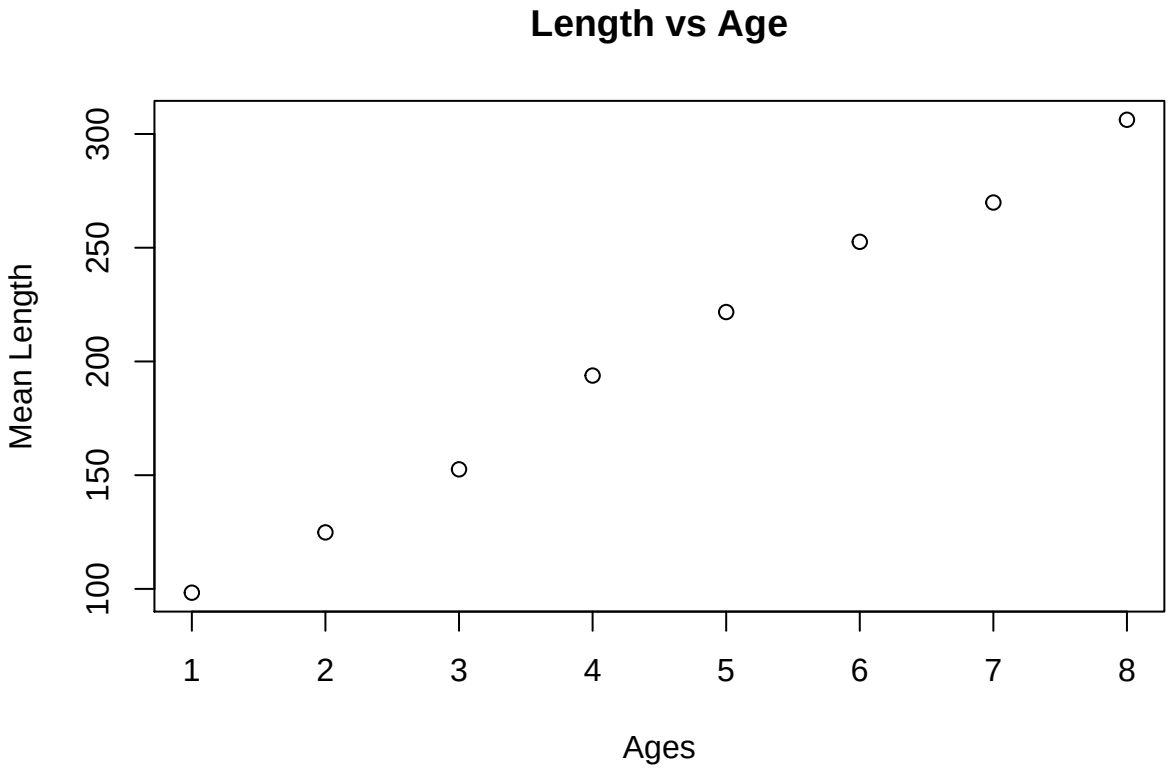
STAT 505: Homework1

Pratima KC

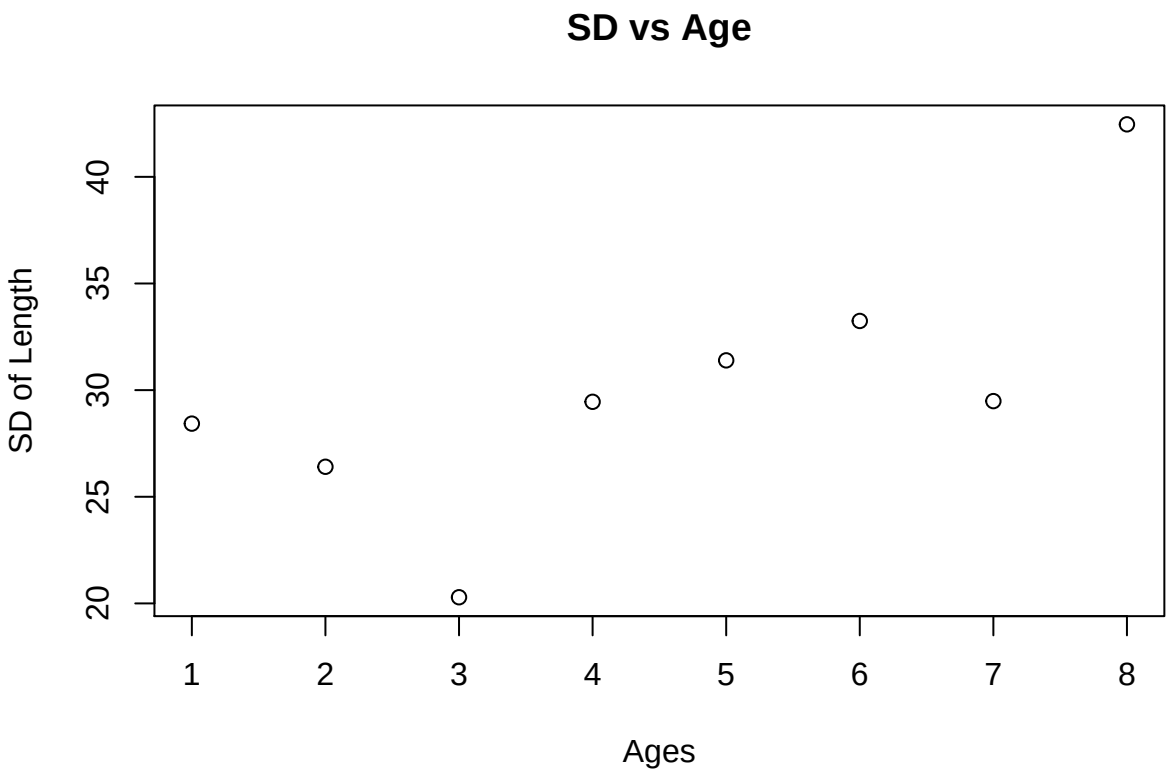
Question 1 a: (1 point) Compute the means for each of the eight (age) subpopulations in the small-mouth bass data. Draw a plot of mean Length versus Age. Is there evidence for a linear relationship?

Answer: Yes, there is a linear relationship between the mean length and age.The plot of mean Length versus Age is given below:



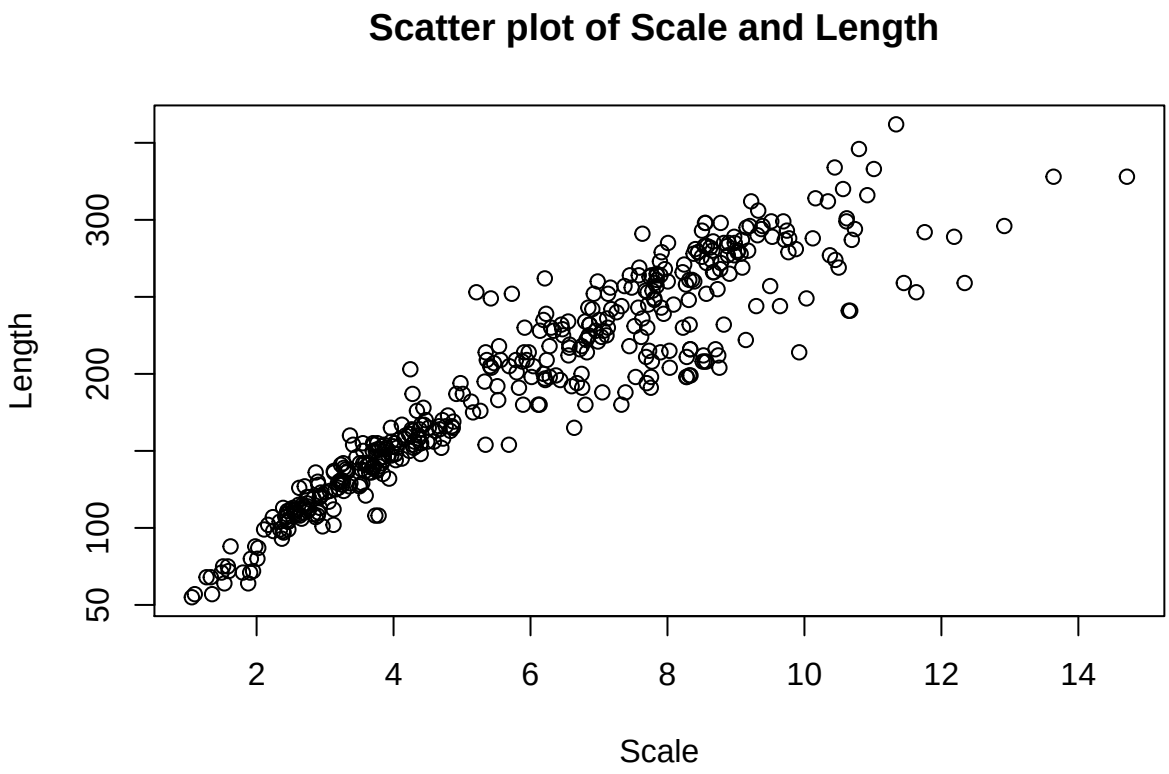
Question 1b: (1 point) Compute the standard deviations for each of the eight (age) subpopulations in the smallmouth bass data. Draw a plot of the standard deviations of Length in each subpopulation versus Age. Does the variance appear constant across the different age populations?

Answer: No the variance is not constant across the different age populations. The plot of the standard deviations of Length in each subpopulation versus Age is given below:



Question 1c: (2 points) Suppose that you want to estimate the relationship between the radius of a key scale (predictor) and the length of the fish (response). Make a scatterplot to investigate the possible linear relationship. Fit a simple linear regression Length ~ Scale and print the R summary. What do you observe?

Answer The scatter plot between a key scale (predictor) and the length of the fish (response) showed a heteroscedisticity. The scatter plot and summary is given below:



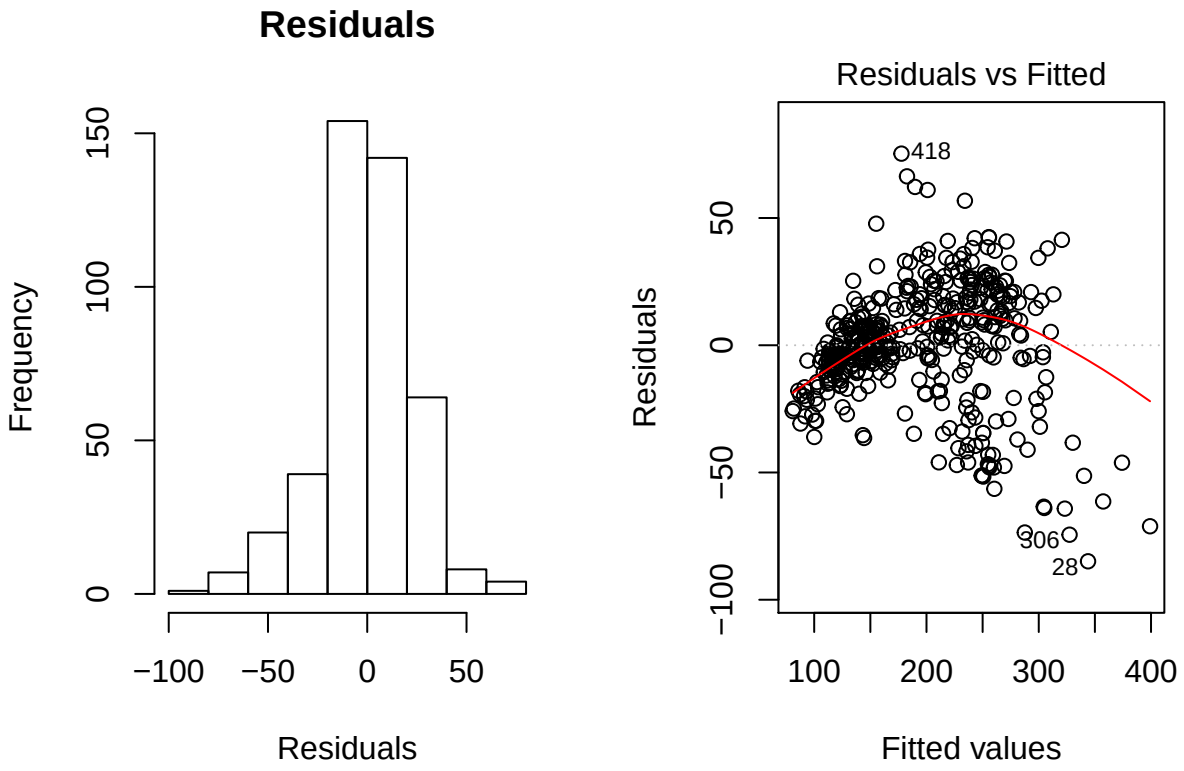

```
fit_wblake<-lm(wblake$Length ~ wblake$Scale)
summary(fit_wblake)
```

```
##
## Call:
## lm(formula = wblake$Length ~ wblake$Scale)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.896  -9.643  -0.021  14.651  75.290
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.2986     2.6423   21.31   <2e-16 ***
## wblake$Scale  23.3068     0.4096   56.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.06 on 437 degrees of freedom
## Multiple R-squared:  0.8811, Adjusted R-squared:  0.8808
## F-statistic:  3237 on 1 and 437 DF,  p-value: < 2.2e-16
```

The observation showed $\hat{\beta}_0$ is 56.29 with standard error of 2.662 and $\hat{\beta}_1$ is 23.306 with standard error of 0.409. This mean for every unit increase in scale there will be 23.306 unit increase in length. The p-value is very small, less than the significant level of 0.01 which mean we can reject the null hypothesis.

Question 1(d) (3 points) Plot the histogram of the residuals from the above regression and the TA plot (Tukey-Anscombe plot, TA plot involves plotting residuals (vertical axis) vs. fitted values (horizontal axis)). Do the normality and the constant variance assumptions appear to hold? Does this linear model seem appropriate? R hint: Use plot(lm(wblake*Length wblake*Scale), which = 1)

Answer The histogram of the residuals and TA plot is given below:



The histogram is almost normal with some slight skewed in left, this shows that it holds the normality. The TA plot showed that the variance is less for the value below 180 and higher for the higher values. This mean the variance is not constant. The linear model does not seem to be appropriate.

Question 1(e) (3 points) Find the fitted value, the 95% confidence interval, and the 95% prediction interval for the new data point Scale = 200.

Answer

```
##        1
## 4717.665
```

```
##        fit      lwr      upr
## 1 4717.665 4561.348 4873.983
```

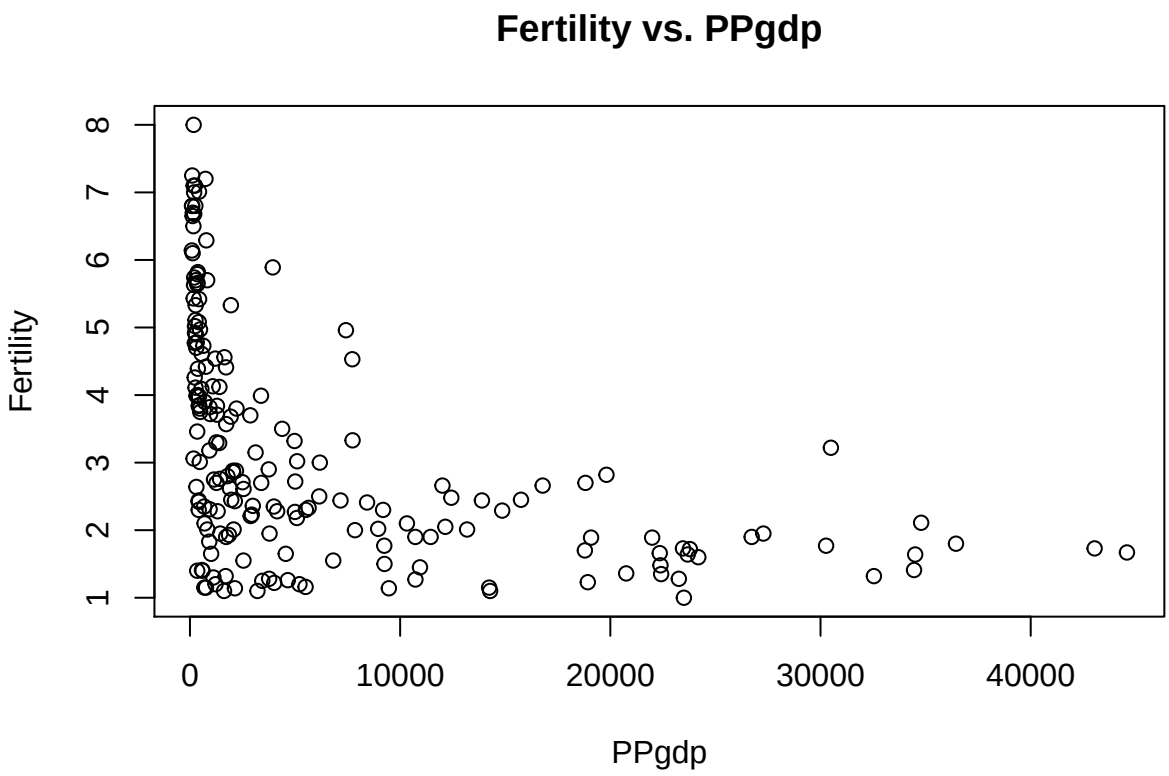
```
##        fit     lwr      upr
## 1 4717.665 4554.91 4880.421
```

Question 2(a) (1 point) Identify the predictor and the response.

Answer PPgdp is predictor and Fertility is response variable.

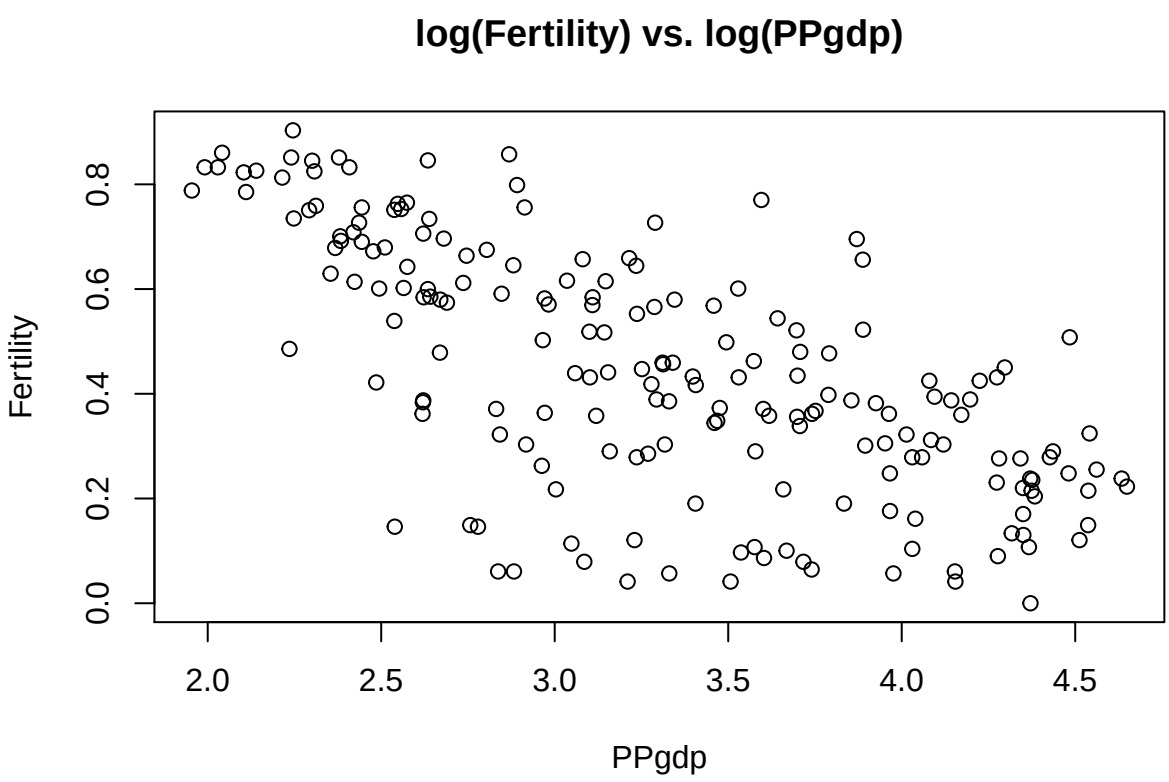
Question 2(b) (1 point) Draw the scatterplot of Fertility on the vertical axis versus PPgdp on the horizontal axis and summarize the information in this graph. Does linear model seem appropriate here?

Answer The scatter plot shows that the fertility is very high in the population with the lower per capita GDP and it drops with in increase in per capita GDP, creating a curve.The linear model doesnot seem to be appropriate. The scatter plot of Fertility on vertical axis verus PPgdp on the horizontal axis is given below:



Question 2(c) (1 point) Draw the scatterplot of log(Fertility) versus log(PPgdp), using the logarithm with base 10. Does the simple linear regression model seem plausible for a summary of this graph?

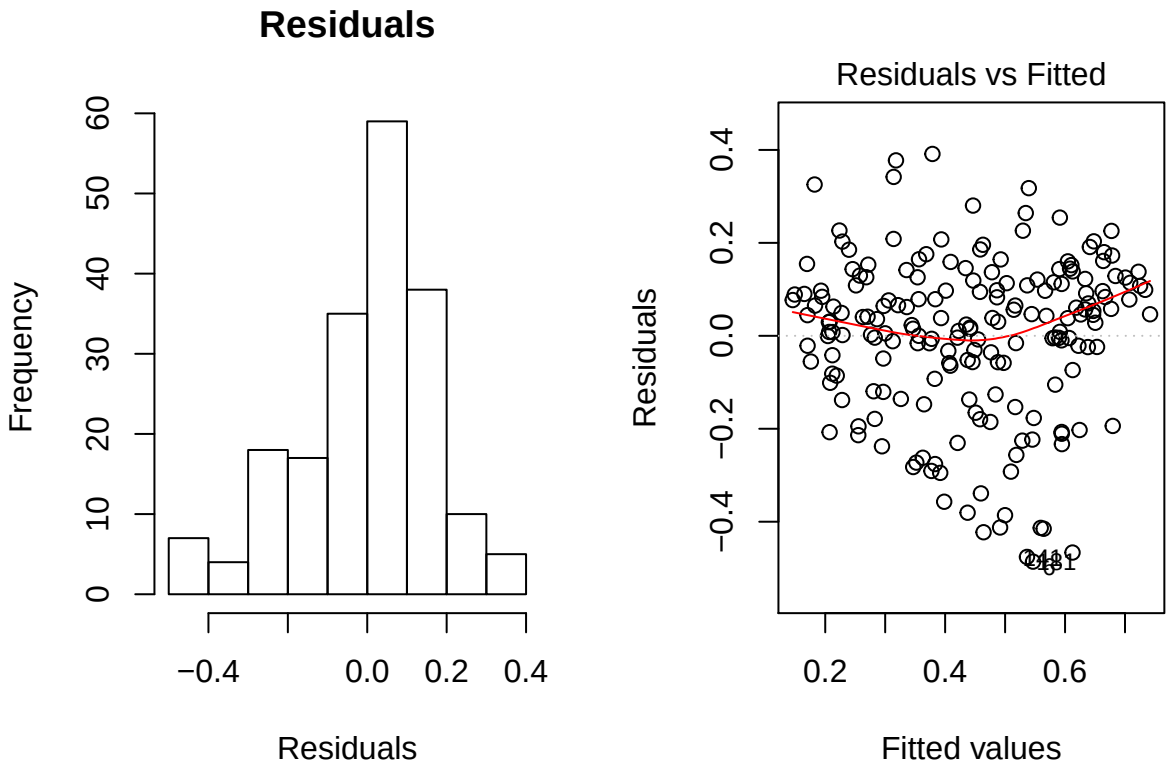
Answer The scatter plot shows that the fertility rate decreases with the increase in the per capita GDP increases. Yes, simple linear regression model seem plausible for a summary of this graph. The scatter plot of log(Fertility) versus log(PPgdp) is given below.



Question 2(d) (1 point) Fit a simple linear regression to the log transformed data from c and print the summary. Answer The fitted simple linear regression to the log transformed summary is given below:

```
##
## Call:
## lm(formula = log10(UN1$Fertility) ~ log10(UN1$PPgdp))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48587 -0.08148  0.03058  0.11327  0.39130
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.17399    0.05879   19.97   <2e-16 ***
## log10(UN1$PPgdp) -0.22116    0.01737  -12.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1721 on 191 degrees of freedom
## Multiple R-squared:  0.4591, Adjusted R-squared:  0.4563
## F-statistic: 162.1 on 1 and 191 DF,  p-value: < 2.2e-16
```

Question 2(e) (3 points) Look at the histogram and TA-plot of the residuals. What can you say about the assumptions on the errors? Answer The histogram and TA-plot of residuals is given below:



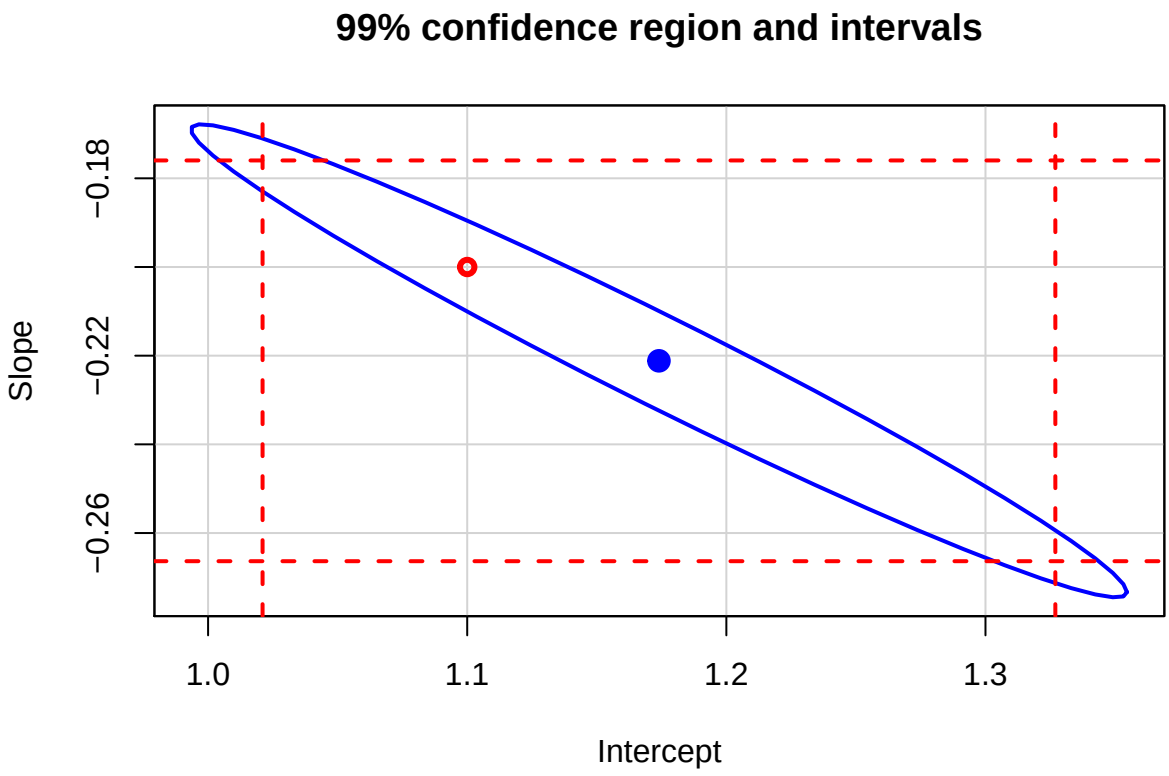
The histogram is normally distributed which shows that it holds the normality. The TA plot shows that the variance is slightly not constant because there is higher variance in the middle right.

Question 2(f) (2 points) Test the null hypothesis that the slope is zero versus the two-sided alternative at the 1% level. Give the t-value and a sentence to summarize the result.

Answer The t-value for slope log10(PPgdp) is -12.73 and the p-value is very small less than the significant level of 0.01 so we reject the null hypothesis.

Question 2(g) (3 points) Plot the marginal 99% confidence intervals for the intercept (Beta0) and slope (Beta1) in the model from c as well as the 99% confidence ellipse for vector (Beta0; Beta1)T . Would you reject the hypothesis (Beta0, Beta1)T = (1.1, -.2) at the 1% level?

Answer



Since the point $(\beta_0, \beta_1)^T = (1.1,-0.2)$ falls inside the confidence region and interval we cannot reject the null hypothesis.

Question 3a (1 point) Find d_i such that: $\hat{\beta}_0 = \sum_{i=1}^{n} d_i y_i$

Answer Here is the solution:

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Substituting $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$ in the above equation

$$= \bar{y} - \sum_{i=1}^{n} c_i y_i \bar{x}$$

$$= \sum_{i=1}^{n} \frac{y_i}{n} - \sum_{i=1}^{n} c_i y_i \bar{x}$$

$$= \sum_{i=1}^{n} y_i \left(\frac{1}{n} - c_i \bar{x} \right)$$

Therefore: $d_i = \left(\frac{1}{n} - c_i \bar{x}\right)$

Question 3b (1 point) Show that $\mathrm{E}\left[\hat{\beta}_0 | X = x\right] = \beta_0$

Answer Here is the solution:

$\mathrm{E}\left[\hat{\beta}_0 | X = x\right]$

Substituting $\hat{\beta}_0 = \sum_{i=1}^{n} d_i y_i$ in the above equation

$= E\left[\sum_{i=1}^{n} diyi | X = x\right]$

$= \sum_{i=1}^{n} di * E[yi|X = x]$

$= \sum_{i=1}^{n} di\left(\beta_0 + \beta_1 x_i\right)$

$= \beta_0 \sum_{i=1}^{n} di + \beta_1 \sum_{i=1}^{n} dix_i$

Now, solving for $= \sum_{i=1}^{n} di = \sum_{i=1}^{n} \left(\frac{1}{n} - c_i\bar{x}\right) = \sum_{i=1}^{n} \frac{1}{n} - \sum_{i=1}^{2} c_i\bar{x} = 1 - \bar{x}\sum_{i=1}^{2} c_i = 1 - \bar{x}\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 1$

Now, solving for $= \sum_{i=1}^{n} d_i x_i = \sum_{i=1}^{n}\left(\frac{1}{n} - c_i\bar{x}\right)x_i = \sum_{i=1}^{n}\left(\frac{x_i}{n}\right) - \sum_{i=1}^{n} c_i x_i\bar{x} = \bar{x} - \bar{x}\sum_{i=1}^{n} c_i x_i = \bar{x} - \bar{x}\sum_{i=1}^{n}\frac{(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \bar{x} - \bar{x}\sum_{i=1}^{n}\frac{(x_i-\bar{x})x_i-\bar{x}+\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \bar{x} - \bar{x}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \bar{x}\sum_{i=1}^{n}\frac{(x_i-\bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \bar{x} - (\bar{x} \times 1) + 0 = 0$

Substituting these values in the equation: $= \beta_0 \sum_{i=1}^{n} di + \beta_1 \sum_{i=1}^{n} dix_i = \beta_0 \times 1 + \beta_1 \times 0 = \beta_0$

Question 3c (2 points) Show that: $\text{Var}\left[\hat{\beta}_0 | X = x\right] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)$

Answer Solution:

$= \text{Var}\left[\hat{\beta}_0 | X = x\right]$

$= \text{Var}\left[\sum_{i=1}^{n} d_i y_i | X = x\right]$

$= \sum_{i=1}^{n} d_i^2 \times \text{Var}\left[y_i | X = x\right]$

Because $\text{Var}[cX] = c^2\,\text{Var}[X]$

$= \sum_{i=1}^{n} d_i^2 \times \sigma^2 = \sigma^2 \sum_{i=1}^{n} d_i^2 = \sigma^2 \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{SXX}\right)^2$

$= \sigma^2 \sum_{i=1}^{n}\left(\frac{1}{n^2} - 2\left(\frac{(x_i-\bar{x})\bar{x}}{SXX}\right)\frac{1}{n} + \left(\frac{(x_i-\bar{x})\bar{x}}{SXX}\right)^2\right)$

$= \sigma^2\left(\left(n \times \frac{1}{n^2}\right) + 0 + \sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})\bar{x}}{SXX}\right)^2\right)$

$= \sigma^2\left(\frac{1}{n} + \sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})\bar{x}}{SXX}\right)^2\right)$

$= \sigma^2\left(\frac{1}{n} + \bar{x}^2\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}\right)$

$= \sigma^2\left(\frac{1}{n} + \bar{x}^2\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$

$= \sigma^2\left(\frac{1}{n} + \bar{x}^2\frac{1}{SXX}\right)$

Question 3d (1 point) Show that: $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = 0$

Answer Solution:

$= \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)$

Substituting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$= \sum_{i=1}^{n}\left(y_i - \hat{B}_0 - \hat{B}_1 x_i\right) = \sum_{i=1}^{n}\hat{\epsilon}_i = 0$

Question 4a (1 point) Show that the least squares estimate of $\hat{\beta}_1$ is: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

Answer Solution: Using RSS equation $RSS = \sum_{i=1}^{n}\left(y_i - \hat{B}_1 x_i\right)^2$

$= \sum_{i=1}^{n} \left(y_i^2 - 2\hat{B}_1 x_i y_i + \left(\hat{B}_1 x_i\right)^2 \right)$

$\frac{\partial RSS}{\partial B_1} = \sum_{i=1}^{n} \left(-2x_i y_i + 2x_i^2 \hat{B}_1 \right) = 0$

$\rightarrow 2\hat{B}_1 \sum_{i=1}^{n} x_i^2 = 2\sum_{i=1}^{n} x_i y_i$

$\hat{B}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

Question 4b (1 point) Show that: $\mathrm{E}\left[\hat{\beta}_1 | X = x\right] = \beta_1$

Answer Solution:

$E\left[\hat{B}_1 | X = x\right] = E\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} | X = x\right]$

$= \frac{1}{\sum_{i=1}^{n} x_i^2} \times E\left[\sum_{i=1}^{n} x_i y_i | X = x\right]$

$= \frac{1}{\sum_{i=1}^{n} x_i^2} \times \sum_{i=1}^{n} E\left[x_i y_i | X = x\right]$

$= \frac{1}{\sum_{i=1}^{n} x_i^2} \times \sum_{i=1}^{n} x_i E\left[y_i | X = x\right]$

$= \frac{1}{\sum_{i=1}^{n} x_i^2} \times \sum_{i=1}^{n} x_i \left(B_1 x_i\right)$

$= B_1 \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = B_1$

Question 4c (2 point) Show that: $\mathrm{Var}\left[\hat{\beta}_1 | X = x\right] = \frac{\sigma^2}{\sum x_i^2}$

Answer Solution: $\mathrm{Var}\left[\hat{\beta}_1 | X = x\right]$

$= \mathrm{Var}\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} | X = x\right] = \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \times \mathrm{Var}\left[\sum_{i=1}^{n} x_i y_i | X = x\right]$

Because $\mathrm{Var}[cX] = c^2 \mathrm{Var}[X]$

$= \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \times \mathrm{Var}\left[y_i | X = x\right] = \frac{1}{\sum_{i=1}^{n} x_i^2} \times \sigma^2$

$= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$