STA 207 HW-4
Due Date: Oct. 10 by 10:20AM in Moodle

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.)
   From hw-3, we have the fitted model
   $$\hat{Y} = -3.42 + 0.15X,$$
   where Y: Textbook price in $ and X is the number of pages.
   To test whether the number of pages is a useful predictor of textbook prices or not, we will set the hypotheses as:
   $$H_0: \beta_1 = 0$$
   $$H_1: \beta_1 \neq 0$$
   Set the level of significance $\alpha = 0.05$
   Test statistic t=7.653
   P-value=0.0000000245
   Since P-value < α, we reject H_0
   Conclusion: At 5% level of significance, we have enough evidence to conclude that number of pages is a statistically significant predictor for the textbook 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%?

The 95% CI for slope is (0.108,0.187) meaning that we are 95% confident that for each additional page, the cost of the textbook is expected to increase somewhere b/w 0.108 to 0.187$.
The 90% CI for slope is (0.115, 0.180) meaning that we are 90% confident that for each additional page, the cost of the textbook is expected to increase somewhere b/w 0.115 to 0.180$.

We notice that as we reduced the confidence from 95% to 90%, the width of the interval reduced.

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

<u>Linearity between children and mid-parent height assumption</u>

To test whether there is a linear relation between the prices and number of pages in a textbook.
Plot a) in Figure 1 below shows a positive, linear relationship between the children and mid-parent height. There is no obvious outlier or non-linearity.
Plot b) in Figure 1 below shows that the estimated value of prices (using model 1) agrees with the observed (sample) prices.
Plot c) in Figure 1 below is the residual plot which doesn't show any clear pattern.
Plot d) in Figure 1 below is Tukey's curve which shows barely any deviation from the flat line, indicating absence of non-linearity.
All plots show that linearity assumption is not failing.
Next, we do the Tukey's curve test:

Step 1: Null hypothesis (Ho): Linearity assumption holds
Alternative Hypothesis (Ha): Linearity assumption fails
Step 2: Set Level of significance α as 0.05 (default value)
Step 3: Test statistic= -0.3114
Step 4: P-value=0.755
Step 5: If P-value $> \alpha$, so we do not reject Ho
Conclusion: At 5% level of significance, linearity assumption holds.

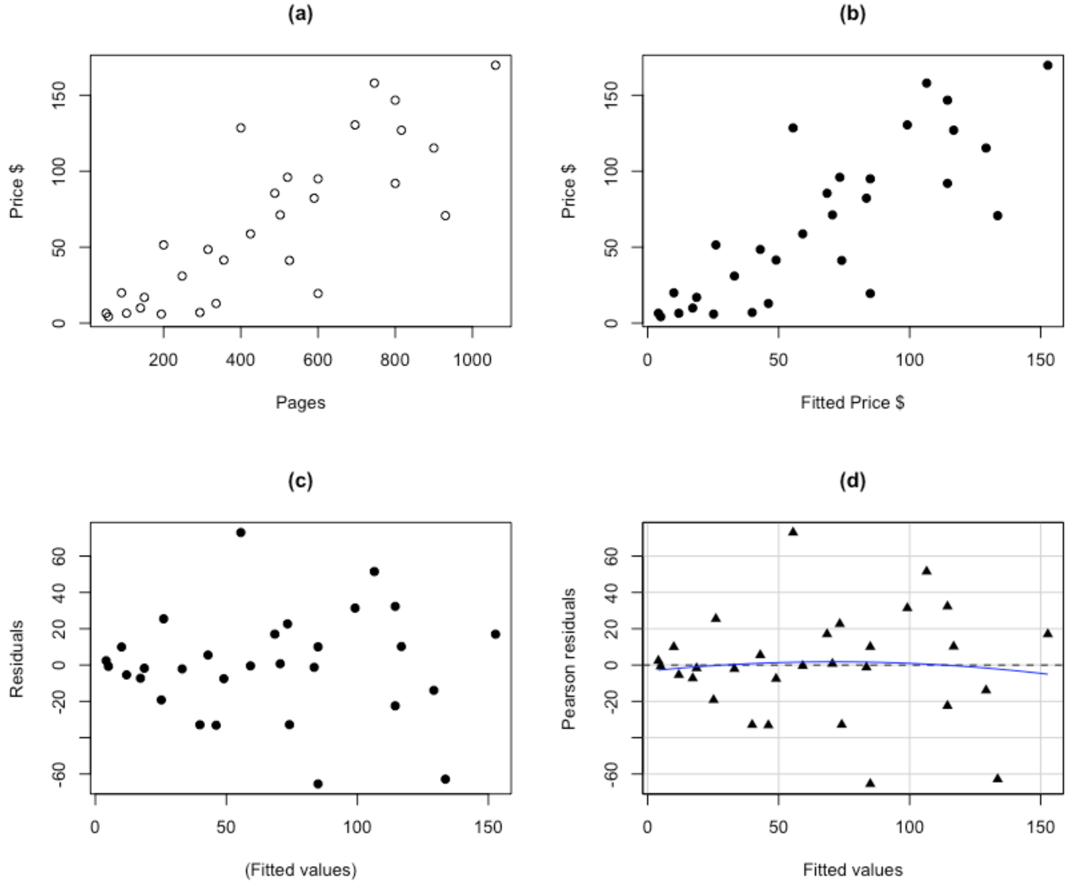


Figure 1: (a) scatterplot of children and mid-parent height, (b): scatterplot of children height and its estimate, (c) Residual plot and (d) Tukey's curve test, basically residual plot with a curve fitted.

Independence of errors assumption

Plot c) in Figure 2 below is the index plot of errors. There is a clear repetitive (cyclic) and positive relation among the errors. Therefore, from this we can see that the independence of errors assumption fails.

Normality of errors assumption

Plot a) in Figure 2 shows histogram of errors which is bell shaped with two peaks in the middle.

Plot b) in Figure 2 is the QQ plot and it shows agreement between the sample and normal quantiles as most values are aligned in a positive linear pattern. Therefore, from plots (a) and (b) in Figure 2, we can say that the normality of errors assumption holds true.

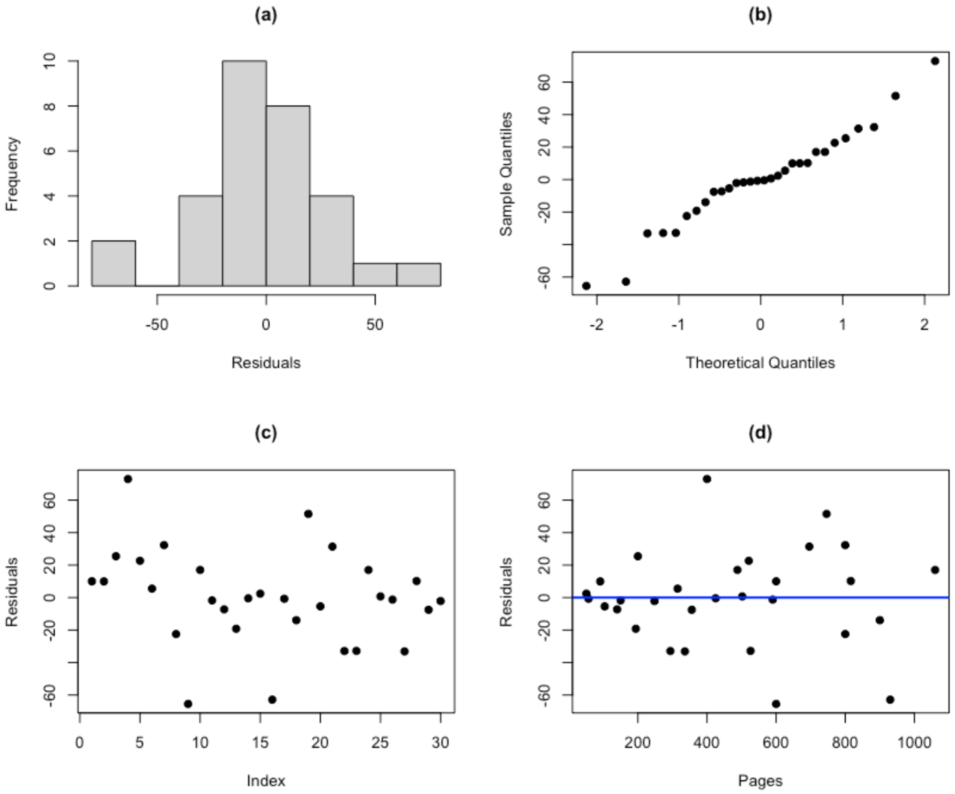
Errors are homoscedastic assumption

Plot d) in Figure 2 shows the residual plot with a mean of errors line. We do not see the variance of errors changing as number of pages increased, except at the beginning when spread is smaller. There is no clear evidence of homoscedasticity failing.

Using the Breusch-Pagan test:

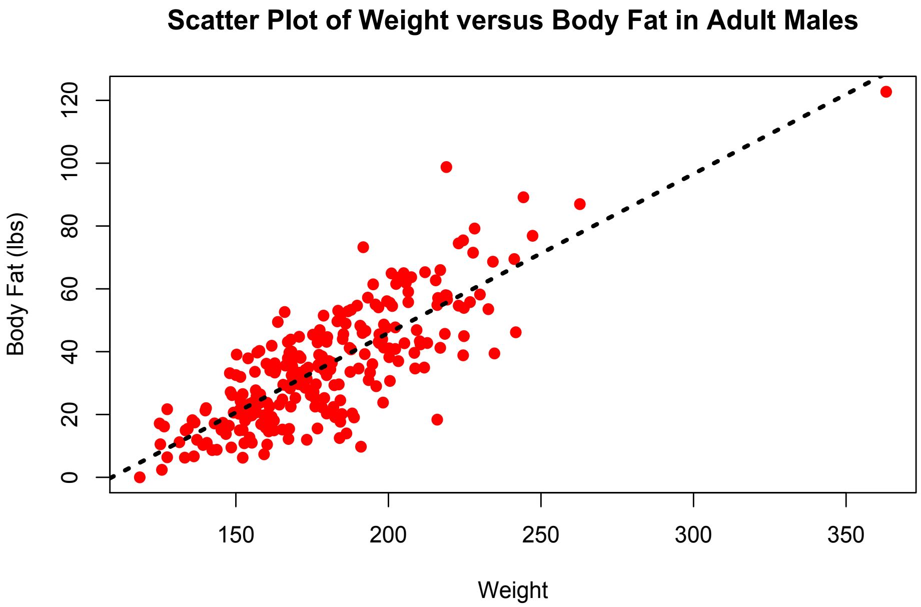
Step 1: H0: Errors are homoscedasticity)
Ha: Errors are heteroscedasticity

Step 2: $\alpha = 0.05$ (default value)
Step 3: Test statistic = 2.451
Step 4: P-value = 0.1174
Step 5: Since P-value > α, so we do not reject the H0.
Conclusion: At 5% level of significance, we have enough evidence that 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).

$\hat{Y} = -55.36 + 0.5066X$ where Y: total body fat (lbs) and X is the weight of adult males.

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

From R output we have Multiple R-squared: 0.6443, that is,
$R^2 = 0.6433$, thus $Cor(X,Y) = \sqrt{0.6443} = \pm 0.803$. Since the slope is positive so correlation must also be positive thus cor(X, Y)=0.803
This means there is a strong, positive relationship between the body fat and weight in adult males.

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

We know that

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

Thus,

$$se(\hat{\beta}_1) = \frac{\hat{\beta}_1}{t} = \frac{0.5066}{21.28} = 0.024$$

The sample-to-sample variability in the slope estimate is ± 0.024 pounds.

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

Using the standard error of residuals, $\hat{\sigma} = 11.08$ pounds it seems like there is on an average 11.08 pounds deviation in data and the fitted line representing a high deviation.

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

The coefficient of determination, R^2 shows that approximately 64% of variability in body fat is explained by the weight of an adult male which seems like a good measure.

(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).

For the hypotheses:

$$H_0: \beta_1 = 0,$$
$$H_a: \beta_1 \neq 0$$

Let $\alpha = 0.05$
t=21.28
P-value=0.00001 <0.05 → Reject H0.
Hence, we have evidence at 5% level of significance to state that the weight is a significant predictor for body fat.