Math 423

Linear Regression

Homework I

Frédéric Boileau

Prof. David A. Stephens

2nd October 2016

1 Data Analysis

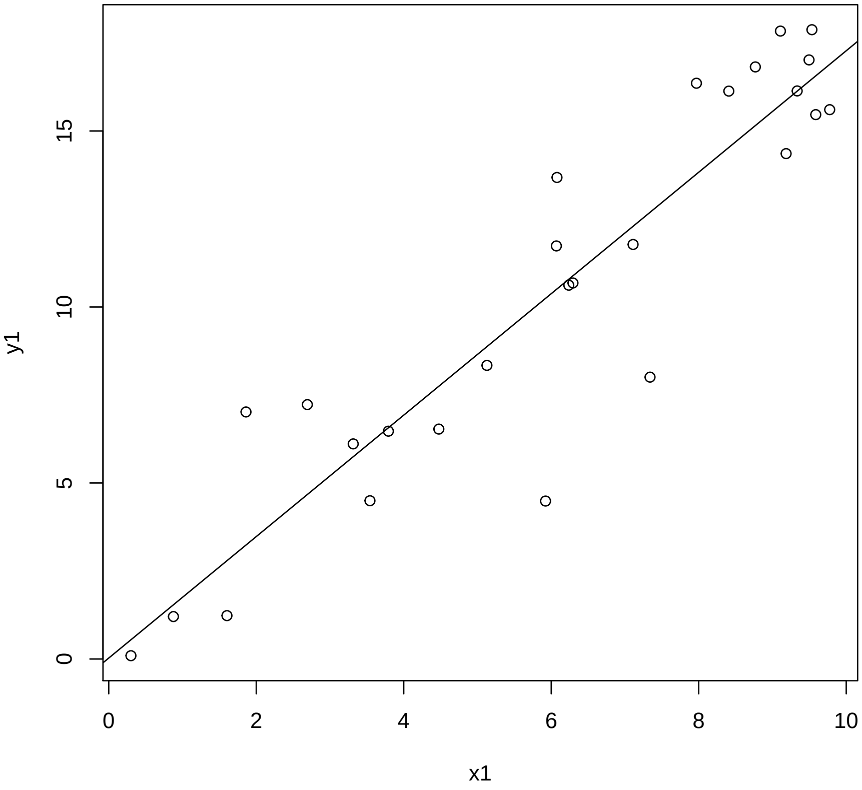
```
#Read in data set 1
library(lmtest)
```

```
## Warning:  package ’lmtest’ was built under R version 3.3.1
## Loading required package:  zoo
## Warning:  package ’zoo’ was built under R version 3.3.1
##
## Attaching package:  ’zoo’
## The following objects are masked from ’package:base’:
##
##    as.Date, as.Date.numeric
```

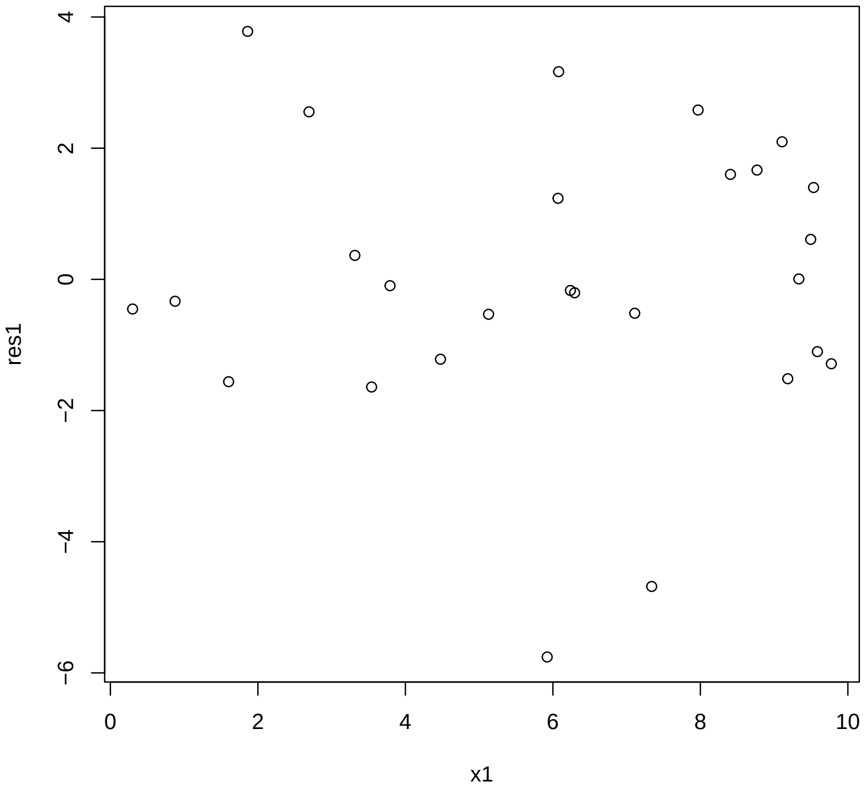
```
file1<-"http://www.math.mcgill.ca/dstephens/Regression/Data/a1-1.txt"
data1<-read.table(file1,header=TRUE)
x1<-data1$x
y1<-data1$y
fit.RP1<-lm(y1~x1)
beta0_1 = fit.RP1$coefficients[1]
beta1_1 = fit.RP1$coefficients[2]
res1 = residuals(fit.RP1)

file2<-"http://www.math.mcgill.ca/dstephens/Regression/Data/a1-2.txt"
data2<-read.table(file2,header=TRUE)
x2<-data2$x
y2<-data2$y
fit.RP2<-lm(y2~x2)
beta0_2 = fit.RP2$coefficients[1]
beta1_2 = fit.RP2$coefficients[2]
res2 = residuals(fit.RP2)

file3<-"http://www.math.mcgill.ca/dstephens/Regression/Data/a1-3.txt"
data3<-read.table(file3,header=TRUE)
x3<-data3$x
y3<-data3$y
fit.RP3<-lm(y3~x3)
beta0_3 = fit.RP3$coefficients[1]
beta1_3 = fit.RP3$coefficients[2]
res3 = residuals(fit.RP3)
```

So we have the parameters $\beta_0 = 0.0267655$ and $\beta_1 = 1.7251209$ from the linear regression. We plot the data with the line of best fit superimposed

```
plot(x1,y1)
abline(beta0_1, beta1_1)
```



```r
plot(x1,res1)
```

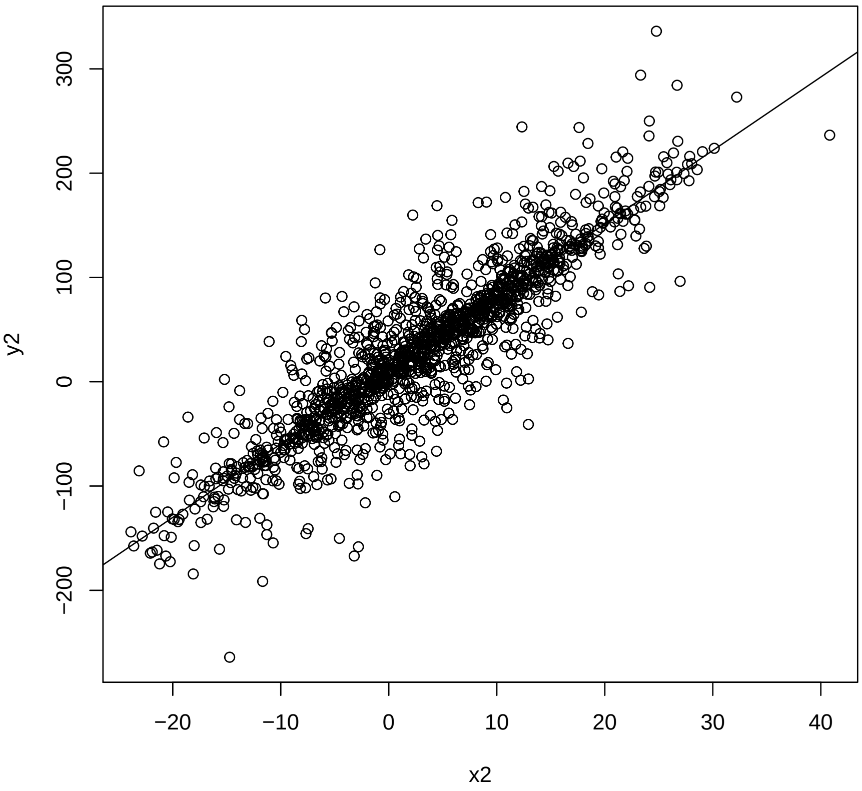


```r
mean(res1)

## [1] 2.664728e-17
```

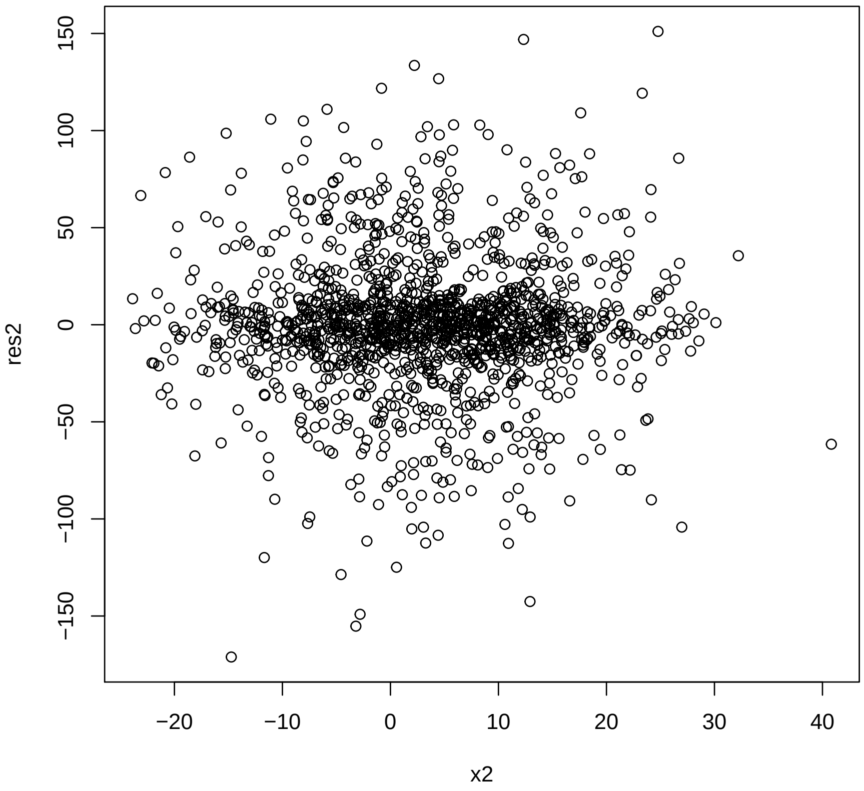
The residual plot seems to display adequate features although hard to tell graphically because of there is not that much data. The mean seemsto be above zero by quick graphicall inspection but is actually virtually null. The variance seems constant. We conclude the linear model used is adequate.

So we have the parameters $\beta_0 = 10.6608534$ and $\beta_1 = 7.0379523$ from the linear regression. We plot the data with the line of best fit superimposed

```
plot(x2,y2)
abline(beta0_2, beta1_2)
```



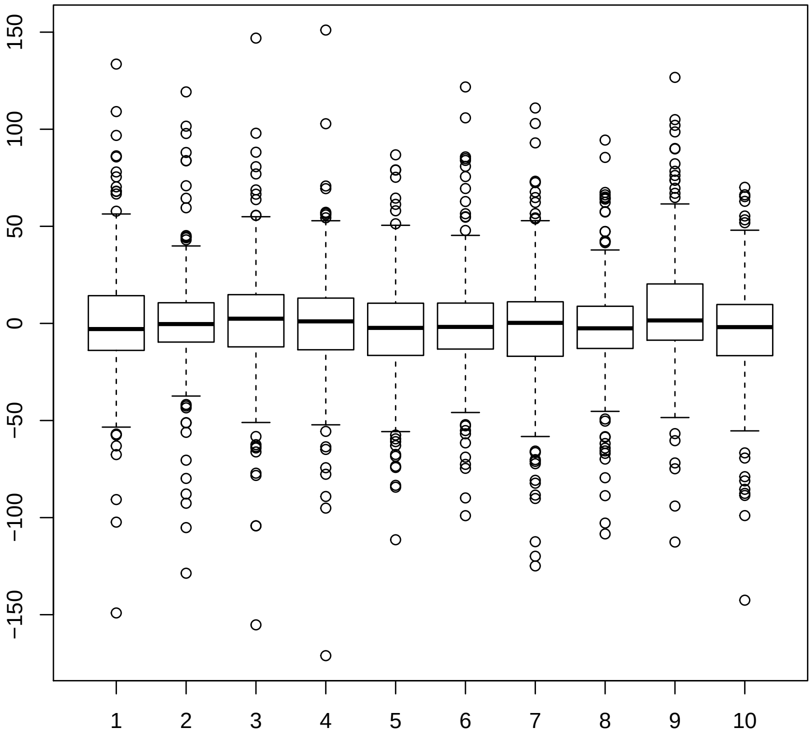
```
plot(x2,res2)
```



```
mean(res2)

## [1] 7.303642e-16
```

We now plot the residuals next to the corresponding xs. The model seems adequate as the residuals are heavily concentrated around the zero line. The computed mean is again virtually zero (7.30e-16). As for the variance it could seem as if the variance changes but we are careful to notice the distribution of the x's doesn't seem uniform and could explain this graphical impression.

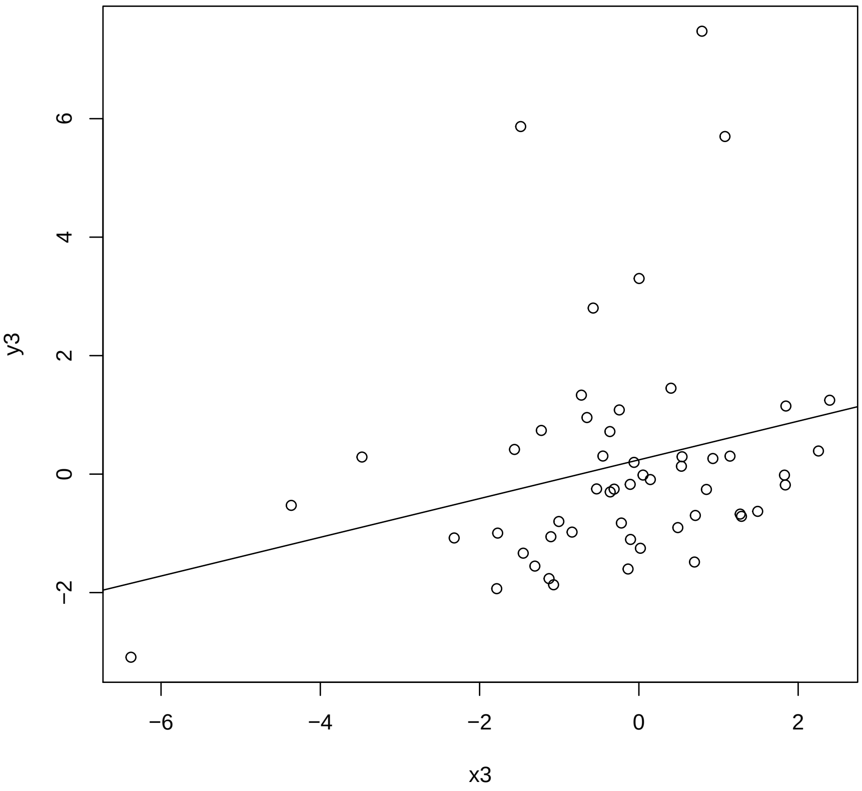
```
boxplot(res2[1:150],res2[151:301],res2[302:452],res2[453:603],
        res2[604:754],res2[755:900],res2[901:1051],res2[1052:1202],
        res2[1203:1353],res2[1354:1445])
```



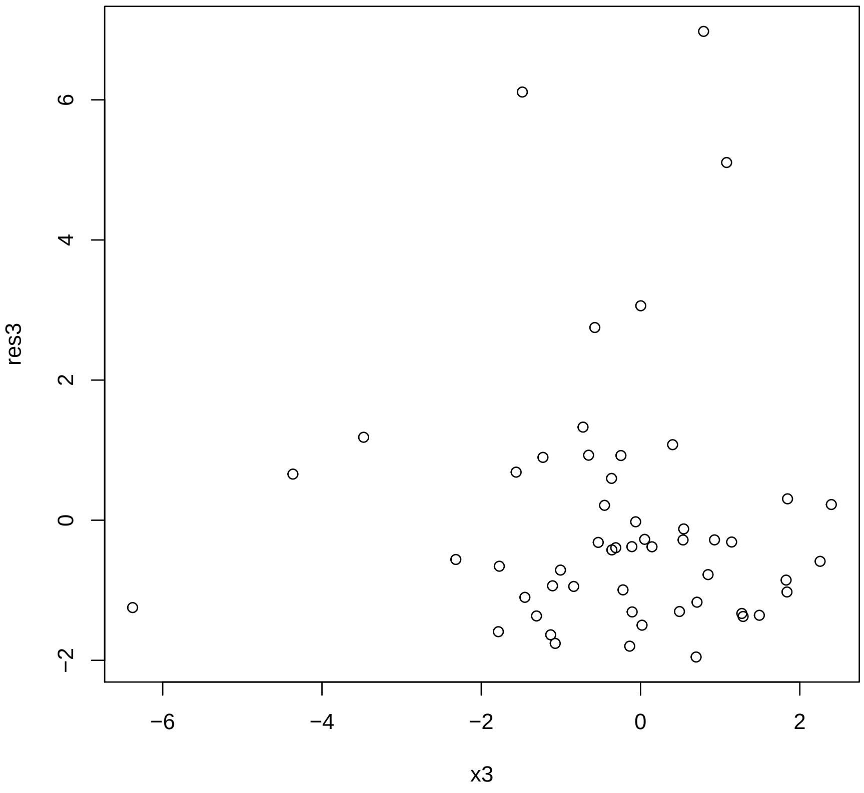
The boxplot confirms there is no obvious change in the variance of the residuals. The straight line model we have chosen is therefore appropriate.

So we have the parameters $\beta_0 = 0.2403328$ and $\beta_1 = 0.3267628$ from the linear regression. We plot the data with the line of best fit superimposed

```
plot(x3,y3)
abline(beta0_3, beta1_3)
```



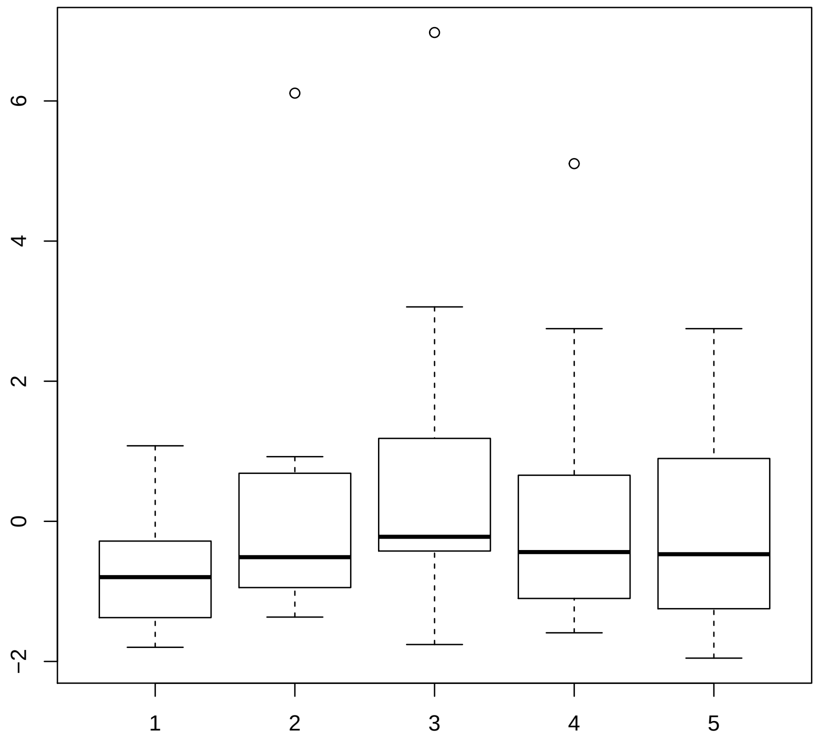
```
plot(x3,res3)
```



```
mean(res3)

## [1] -4.399815e-17
```

We now plot the residuals next to the corresponding xs. The mean of the residuals is again virtually zero eventhough graphically it seems it could be otherwise. Once again the data in the x's isn't evenly distributed and so is hard to analyze graphically. Even the box plot does not reveal much. Nothing contradicts the model specifically but it looks as if more data would be needed to use it with confidence.

```
boxplot(res3[1:10],res3[11:20],res3[21:30],res3[31:40],res3[40:53])
```



2 Theoretical demonstrations

2.1 Location Shift

$$
\begin{aligned}
S^*_{xx} &= \sum_{i=1}^{n}\left[(x_i - m) - (\frac{1}{n}\sum_{i=1}^{n}(x_i - m))\right]^2 \\
&= \sum_{i=1}^{n}((x_i - m) - (\bar{x} - m))^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= S_{xx}
\end{aligned}
$$

$$
\begin{aligned}
S^*_{xy} &= \sum_{i=1}^{n} y_i\left[(x_i - m) - (\frac{1}{n}\sum_{i=1}^{n}(x_i - m))\right] \\
&= \sum_{i=1}^{n} y_i(x_i - \bar{x}) \\
&= S_{xy}
\end{aligned}
$$

$$
\hat{\beta}^*_1 = \frac{S^*_{xy}}{S^*_{xx}} = \hat{\beta}_1 \quad \hat{\beta}^*_0 = \bar{y} - \hat{\beta}_1(\bar{x} - m)
$$

As for the properties, since $\hat{\beta}_1$ doesn't change with scaling it has the same as the ones for the original estimator, namely it is unbiased and we have variance of $\frac{\sigma^2}{S_{xx}}$.

$$
\begin{aligned}
E(\hat{\beta}^*_0) &= E(\bar{y} - \hat{\beta}_1(\bar{x} - m)) \\
&= E(\bar{y}) - \beta_1(\bar{x} - m) \\
&= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} - \beta_1 m \\
&= \beta_0 - \beta_1 m
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}^*_0) &= \mathrm{Var}(\bar{y} - \hat{\beta}_1(\bar{x} - m)) \\
&= \mathrm{Var}(\hat{y}) + (\bar{x} - m)^2\mathrm{Var}(\hat{\beta}_1) - 2(\bar{x} - m)\mathrm{Cov}(\bar{y}, \hat{\beta}_1) \\
&= \sigma^2\left(\frac{1}{n} + \frac{(\bar{x} - m)^2}{S_{xx}}\right)
\end{aligned}
$$

2.2 Scaling

$$
\begin{aligned}
S^*_{xx} &= \sum_{i=1}^{n}(l(x_i - \bar{x}))^2 \\
&= l^2 S_{xx}
\end{aligned}
$$

$$
\begin{aligned}
S^*_{xy} &= \sum_{i=1}^{n} y_i(l(x_i - \bar{x})) \\
&= lS_{xy}
\end{aligned}
$$

$$
\hat{\beta}^*_1 = \frac{S^*_{xy}}{S^*_{xx}} = \frac{lS^*_{xy}}{l^2 S^*_{xx}} = \frac{\hat{\beta}_1}{l} \qquad \hat{\beta}^*_0 = \bar{y} - \hat{\beta}^*_1(l\bar{x}) = \bar{y} - \hat{\beta}_1\bar{x}
$$

$$
\begin{aligned}
\mathrm{E}(\hat{\beta}^*_1) &= \frac{1}{l}\beta_1 \\
\mathrm{Var}(\hat{\beta}^*_1) &= \frac{1}{l^2}\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{l^2 S_{xx}} \\
\mathrm{E}(\hat{\beta}^*_0) &= E(\bar{y} - \frac{\hat{\beta}_1}{l}\bar{x}) \\
&= \beta_0 + \beta_1\bar{x} - \frac{\beta_1}{l}\bar{x} \\
&= \beta_0 + \bar{x}\beta_1(1 - \frac{1}{l}) \\
\mathrm{Var}(\beta^*_0) &= \mathrm{Var}(\bar{y}) + (\frac{\bar{x}}{l})^2\mathrm{Var}(\hat{\beta}_1) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{l^2 S_{xx}}\right)
\end{aligned}
$$

Note: In all derivations of variance we have assumed the covariance between the sample mean and the slope is zero which is easy to see by the fact that the errors are independent.