Math 423 Assignment 1

Jonathan Pearce 260672004

10/9/2017

Question a.

Data set 1

```
#starter code, read data
file1<-"http://www.math.mcgill.ca/yyang/regression/data/a1-1.txt"
data1<-read.table(file1,header=TRUE)
x1<-data1$x
y<-data1$y
```

(i) Parameter Estimates

```
fit.ds1<-lm(y~x1)
#summary(fit.ds1)
#save to variable so I can embed values in text in document below
z1<-coef(fit.ds1)
z1
```

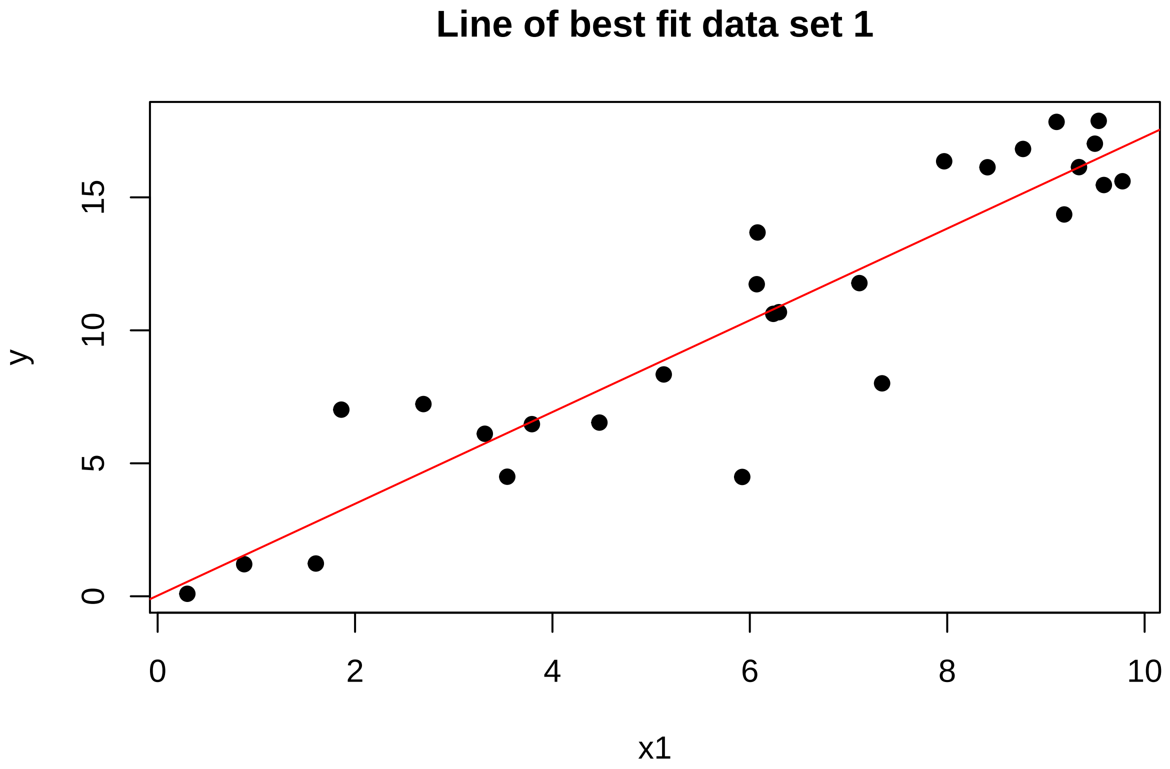
```
## (Intercept)          x1
##  0.02676552  1.72512091
```

The lm function calculates the following parameter estimates. The intercept $\widehat{\beta}_0 = 0.0267655$ and the slope $\widehat{\beta}_1 = 1.7251209$. It follows that the line of best fit is

$$y = 0.027 + 1.725x_1.$$

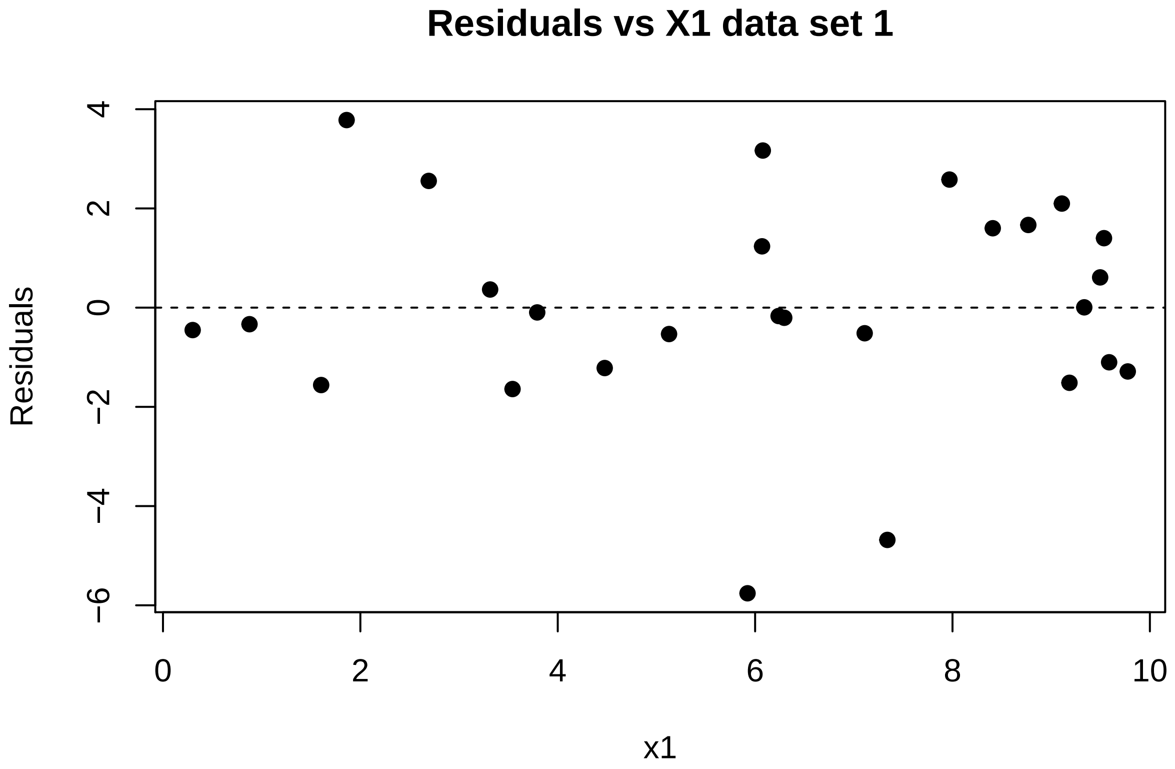
(ii) Line of Best Fit Plot

```
#Plot data and line of best fit
plot(x1,y,pch=19,xlab='x1',ylab='y')
abline(coef(fit.ds1),col='red')
title('Line of best fit data set 1')
```



(iii) Residual Plot

```r
RP.resids<-residuals(fit.ds1)
plot(x1,RP.resids,xlab='x1',ylab='Residuals',pch=19)
abline(h=0,lty=2)
title('Residuals vs X1 data set 1')
```



```
#sum(RP.resids)
length(x1)
```

```
## [1] 27
```

(iv) There are only 27 observations in our data set which makes it difficult to confidently determine whether the assumptions of least squares fitting are met. Withstanding the limited data, it seems that $E(\epsilon) = 0$, there are no outliers and the observations seem relativly balanced above and below the line $y = 0$. Further it appears that $Var(\epsilon)$ is constant, the variance of the residual values seems constant across the entire range of x_1 values, therefore the assumptions are met.

Data set 2

```
#starter code, read data
file1<-"http://www.math.mcgill.ca/yyang/regression/data/a1-2.txt"
data2<-read.table(file1,header=TRUE)
x1<-data2$x
y<-data2$y
```

(i) Parameter Estimates

```
fit.ds2<-lm(y~x1)
#summary(fit.ds2)
#save to variable so I can embed values in text in document below
z<-coef(fit.ds2)
z
```

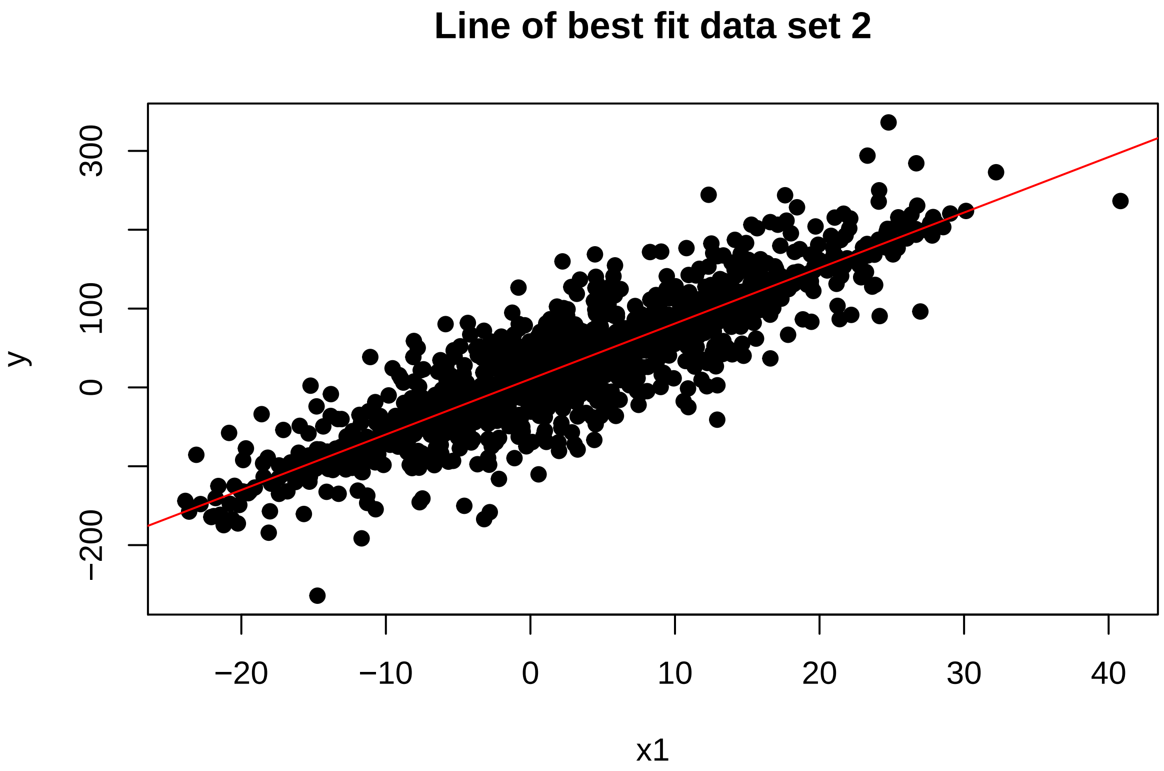
```
## (Intercept)          x1
##   10.660853    7.037952
```

The lm function calculates the following parameter estimates. The intercept $\widehat{\beta}_0 = 10.6608534$ and the slope $\widehat{\beta}_1 = 7.0379523$. It follows that the line of best fit is

$$y = 10.661 + 7.038x_1.$$

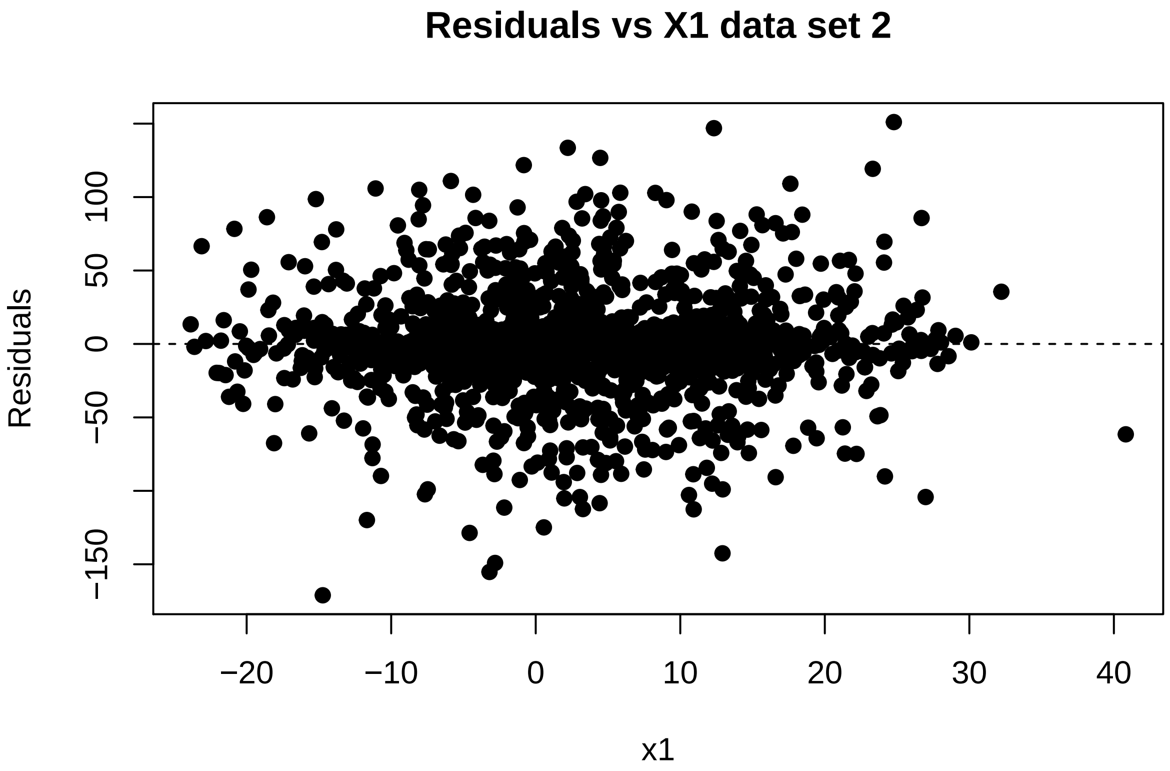
(ii) Line of Best Fit Plot

```
#Plot data and line of best fit
plot(x1,y,pch=19,xlab='x1',ylab='y')
abline(coef(fit.ds2),col='red')
title('Line of best fit data set 2')
```



(iii) Residual Plot

```
RP.resids<-residuals(fit.ds2)
plot(x1,RP.resids,xlab='x1',ylab='Residuals',pch=19)
abline(h=0,lty=2)
title('Residuals vs X1 data set 2')
```



```
#sum(RP.resids)
#length(x1)
```

(iv) Observing the residual plot, it is safe to conclude that $E(\epsilon) = 0$ since the residual points are densely populated around the line $y = 0$ and appear to be balanced above and below the line. There is a minor concern regarding the variance of the residuals, from $x_1 = -25$ to $x_1 = -15$ it appears that the variance of the residual values is slightly less than the other residual values of the graph, since very few of these residual values go above 100 or below -50. However, it appears the residual values on the interval $[-15, 40]$ have constant variance and since the interval $[-25, -15]$ has a relativly low density of data points and the variance appears to be only slightly smaller I would be inclined to conclude that the residual values have a constant variance and therefore all assumptions are met.

Data set 3

```
#starter code, read data
file1<-"http://www.math.mcgill.ca/yyang/regression/data/a1-3.txt"
data3<-read.table(file1,header=TRUE)
x1<-data3$x
y<-data3$y
```

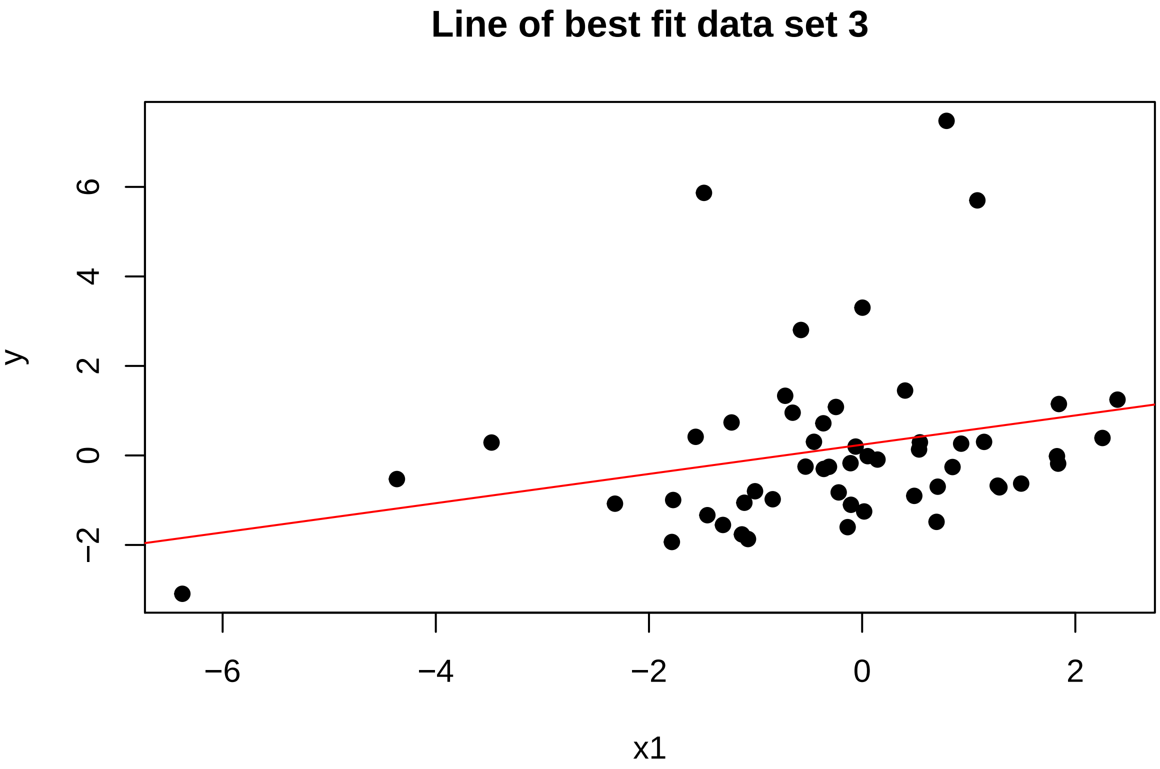
(i) Parameter Estimates

```
fit.ds3<-lm(y~x1)
#summary(fit.ds3)
#save to variable so I can embed values in text in document below
z<-coef(fit.ds3)
z
```

```
## (Intercept)          x1
##   0.2403328   0.3267628
```

(ii) Line of Best Fit Plot

```
#Plot data and line of best fit
plot(x1,y,pch=19,xlab='x1',ylab='y')
abline(coef(fit.ds3),col='red')
title('Line of best fit data set 3')
```

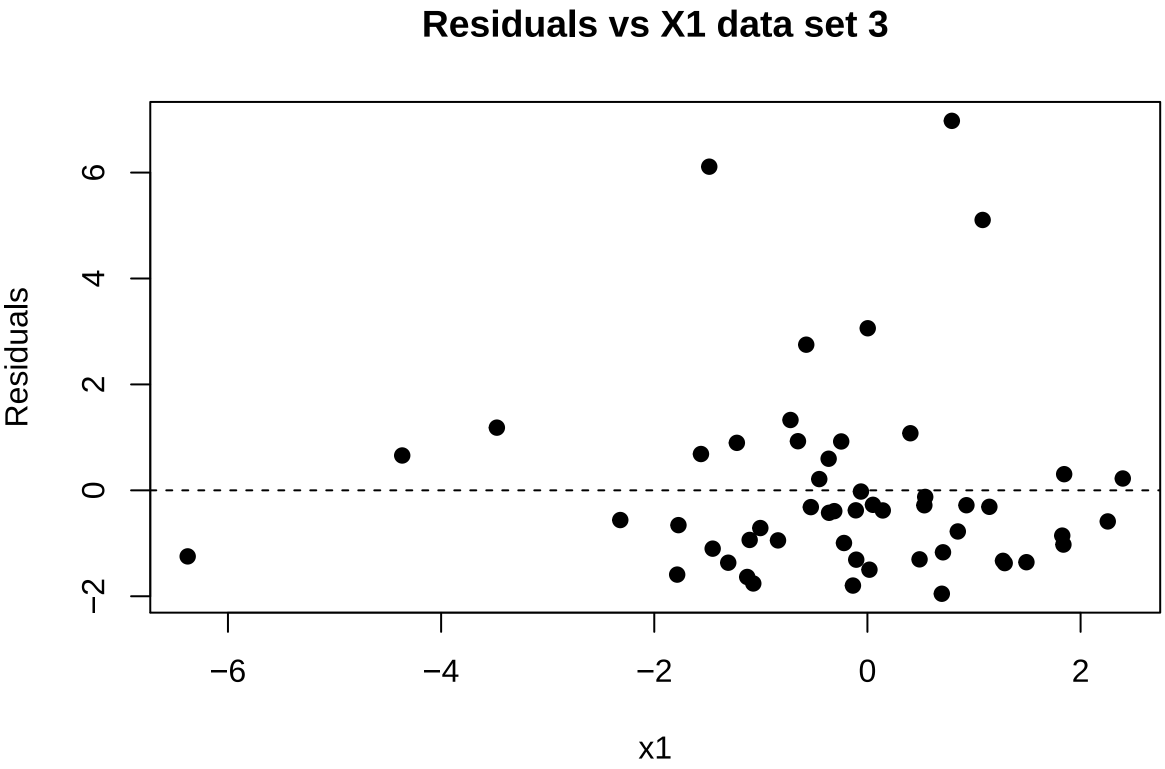


The lm function calculates the following parameter estimates. The intercept $\widehat{\beta}_0 = 0.2403328$ and the slope $\widehat{\beta}_1 = 0.3267628$. It follows that the line of best fit is

$$y = 0.24 + 0.327x_1.$$

(iii) Residual Plot

```
RP.resids<-residuals(fit.ds3)
plot(x1,RP.resids,xlab='x1',ylab='Residuals',pch=19)
abline(h=0,lty=2)
title('Residuals vs X1 data set 3')
```



```
#sum(RP.resids)
#length(x1)
```

(iv) In this data set the issue arises with the assumption of constant variance among the residual values. The sparsity of the 5 residuals greater than 2 (especially the 3 residuals greater than 4) creates too much uncertainity with regards to concluding a constant variance. The 3 points with residual values greater than 4 are clear outliers and violate the assumption. As a note it is very possible that $E(\epsilon) = 0$.

Question b.

I will demonstrate the tranformations in (i) and (ii) numerically using data set 1. Recall for data set 1 we had the intercept $\widehat{\beta}_0 = 0.0267655$ and the slop $\widehat{\beta}_1 = 1.7251209$.

Part i.

let m = 5.

```
x1<-data1$x
y<-data1$y
m<-5
#Shift x1 data
x1<-x1-m
#refit line of best fit with data shifted
fit.ds1<-lm(y~x1)
#summary(fit.ds1)
ls<-coef(fit.ds1)
ls
```

```
## (Intercept)          x1
##    8.652370    1.725121
```

$$\widehat{\beta}_0^{new} = 8.6523701 = 0.0267655 + 5 * 1.7251209 = \widehat{\beta}_0 + m * \widehat{\beta}_1$$
$$\widehat{\beta}_1^{new} = 1.7251209 = \widehat{\beta}_1$$

We conclude numerically, $\widehat{\beta}_0^{new} = \widehat{\beta}_0 + m * \widehat{\beta}_1$ and $\widehat{\beta}_1^{new} = \widehat{\beta}_1$ for data set 1

Part ii.

let l = 2.2.

```
x1<-data1$x
y<-data1$y
l<-2.2
#Shift x1 data
x1<-x1*l
#refit line of best fit with data shifted
fit.ds1<-lm(y~x1)
#summary(fit.ds1)
ss<-coef(fit.ds1)
ss
```

```
## (Intercept)          x1
##  0.02676552  0.78414587
```

$$\widehat{\beta}_0^{new} = 0.0267655 = \widehat{\beta}_0$$

$$\widehat{\beta}_1^{new} = 0.7841459 = \frac{1.7251209}{2.2} = \frac{1}{l}\widehat{\beta}_1$$

We conclude numerically, $\widehat{\beta}_0^{new} = \widehat{\beta}_0$ and $\widehat{\beta}_1^{new} = \frac{1}{l}\widehat{\beta}_1$ for data set 1

Part iii.

Location Shift Data

From the question we have $x_{i1}^{new} = x_{i1} - m$ and we also have our model $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$. Expressing the model in terms of the new transformed data we get;

$$Y_i = \beta_0^{new} + \beta_1^{new} x_{i1}^{new} + \epsilon_i = \beta_0^{new} + \beta_1^{new}(x_{i1} - m) + \epsilon_i = (\beta_0^{new} - m\beta_1^{new}) + \beta_1^{new} x_{i1} + \epsilon_i$$

Therefore we know,

$$\beta_1^{new} = \beta_1$$

and

$$\beta_0 = \beta_0^{new} - m\beta_1^{new} \implies \beta_0^{new} = \beta_0 + m\beta_1$$

let $M = \begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix}$, we now have the following,

$$\begin{bmatrix} \beta_0^{new} \\ \beta_1^{new} \end{bmatrix} = \begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = M \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Now we have $\widehat{\beta}^{new} = M\widehat{\beta}$ it follows,

$$E[\widehat{\beta}^{new}|X] = E[M\widehat{\beta}|X] = ME[\widehat{\beta}|X] = M\widehat{\beta}X$$

Therefore we conclude $\widehat{\beta}_1^{new} = \widehat{\beta}_1$ and $\widehat{\beta}_0^{new} = \widehat{\beta}_0 + m\widehat{\beta}_1$. This agrees with what I showed numerically in Part i. Now we will compute the variance.

$$Var(\widehat{\beta}^{new}|X) = Var(M\widehat{\beta}|X) = MVar(\widehat{\beta}|X)M^T = M\sigma^2(X^TX)^{-1}M^T$$

$$Var(\widehat{\beta}_1^{new}|X) = Var(\widehat{\beta}_1|X) = \frac{\sigma^2}{S_{xx}}$$

$$Var(\widehat{\beta}_0^{new}|X) = Var(\widehat{\beta}_0 + m\widehat{\beta}_1|X) = Var(\widehat{\beta}_0|X) + m^2Var(\widehat{\beta}_1|X) + 2mCov(\widehat{\beta}_0, \widehat{\beta}_1|X)$$

$$= \frac{\sigma^2 \sum x_{i1}^2}{nS_{xx}} + \frac{m^2\sigma^2}{S_{xx}} + \frac{\bar{x}_i\sigma^2}{S_{xx}} = \frac{\sigma^2 \sum x_{i1}^2 + nm^2\sigma^2 + n\bar{x}_i\sigma^2}{nS_{xx}}$$

For the location shift data, the estimators remain unbiased, the variance for $\widehat{\beta}_1$ is the same and the adjusted variance for $\widehat{\beta}_0$ is expressed above.

Rescaled Data

From the question we have $x_{i1}^{new} = lx_{i1}$ and we also have our model $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$. Expressing the model in terms of the new transformed data we get;

$$Y_i = \beta_0^{new} + \beta_1^{new}x_{i1}^{new} + \epsilon_i = \beta_0^{new} + \beta_1^{new}(lx_{i1}) + \epsilon_i = \beta_0^{new} + l\beta_1^{new}x_{i1} + \epsilon_i$$

Therefore we know,

$$\beta_0^{new} = \beta_0$$

and

$$\beta_1 = l\beta_1^{new} \implies \beta_1^{new} = \frac{1}{l}\beta_1$$

let $L = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{l} \end{bmatrix}$, we now have the following,

$$\begin{bmatrix} \beta_0^{new} \\ \beta_1^{new} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{l} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = L \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Now we have $\widehat{\beta}^{new} = L\widehat{\beta}$ it follows,

$$E[\widehat{\beta}^{new}|X] = E[L\widehat{\beta}|X] = LE[\widehat{\beta}|X] = L\widehat{\beta}X$$

Therefore we conclude $\widehat{\beta}_0^{new} = \widehat{\beta}_0$ and $\widehat{\beta}_1^{new} = \frac{1}{l}\widehat{\beta}_1$. This agrees with what I showed numerically in Part ii. Now we will compute the variance.

$$Var(\widehat{\beta}^{new}|X) = Var(L\widehat{\beta}|X) = LVar(\widehat{\beta}|X)L^T = L\sigma^2(X^TX)^{-1}L^T$$

$$Var(\widehat{\beta}_0^{new}|X) = Var(\widehat{\beta}_0|X) = \frac{\sigma^2 \sum x_{i1}^2}{nS_{xx}}$$

$$Var(\widehat{\beta}_1^{new}|X) = Var(\frac{1}{l}\widehat{\beta}_1|X) = \frac{1}{l^2}Var(\widehat{\beta}_1|X) = \frac{\sigma^2}{l^2S_{xx}}$$

For the rescaled data, the estimators remain unbiased, the variance for $\widehat{\beta}_0$ is the same and the adjusted variance for $\widehat{\beta}_1$ is expressed above.