HW 02

Group 10

2022-10-03

Homework 2

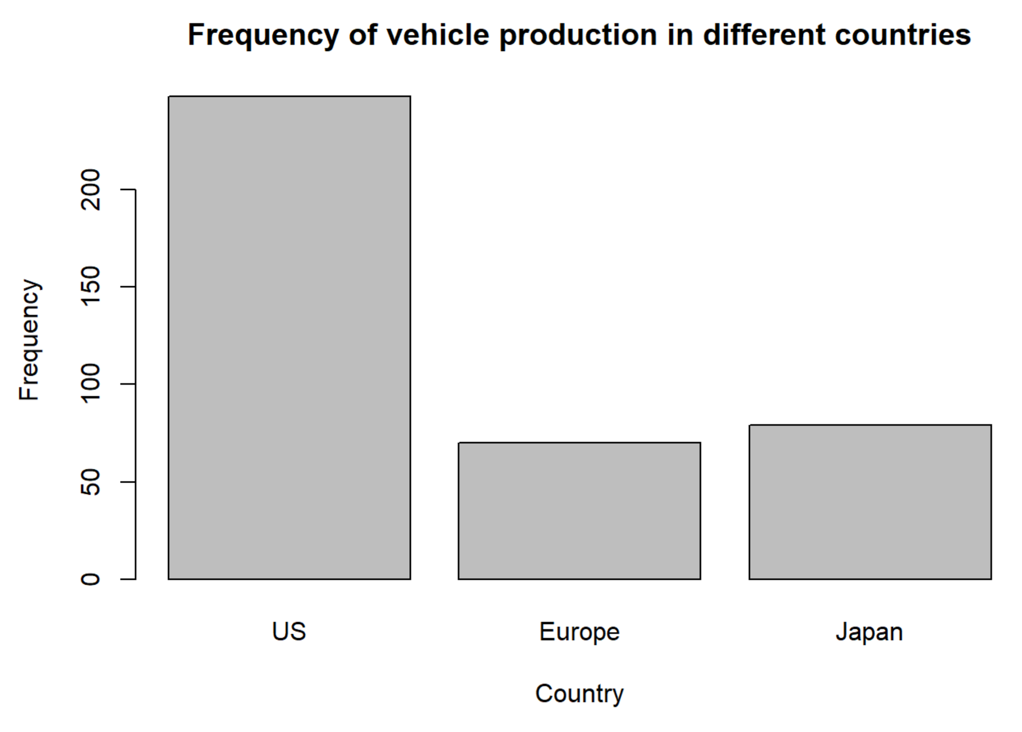
Read csv

```
auto = read.csv("Auto.csv", na.strings = "NA")
```

1(a)

```
auto$origin = factor(auto$origin, 1:3, c("US", "Europe", "Japan"))

freq <- table(auto$origin)
barplot(freq, main = "Frequency of vehicle production in different countries", xlab = "Country", ylab = "Frequenc
y")
```



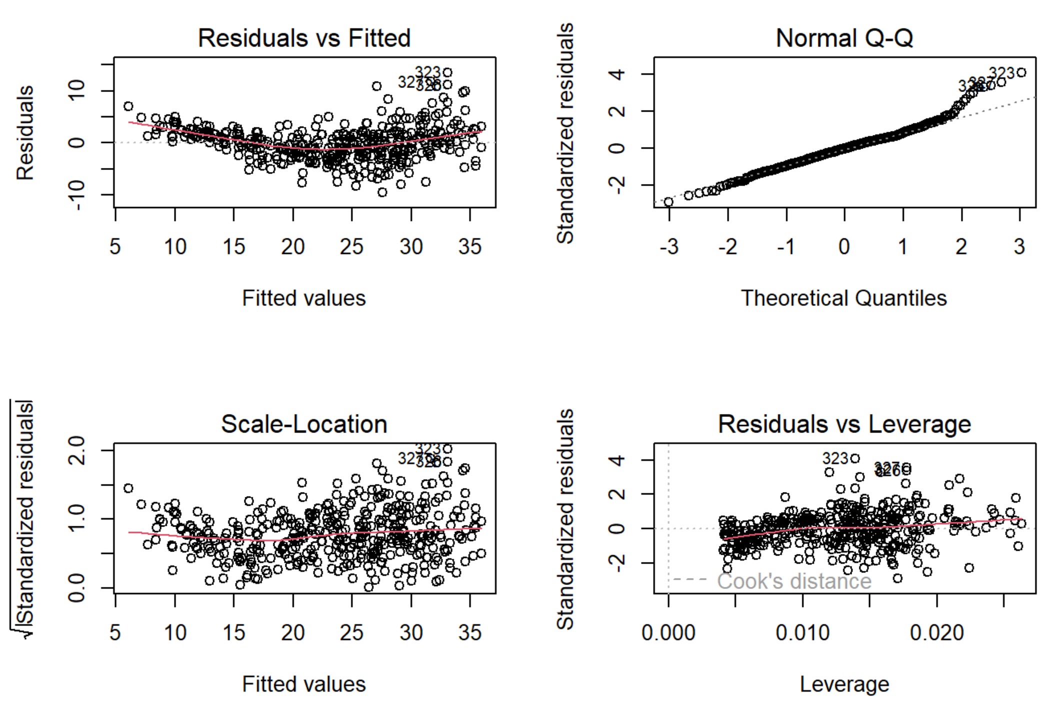

```
prop.table(table(auto$origin))
```

```
##
##        US    Europe     Japan
## 0.6246851 0.1763224 0.1989924
```

Above you can see the frequency is much greater in the US.

1(b)

```
lmb <- lm(mpg ~ origin + weight + year, data = auto)
par(mfrow=c(2,2))
plot(lmb)
```



The plots indicate that: - The error term is not normally distributed - The variance is not constant, thus heteroskedasticity is present - The data is not normally distributed

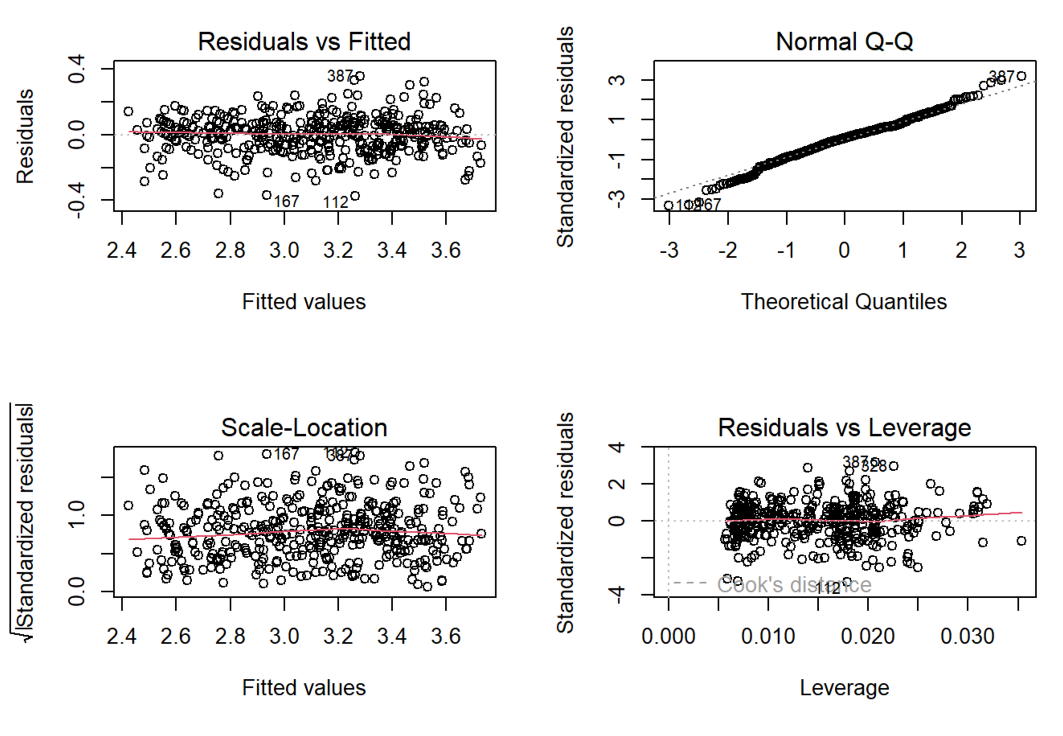
We can tell that the The error term is not normally distributed by looking at the Residuals vs Fitted plot. The red line doesn't move along 0 at all. It looks more like a quadratic function.

The variance as x increases on the Residuals vs Fitted plot which is an indicator that heteroskedasticity is present.

We can see from the QQ plot that the data doesn't follow the line. Towards the top the data skews upwards.

1(c)

```
lmc <- lm(log(mpg) ~ origin + log(weight) + year + I(year^2), data = auto)
par(mfrow=c(2,2))
plot(lmc)
```




```
summary(lmc)
```

```
##
## Call:
## lm(formula = log(mpg) ~ origin + log(weight) + year + I(year^2),
##     data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37408 -0.06782  0.00899  0.06903  0.35766
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.4693014  2.6833895   6.883 2.34e-11 ***
## originEurope  0.0668291  0.0176293   3.791 0.000174 ***
## originJapan   0.0319711  0.0179382   1.782 0.075477 .
## log(weight)  -0.8750305  0.0270390 -32.362  < 2e-16 ***
## year         -0.2559684  0.0712094  -3.595 0.000366 ***
## I(year^2)     0.0019051  0.0004687   4.065 5.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1136 on 391 degrees of freedom
## Multiple R-squared:  0.8898, Adjusted R-squared:  0.8884
## F-statistic: 631.7 on 5 and 391 DF,  p-value: < 2.2e-16
```

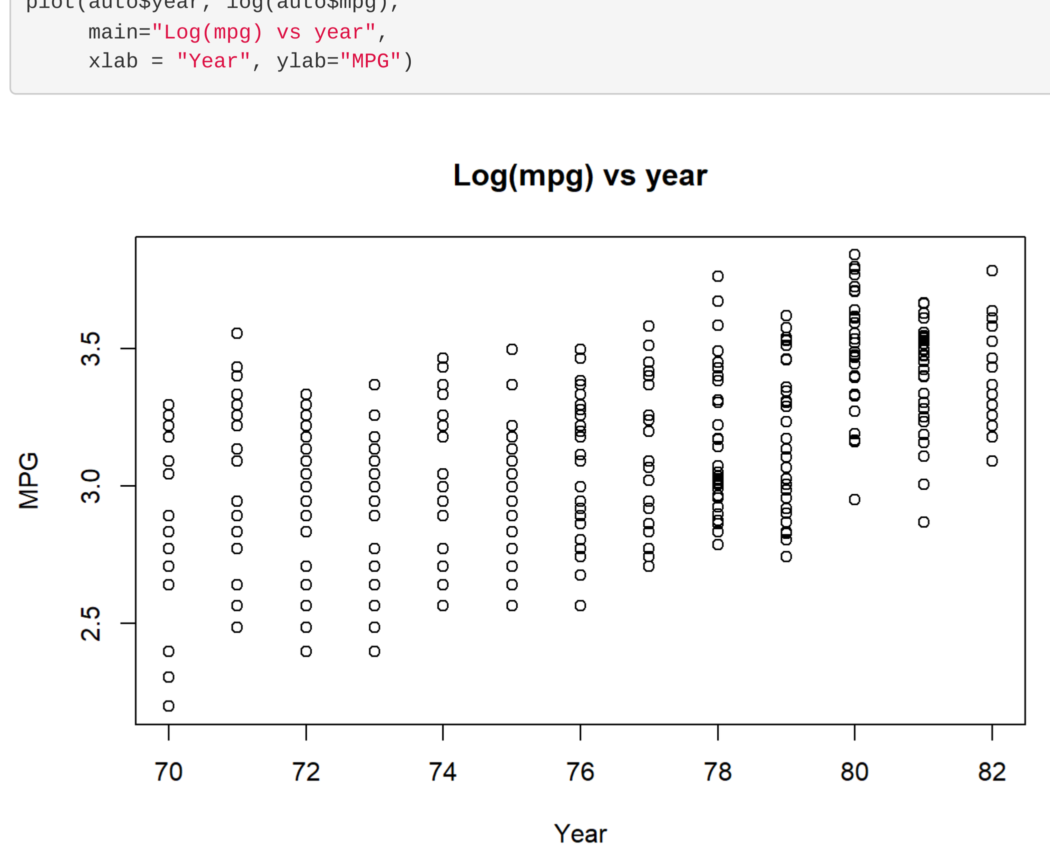
The model assumptions seem to have been roughly satisfied now.

The previously unsatisfied assumptions: - Heteroskedacicity - Error term is not normally distributed - Data is not normally distributed

When looking at the Residuals vs Fitted plot, we see the line follows 0 well and the variance is pretty much constant for each x value. We also see from the QQ plot that the data is more normally distributed now.

1(d)

```
plot(auto$year, log(auto$mpg),
     main="Log(mpg) vs year",
     xlab = "Year", ylab="MPG")
```




```
min = -coef(lmc)[5] / (2 * coef(lmc)[6])
```

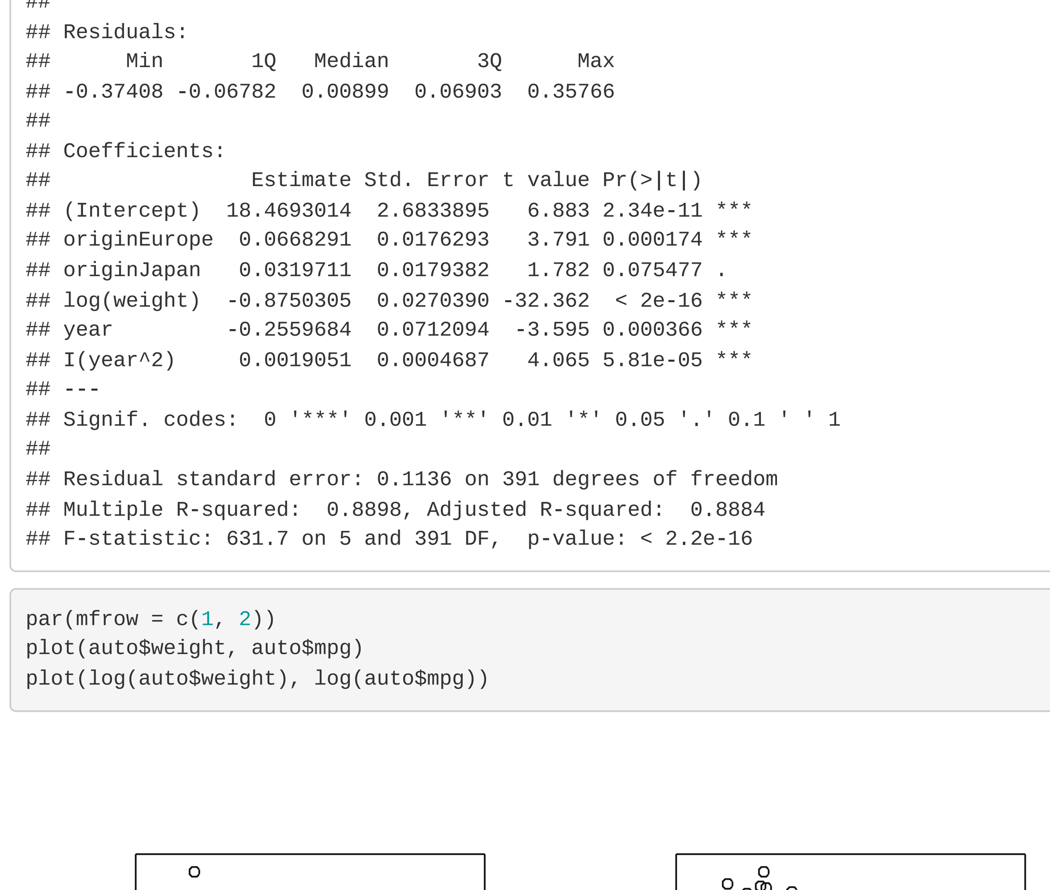
The relationship appears U-shaped based on the plot above. The minimum is 67.1781178.

1(e)

```
summary(lmc)
```

```
##
## Call:
## lm(formula = log(mpg) ~ origin + log(weight) + year + I(year^2),
##     data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37408 -0.06782  0.00899  0.06903  0.35766
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.4693014  2.6833895   6.883 2.34e-11 ***
## originEurope  0.0668291  0.0176293   3.791 0.000174 ***
## originJapan   0.0319711  0.0179382   1.782 0.075477 .
## log(weight)  -0.8750305  0.0270390 -32.362  < 2e-16 ***
## year         -0.2559684  0.0712094  -3.595 0.000366 ***
## I(year^2)     0.0019051  0.0004687   4.065 5.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1136 on 391 degrees of freedom
## Multiple R-squared:  0.8898, Adjusted R-squared:  0.8884
## F-statistic: 631.7 on 5 and 391 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1, 2))
plot(auto$weight, auto$mpg)
plot(log(auto$weight), log(auto$mpg))
```



It tells us that as you increase the weight the mpg falls. The relationship for the unlogged version is similar, less linear, but still negative.

2(a)

$$
\begin{gathered}
y_i = \gamma_0 + \gamma_1(x_i - \bar{x}) + \gamma_2(x_i - \bar{x})^2 + e_i = \\
\gamma_0 + y_1 x_i - y_1\bar{x} + y_2 x_i^2 - 2y_2 x_i \bar{x} + \gamma_2\bar{x}^2 + e_i = \\
(\gamma_0 - \gamma_1\bar{x} + \gamma_2\bar{x}^2) + (\gamma_1 - 2\gamma_2\bar{x})x_i + \gamma_2 x_i^2 + e_i \\
\therefore \beta_0 = \gamma_0 - \gamma_1\bar{x} + \gamma_2\bar{x}^2 \\
\beta_1 = \gamma_1 - 2\gamma_2\bar{x} \\
\beta_2 = \gamma_2
\end{gathered}
$$

2(b)

```
auto$year_squared <- auto$year^2
summary(lm(mpg ~ year + year_squared, data = auto))
```

```
##
## Call:
## lm(formula = mpg ~ year + year_squared, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.349  -5.109  -0.878   4.587  18.196
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  577.25230  146.67144   3.936 9.81e-05 ***
## year         -15.84090    3.86508  -4.098 5.05e-05 ***
## year_squared   0.11230    0.02542   4.419 1.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.23 on 394 degrees of freedom
## Multiple R-squared:  0.3694, Adjusted R-squared:  0.3662
## F-statistic: 115.4 on 2 and 394 DF,  p-value: < 2.2e-16
```

2(c)

The correlation between year and year squared is 0.999759.

2(d)

The mean of year is 75.9949622.

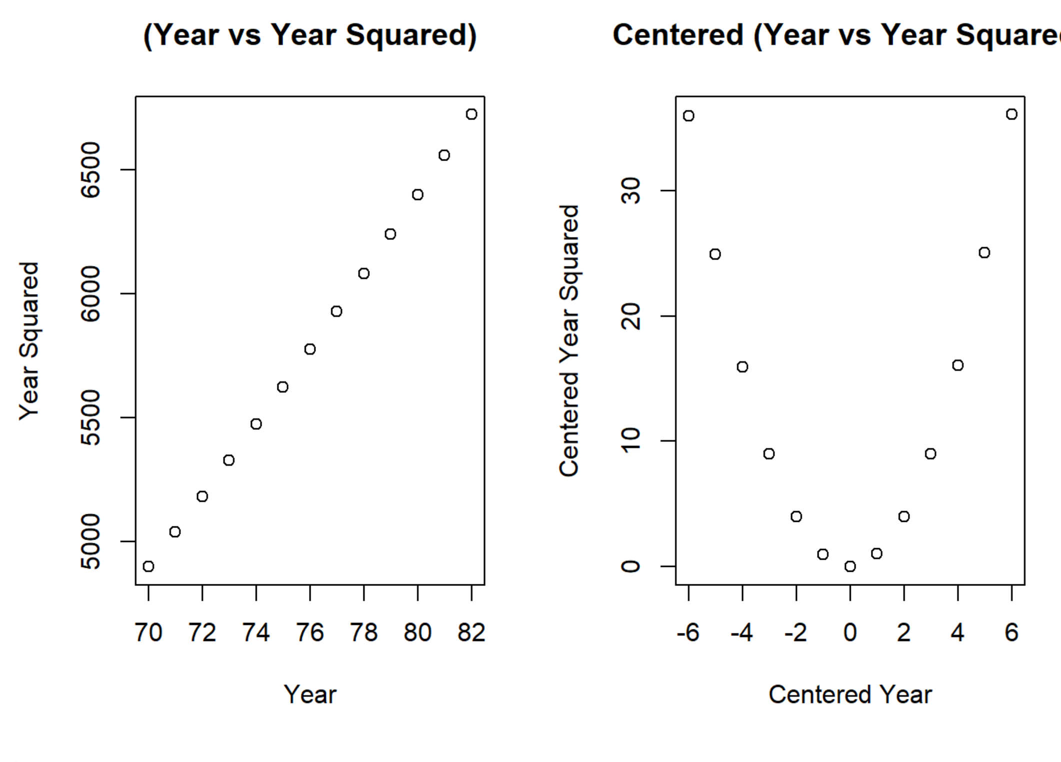
2(e)

```
auto$year_centered <- auto$year - mean(auto$year)
auto$year_centered_squared <- (auto$year_centered)^2
```

The correlation between centered year and centered year squared is 0.014414

2(f)

```
par(mfrow = c(1, 2))
plot(auto$year, auto$year_squared, main = "(Year vs Year Squared)", xlab = "Year", ylab = "Year Squared")
plot(auto$year_centered, auto$year_centered_squared, main = "Centered (Year vs Year Squared)", xlab = "Centered Y
ear", ylab = "Centered Year Squared")
```



2(g)

```
lm_2g <- lm(mpg ~ year_centered + year_centered_squared, data = auto)
summary(lm_2g)
```

```
##
## Call:
## lm(formula = mpg ~ year_centered + year_centered_squared, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.349  -5.109  -0.878   4.587  18.196
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           21.99061    0.46577  47.214  < 2e-16 ***
## year_centered          1.22778    0.08486  14.469  < 2e-16 ***
## year_centered_squared  0.11230    0.02542   4.419 1.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.23 on 394 degrees of freedom
## Multiple R-squared:  0.3694, Adjusted R-squared:  0.3662
## F-statistic: 115.4 on 2 and 394 DF,  p-value: < 2.2e-16
```

2(h)

$$
\begin{gathered}
\beta_0 = \gamma_0 - \gamma_1\bar{x} + \gamma_2\bar{x}^2 \\
\beta_1 = \gamma_1 - 2\gamma_2\bar{x} \\
\beta_2 = \gamma_2
\end{gathered}
$$

```
gamma_0 <- coef(lm_2g)[1]
gamma_1 <- coef(lm_2g)[2]
gamma_2 <- coef(lm_2g)[3]
mean_year <- mean(auto$year)
```

$$
\begin{gathered}
\beta_0 = 577.2522975 \\
\beta_1 = -15.8409008 \\
\beta_2 = 0.1123014
\end{gathered}
$$

Problem 3

① Show $h_{ij} = \frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{S_{xx}}$ and $h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}$

$H = X(X^TX)^{-1}X^T$

In simple linear regression, $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$.

$h_{ii} = (1 \ \ x_i)\,(X^TX)^{-1}\begin{pmatrix} 1 \\ x_i \end{pmatrix}$

$X^TX = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}$ $\quad (X^TX)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - n^2\bar{x}^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}$

$= \frac{1}{n\sum_{i=1}^n (x_i-\bar{x})^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}$

$= \frac{1}{S_{xx}}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$

$\therefore h_{ii} = (1 \ \ x_i)\cdot\frac{1}{S_{xx}}\cdot\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}\cdot\begin{pmatrix} 1 \\ x_i \end{pmatrix}$

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

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

$= \frac{1}{n} + (x_i-\bar{x})^2/S_{xx}$

$\therefore h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}$

Similarly, for h_{ij}. $h_{ij} = \frac{1}{S_{xx}}(1 \ \ x_i)\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}\begin{pmatrix} 1 \\ x_j \end{pmatrix}$

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

$= \frac{1}{S_{xx}}\left[\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2\right) + (x_ix_j - \bar{x}x_i - \bar{x}x_j + \bar{x}^2)\right]$

$= \frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{S_{xx}}$

② Show $\sum_{j=1}^{n} h_{ij} = 1$

$$\sum_{j=1}^{n} h_{ij} = \sum_{j=1}^{n} \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}$$

$$= 1 + \frac{(x_i - \bar{x})}{S_{xx}} \cdot \sum_{j=1}^{n} (x_j - \bar{x})$$

$$= 1 + \frac{(x_i - \bar{x})}{S_{xx}} \cdot \left(\sum_{j=1}^{n} x_j - n\bar{x} \right)$$

$$= 1 + \frac{(x_i - \bar{x})}{S_{xx}} \cdot 0$$

$$= 1.$$

Q4

a. if $\begin{cases} x_i = x' \text{ for } i=1,\cdots,n-1 \\ x_i = x'' \text{ for } i=n \end{cases}$, then $\bar{x} = \dfrac{(n-1)x'+x''}{n}$

$$S_{xx} = \Sigma(x_i-\bar{x})^2 = \Sigma\, x_i^2 - 2\bar{x}x_i + \bar{x}^2 = \Sigma x_i^2 - 2\bar{x}\Sigma x_i + n\bar{x}^2 = \Sigma x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \Sigma x_i^2 - n\bar{x}^2$$

$$\therefore S_{xx} = \sum_{i=1}^{n-1} x'^2 + x''^2 - n\cdot\left[\frac{(n-1)x'+x''}{n}\right]^2$$

$$= (n-1)x'^2 + x''^2 - \frac{1}{n}\left[(n-1)x'+x''\right]^2$$

$$= (n-1)x'^2 + x''^2 - \frac{1}{n}\left[(n-1)^2x'^2 + 2(n-1)x'x'' + x''^2\right]$$

$$= \frac{n-1}{n}\left(nx'^2 + \frac{n}{n-1}x''^2 - (n-1)x'^2 - 2x'x'' - \frac{1}{n-1}x''^2\right]$$

$$= \frac{n-1}{n}\left(x''^2 + x'^2 - 2x'x''\right)$$

$$= \frac{n-1}{n}\cdot(x'-x'')^2$$

b. $(x_i-\bar{x})(x_n-\bar{x}) = (x'-\bar{x})(x''-\bar{x})$

$$= \left(x' - \frac{(n-1)x'+x''}{n}\right)\left(x'' - \frac{(n-1)x'+x''}{n}\right)$$

$$= \frac{(nx'-nx'+x'-x'')(nx''-nx'+x'-x'')}{n^2}$$

$$= \frac{-(x'-x'')\left[(n-1)(x'-x'')\right]}{n^2}$$

$$= \frac{-(n-1)(x'-x'')^2}{n^2}$$

$$(x_n-\bar{x})^2 = \left(x'' - \frac{(n-1)x'+x''}{n}\right)^2$$

$$= \left(\frac{nx''-nx'+x'-x''}{n}\right)^2$$

$$= \left[\frac{(1-n)}{n}(x'-x'')\right]^2$$

$$= \left(\frac{n-1}{n}\right)^2(x'-x'')^2$$

(c) From Q3, $h_{in} = \frac{1}{n} + \frac{(x_i-\bar{x})(x_n-\bar{x})}{S_{xx}}$

$$= \frac{1}{n} + \frac{(n-1)(x'-x'')^2}{n^2} \cdot \frac{1}{(x'-x'')^2\left(\frac{n-1}{n}\right)}$$

$$= \frac{1}{n} - \left(\frac{n-1}{n^2}\right)\left(\frac{n}{n-1}\right)$$

$$= \frac{1}{n} - \frac{1}{n}$$

$$= 0.$$

$$h_{nn} = \frac{1}{n} + \frac{(x_n-\bar{x})^2}{S_{xx}}$$

$$= \frac{1}{n} + \left(\frac{n-1}{n}\right)(x'-x'')^2 \cdot \frac{1}{(x'-x'')^2 \cdot \left(\frac{n-1}{n}\right)}$$

$$= \frac{1}{n} + \left(\frac{n-1}{n}\right)^2 \cdot \left(\frac{n}{n-1}\right)$$

$$= \frac{1}{n} + \frac{n-1}{n}$$

$$= 1$$

Q5

Proof : $E(\hat{\beta}_{GLS}) = \beta$ and $Cov(\hat{\beta}_{GLS}) = (X'\Sigma^{-1}X)^{-1} = \sigma^2 (X'WX)^{-1}$

$$\hat{\beta}_{GLS} = (X'WX)^{-1} X'WY$$

$$\Rightarrow \hat{\beta}_{GLS} = (X'WX)^{-1} X'W(\beta X + \epsilon)$$

$$= \beta \cancel{(X'WX)^{-1}(X'WX)} + (X'WX)^{-1} X'W \cdot \epsilon$$

$$= \beta + (X'WX)^{-1} X'W \epsilon$$

$$E[\hat{\beta}_{GLS}] = E[\beta] + E[(X'WX)^{-1} X'W \epsilon]$$

$$= \beta + (X'WX)^{-1} X'W \cancel{E[\epsilon]}^{0}$$ by assumption

$$= \beta$$

$\Rightarrow$ given $Var(AX) = A\,Var(X)\,A'$, we let.

$$A = (X'WX)^{-1} X'W$$

then: $Var(\hat{\beta}_{GLS}) = (X'WX)^{-1} X'W\,Var(Y|X)\,WX(X'WX)^{-1}$

$$\because Var(Y|X) = \sigma^2 W^{-1}$$

$$\therefore Var(\hat{\beta}_{GLS}|X) = \sigma^2 (X'WX)^{-1} X'WX (X'WX)^{-1}$$

$$= \sigma^2 (X'WX)^{-1}$$

$$\because Cov(X, X) = Var(X)$$

$$\therefore Cov(\hat{\beta}_{GLS}) = \sigma^2 (X'WX)^{-1}$$

Q6 (a) $\bar{y}_w$ will be an unbiased estimator for μ when:

$$E(\bar{y}_w) = \mu$$

$$E(\bar{y}_w) = E(w_1 Y_1 + w_2 Y_2) = E(w_1 Y_1) + E(w_2 Y_2)$$

$$= w_1 E(Y_1) + w_2 E(Y_2)$$

$$= w_1 \mu + w_2 \mu$$

$$= (w_1 + w_2)\mu$$

when $w_1 + w_2 = 1$, $E(\bar{y}_w) = \mu$, therefore is unbiased

(b) $Var(\bar{y}_w) = Var(w_1 Y_1 + w_2 Y_2)$

$$= w_1^2 Var(Y_1) + w_2^2 Var(Y_2)$$

$$= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 \xcancel{Cov(Y_1, Y_2)} \rightarrow 0$$

$$= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2$$

(c) $Var(\bar{y}_w)$ is minimized when $\dfrac{\partial V(\bar{y}_w)}{\partial w_i} = 0$

$$\frac{\partial V(\bar{y}_w)}{\partial w_i} = \frac{\partial (w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2)}{\partial w_1} + \frac{\partial (w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2)}{\partial w_2}$$

$$= 2\sigma_1^2 w_1 + 2\sigma_2^2 w_2$$

let $2\sigma_1^2 w_1 + 2\sigma_2^2 w_2 = 0$

$$\sigma_1^2 w_1 = -\sigma_2^2 w_2 \Rightarrow \begin{array}{l} w_1 = -(\sigma_2^2 w_2) \cdot \dfrac{1}{\sigma_1^2} \\ w_2 = -(\sigma_1^2 w_1) \cdot \dfrac{1}{\sigma_2^2} \end{array}$$

$\Rightarrow$ $Var(\bar{y}_w)$ is minimized when $w_i \propto 1/\sigma_i^2$

Question 7

If we imagine a plot of the data, we think it will resemble f(x) = 1/x. Quickly decreasing at first, then slowly, then plateauing. We think: theft, battery, assault, narcotics, and homicide will affect the demand whereas deceptive practice, burglary, and criminal trespassing will be somewhat or completely independent. We came up with this list because we think all crimes that're committed near the bike station will have a much stronger effect on the demand for bikes. Theft, battery, assault, narcotics, and homicide will make the renter feel more concerned about their personal safety. This will decrease their willingness to go to the station and rent a bike. The other crimes wouldn't occur near the bike station. For this reason, we think they won't have much of and effect on bike rentals.

Some crimes may have association. For example, a theft may escalate to an assault or even homicide because the person being stolen from might try to defend themselves and get hurt. Another example could be; someone taking narcotics would logically be more likely to be involved more crimes of any type. Some might be independent because there's no way for the situation to escalate. For example, deceptive practice probably won't be related to theft, homicide, or criminal trespassing.

The actual results could be different. We are thinking logically. People renting bikes won't have perfect information, think perfectly logically, and we have preconceived notions about people we don't understand. There are many other factors besides crime that affect bike demand such as geography, weather, etc. We aren't accounting for any of these.