Homework-6-Q2

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: April 7 2025

Problem 2.

A researcher is interested in understanding the relationship between income (measured in thousands of dollars per year) and the amount of money spent yearly on food (also measured in thousands of dollars). The researcher has collected data on these two variables, see income.csv. Conduct an analysis to model the relationship between these two variables, using the food expenditures the outcome. Be sure to state all assumptions, check their validity, describe any statistical tests you perform, and perform any required remedial measures.

- **Y (response)**: Food Expenditures in thousands (food_expenditures)
- **X1 (predictor)**: Income in thousands (income_thousands)

```
data_path = "/Users/elizabethnemeti/Documents/GitHub/BIOS507-Coursework/Homeworks/Homework-6/"
data_file <- file.path(data_path, "food_income.csv")
FOODINCOME_data <- read.csv(data_file, header = TRUE)

str(FOODINCOME_data)
```

```
## 'data.frame':    100 obs. of  2 variables:
##  $ income_thousands : num  57.7 53.2 79.7 35.1 72.2 ...
##  $ food_expendatures: num  25.86 7.84 42.48 42.16 5 ...
```

```
head(FOODINCOME_data)
```

```
##   income_thousands food_expendatures
## 1         57.69895         25.860930
## 2         53.19053          7.837923
## 3         79.70902         42.476991
## 4         35.07448         42.160502
## 5         72.16944          5.000000
## 6         73.53937         55.087488
```

Regression model: E[food_expendatures] = ß0 + ß1(income_thousands)

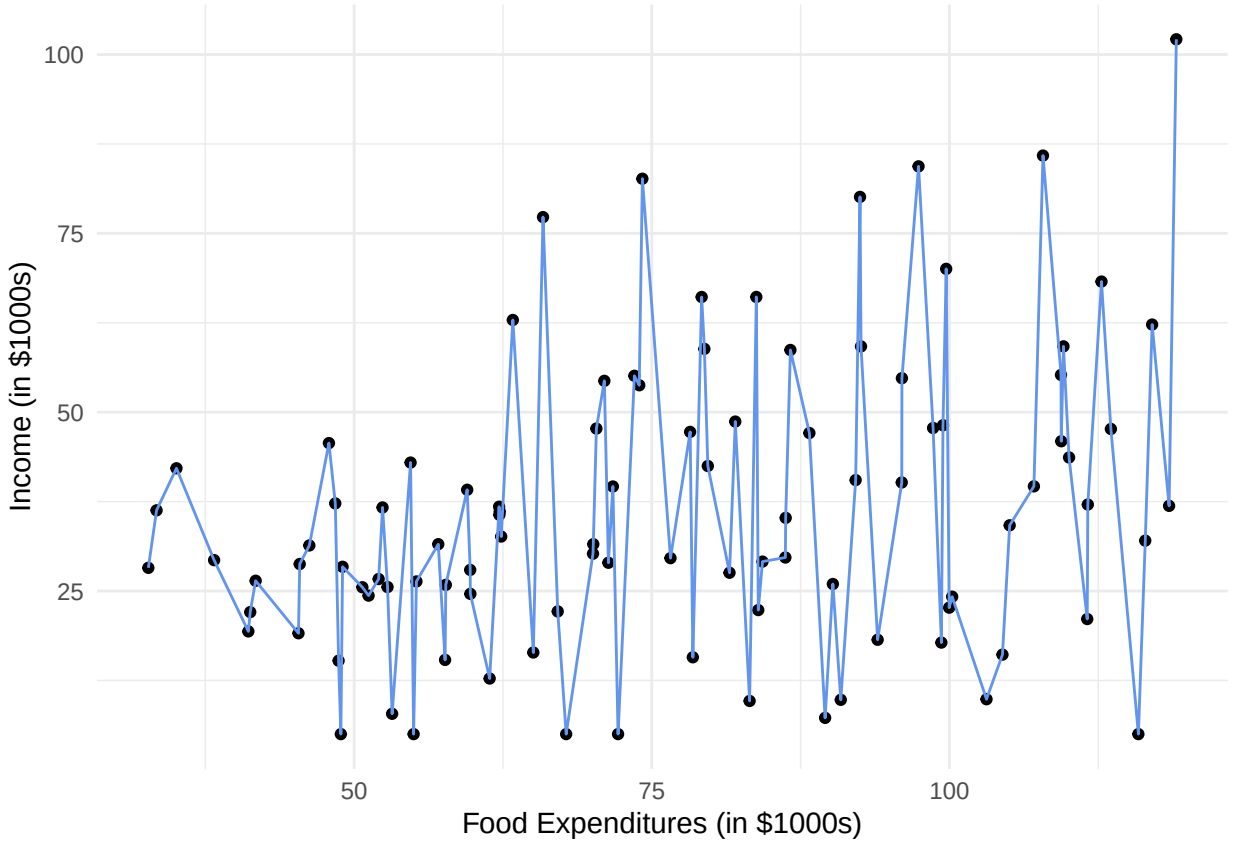
First, let's fit a simple linear regression model.

```
OLS_model <- lm(food_expendatures ~ income_thousands, data = FOODINCOME_data)
summary(OLS_model)
```

```
##
## Call:
## lm(formula = food_expendatures ~ income_thousands, data = FOODINCOME_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.500 -11.404  -1.056  10.990  53.734
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      15.00583    6.59333   2.276 0.025028 *
## income_thousands  0.28043    0.08212   3.415 0.000929 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.27 on 98 degrees of freedom
## Multiple R-squared:  0.1063, Adjusted R-squared:  0.09723
## F-statistic: 11.66 on 1 and 98 DF,  p-value: 0.0009293
```

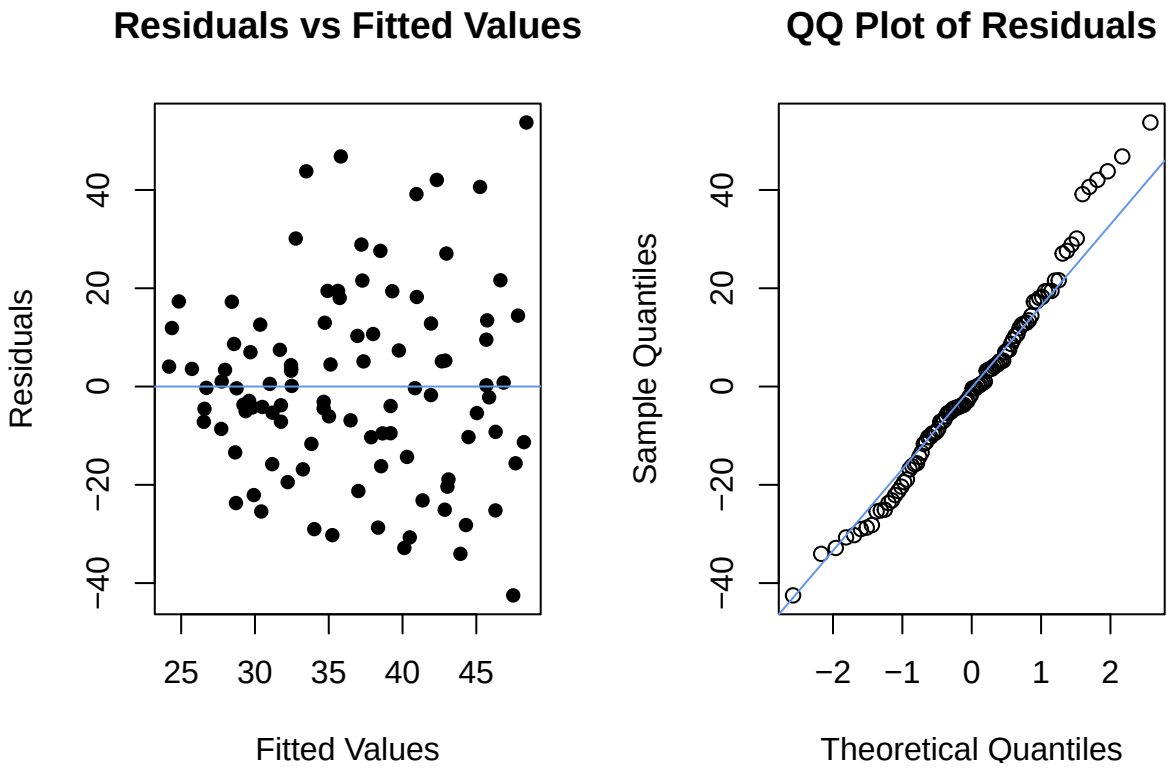
```
ggplot(FOODINCOME_data, aes(x = income_thousands, y = food_expendatures)) +
  geom_point(color = "black") +
  geom_line(aes(y = food_expendatures), color = "cornflowerblue") +
  labs(x = "Food Expenditures (in $1000s)",
       y = "Income (in $1000s)") +
  theme_minimal()
```



Looking at the summary() output, the estimated intercept is 15.01, meaning that when income is $0, the expected food expenditure is still about $15,010.ß1 is 0.28, meaning that for each $1000 increase in income, food expenditure increases by about $280. We can fill in our model: E[Y] = 15.01 + 0.28(income in $1000s). R^2 is 0.1063, indicating the model explains only about 10.63% of variance in food expenditure. However, the p value is 0.0009293, where $p < 0.05$. Therefore, the slope is statistically significant, though R^2 explains little of the variance in Y. This suggests, that a different model may explain the relationship between income and food expenditure better.

Assumptions of the model: - The relationship between income and food expenditures is linear.
- Residuals are normally distributed.
- Observations are independent (iid).
- The variance of residuals is constant across all X values (constant variance - homoscedasticity).

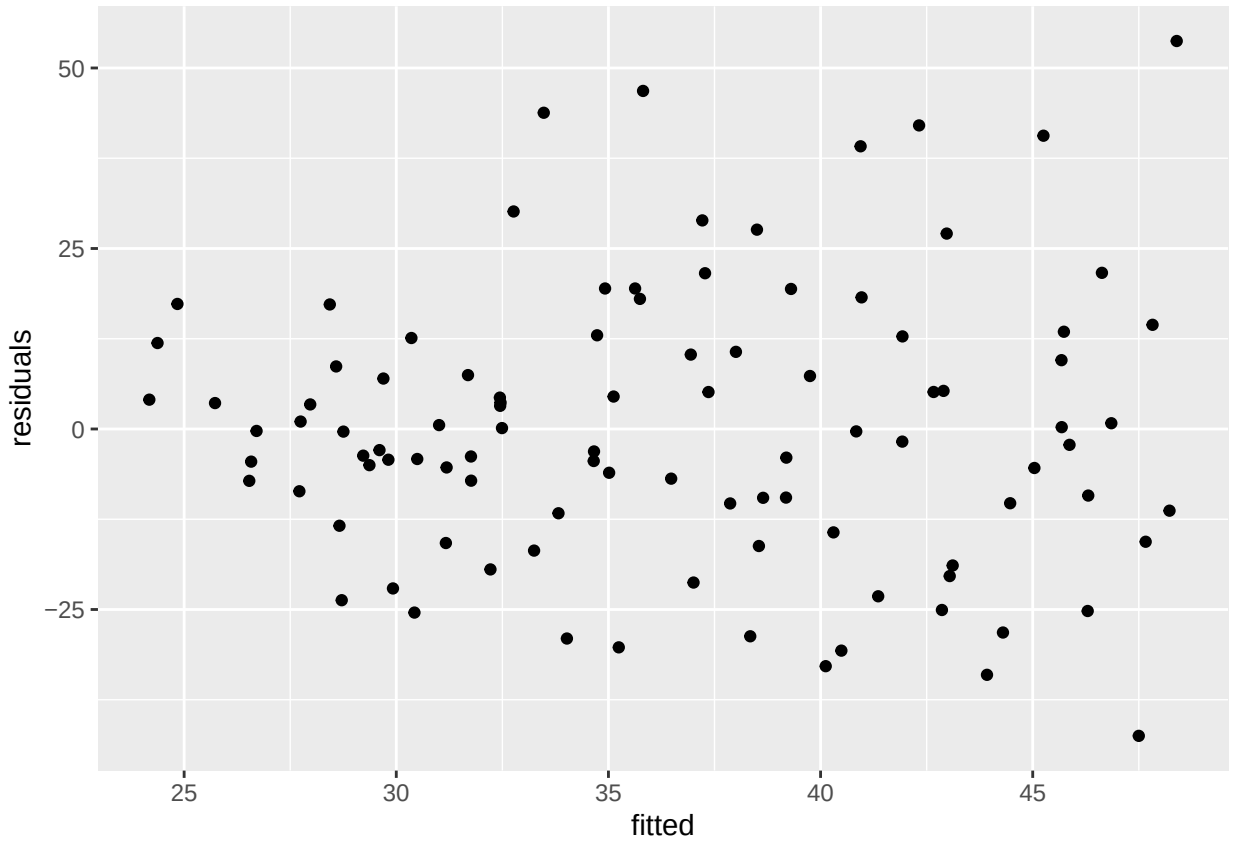
```r
par(mfrow = c(1, 2))

plot(OLS_model$fitted.values, residuals(OLS_model),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values",
     pch = 16)
abline(h = 0, col = "cornflowerblue")

qqnorm(residuals(OLS_model), main = "QQ Plot of Residuals")
qqline(residuals(OLS_model), col = "cornflowerblue")
```



Looking at the fitted values vs residuals, we can see a clear fanning pattern, where the spread of residuals increases as fitted values increase. This indicates that homoscedasticity has been violated. Additonally, the QQ plot of residuals shows a heavy right tail, again indicating violation of normality.

To deeper examine the assumption of constant variance, we can do a residuals vs fitted plot.

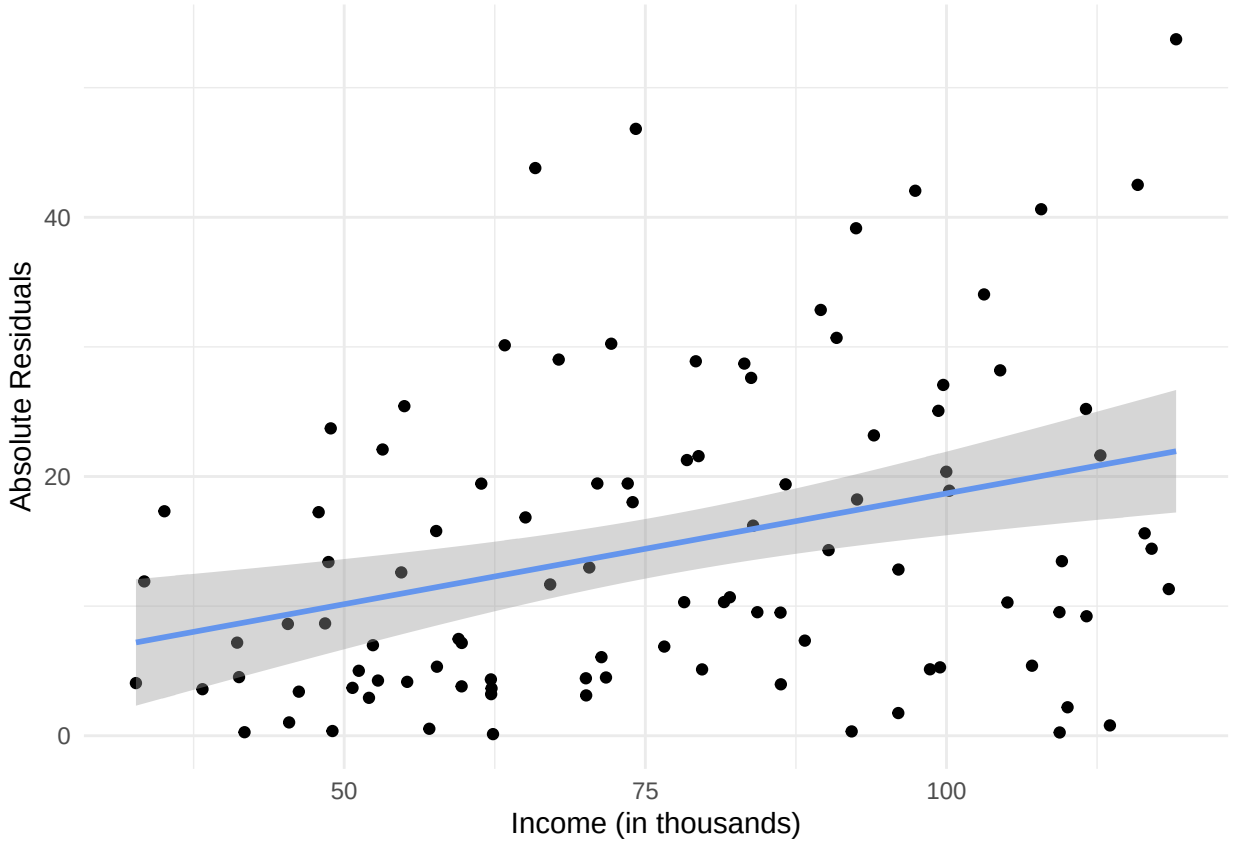
```r
# extracting residuals and fitted vals
res_fit_data <- tibble(
  residuals = OLS_model$residuals,
  fitted = OLS_model$fitted.values,
  income = FOODINCOME_data$income_thousands
)

# generating residuals vs fitted plot
res_fit_plot <- res_fit_data %>%
  ggplot(aes(x = fitted, y = residuals)) +
  geom_point()

res_fit_plot
```



The fan-like pattern violates the homoscedasticity assumption, and is evidence that the variance is not constant. We can see that the spread of residuals increases as fitted values increase, suggesting that as income increases, variability in food expenditures increases. This is important to remediate, otherwise our interpretation of the p value, and inferences from the model output will likely be incorrect. Therefore, we are going to perform weighted least squares (WLS). The goal of WLS will be to stabilize variance by doing regression of absolute residuals on the predictor of interest (Income). The fitted values of this model will be the weights.

```
# estimate a function for the standard deviations (dont forget the absolute values)
abs_res_model <- lm(abs(residuals) ~ income, data = res_fit_data)

# use the fitted values from this standard deviation fit
predicted_sd <- abs_res_model$fitted.values

# add weights
# gives weights where observations with higher variance get less weight (more stable ones get more infl
weights <- 1 / predicted_sd^2

# adding a visual
abs_res_model <- lm(abs(residuals) ~ income, data = res_fit_data)

ggplot(res_fit_data, aes(x = income, y = abs(residuals))) +
  geom_point() +
  geom_smooth(method = "lm", color = "cornflowerblue") +
  labs(x = "Income (in thousands)",
       y = "Absolute Residuals") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



In the plot above, we can see how the variance of the residuals increases as income increases. The upward trend in the absolute residuals confirms that the higher the income observations, the greater the variability in food expenditure, which will need remedial measures.

Now that we have our standard deviation estimates we can fit the WLS model.

```
WLS_model <- lm(food_expendatures ~ income_thousands,
                data = FOODINCOME_data,
                weights = weights) # key is to use optional weights arg

summary(WLS_model)
```

```
##
## Call:
## lm(formula = food_expendatures ~ income_thousands, data = FOODINCOME_data,
##     weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4379 -0.8371 -0.0787  0.7103  3.3936
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      16.76691    5.03497   3.330 0.001225 **
## income_thousands  0.25600    0.07505   3.411 0.000942 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.268 on 98 degrees of freedom
## Multiple R-squared:  0.1061, Adjusted R-squared:  0.097
## F-statistic: 11.64 on 1 and 98 DF,  p-value: 0.0009415
```

Now that we have two sets of estimates (ols and wls) we need to compare them. They should be similar if everything is working.

```
print(cbind(OLS_model$coefficients , WLS_model$coefficients))
```

```
##                       [,1]       [,2]
## (Intercept)      15.005829 16.7669104
## income_thousands  0.280427  0.2559979
```

OLS output is column [,1], and WLS output is [,2]. In the OLS model, the intecept was 15.01 and the slope (ß1) was 0.280. After applying WLS to address the heteroscedasticity, the intercept increased slightly to 16.77, and the slope (ß1) decreased slightly to 0.256. The WLS model is E[food_expendatures] = 16.77 + 0.256(income). This means, that after using WLS, each $1000 increase in income sees a $256 increase in food expenditure, comapred to the $280 increase with OLS. As the estimates are fairly similar, we do not need to continue iterating with WLS.