Banks often use credit scores to determine if to offer someone a loan or mortgage. Suppose that a bank wants to predict the credit score of a person using other metrics. In a random sample of 200 credit card applicants, they recorded 4 variables:

- Credit Score (points)
- Income ($10,000s)
- Age (years)
- Education (numbers of years)

The dataset is called "STA4163 Mini Project 3 Dataset".

Part I: Simple Linear Regression

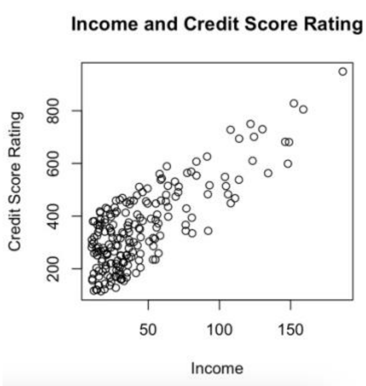
First, they want to see how well income can perform as a sole predictor of credit score using simple linear regression.

(a) Identify the independent and dependent variable.

Independent Variable: Income

Dependent Variable: Credit Score

(b) Create a scatterplot for this data. Describe the relationship between income and credit score.



The relationship between income and credit score ratings appear to be positive. As income increases, credit score ratings tend to increase as well.

(c) State the least-squares estimate of the regression line.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 193.2599    11.1083   17.40   <2e-16 ***
Income        3.5573     0.1982   17.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\boldsymbol{\hat{Y} = 193.2599 + 3.5573X_1}$

(d) Give practical interpretations of the y-intercept and the slope.

$\widehat{\boldsymbol{B_0}}$: It is an estimate of the y-intercept and there is no practical interpretation since an income of 0 was not sampled.

$\widehat{\boldsymbol{B_1}}$: For every $1 increase in income, credit score rating is expected to increase by 3.5573 points.

(e) Find the SSE.

```
Analysis of Variance Table

Response: Rating
           Df  Sum Sq Mean Sq F value    Pr(>F)
Income      1 2929736 2929736  322.28 < 2.2e-16 ***
Residuals 198 1799924    9091
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SSE = 1,799,924

(f) Find and estimate of σ and give practical interpretation of the estimate.

SSE = 1,799,924

n = 200

$$s^2 = \frac{SSE}{n-2} = \frac{1{,}799{,}924}{198} = 9090.\overline{52}$$

$$s = \sqrt{s^2} = \sqrt{9090.\overline{52}} = 95.34424604$$

$\boldsymbol{s \approx 95.344}$

We expected most roughly 95% of observations to lie within a credit score rating of 2($\mathbf{95.344}$) = 190.688 of the regression line.

(g) Conduct a hypothesis test at α = 5% to determine whether there is a significant linear relationship between income and credit score.

Step 0: Assumptions

- Assume that for a fixed value of income, the errors are normally distributed with a mean of zero and a constant variance of σ^2 regardless of income. The errors associated with income and credit score ratings are also independent.

Step 1: Null Hypothesis

- $H_0 : \beta_1 = 0$

Step 2: Alternative Hypothesis

- $H_\alpha : \beta_1 \neq 0$

Step 3: Test Statistic

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 193.2599    11.1083   17.40   <2e-16 ***
Income        3.5573     0.1982   17.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $t = 17.95$

Step 4: Rejection Region

- $p - value = < \ 2 \times 10^{-16}$

Step 5: Conclusion

- At α = 5%, we reject H_0 and conclude that there is a significant relationship between income and credit score rating.

(h) Construct a 95% confidence interval for the slope. Interpret the interval.

```
                 2.5 %     97.5 %
(Intercept) 171.354193 215.165557
Income        3.166568   3.948098
```

We are 95% confident that for every $1 increase in income, credit score rating is expected to increase by 3.167 to 3.948 points.

(i) Find and interpret the coefficient of correlation.

```
Residual standard error: 95.34 on 198 degrees of freedom
Multiple R-squared:  0.6194,    Adjusted R-squared:  0.6175 
F-statistic: 322.3 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
> cor(Income, Rating)
[1] 0.7870445
```

$r^2 = 0.6194$

$r = \sqrt{0.6194}$

$r = 0.7870196948$

$\boldsymbol{r \approx 0.787}$

There is a strong, positive linear relationship between income and credit score ratings.

(j) Find and interpret the coefficient of determination.

```
Residual standard error: 95.34 on 198 degrees of freedom
Multiple R-squared:  0.6194,    Adjusted R-squared:  0.6175
F-statistic: 322.3 on 1 and 198 DF,  p-value: < 2.2e-16
```

$\boldsymbol{r^2 = 0.6194}$

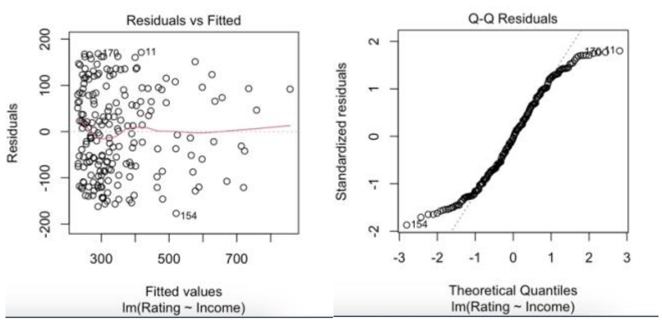
About 61.94% of the variation in credit score ratings are explained by income.

(k) Analyze the appropriate plots to check the assumptions of the errors. Make sure to:

a. State all the assumptions.

1. **The mean error is 0.**
2. **The errors have a constant variance.**
3. **The errors are normally distributed.**
4. **The errors are independent.**

b. Show all your plots and describe which assumption you are analyzing, and how the plot shows that assumption is or isn't violated.



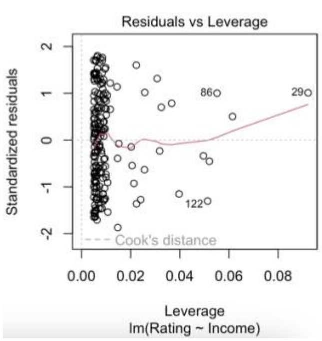
- **The "Residuals vs Fitted" plot tests the assumption 1 (the mean error is 0) and isn't violated because the plot shows that the points are not randomly distributed about the y = 0 line. The plot's distribution displays more points on the left side of the plot than the right side.**
- **The "Residuals vs Fitted" plot also tests assumption 2 (the errors have a constant variance) and isn't violated. Although there are some residuals that are close to the y = 0 line, majority of the residuals are spread out.**
- **The "Q-Q Residuals" plot tests the assumption 3 (the errors are normally distributed) and when conducting the Shapiro-Wilk Test, the test comes back significant since the p-value is 0.00001294. Therefore, we reject $\boldsymbol{H_0}$ and conclude that the errors do not follow a normal distribution. This also means that from the Shapiro-Wilk Test, this assumption is violated.**

- **Assumption 4 is not violated because the data does not depend on time, demonstrating that the residuals are independent of one another.**

Overall, does it appear that the assumptions have been met?

The assumptions have not been met because assumption 3 is violated.

(l) Analyze for outliers. Should any observations be removed? Explain why.



The absolute value of all standardized residuals is less than 3, so there are no outliers, and no observations should be removed.

(m) Is the model appropriate to use? State why or why not.

The model is not appropriate to use because the assumption of errors was violated. Specifically, assumption 1, assumption 2, and assumption 3 were violated.

(n) Regardless of your previous answer, find the predicted credit score for someone who makes $56,000 a year.

```
       fit      lwr      upr
1 392.4705 203.9275 581.0136
```

We are 95% confident that the predicted credit score rating when someone's income is $56,000 a year is between 203.9275 and 581.0136.

Part II: Multiple Linear Regression

(a) Build a model using income, age, and education to predict credit score. Write down the least-squares estimate of the regression equation.

$\hat{Y} = \mathbf{247.4118 + 3.5768}X_1 - \mathbf{0.3550}X_2 - \mathbf{2.6173}X_3$

(b) Find and interpret r^2 and adjusted r^2. Use one of these metrics to compare this model to the model in Part I. Which model appears to be better?

Part I:

```
Residual standard error: 95.34 on 198 degrees of freedom
Multiple R-squared:  0.6194,    Adjusted R-squared:  0.6175
F-statistic: 322.3 on 1 and 198 DF,  p-value: < 2.2e-16
```

Part II:

```
Residual standard error: 95.31 on 196 degrees of freedom
Multiple R-squared:  0.6236,    Adjusted R-squared:  0.6178
F-statistic: 108.2 on 3 and 196 DF,  p-value: < 2.2e-16
```

$\boldsymbol{r^2 = 0.6236}$ **and** $\boldsymbol{r^2{}_a}$ **=** $\boldsymbol{0.6178}$

The multiple linear regression model appears to be better than the simple linear regression model in Part I because the value of r^2 and $r^2{}_a$ in Part II is larger than the value of r^2 and $r^2{}_a$ in Part I. In Part I, about 61.94% of the variation in credit score ratings are explained by income whereas in Part II, 62.36% of the variation in credit score ratings are explained by income.

(c) Conduct an overall F-test.

Step 0: Assumptions

- $\varepsilon = 0$ is valid, no clear pattern and constant variance is also valid, no large cone pattern is present. The points roughly follow a straight line, so normality is valid. Our data doesn't depend on time, so independence is valid.

Step 1: Null Hypothesis

- H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

Step 2: Alternative Hypothesis

- H_α: At least one $\beta_i \neq 0$

Step 3: Test Statistic

```
Residual standard error: 95.31 on 196 degrees of freedom
Multiple R-squared:  0.6236,    Adjusted R-squared:  0.6178
F-statistic: 108.2 on 3 and 196 DF,  p-value: < 2.2e-16
```

- F = 108.2

Step: 4 Rejection Region

- P-value = $< 2.2 \times 10^{-16}$

Step 5: Conclusion

- At α = 5%, we reject H_0 and conclude that at least 1 of income, age, and education contributes to the credit score rating.

(d) The bank is particularly interested in using education level to predict credit scores. Conduct a hypothesis test and make a recommendation on whether education level should be used.

Step 0: Assumptions

- $\varepsilon = 0$ is valid, no clear pattern and constant variance is also valid, no large cone pattern is present. The points roughly follow a straight line, so normality is valid. Our data doesn't depend on time, so independence is valid.

Step 1: Null Hypothesis

- $H_0: \beta_i = 0$

Step 2: Alternative Hypothesis

- $H_\alpha: \beta_i \neq 0$

Step 3: Test Statistic

```
Analysis of Variance Table

Response: Rating
           Df  Sum Sq Mean Sq  F value Pr(>F)
Income      1 2929736 2929736 322.5278 <2e-16 ***
Age         1    7394    7394   0.8139 0.3681
Education   1   12131   12131   1.3355 0.2492
Residuals 196 1780399    9084
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- F = 1.3355

Step: 4 Rejection Region

- P-value = 0.2492

Step 5: Conclusion

- At α = 5%, we fail to reject H_0 and cannot conclude that education significantly contributes to the prediction of credit score ratings, given all other variables.