Project 2

Chloe Chen & Kangrui Liu

9/12/2023

1. JWHT Chapter 2. Exercise 5.

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification?

High Flexibility Advantages(less flexible - opposite of flexible):

1) High Predictive Capacity: Can capture complex patterns and non-linear relationships in data, which may more accurately model the true function.
2) Effective for Complex Data: Suitable for complex datasets with intricate relationships.
3) Overfitting Control: Overfitting can be mitigated with proper regularization techniques.

High Flexibility Disadvantages(less flexible - opposite of flexible):

1) Complexity and Interpretability: Often challenging to interpret due to a large number of parameters.
2) Data Requirements: Requires substantial data for accurate parameter estimation.
3) Computational Complexity: May be computationally intensive.

Under what circumstances might a more flexible approach be preferred to a less flexible approach?

1) The dataset is large and diverse with different groups, providing ample data for model training.
2) The relationship between predictors and the response variable is complex and non-linear.
3) Achieving the highest predictive accuracy is a top priority.
4) Adequate regularization techniques can be applied to control overfitting.
5) Computational resources are available for training and inference.

When might a less flexible approach be preferred?

1) The dataset is small or has limited variability, making overfitting a concern.
2) Interpretability of the model is essential for understanding the relationships between variables.
3) Computational resources are limited or real-time processing is required.
4) The underlying data relationships are believed to be relatively simple and linear.
5) The model needs to be robust and less sensitive to outliers or noisy data.

2. Faraway Chapter 2. Exercise 2.

The dataset uswages is drawn as a sample from the Current Population Survey in 1988.

Fit a model with weekly wages as the response and years of education and experience as predictors in linear regression.

```
library(faraway)
data("uswages")
pred_wage<-lm(wage~educ+exper,uswages)
summary(pred_wage)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = uswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1018.2  -237.9   -50.9   149.9  7228.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -242.7994    50.6816  -4.791 1.78e-06 ***
## educ          51.1753     3.3419  15.313  < 2e-16 ***
## exper          9.7748     0.7506  13.023  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 427.9 on 1997 degrees of freedom
## Multiple R-squared:  0.1351, Adjusted R-squared:  0.1343
## F-statistic:   156 on 2 and 1997 DF,  p-value: < 2.2e-16
```

Report and give a simple interpretation to the regression coefficient for years of education.

- $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_1 + \epsilon_i$, where $X1$ is *educ*, $X2$ is *exper*.
- The regression coefficient for the years of education in the linear regression model is approximately 51.1753. This coefficient represents the estimated change in the response variable *wage* for a one-unit change in the predictor variable *educ*, while holding all other predictors constant.
- For every additional year of education, there is an estimated increase of approximately 51.1753 units in the weekly wage, while controlling for the effect of the exper variable.
- In other words, individuals with higher levels of education tend to earn higher weekly wages, and the coefficient quantifies the expected increase in wages associated with each additional year of education when other factors are held constant.
- Additionally, the significance of the coefficient (indicated by the *** in the output) suggests that the relationship between education and wages is statistically significant.

Now fit the same model but with logged weekly wages.

```
uswages$log_wage <- log(uswages$wage)
log_pred_wage <- lm(log_wage ~ educ + exper, data = uswages)
summary(log_pred_wage)
```

```
##
## Call:
## lm(formula = log_wage ~ educ + exper, data = uswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7533 -0.3495  0.1068  0.4381  3.5699
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.650319   0.078354   59.35   <2e-16 ***
## educ        0.090506   0.005167   17.52   <2e-16 ***
## exper       0.018079   0.001160   15.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6615 on 1997 degrees of freedom
## Multiple R-squared:  0.1749, Adjusted R-squared:  0.174
## F-statistic: 211.6 on 2 and 1997 DF,  p-value: < 2.2e-16
```

Give an interpretation to the regression coefficient for years of education.

- In this regression model, the regression coefficient for the educ variable is approximately 0.090506. This coefficient represents the estimated change in the *log_wage* for a one-unit change in the predictor variable *educ*, while holding all other predictors constant.

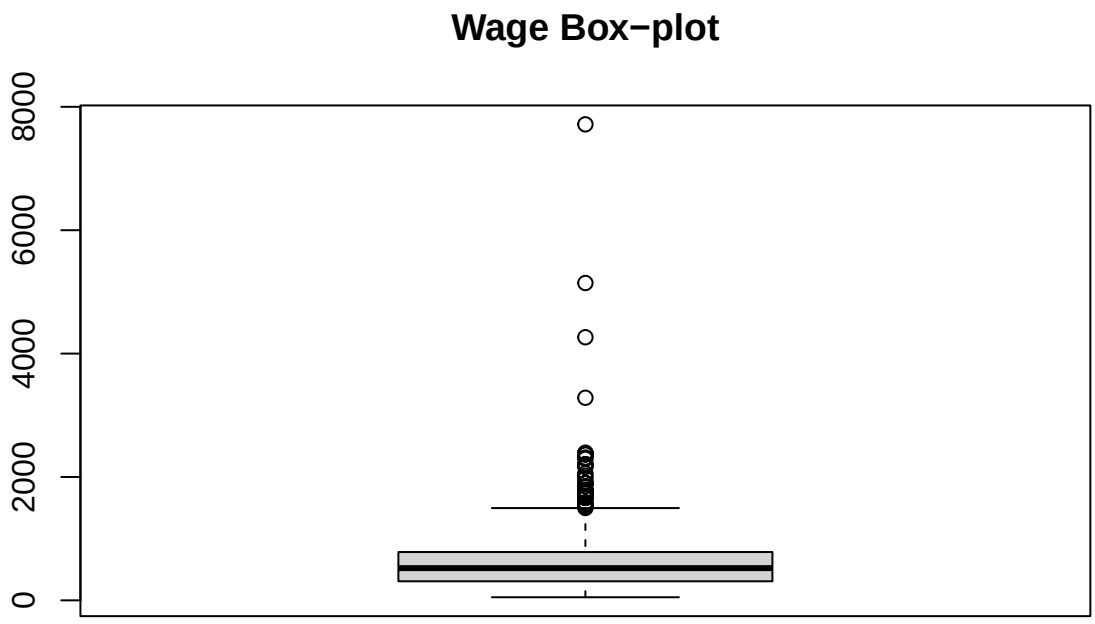
```
increase <- (exp(0.090506) - 1) * 100  #calculate the percaentage(?)
increase
```

```
## [1] 9.472808
```

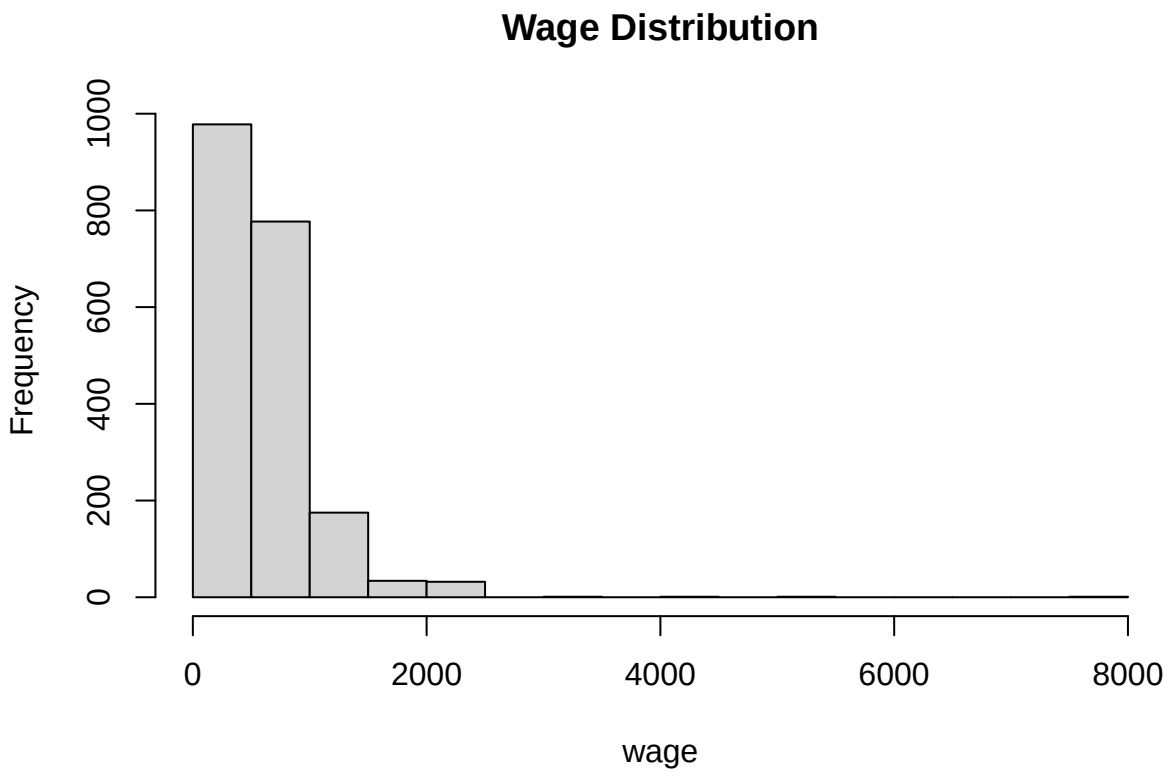
- For every additional year of education, there is an estimated multiplicative effect on the log_wage of approximately 0.090506. In other words, each additional year of education is associated with an approximately 9.47 increase in the weekly wage, on average, when controlling for the effect of the *exper* variable.

Which interpretation is more natural?

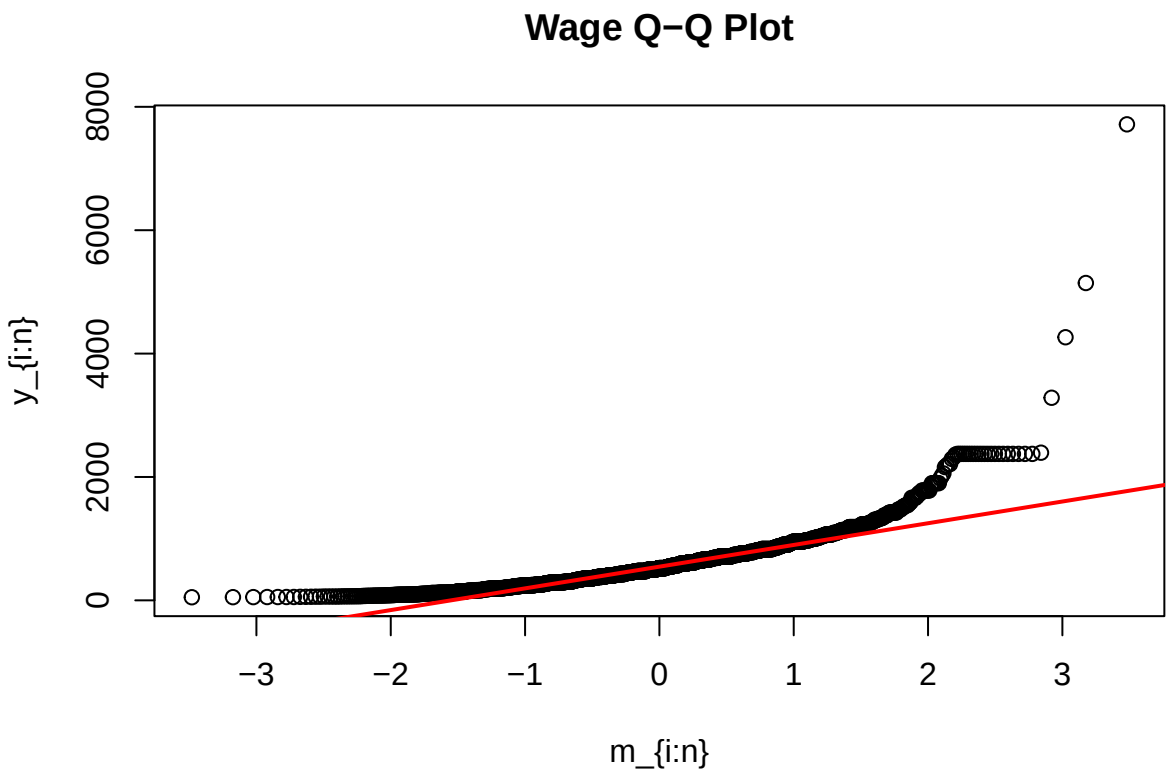
```
library(ggplot2)
#boxplot
boxplot(uswages$wage, main="Wage Box-plot")
```



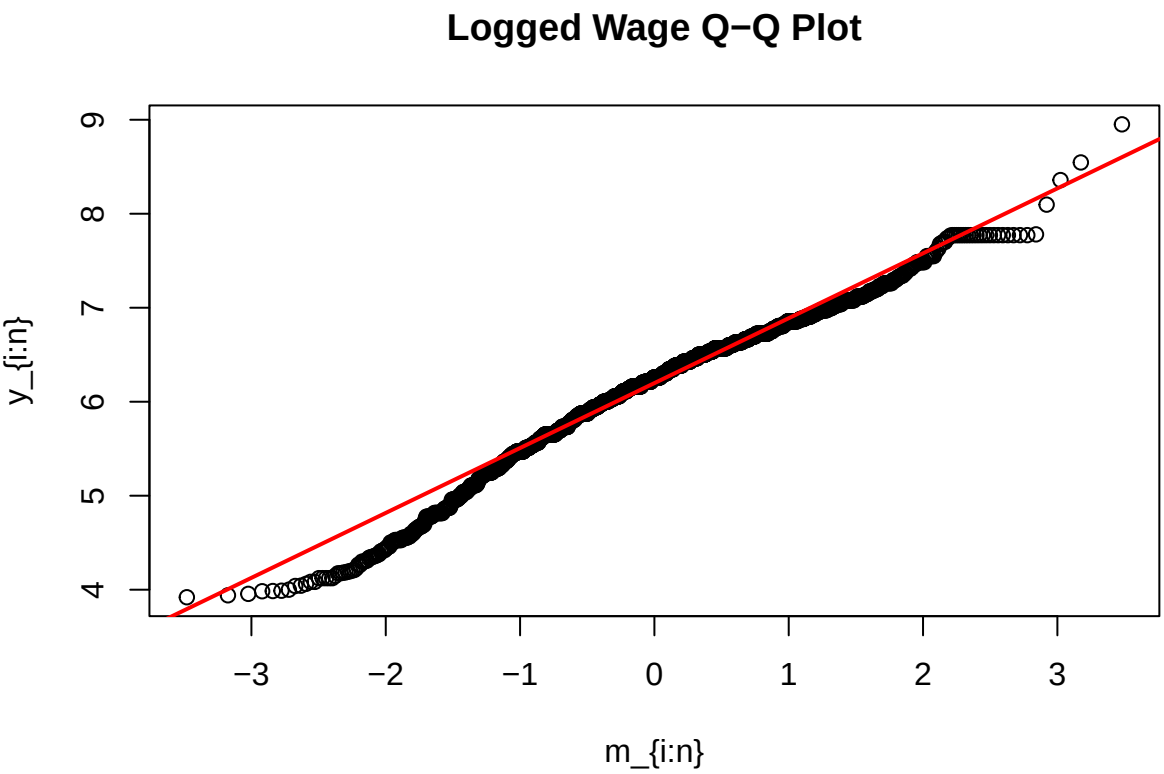

```r
#histagram
hist(uswages$wage, main = "Wage Distribution", xlab = "wage", ylab = "Frequency")
```




```
#Q-Q plot
qqnorm(uswages$wage, main="Wage Q-Q Plot", ylab="y_{i:n}", xlab="m_{i:n}")
qqline(uswages$wage, col="red",lwd=2)
```




```
qqnorm(uswages$log_wage, main="Logged Wage Q-Q Plot", ylab="y_{i:n}", xlab="m_{i:n}")
qqline(uswages$log_wage, col="red",lwd=2)
```



- The interpretation of the coefficient for educ in this **logged model** is more natural.
- *Reasons*

1) **Conformity with Reality**: In real life, wage(y) can only take non-negative values. If we do a linear regression directly on y and x, we get an estimate of the model, and then bring in an x to calculate $\hat{y}$, which may take a negative number that doesn't make sense in real life, whereas taking a logged y solves this problem.

2) **Interpretability**: Because the coefficient in the logged model represents a multiplicative effect, it reflects a percentage change in wages associated with an additional year of education which is often more meaningful when working with something like wages.

3) **Outlier Mitigation**: From box plot we can see there are outliers in wage data that can have a disproportionate influence on the results. Taking the log can help reduce the impact of extreme values and make the model more robust to outliers.

4) **Normalization of Skewed Data**: As we can see from the histogram and the Q-Q plot, the Wage data exhibits right-skewness, meaning that there may be a few high earners that create a long right tail in the distribution. Taking the log of wages helps to reduce the skewness, making the data more symmetrically distributed and conforming more closely to the assumptions of linear regression, which assumes that the residuals are normally distributed.