result shows that the food condiment does have significant influence. Therefore, E is the most likely plot.



**NOTE: While (E) was supposed to be the logically correct answer, purely based on the visual aspect (as was intended while building the question), we understand that based on the numerical values, plot D is closer to the correct plot. Thus, both plot (D) and (E) will be considered the right answers. The plot based on numerical values:**

**Code for Question 14 to 18**

> library(lmtest)

> library(MASS)

>

> mydata = Boston

> OLSModel = lm(medv ~ ., mydata)

>

> summary(OLSModel)

Call:

lm(formula = medv ~ ., data = mydata)

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(OLSModel)

>

> shapiro.test(residuals(OLSModel))

Shapiro-Wilk normality test

data: residuals(OLSModel)

W = 0.90138, p-value < 2.2e-16

> dwtest(OLSModel)

Durbin-Watson test

data: OLSModel

DW = 1.0784, p-value < 2.2e-16

alternative hypothesis: true autocorrelation is greater than 0

> sqrt\_OLSModel = lm(sqrt(medv) ~ ., mydata)

> summary(sqrt\_OLSModel)

Call:

lm(formula = sqrt(medv) ~ ., data = mydata)

Residuals:

Min 1Q Median 3Q Max

-1.35292 -0.26646 -0.04765 0.20693 2.20133

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.6171186 0.4830511 13.699 < 2e-16 \*\*\*

crim -0.0166326 0.0031107 -5.347 1.37e-07 \*\*\*

zn 0.0037856 0.0012993 2.913 0.003737 \*\*

indus 0.0035719 0.0058207 0.614 0.539728

chas 0.2546251 0.0815500 3.122 0.001900 \*\*

nox -1.8316237 0.3615452 -5.066 5.75e-07 \*\*\*

rm 0.3045221 0.0395573 7.698 7.62e-14 \*\*\*

age 0.0001287 0.0012503 0.103 0.918076

dis -0.1345203 0.0188787 -7.125 3.71e-12 \*\*\*

rad 0.0322156 0.0062798 5.130 4.18e-07 \*\*\*

tax -0.0013597 0.0003559 -3.820 0.000151 \*\*\*

ptratio -0.0941484 0.0123830 -7.603 1.48e-13 \*\*\*

black 0.0009775 0.0002542 3.845 0.000136 \*\*\*

lstat -0.0603341 0.0048003 -12.569 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4492 on 492 degrees of freedom

Multiple R-squared: 0.7758, Adjusted R-squared: 0.7698

F-statistic: 130.9 on 13 and 492 DF, p-value: < 2.2e-16

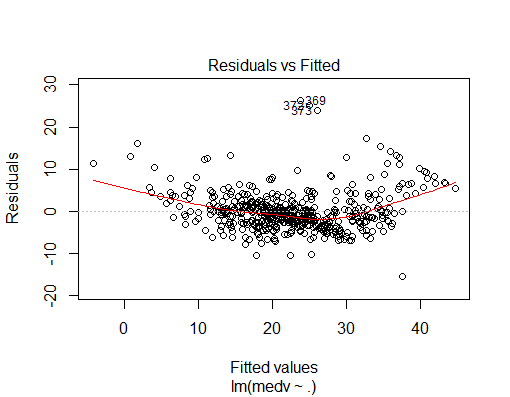
> plot(sqrt\_OLSModel)

**Question 14. Let’s check for non-constant variance using visual inspection of a diagnostic plot (Residuals vs Fitted). What do you observe?**

Answer 14. (A) There seems to be homoskedasticity but there is clearly a non-linear pattern.

For the majority part (between fitted values 5 to 35, where more than 95% of the data points are observed), the residuals are not aggressively diverging out of or converging into the fitted line (which is usually seen as a cone formation that suggests that the variation of residuals changes with fitted values). Thus, we infer that Homoskedasticity is observed for the most part of it.

However, there is clearly a pattern in the residuals. We observe that towards the either ends (values less than 5 and greater than 35) of the fitted plot (the red line), there are no residuals on the negative side of the axis. This is indicative of the fact that a non-linear data set is being fitted to a linear model.



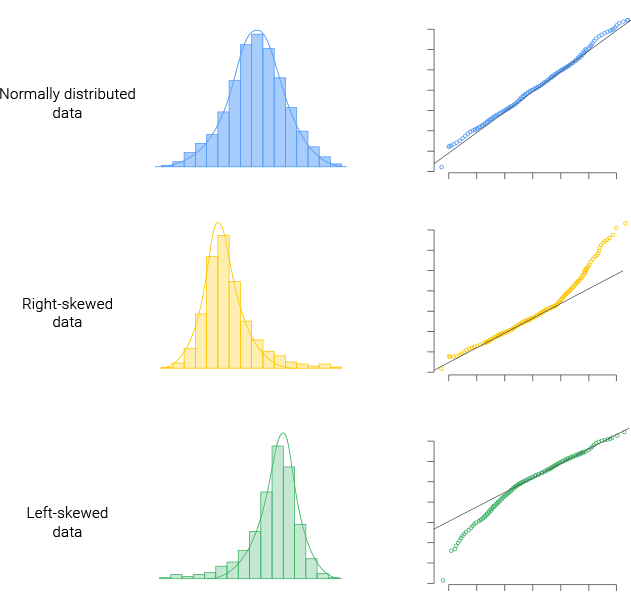
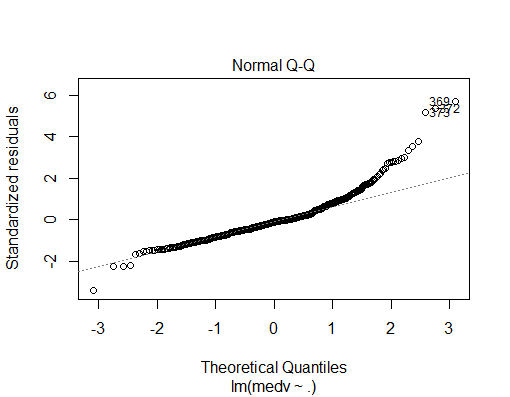
**NOTE: We understand that a lot of students faced problem judging Homoskedasticity or Heteroskedasticity for this particular problem. Usually, visual interpretation is a supportive factor to statistical tests, and we this problem was particularly designed for the very reason. In light of this, both (A) and (C) will be regarded as correct.**

**Question 15. Let’s check if our residuals are normal by doing visual inspection of a diagnostic plot. (Normal Q-Q plot). What do you observe?**

Answer 15. (C) There seems to be non-normality with distribution being right-skewed.

As can be seen in the plot, towards the right end of Theoretical Quantiles, the Standardized residuals of the data tend to shift towards the 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 Standardized ones on the Y axis.



**Question 16. 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 as follows:**

> 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 Question 15?**

Answer 16. (B) Yes, we get a low p-value in Shapiro-Wilk test which means the residuals are not normally distributed and visual inspection in Question 15 also led to the conclusion that there is a non-normal distribution of residuals.

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 17. 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?**

Answer 17. (B) The small p-value indicates that there might be autocorrelation.

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 18. Let’s try to correct for some violations in assumptions. What’s the impact of applying a square-root transformation on the response variable? Please note that independent variables are all the variables except age and indus.**

Answer 18. (D) The non-linear pattern earlier visible in residuals in Question 14 becomes less prominent and the residuals are still not normally distributed.

This is reinforced by the fact that there seems to be a non-linear relation in the response and predictor variables. Here, we need to check two aspects of the residuals :- how are the residuals distributed and what is the effect on the residuals when the square-root transformation is applied.

We can simply check the residuals vs fitted line plot using the plot() function which shows as below:

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

The plot is similar in nature with the one in Question 14, but with slightly more values dropping below the x-axis. The QQ plot further shows that the residuals are not normally distributed and thus we should expected uneven distribution of residuals around the fitted line. Thus the non-linear trend is still expected. But to truly appreciate the effect of the square-root transformation, we should compare the residuals on the same range for the Y-axis.

The following piece of code should help visualize:

> mydata = Boston

> OLSModel = lm(medv ~ ., mydata)

> sqrt\_OLSModel = lm(sqrt(medv) ~ ., mydata)

> plot(fitted.values(OLSModel), residuals(OLSModel), ylim = c(-15,20))

> plot(fitted.values(sqrt\_OLSModel), residuals(sqrt\_OLSModel), ylim = c(-15,20))

A close up of a logo

Description automatically generatedA screenshot of a cell phone

Description automatically generated

As can be seen, the residuals have dropped significantly closer to 0, and with this, we can say that the non-linear pattern is less prominent.

**Code for Questions 19 to 21**

> edsal = read.csv('C:\\Users\\anmol\\Desktop\\TA Work\\EDSAL.csv')

> edsal$LogExperience = log(edsal$Experience)

> edsal$LogSalary = log(edsal$Salary)

>

> linlin = lm(Salary ~ Experience, edsal)

> linlog = lm(Salary ~ LogExperience, edsal)

> loglin = lm(LogSalary ~ Experience, edsal)

> loglog = lm(LogSalary ~ LogExperience, edsal)

> summary(linlin)

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

> summary(linlog)

Call:

lm(formula = Salary ~ LogExperience, 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

LogExperience 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

> summary(loglin)

Call:

lm(formula = LogSalary ~ 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

> summary(loglog)

Call:

lm(formula = LogSalary ~ LogExperience, 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 \*\*\*

LogExperience 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

**Question 19. Which of the 4 fitted models has the highest R-square value?**

Answer 19. (C) Log-Lin

The Log-Linear model has the highest R-Square value of 0.7434 and hence is the correct answer.

**Question 20. Which is the interpretation of the slope coefficient for the Log-Lin model?**

Answer 20. (B) Increasing Experience by 1 unit leads to ((e^0.037087)-1) \*100% increase in Salary

We are considering the following model:

→

Increasing Experience by one unit will increase log(Salary) by 0.37087 units.

However, the model can also be written as

The term in the equation above represents the percentage change in Y. However both and are differentials (infinitesimally small) and thus for our purpose need to be changed to and .

For unit change, . We thus consider Y at X=0 and X=1 to compute .

**Question 21. Which is the interpretation of the slope coefficient for the Lin-Log model?**

Answer 21. (C) Increasing Experience by 1% leads to 0.01\*34.985 units increase in Salary

We are considering the following model:

→

Increasing Log (Experience) by one unit will increase Salary by 34.985 units.

However, the model can also be written as

The term in the equation above represents the percentage change in X. Thus for 1% change in Experience, Y changes by a factor of