STAT 423 Homework 1

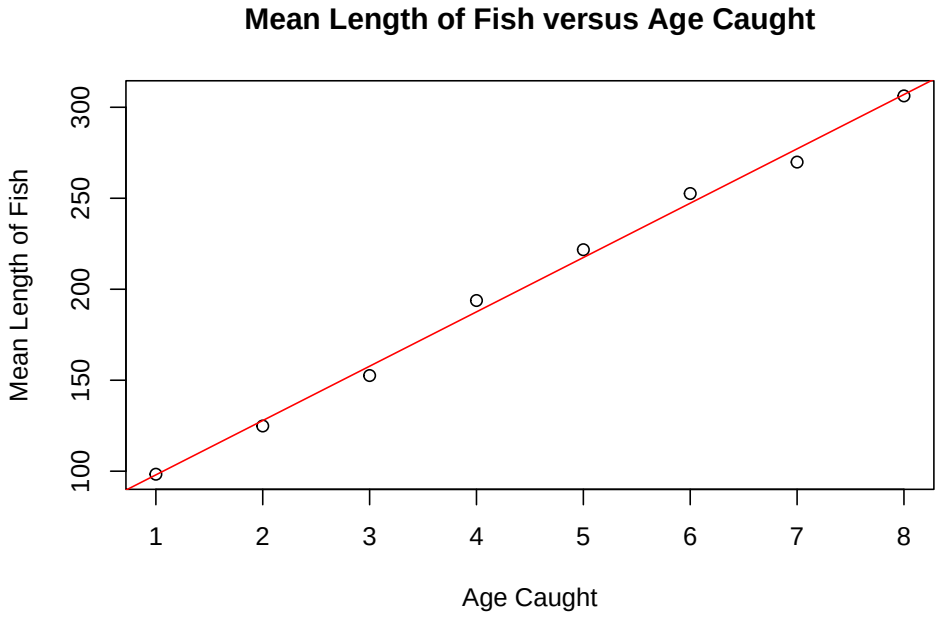
1. Use data `wblake` in R package `alr4`. You can obtain a description of the data set by loading package `alr4` and executing `?wblake`.

a. Compute the means for each of the eight (age) subpopulations in the smallmouth bass data. Draw a plot of mean `Length` versus `Age`. Is there evidence for a linear relationship?

We will start by using the `aggregate` function in base `R` to calculate the mean of the `Length` variable for each of the eight `Age` subpopulations. The resulting data frame is printed below.

```
##   Age Mean Length
## 1   1    98.34211
## 2   2   124.84722
## 3   3   152.56383
## 4   4   193.80000
## 5   5   221.72059
## 6   6   252.59770
## 7   7   269.86885
## 8   8   306.25000
```

We will now draw a plot of `Mean Length` versus `Age`, and determine if there is evidence for a linear relationship between the two variables. The resulting plot is printed below.



As seen in the above scatterplot of `Mean Length` versus `Age`, with the least squares regression line interpolated on the plot, there appears to be a strong positive linear relationship between the mean length of a fish when

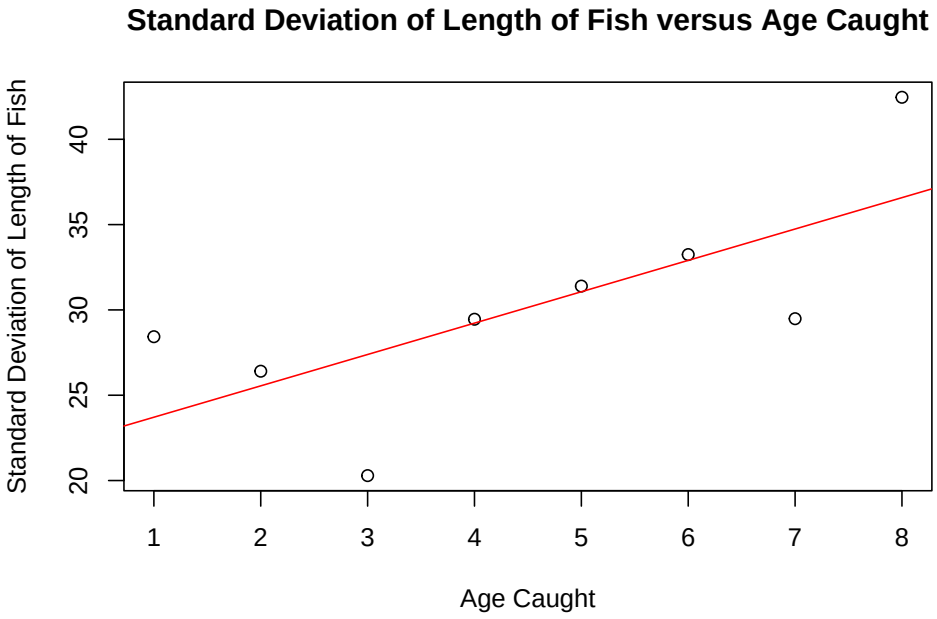
it was first caught, and the age of a fish when it was first caught. It is important to note that there is a small sample size due to the fact that there is only one value per `Age` category, so it is hard to analyze the true relationship between the two variables.

b. 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?

We will start by using the `aggregate` function in base `R` to calculate the standard deviation of the `Length` variable for each of the eight `Age` subpopulations. The resulting data frame is printed below.

```
##   Age SD of Length
## 1   1     28.42941
## 2   2     26.40618
## 3   3     20.28960
## 4   4     29.45263
## 5   5     31.39581
## 6   6     33.24275
## 7   7     29.48529
## 8   8     42.46077
```

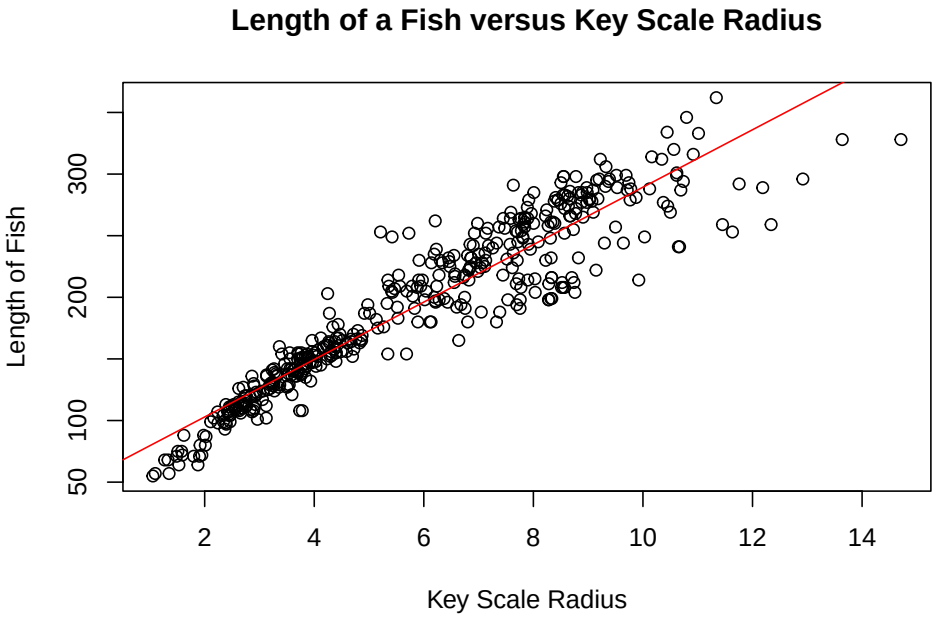
We will now draw a plot of `SD of Length` versus `Age`, and determine if there is evidence for a linear relationship between the two variables. The resulting plot is printed below.



As seen in the above scatterplot of `SD of Length` versus `Age`, with the least squares regression line interpolated on the plot, there appears to be no linear relationship between the two variables, or at the very most, a weak positive linear relationship between the standard deviation of the length of a fish when it was first caught, and the age of a fish when it was first caught. It is important to note that there is a small sample size due to the fact that there is only one value per `Age` category, so it is hard to analyze the true relationship between the two variables. Due to the violation of the linearity assumption, it appears as if the variance is not constant across the different age populations.

c. 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 $Length \sim Scale$ and print the `R` summary. What do you observe?

We will start by making a scatterplot of `Length` versus `Scale` (with the least squares line interpolated) to investigate the possible linear relationship. The resulting plot is printed below.



As seen in the above scatterplot of `Length` versus `Scale`, there seems to be a moderate to strong linear relationship between the length of a fish when it was first caught, and the key scale radius of a fish when it was first caught. However, as the key scale radius increases, it appears as if the length of the fish becomes more variable. Thus, we have evidence of heteroscedasticity in the data.

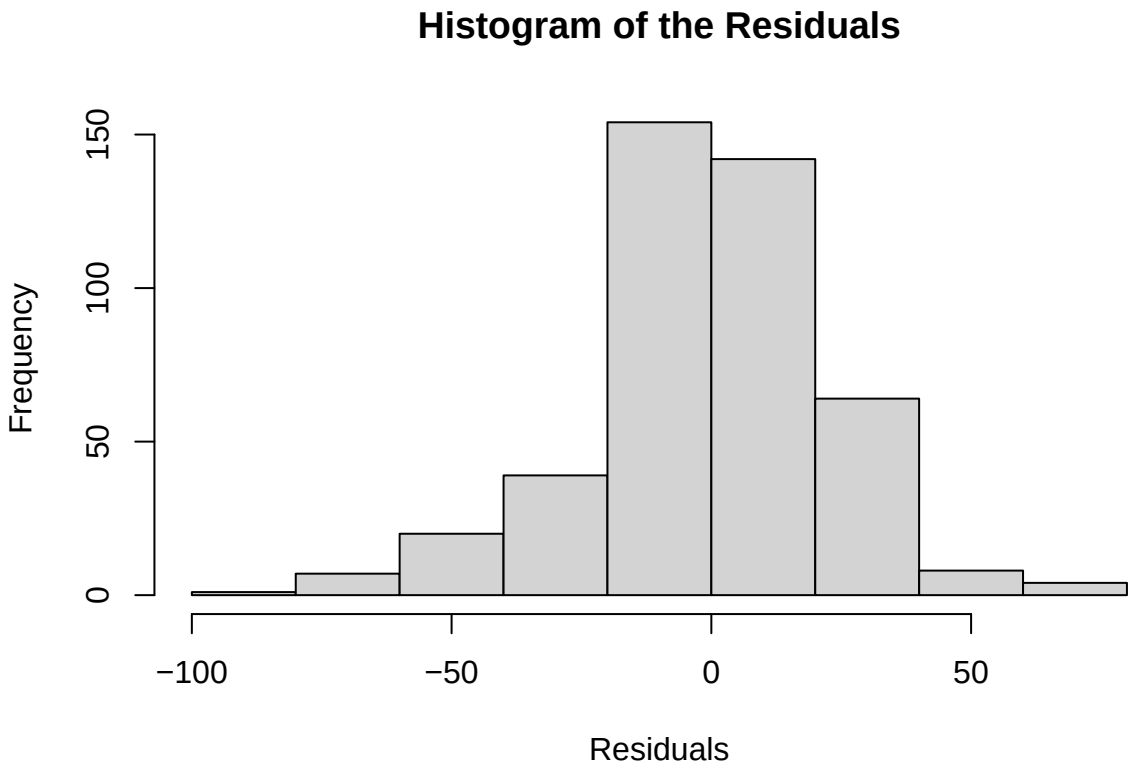
We will now fit a simple linear regression $Length \sim Scale$ and print the corresponding `R` summary. This output is shown below.

```
##
## Call:
## lm(formula = Length ~ Scale, data = wblake)
##
## 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 ***
## 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
```

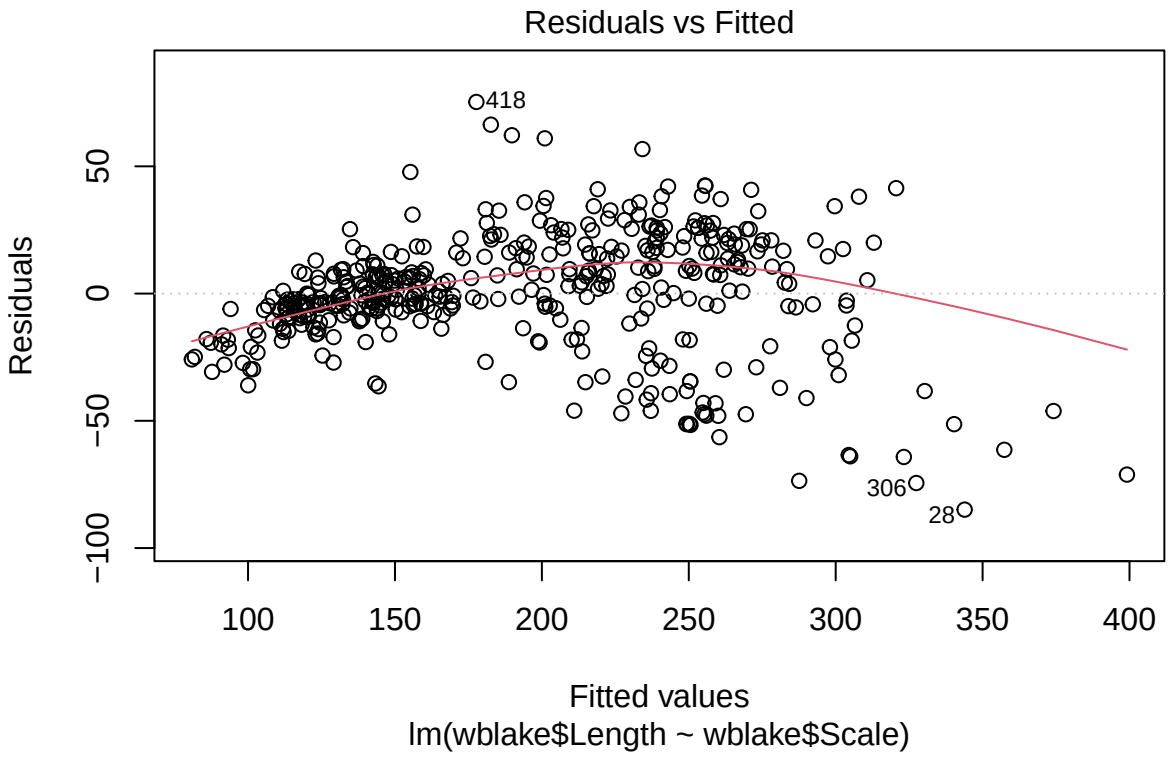
As can be seen in the above `R` output from fitting a simple linear regression model `Length` on `Scale`, we see that the values of β_0 and β_1 are significant, as well as the F-statistic. This means that the model as a whole is significant, as well as telling us that the key scale radius of a fish when it is first caught is a significant predictor of the length of a fish when it is first caught. Furthermore, the high R^2 value tells us that approximately 88.11% of the variation in the length of a fish when it is first caught can be explained by the key scale radius of a fish when it is first caught. This also implies that there is a high correlation, which backs up our claim of a strong linear relationship.

d. 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)`.

We will start by plotting the histogram of the residuals from the above regression and the Tukey-Anscombe plot. The resulting plots are printed below.



As seen in the above histogram, the residuals are roughly symmetrically distributed about a value close to zero. However, there seems to be small left skew. Altogether, one could assume that the residuals are normally distributed, although this may need to be further analyzed before a decision is made.



Moving on to the Tukey-Anscombe plot, there seem to be some issues. First off, due to the curvilinear pattern of the loess curve, we seem to have even more evidence of non-constant variance. Also, due to the fact that the residuals don't seem to "bounce randomly" around the x-axis, there is evidence that the relationship between these two variables is non-linear. Overall, it seems as if the simple linear regression model isn't the most appropriate fit for this data, however, more plots and better analysis techniques would need to be used in order to completely back up this statement.

e. Find the fitted value, the 95% confidence interval, and the 95% prediction interval for the new data point `Scale = 200`.

We will start off by finding the fitted value for the new data point `Scale = 200`. We will do this by using the `predict` function in base `R`. This fitted value is shown below.

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

As computed in `R`, and shown in the output above, the fitted length value for the new data point `Scale = 200` is 4717.6654068.

We will now find the 95% confidence interval for the new data point `Scale = 200`. We will do this by using the `predict` function in base `R`. This confidence interval is shown below.

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

As computed in `R`, and shown in the output above, the 95% confidence interval for the new data point `Scale = 200` is $[4561.348, 4873.983]$.

We will now find the 95% prediction interval for the new data point `Scale = 200`. We will do this by using the `predict` function in base `R`. This prediction interval is shown below.

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

As computed in `R`, and shown in the output above, the 95% prediction interval for the new data point `Scale = 200` is $[4554.91, 4880.421]$. It is wider, as expected.

2. The data in the file UN1 in R package `alr4` contains the data on:

- `PPgdp` - the 2001 gross national product per person in US dollars,
- `Fertility` - the birth rate per 1000 women in the population in the year 2000,
- `locality` - Place where the data was collected. (This variable is unlabeled in the data, but is listed as the row name.)

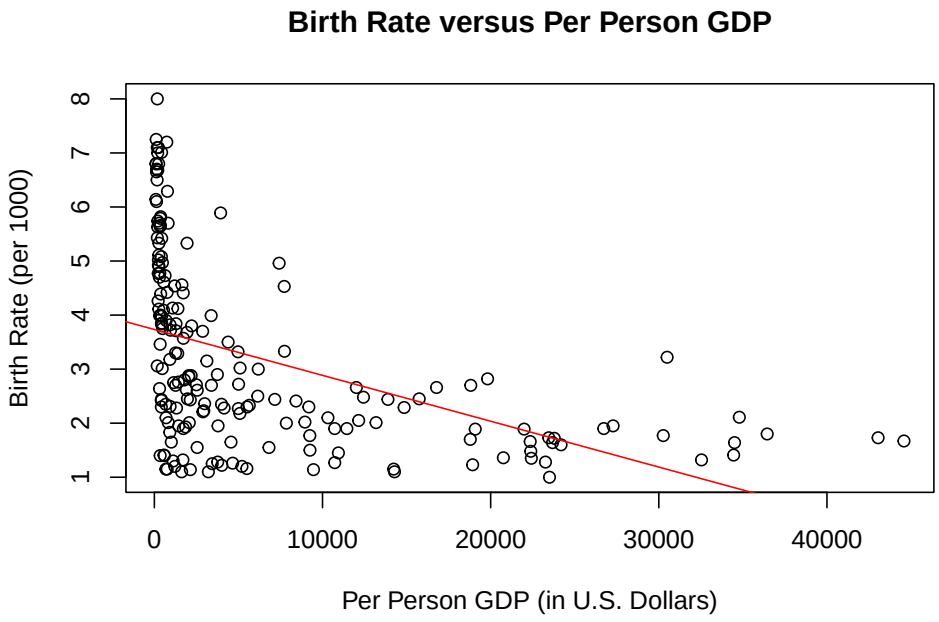
The data are collected from 193 localities, mostly UN member countries. In this problem, we will study the conditional distribution of `Fertility` given `PPgdp`.

a. Identify the predictor and the response.

As denoted in the problem statement, we want to study the conditional distribution of `Fertility` given `PPgdp`, therefore, it follows that the predictor variable is `PPgdp`, and the response variable in `Fertility`.

b. Draw the scatterplot of `Fertility` on the vertical axis versus `PPgdp` on the horizontal axis and summarize the information in this graph. Does a linear model seem appropriate here?

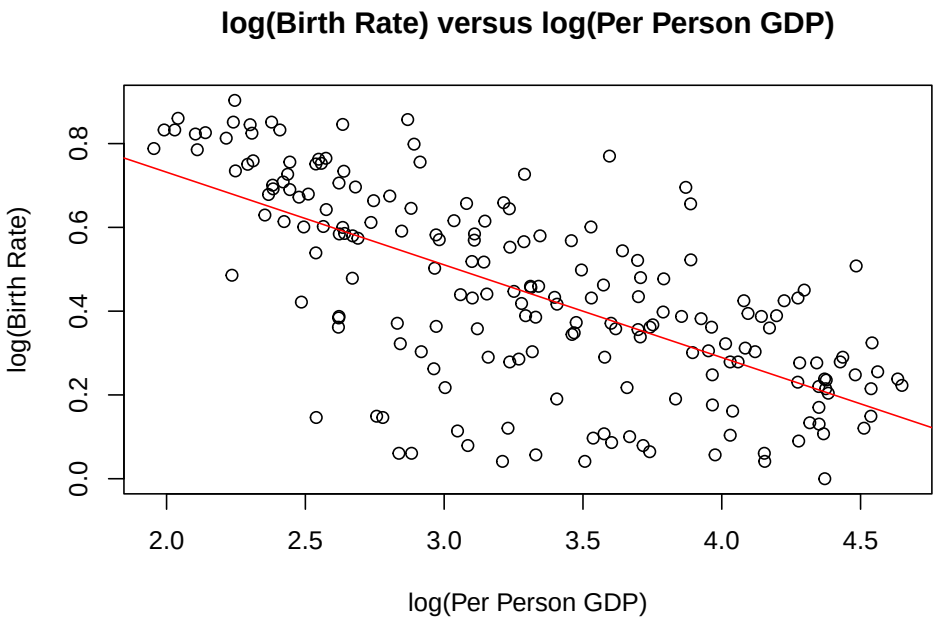
We will start off this problem by drawing a scatterplot of `Fertility` versus `PPgdp`, and determine if there is evidence for a linear relationship between the two variables. The resulting plot is printed below.



As can be seen in the above scatterplot of `Fertility` versus `PPgdp`, with the least squares regression line interpolated on the plot, there appears to be no linear relationship between birth rate and per person GDP. In particular, for low values of per person GDP, there is a lot of spread in the values that birth rate takes on. Furthermore, once per person GDP increases a little bit, the variance seems to slowly start decreasing. With all of that being said, fitting a linear model in this case is not appropriate.

c. 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?

We will start off this problem by drawing a scatterplot of `log(Fertility)` versus `log(PPgdp)`, and determine if there is evidence for a linear relationship between the two variables. The resulting plot is printed below.



As seen in the above scatterplot of `log(Fertility)` versus `log(PPgdp)`, with the least squares regression line interpolated on the plot, there appears to be a moderate negative linear relationship between the log birth rate and the log per person GDP. Furthermore, there seems to be somewhat constant variability, with less variability at the ends. Although not a perfect fit by any means, it could be justified that a linear regression model is appropriate in this case.

d. Fit a simple linear regression to the log transformed data from c and print the summary.

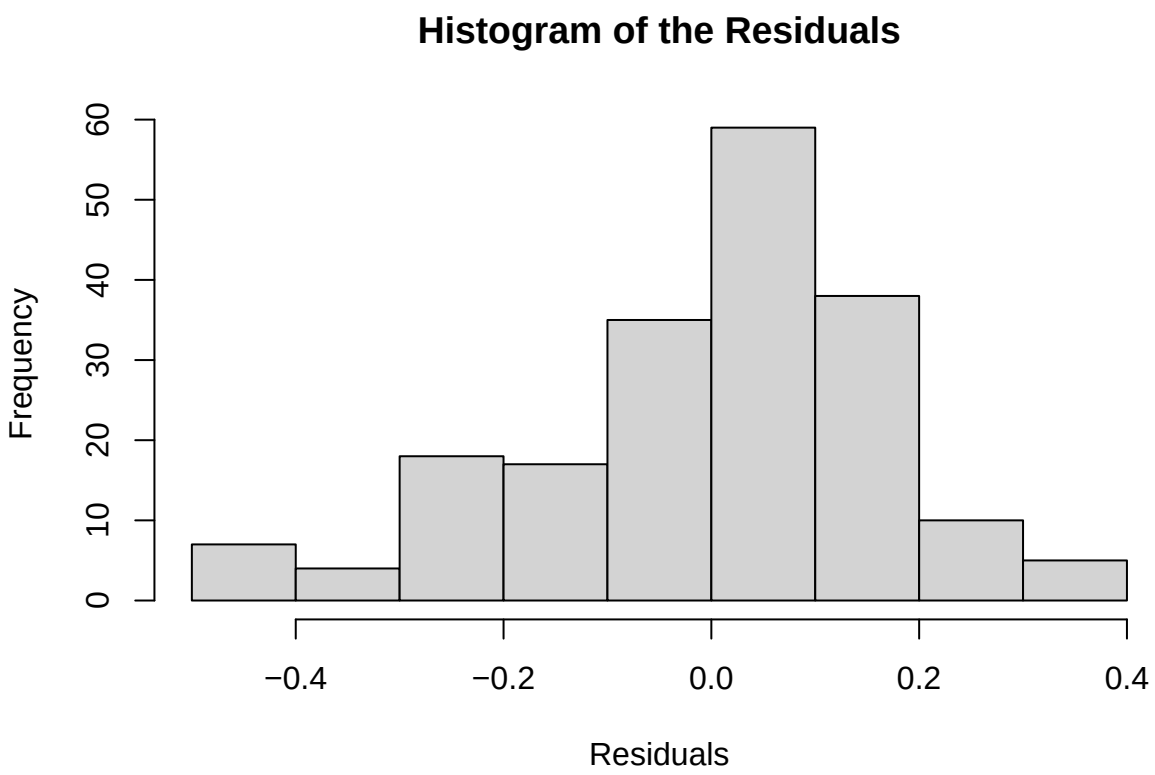
We will start this problem by fitting a simple linear regression to the data from part c and print the corresponding `R` summary. This output is shown below.

```
##
## Call:
## lm(formula = log_Fertility ~ log_PPgdp, data = log_data)
##
## 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 ***
## log_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
```

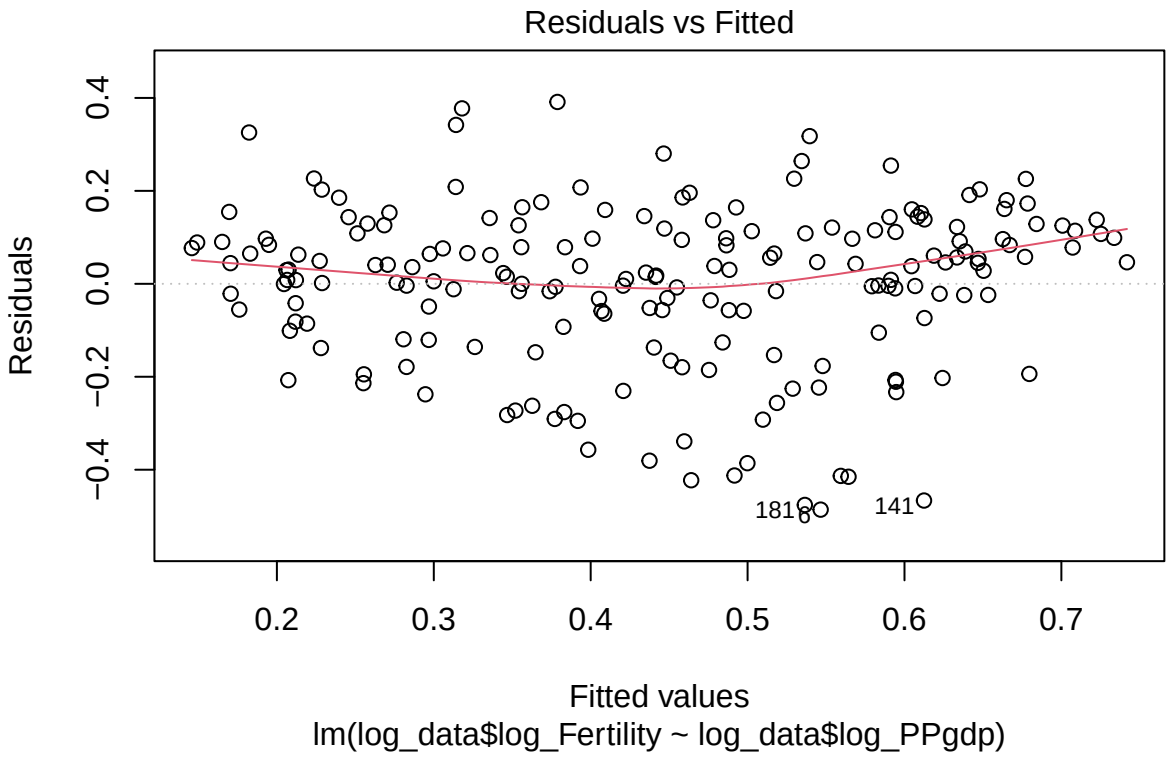
As can be seen in the above `R` output from fitting a simple linear regression model on the data from part c, we see that the values of β_0 and β_1 are significant, as well as the F-statistic. This means that the model as a whole is significant, as well as telling us that the log of per person GDP is a significant predictor of the log of birth rate. Furthermore, the moderate R^2 value tells us that approximately 45.91% of the variation in the log birth rate can be explained by the log of per person GDP. This also implies that there is a moderate correlation, which backs up our claim of a moderate linear relationship.

e. Look at the histogram and TA plot of the residuals. What can you say about the assumptions on the errors?

We will start by plotting the histogram of the residuals from the above regression and the Tukey-Anscombe plot. The resulting plots are printed below.



As seen in the above histogram, the residuals are roughly symmetrically distributed about a value close to zero. However, there seems to be noticeable left skew. Unlike the last histogram of the residuals, this one has a more noticeable left skew, however, the rest of the histogram (apart from the skew) seems approximately normal. Altogether, one could probably assume that the residuals are normally distributed, although this may need to be further analyzed before a decision is made.



Moving on to the Tukey-Anscombe plot, due to the minimal curvilinear pattern in the residuals, the constant variance assumption seems to be upheld (or at the very most slightly violate, as mentioned in the description of the histogram). Furthermore, due to the fact that the residuals appear evenly on both sides of the loess curve, as well as appearing to "bounce randomly," we have evidence that the linearity assumption is also upheld. Overall, it seems as if the simple linear regression model provides a decent fit to that data, however, more plots and better analysis techniques would need to be used in order to completely back up this statement. It is important to note that even though the assumptions weren't perfectly upheld, these assumptions weren't severely violated, thus fitting a simple linear regression model as a "starting" model might not be a bad idea.

f. 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 results.

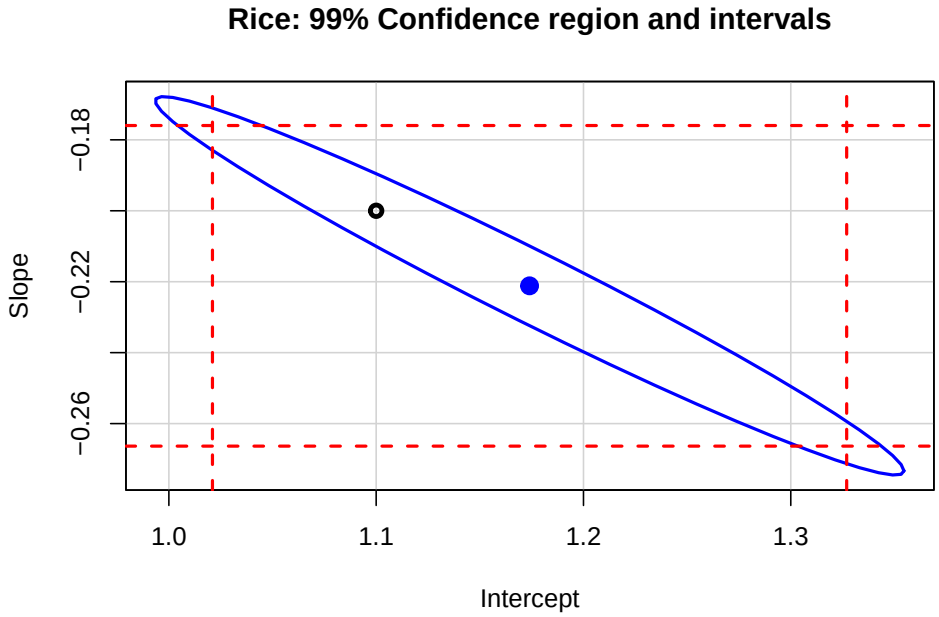
In this subpart, we will test $H_0 : \hat{\beta}_1 = 0$ versus $H_1 : \hat{\beta}_1 \neq 0$. In particular, we will give the t-value and a sentence to summarize the results. These calculations are done below.

```
##     beta_1   t_value      p_value
## 1 -0.22116 -12.73366 2.731002e-27
```

As calculated in `R` and shown in the above output, the estimated slope value using least squares regression is $\hat{\beta}_1 = -0.22116$. The corresponding t-statistic value was calculated as $T = -12.73366$, which led to a p-value of $2.7310018 \times 10^{-27}$. Hence at the significance level of $\alpha = 0.01$ we reject the null hypothesis in favor of the alternative and conclude that `log(PPgdp)` is a significant predictor of `log(Fertility)`.

g. Plot the marginal 99% confidence intervals for the intercept (β_0) and slope (β_1) in the model from c as well as the 99% confidence ellipse for the vector $(\beta_0, \beta_1)^T$. Would you reject the hypothesis $(\beta_0, \beta_1)^T = (1.1, -.2)$ at the 1% level?

In this subpart, we will plot the marginal 99% confidence intervals for the intercept and slope in the model from c as well as the 99% confidence ellipse for the vector $(\beta_0, \beta_1)^T$. We will also see if we can reject the hypothesis $(\beta_0, \beta_1)^T = (1.1, -.2)$ at the 1% level. These intervals and ellipse are shown below.



As can be seen above, since the estimates of the parameters fall within the bounds of both confidence intervals, at the 1% we fail to reject the null hypothesis that $(\beta_0, \beta_1)^T = (1.1, -.2)$. Furthermore, since $(\beta_0, \beta_1)^T = (1.1, -.2)$ falls within the estimated confidence ellipse, we fail to reject the to reject the null hypothesis that $(\beta_0, \beta_1)^T = (1.1, -.2)$.

3. Suppose that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{and} \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

and let $\hat{\beta}_0$ and $\hat{\beta}_1$ be OLS estimates of β_0 and β_1 based on n samples.

a. Find d_i such that

$$\hat{\beta}_0 = \sum_{i=1}^{n} d_i y_i.$$

In this problem, our goal is to find d_i such that $\hat{\beta}_0 = \sum_{i=1}^{n} d_i y_i$. In this calculation we will use the following facts: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$, and $c_i = \frac{x_i - \bar{x}}{SXX}$. This calculation is done below.

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&= \bar{y} - \sum_{i=1}^{n} c_i y_i \bar{x} \\
&= \sum_{i=1}^{n} \frac{1}{n} y_i - \sum_{i=1}^{n} \frac{\bar{x}(x_i - \bar{x})}{SXX} y_i \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX} \right) y_i
\end{aligned}$$

Hence we have shown that we can write $\hat{\beta}_0$ as $\sum_{i=1}^{n} d_i y_i$, with $d_i = \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX}$. Furthermore, the d_i can also be written as $d_i = \frac{1}{n} - c_i \bar{x}$.

b. Show that

$$E[\hat{\beta}_0 | X = x] = \beta_0$$

Using our result from part a, we will show that $\hat{\beta}_0$ is an unbiased estimator of β_0. That is, we will show that $E[\hat{\beta}_0 | X = x] = \beta_0$. This calculation is done below.

$$\begin{aligned}
E\left[\hat{\beta}_0 | X = x\right] &= E\left[\sum_{i=1}^{n} d_i y_i | X = x\right] \\
&= E\left[\sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX} \right) y_i | X = x\right] \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX} \right) E\left[y_i | X = x\right] \quad \text{(Linearity of expectation)} \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX} \right) (\beta_0 + \beta_1 x_i) \quad \text{(Since } E\left[y_i | X = x\right] = \beta_0 + \beta_1 x_i) \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX} \right) \beta_0 + \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX} \right) \beta_1 x_i \\
&= \sum_{i=1}^{n} \left(\frac{\beta_0}{n} - \frac{\beta_0 \bar{x}(x_i - \bar{x})}{SXX} \right) + \sum_{i=1}^{n} \left(\frac{\beta_1 x_i}{n} - \frac{\beta_1 x_i \bar{x}(x_i - \bar{x})}{SXX} \right) \\
&= \sum_{i=1}^{n} \frac{\beta_0}{n} - \sum_{i=1}^{n} \frac{\beta_0 \bar{x}(x_i - \bar{x})}{SXX} + \sum_{i=1}^{n} \frac{\beta_1 x_i}{n} - \sum_{i=1}^{n} \frac{\beta_1 x_i \bar{x}(x_i - \bar{x})}{SXX} \\
&= \frac{\beta_0}{n} \sum_{i=1}^{n} - \frac{\beta_0 \bar{x}}{SXX} \sum_{i=1}^{n} (x_i - \bar{x}) + \frac{\beta_1}{n} \sum_{i=1}^{n} x_i - \beta_1 \bar{x} \sum_{i=1}^{n} \frac{x_i (x_i - \bar{x})}{SXX}
\end{aligned}$$

Since $\sum_{i=1}^n (x_i - \bar{x}) = 0$, $\sum_{i=1}^n x_i = n\bar{x}$, $\sum_{i=1}^n 1 = n$, and $SXX = \sum_{i=1}^n (x_i - \bar{x})^2$, we can simplify the above expression and continue the calculation. This is done below.

$$\begin{aligned}
&= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x}\sum_{i=1}^n \frac{x_i(x_i - \bar{x}) - \bar{x}(x_i - \bar{x}) + \bar{x}(x_i - \bar{x})}{SXX} \\
&= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x}\sum_{i=1}^n \frac{(x_i - \bar{x})^2 + \bar{x}(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x}\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1\bar{x}\frac{\bar{x}\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} \\
&= \beta_0
\end{aligned}$$

Hence, we have shown that $\hat{\beta}_0$ is an unbiased estimator of β_0. That is, we have shown that $E[\hat{\beta}_0 | X = x] = \beta_0$.

c. Show that

$$Var[\hat{\beta}_0 | X = x] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right)$$

Using our result from part a, we will show that $Var[\hat{\beta}_0 | X = x] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right)$. The main fact that will be used to solve this problem is $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \implies y_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. More importantly, this means that the y_i's are independent. This calculation is done below.

$$\begin{aligned}
Var\left[\hat{\beta}_0 | X = x\right] &= Var\left[\sum_{i=1}^n d_i y_i | X = x\right] \\
&= Var\left[\sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX}\right) y_i | X = x\right] \\
&= \sum_{i=1}^n Var\left[\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX}\right) y_i | X = x\right] \quad \text{(Since the } y_i\text{'s are independent)} \\
&= \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX}\right)^2 Var\left[y_i | X = x\right] \quad \text{(Non-linearity of variance)} \\
&= \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SXX}\right)^2 \sigma^2 \quad \text{(Since } Var[y_i | X = x] = \sigma^2\text{)} \\
&= \sigma^2 \sum_{i=1}^n \left(\frac{1}{n^2} + \frac{\bar{x}^2(x_i - \bar{x})^2}{SXX^2} - \frac{\bar{x}(x_i - \bar{x})}{nSXX}\right) \\
&= \sigma^2 \left(\frac{1}{n^2}\sum_{i=1}^n + \frac{\bar{x}^2\sum_{i=1}^n (x_i - \bar{x})^2}{SXX^2} - \frac{\bar{x}\sum_{i=1}^n (x_i - \bar{x})}{nSXX}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2 SXX}{SXX^2}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)
\end{aligned}$$

Hence we have shown that $Var[\hat{\beta}_0 | X = x] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right)$.

d. Show that

$$\sum_{i=1}^n (y_i - \hat{y}_i) = 0$$

In this problem, we will show that the sum of the residuals is zero. That is, we will show that $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$. In particular, we will use the fact that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, and $\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x}$. This calculation is done below.

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i) &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i \\
&= n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x} \\
&= n\bar{y} - n(\bar{y} - \hat{\beta}_1 \bar{x}) - n\hat{\beta}_1 \bar{x} \\
&= n\bar{y} - n\bar{y} + n\hat{\beta}_1 \bar{x} - n\hat{\beta}_1 \bar{x} \\
&= 0
\end{aligned}$$

Hence we have shown that the sum of the residuals is zero. That is, we have shown that $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$.

4. Occasionally, a mean function in which the intercept is known apriori to be zero may be fit. This mean function is given by

$$E[Y|X = x] = \beta_1 x,$$

The residual sum of squares for this model, assuming the errors are independent with common variance σ^2, is

$$RSS = \sum_{i=1}^{n}(y_i - \beta_1 x_i)^2.$$

a. Show that the least squares estimate of β_1 is

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

In this problem, we will show that the least squares estimate of β_1 is $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$. We will start the process of finding the least squares estimate of β_1, by differentiating RSS with respect to β_1. This is done below.

$$\begin{aligned}
\frac{d}{d\beta_1} RSS &= \frac{d}{d\beta_1} \sum_{i=1}^{n}(y_i - \beta_1 x_i)^2 \\
&= 2\sum_{i=1}^{n}(y_i - \beta_1 x_i) \cdot (-x_i) \\
&= -2\sum_{i=1}^{n}(x_i y_i - \beta_1 x_i^2)
\end{aligned}$$

As shown above, the derivative of RSS with respect to β_1 is $-2\sum_{i=1}^n(x_i y_i - \beta_1 x_i^2)$. We will now set this equation to zero and solve for β_1. This will be the value of $\hat{\beta}_1$ that minimizes RSS, we will omit the second derivative test for the sake of space. This is done below.

$$\begin{aligned}
0 &= -2\sum_{i=1}^{n}(x_i y_i - \beta_1 x_i^2) \\
&= \sum_{i=1}^{n}(x_i y_i - \beta_1 x_i^2) \\
&= \sum_{i=1}^{n} x_i y_i - \beta_1 \sum_{i=1}^{n} x_i^2 \\
\beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}
\end{aligned}$$

Hence, as shown above, the least squares estimate of β_1 is $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$.

b. Show that

$$E[\hat{\beta}_1|X = x] = \beta_1.$$

Using our result from part a, we will show that $\hat{\beta}_1$ is an unbiased estimator of β_1. That is, we will show that $E[\hat{\beta}_1|X = x] = \beta_1$. This calculation is done below.

$$\begin{aligned}
E[\hat{\beta}_1|X = x] &= E\left[\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}|X = x\right] \\
&= \frac{E\left[\sum_{i=1}^n x_i y_i|X = x\right]}{\sum_{i=1}^n x_i^2} \quad \text{(Linearity of expectation)} \\
&= \frac{\sum_{i=1}^n x_i E\left[y_i|X = x\right]}{\sum_{i=1}^n x_i^2} \quad \text{(Linearity of expectation)} \\
&= \frac{\sum_{i=1}^n x_i \beta_1 x_i}{\sum_{i=1}^n x_i^2} \quad \text{(Since } E\left[y_i|X = x\right] = \beta_1 x_i\text{)} \\
&= \frac{\beta_1 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} \\
&= \beta_1
\end{aligned}$$

Hence, we have shown that $\hat{\beta}_1$ is an unbiased estimator of β_1. That is, we have shown that $E[\hat{\beta}_1|X = x] = \beta_1$.

c. Show that

$$Var[\hat{\beta}_1|X = x] = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

Using our result from part a, we will show that $Var[\hat{\beta}_1|X = x] = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$. The main fact that will be used to solve this problem is $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \implies y_i \overset{iid}{\sim} \mathcal{N}(\beta_1 x_i, \sigma^2)$. More importantly, this means that the y_i's are independent. This calculation is done below.

$$\begin{aligned}
Var[\hat{\beta}_1|X = x] &= Var\left[\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}|X = x\right] \\
&= \frac{1}{\left(\sum_{i=1}^n x_i^2\right)^2} Var\left[\sum_{i=1}^n x_i y_i|X = x\right] \quad \text{(Non-linearity of variance)} \\
&= \frac{1}{\left(\sum_{i=1}^n x_i^2\right)^2} \sum_{i=1}^n Var\left[x_i y_i|X = x\right] \quad \text{(Since the } y_i\text{'s are independent)} \\
&= \frac{1}{\left(\sum_{i=1}^n x_i^2\right)^2} \sum_{i=1}^n x_i^2 Var\left[y_i|X = x\right] \quad \text{(Non-linearity of variance)} \\
&= \frac{1}{\sum_{i=1}^n x_i^2} \sigma^2 \quad \text{(Since } Var[y_i|X = x] = \sigma^2\text{)}
\end{aligned}$$

Hence we have shown that $Var[\hat{\beta}_1|X = x] = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$.