MATH 8050 Homework 3

Parampreet Singh

Due March 15 at 11:59pm

Instructions: For the problems that follow, please show all of your work. For problems that you solve using software, please provide the code you used either with the problem or in an appendix at the end of your homework document. To submit your answers for homework, please upload a **single** PDF to Canvas.

Problem 1

We will study the the `trees` data set contained in R. The `trees` data set contains three measurements taken on $n = 31$ cherry trees; we will study how the `Height` and `Volume` variables are related to each other. To load this data into R, run `data("trees")` in your R console.

(a) Treating `Height` as the dependent (response) variable, plot an appropriate scatterplot of `Volume` versus `Height`.
(b) Fit a simple linear regression model to the data, treating `Height` as the dependent variable. Plot the line from this model over your scatterplot.

(c) What are the values of the least squares estimates of β_0 and β_1? Interpret these estimates.
(d) What is the value of the least squares estimate for σ^2 for this regression fit?
(e) What is the value of the maximum likelihood estimate for σ^2 for this regression fit?
(f) Provide 90% confidence intervals for β_0 and β_1.
(g) At the $\alpha = 0.05$ level, perform a test for the following hypotheses:

$$H_0 : \beta_1 \geq 0.5$$

$$H_A : \beta_1 < 0.5$$

Problem 2

Recall that the least squares estiamtes for β_0 and β_1 are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \qquad\qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Since $\hat{\beta}_0$ depends on $\hat{\beta}_1$, the two estimates are not independent. Find the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$, where:

$$\begin{aligned} \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= E[(\hat{\beta}_0 - E(\hat{\beta}_0))(\hat{\beta}_1 - E(\hat{\beta}_1))] \\ &= E(\hat{\beta}_0\hat{\beta}_1) - E(\hat{\beta}_0)E(\hat{\beta}_1) \end{aligned}$$

Problem 3

We will study a data set containing the heights (in inches) of 898 people and the heights of each of their biological parents. To get the data into R, download `family_heights.txt` from Canvas and run the following (you will need to replace ~/ with the file path of the location where you saved the data)

```r
dat.heights <- read.table("~/family_heights.txt", header = T)
```

(a) Provide a scatterplot with the simple linear regression line for the family heights data using only fathers' heights versus sons' heights (as the response variable). To more easily work with only the son' heights, you can subset the data set with the following:

```r
son.heights <- dat.heights[dat.heights$Gender=="M",]
```

(b) List and interpret the slope estimate in the simple linear regression model with sons' heights as the response variable and fathers' heights as the independent variable.

(c) For the same subset of data, and using your SLR fit, plot the residuals. Do you see any pattern in the residuals?

(d) For the same subset of data, and using your SLR fit, plot a Normal QQ plot of the residuals. Does the SLR model seem reasonable?

(e) Provide an estimate and 95% confidence interval for the average height of sons whose fathers are 70 inches tall, i.e. for $E(y|x = 70)$.

(f) Using the SLR model you found above, predict the height of a son whose father is 70 inches. Provide a 95% prediction interval.

Problem 4

Suppose we collect n pairs of observations (x_i, y_i), fit a simple linear regression model, and obtain estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2 = \text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$. It is often necessary to scale one of the variables, e.g. to change the units from feet to meters.

(a) Suppose we replace each x_i with $c \times x_i$, where c is some constant. How do the regression estimates change with respect to the original $\hat{\beta}_0$, $\hat{\beta}_1$, and MSE?

(b) Now suppose that we instead scale the response variable, replacing y_i with $a \times y_i$, where a is some constant. Use the original unscaled x_i's. How do the regression estimates change with respect to the original $\hat{\beta}_0$, $\hat{\beta}_1$, and MSE now?

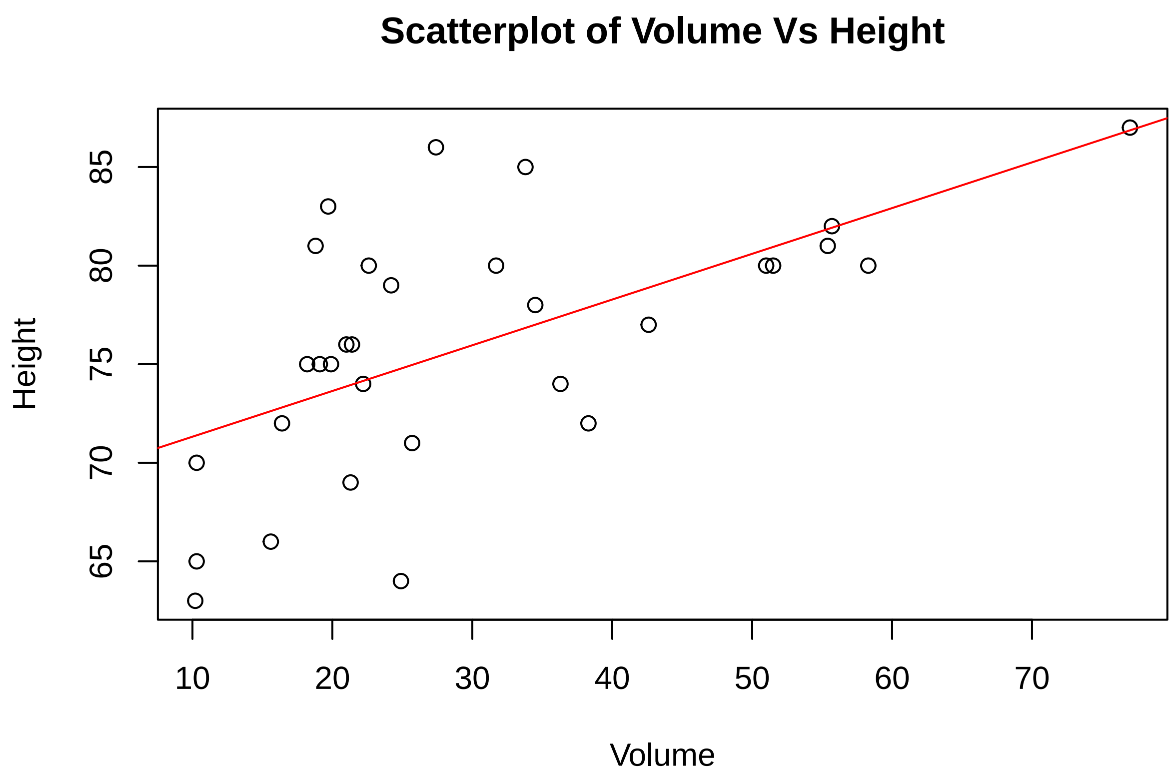
Data Analysis - HW3

Parampreet Singh

2024-03-14

Problem 1

```r
# Load the trees dataset
data("trees")

# A: Scatter Plot of Volume Vs Height
plot(trees$Volume, trees$Height,
     main = "Scatterplot of Volume Vs Height", xlab = "Volume", ylab = "Height")

# B: Fit SLR. Plot model over the scatter plot
model <- lm(Height ~ Volume, data = trees)
abline(model, col="red")
```




```r
# C: Least squares estimates of  0 and  1
beta_estimates <- coef(model)
print(beta_estimates)
```

```
## (Intercept)      Volume
##  69.0033557   0.2318999
```

```r
# D: LSE for sigma_squared
lse_sigma_squared <- sum(model$residuals^2) / model$df.residual
print(paste("Least squares estimate for  2: ", lse_sigma_squared))
```

```
## [1] "Least squares estimate for  2:  26.9680888634519"
```

```r
# E: MLE for sigma_squared
mle_sigma_squared <- sum(model$residuals^2) / (model$df.residual + 2)
print(paste("Maximum likelihood estimate for  2: ", mle_sigma_squared))
```

```
## [1] "Maximum likelihood estimate for  2:  25.228212162584"
```

```r
# F: 90% Confidence intervals for  0 and  1
print("90% Confidence intervals for  0 and  1:")
```

```
## [1] "90% Confidence intervals for  0 and  1:"
```

```r
confint(model, level = 0.90)
```

```
##                    5 %       95 %
## (Intercept) 65.6485507 72.3581608
## Volume       0.1338955  0.3299043
```

```r
# G: Hypothesis test for  1

# Null hypothesis:  1   0.5
# Alternative hypothesis:  1 < 0.5
# This will be a one-tailed test

hypothesis <- summary(model)$coefficients[2,]
p_value <- if (hypothesis[1] < 0.5) hypothesis[4]/2 else 1 - hypothesis[4]/2
result <- if (p_value< 0.05) "Reject Null Hypothesis" else "FTR Null Hypothesis"

print(paste("P-value: ", p_value))
```

```
## [1] "P-value:  0.000189191173959245"
```

```r
print(paste("Hypothesis Decision: ", result))
```

```
## [1] "Hypothesis Decision:  Reject Null Hypothesis"
```

Problem 2

we have, $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = E[(\hat{\beta}_0 - E(\hat{\beta}_0))(\hat{\beta}_1 - E(\hat{\beta}_1))]$$

$\cancel{= E[\hat{\beta}_0 \hat{\beta}_1 - \beta_0 \beta_1]}$ $\quad E(\hat{\beta}_0) = \beta_0 \quad E(\hat{\beta}_1) = \beta_1$

$$= E[(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)]$$

Substitute $\hat{\beta}_0$,

$$= E[(\bar{y} - \hat{\beta}_1 \bar{x} - \beta_0)(\hat{\beta}_1 - \beta_1)]$$

$$= E[\bar{y}\hat{\beta}_1 - \hat{\beta}_1^2 \bar{x} - \beta_0 \hat{\beta}_1 - \bar{y}\beta_1 + \beta_1 \hat{\beta}_1 \bar{x} + \beta_0 \beta_1]$$

Using linearity of expectation,

$$= E(\bar{y}\hat{\beta}_1) - E(\hat{\beta}_1^2)\bar{x} - \beta_0 E(\hat{\beta}_1) - \bar{y}\beta_1 + \beta \bar{x} E(\hat{\beta}_1) + \beta_0 \beta_1$$

$$= \cancel{\bar{y}\beta_1} - E(\hat{\beta}_1^2)\bar{x} - \cancel{\beta_0 \beta_1} - \cancel{\bar{y}\beta_1} + \beta_1^2 \bar{x} + \cancel{\beta_0 \beta_1}$$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -E(\hat{\beta}_1^2)\bar{x} + \beta_1^2 \bar{x}$$

Now we know,

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \quad \text{and} \quad E(\hat{\beta}_1^2) = \text{Var}(\hat{\beta}_1) + [E(\hat{\beta}_1)]^2$$

$$E(\hat{\beta}_1^2) = \frac{\sigma^2}{S_{xx}} + \hat{\beta}_1^2$$

Scanned with CamScanner

substituting our $E(\hat{\beta}_1^2)$ into our covariance equation

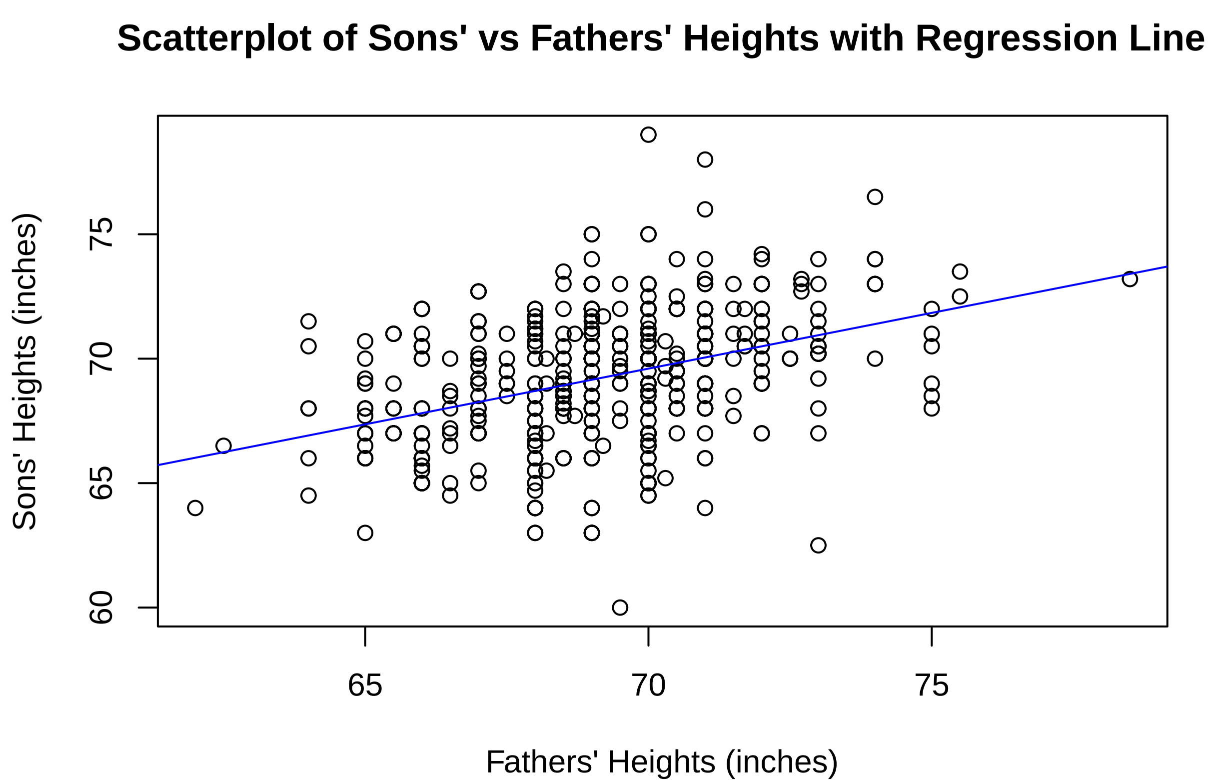
$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\left(\frac{\sigma^2}{S_{pp}} + \beta_1^2\right)\bar{x} + \beta_1^2 \bar{x}$$

$$= -\frac{\sigma^2}{S_{pp}}\bar{x} - \cancel{\beta_1^2 \bar{x}} + \cancel{\beta_1^2 \bar{x}}$$

$$\boxed{\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\left(\frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)\bar{x}}$$

PROBLEM 3

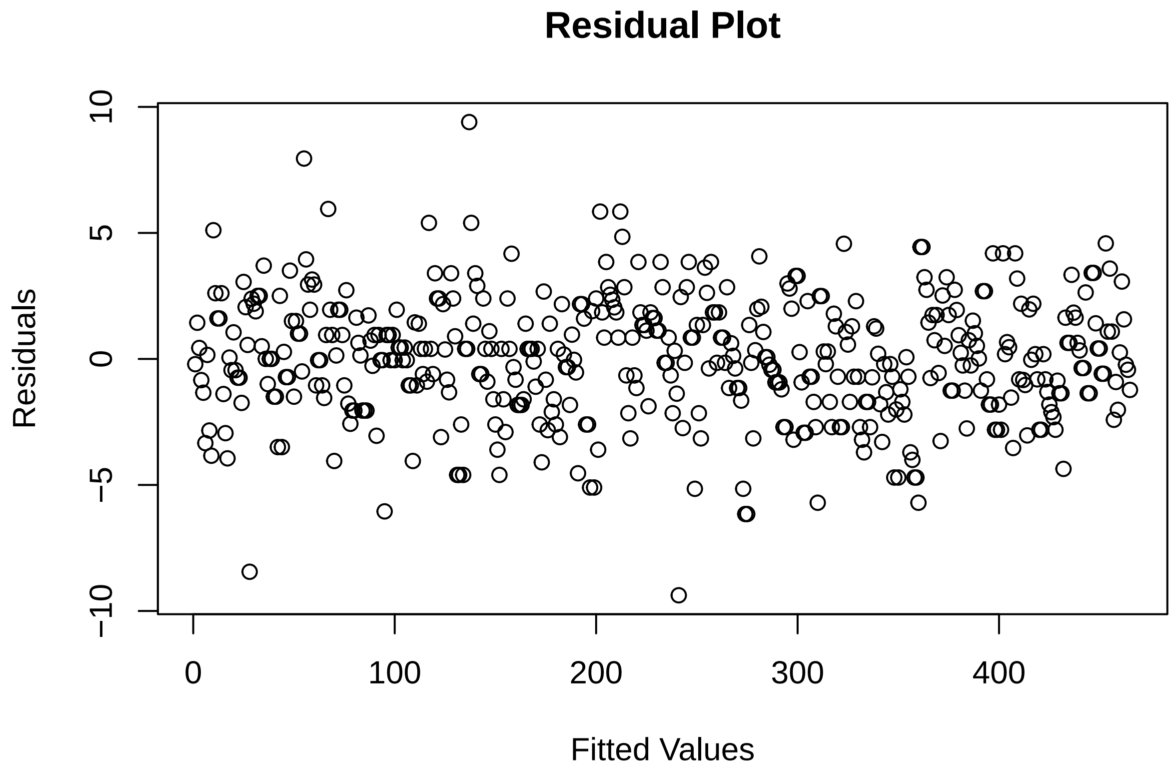
```
# Read the data into R
dat.heights <- read.table("~/family_heights.txt",
                          header = T)

# Subset the data for sons only
son.heights <- dat.heights[dat.heights$Gender == "M", ]

# A: Scatter plot with simple linear regression line for sons'
# heights based on fathers' heights
model <- lm(Height ~ Father, data = son.heights)
plot(son.heights$Father, son.heights$Height,
     main = "Scatterplot of Sons' vs Fathers' Heights with Regression Line",
     xlab = "Fathers' Heights (inches)",
     ylab = "Sons' Heights (inches)")
abline(model, col = 'blue')
```



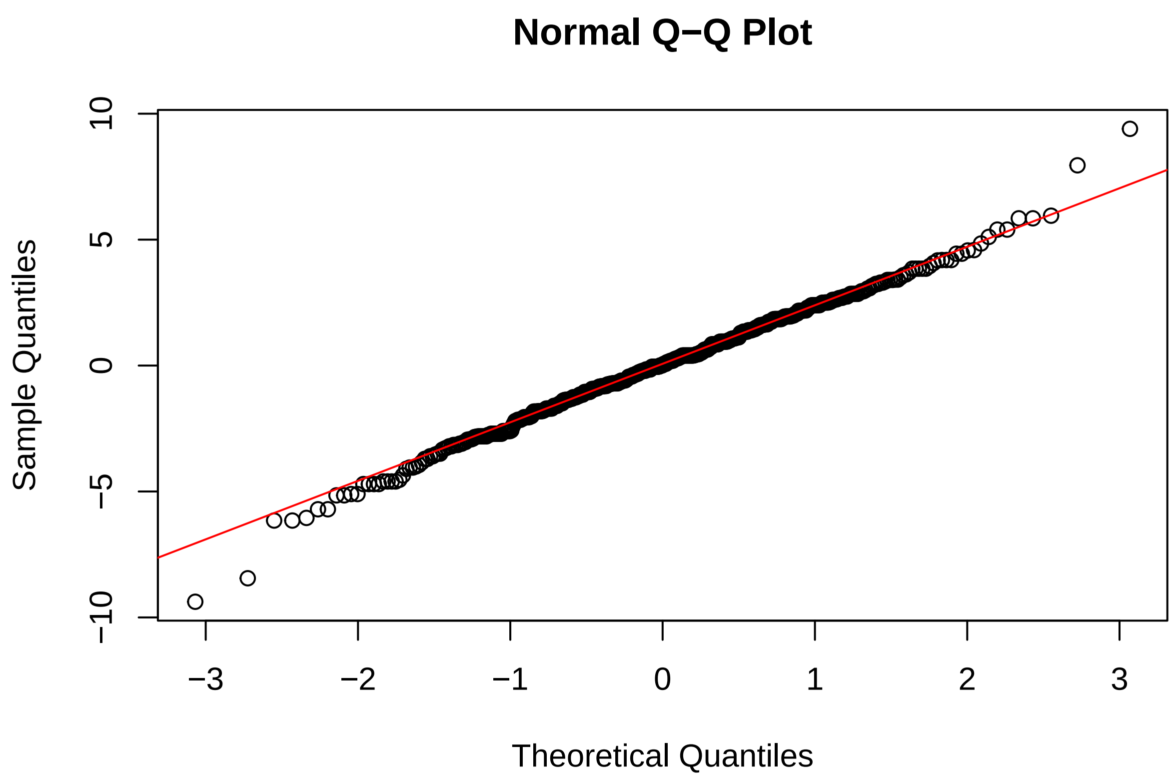

```
# B: List and interpret the slope estimate
# Interpretation: The slope estimate represents the expected change in sons'
# height for a one-inch increase in fathers' height.
slope <- coef(model)['Father']
cat("Slope Estimate: ", slope)
```

```
## Slope Estimate:  0.4477479
```

```r
# C: Plot the residuals and check for any pattern
residuals <- residuals(model)
plot(residuals, main = "Residual Plot",
     xlab = "Fitted Values", ylab = "Residuals")
```




```r
# Interpretation: The residuals bounce randomly across 0. This implies our
# linearity assumption holds. Also, the residuals form a horizontal band around
# 0, suggesting the error variances are equal.

# D: Normal QQ plot and its interpretation
qqnorm(residuals)
qqline(residuals, col = "red")
```




```
# Interpretation: The points lie closely to the qq reference line. Hence, the
# residuals follow a normal distribution and our normality of errors assumption
# holds.

# E: Provide an estimate and 95% confidence interval for the average height of
# sons whose fathers are 70 inches tall.
new.data <- data.frame(Father = 70)
confidence_prediction <- predict(model, new.data, interval = "confidence")
cat("95% CI for the average height of sons with father's height 70 inches:", confidence_prediction, "\n
```

```
## 95% CI for the average height of sons with father's height 70 inches: 69.60127 69.3663 69.83623
```

```
# F: Predict the height of a son whose father is 70 inches tall with a 95%
# prediction interval
prediction_interval <- predict(model, new.data, interval = "prediction")
cat("95% PI for the height of a son with father's height 70 inches:",
    prediction_interval, "\n")
```

```
## 95% PI for the height of a son with father's height 70 inches: 69.60127 64.83138 74.37116
```

Problem 4

For a simple linear regression model, we have

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where,
$y_i \to$ Response variable
$x_i \to$ Independent variable
$\beta_0 \to$ Intercept
$\beta_1 \to$ slope
$\varepsilon_i \to$ Random Error

The estimates for β_0, β_1 and MSE are

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$MSE = \frac{\Sigma (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \varepsilon_i)}{n-2}$$

(a) If we replace x_i with $c \times x_i$, we have

$$y_i = \beta_0 + \beta_1 (c x_i) + \varepsilon_i$$

$$y_i = \beta_0 + (c\beta_1) x_i + \varepsilon_i$$

We can write $\beta_1' = c\beta_1$

we have, $y_i = \beta_0 + \beta_1' x_i + \varepsilon_i$

Therefore, we have no affect on $\hat{\beta}_0$.

The new $\hat{\beta}_1' = \frac{\hat{\beta}_1}{c}$, to counter the affect in x_i and hence maintain the same proportional relationship of model

Since both $\hat{\beta}_1$ & x_i are scaled with c, there is no effect on the spread of residuals and hence there is no change in MSE.

(b) As we know,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Now, multiply LHS with a

$$a \times y_i$$

To maintain the proportional relationship we do the same with R.H.S

$$R.H.S: \quad a(\beta_0 + \beta_1 x_i) + a\varepsilon_i$$

$$\therefore \boxed{\begin{aligned} \hat{\beta}_0' &= a \cdot \hat{\beta}_0 \\ \hat{\beta}_1' &= a \cdot \hat{\beta}_1 \end{aligned}}$$

for MSE, which is a measure of variance we have

$$MSE' = \frac{\Sigma(a y_i - a(\beta_0 + \beta_1 x_i))^2}{n-2}$$

$$\boxed{MSE' = a^2 . MSE}$$

These results show that scaling the predictor variable affects only the slope estimate, while scaling the response variable affects all parameters and MSE.