98

STA 207 HW-4

Due Date: Oct. 10 by 10:20AM in Moodle
Derin Gezgin | Camel ID: 00468038

Problem 1: (50 points)

For the textbook prices question in HW 3 you analyzed the data TextPrices. Continue the previous analysis by answering the following questions:

1. ***[5 points] Perform a hypothesis test to address the students' question of whether the number of pages is a useful predictor of a textbook's price. Report all steps (the hypotheses, level of significance, test statistic, and p-value, along with your conclusion within the context.)***

   The Hypothesis
   H_0 = Number of pages is not a useful predictor of a textbook's price. ✓
   H_1 = Number of pages is a useful predictor of a textbook's price.

   Level of Significance → 0.05
   Test Statistic → 7.653 ✓
   p-Value → 2.45e-08

   Conclusion
   p-Value is less than the level of significance (0.05), this means that we have enough evidence to reject the null hypothesis and we can conclude that there's a linear relationship between the number of pages and a textbook's price. ✓

2. ***[15 points] Determine a 95% confidence interval (CI) for the population slope coefficient. Interpret the CI in the context of the data. Determine a 90% CI for the population slope coefficient. Interpret the CI in the context of the data. What happened when we reduced the confidence level from 95% to 90%?***

   We're %95 confident that the actual difference in the estimated textbook price for each 1-page increase in the number of pages is between 0.108 and 0.187$ ✓

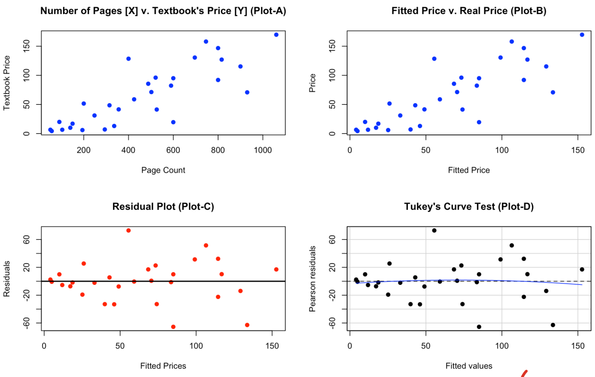
   We're %90 confident that the actual difference in the estimated textbook price for each 1-page increase in the number of pages is between 0.115 and 0.18.$ ✓

   When we reduced the confidence level from 95% to 90%, we have a tighter value bound as we're less confident about our prediction. When we're less confident, we can reduce the value range and make more precise (and less likely) predictions. ✓

3. ***[30 points] For the model you fitted in homework-3, test the LINE conditions. Report all steps and discuss findings.***

L) There is a linear relationship between X and Y

For this condition, we need four different plots: Scatterplot of Y versus X in SLR, Scatterplot of $\hat{Y}$ versus Y, Residual plot, and Tukey's curve test.



From Plot-A, and B, we can see that there's a somewhat strong linear relationship in the positive direction. There are a few outliers. Plot-B also shows that the estimated value of price agrees with the real price with a few outliers.

Plot-C shows us that the residuals are distributed randomly around the 0 line.

Plot-D shows us that the overall trend in the residuals is very close to 0 which shows that there's no clear trend in residuals.

Tukey Test

H_0 = The textbook price and the textbook page count has a linear relationship.
H_A = The textbook price and the page count does not have a linear relationship.

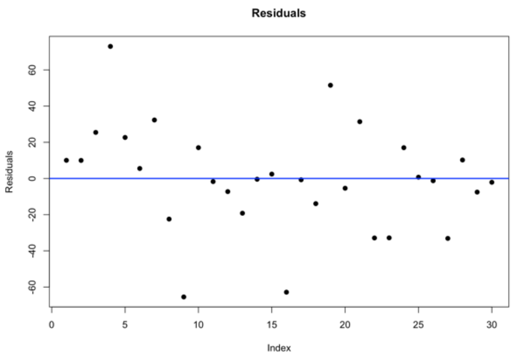
Level of Significance → 0.05
Test Statistic → -0.3114
p-Value → 0.7555

Conclusion
p-Value is larger than the level of significance (0.05), this means that we don't have enough evidence to reject the null hypothesis and we can conclude that there's a linear relationship between the number of pages and a textbook's price.

I) Independent errors

In this case, we have to make an index plot of the residuals.

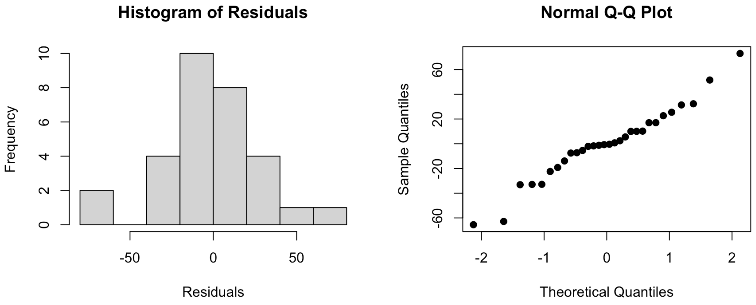
If we check this index plot of residuals, we can see that there's not a clear pattern among the errors which supports the independence of errors.



-1, There is a repetitive pattern so assumption fails

N) Normally distributed errors

For this test, we have to check the histogram of errors and also the QQ plot.



Put your graphs all together like above

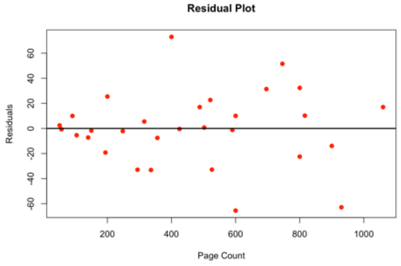
If we check the histogram of residuals, we can see that there's a distribution that is similar to normal distribution. This distribution is not exactly a normal distribution but it seems like it.

On the other hand, if we check the normal Q-Q plot, it's showing a slightly strong positive linear relationship, I won't say this is very strong. As the degree of the line is close to 45 degrees, we can conclude that the normality assumption holds true.

E) Errors are homoscedastic

For this, we have to check the residual plot with the mean of errors values line.

We can see that the errors do not change as the page count increases. So, we can conclude that the errors are homoscedastic.



Using the Breusch-Pagan test:

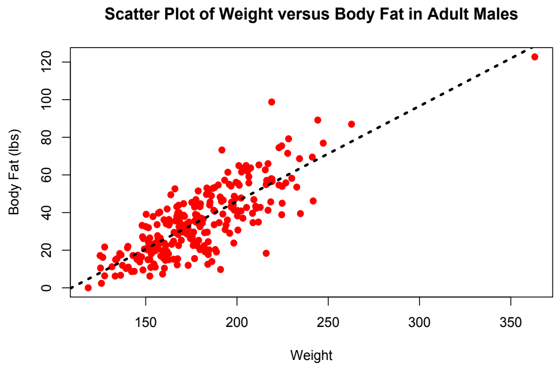
The Hypothesis
H_0 = Errors are homoscedastic
H_A = Errors are heteroscedastic

Level of Significance → 0.05
Test Statistic → 2.4513
p-Value → 0.1174

Conclusion
p-Value is more than the level of significance (0.05), this means that we don't have enough evidence to reject the null hypothesis and we can conclude that the errors are homoscedastic.

Problem 2: (50 points)

Using a sample of 252 adult males, a study would like to establish if there is a linear relationship between a man's percent body fat and his weight based on a simple linear regression. The regression model R-output is given below. The scatter plot of total body fat and weight along with the fitted linear regression line is shown in the Figure below.



R-output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) _______     4.3161  -12.83   0.00001 ***
X             0.5066    _____    21.28   0.00001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 11.08 on 250 degrees of freedom
Multiple R-squared:  0.6443,    Adjusted R-squared:  0.6429
F-statistic: 452.9 on 1 and 250 DF,  p-value: < 2.2e-16
```

Using the R-output and scatterplot answer the following questions:

(a) Report the fitted regression model (equation).

The formula of the $t\ value$ is $\frac{\widehat{B_\iota}-B_{i,H_0}}{se(\widehat{B_\iota})}$. In our case as we say that for H_0, B_i is also 0 our formula would be t =$\frac{\widehat{B_\iota}}{se(\widehat{B_\iota})}$. From the given output we already know that t value is -12.83. At the same time, we know that $se(\widehat{B_\iota})$ is 4.3161. If we do the multiplication B_i would be **-55.375563**

So, the fitted model equation would be:

$$\hat{Y} = 0.5066x - 55.3766$$

where $\hat{Y}$ is the body fat and the x is the wight.

(b) Find the correlation between the weight and percent body fat of adult males (justify both magnitude and direction). Interpret the correlation.

If we look at the slope and the plot, we can see that there's a positive correlation between the variables, and it seems fairly strong.

If we check the square root of the multiple r squared value (0.6443), it is +/- 0.803. As the graph shows positive relationship, we can say the R-Value is also positive. Considering it's close to 1, it can be said that it's a strong relationship.

(c) Report the standard error in slope estimate and interpret it.

The formula of the $t\ value$ is $\frac{\widehat{B_i}-B_{i,H_0}}{se(\widehat{B_i})}$. In our case as we say that for H$_0$, B$_i$ is also 0 our formula would be t $=\frac{\widehat{B_i}}{se(\widehat{B_i})}$. From the given output we already know that t value is 21.28. At the same time, we know that slope estimate is 0.5066. So, the equation would be $21.28 = \frac{0.5066}{se(\widehat{B_i})}$. If we do the multiplication $se(\widehat{B_i})$ x $21.28 = 0.5066$ would be **0.0238.**

So, the sample-to-sample variability in the slope estimate is 0.0238.

(d) Report the standard error of residuals and interpret it.

The residual standard error is 11.08.

The average distance that the observed values of body fat are different from the fitted values by 11.08 lbs.

(e) Report the coefficient of determination (R^2) and interpret it.

The coefficient of determination is 0.6443, this means that the proportion of variation in body fat estimates explained by weight is 64.4%.

(f) It is hypothesized that weight is a significant predictor for the body fat. Use the given output to find out if this true or not. (show all steps: hypotheses, level of significance, test statistic, P-value and conclusion in context of the problem).

The Hypothesis

H$_0$ = Weight of an individual is not a useful predictor of body fat.
H$_1$ = Weight of an individual is a useful predictor of body fat.

Level of Significance → 0.05
Test Statistic → ~~-12.83~~ -1, 21.28
p-Value → 0.0001

Conclusion

p-Value is less than the level of significance (0.05), this means that we have enough evidence to reject the null hypothesis and we can conclude that there's a linear relationship between the weight and the body fat percentage of an individual.