Math 423

Linear Regression

Homework II

Frédéric Boileau

Prof. David A. Stephens

16th October 2016

a

```
require(printr)

## Loading required package:  printr

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

## Error:  RStudio not running

salary<-read.csv("salary.csv",header=TRUE)
x1<-salary$SPENDING/1000
y<-salary$SALARY
```

we want to estimate the parameter β_1 and β_0, namely the slope and the intercept. We use the least square estimators. This is a case of simple linear regression so we can use the following equations:

$$\hat{\beta}_1 = \frac{S_{xx}}{S_{xy}} \tag{1}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2}$$

$$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \hat{x}) \tag{3}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \hat{x} \tag{4}$$

```
xbar = mean(x1)
ybar = mean(y)
Sxx = sum((x1 - xbar)^2)
Sxy = sum(y*(x1 - xbar))
slope = Sxy/Sxx
intercept = ybar - slope*xbar
print(slope)

## [1] 3307.585

print(intercept)

## [1] 12129.37
```

```r
fit.RP1 = lm(y~x1)
print(coef(fit.RP1))

## (Intercept)          x1
##   12129.371    3307.585
```

b and c

The residual standard error is given by

$$\hat{\sigma}^2 = \frac{SS_{\text{Res}}}{n-2} \tag{5}$$

MoreoverSS_{Res} is the sum of squares of error:

$$\begin{align} SS_{\text{Res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &= \sum_{i=1}^{n} y_i^2 - n\hat{y}^2 - \hat{\beta}_1 S_{xy} \tag{6} \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1 S_{xy} \tag{7} \end{align}$$

```
SSRes = sum((y - ybar)^2) - slope*Sxy
n = length(x1)
residualStdError = sqrt(SSRes/(n-2))
print(residualStdError)

## [1] 2324.779
```

d

We wish to compute the standard error with the values in the table already given. This table gives us the degrees of freedom (49) and the t value.

$$t_0 = \frac{\hat{\beta}_0}{\text{se}(\hat{\beta}_0)} \Rightarrow \text{se}(\hat{\beta}_0) = \frac{\hat{\beta}_0}{t_0} = \frac{12129.4}{10.13} \tag{8}$$

Now to do the computation directly from the data we use the actual formula for the standard error which is given by

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{MS_{\text{Res}}}{S_{xx}}} \qquad MS_{\text{Res}} = \frac{SS_{\text{Res}}}{n-2} = \hat{\sigma}^2 \tag{9}$$

```
MSRes = residualStdError / sqrt(n-2)
print(MSRes)

## [1] 332.1113
```

e

We will derive simple expressions from known relationships

$$
\begin{aligned}
SS_{\text{Res}} &= SS_{\text{T}} - \hat{\beta}_1 S_{xy} \\
SS_{\text{T}} &= SS_{\text{R}} + SS_{\text{Res}}
\end{aligned}
$$

It is easy to see then that

$$
\begin{aligned}
&SS_{\text{T}} = SS_{\text{Res}} + \hat{\beta}_1 S_{xy} \\
&SS_{\text{R}} = \hat{\beta}_1 S_{xy}
\end{aligned}
$$

```
SST = SSRes + slope*Sxy
SSR = slope*Sxy
Rsqrd = SSR/SST
print(Rsqrd)

## [1] 0.6967813
```

f

```
p = 2
Fstat = (SSR/(p-1))/(SSRes/(n-p))
print(Fstat)

## [1] 112.5995
```

g

$$y^\top (I_n H_1) y = y^\top (I_n - H) y + y^\top (H - H_1) y$$

The first statement we want to show is

$$\mathrm{trace}(I_n - H_1) = n - 1$$

Well the matrix I_n has $a_{ii} = 1 \,\forall\, i \in [1, n]$ and $h_{ii} = 1/n \,\forall\, i \in [1, n]$ By definitions:

$$\begin{aligned}
\mathrm{trace}(I_n - H_1) &= \sum_{i=1}^{n} (a_{ii} - h_{ii}) && (10) \\
&= \sum_{i=1}^{n} (1 - 1/n) && (11) \\
&= n(1 - 1/n) = n - 1 && (12)
\end{aligned}$$

The second statement we need to prove is that:

$$\mathrm{trace}(H - H_1) = p - 1 \tag{13}$$

We use the properties of the trace operator:

$$\mathrm{trace}(H - H_1) = \mathrm{trace}(H) - \mathrm{trace}(H_1) \tag{14}$$

$$\begin{aligned}
\mathrm{trace}(H) &= \mathrm{trace}(X(X^\top X)^{-1} X^\top) && (15) \\
&= \mathrm{trace}(X^\top X (X^\top X)^{-} 1) && (16) \\
&= \mathrm{trace}(I_p) \qquad \text{since} \quad X^\top X \in \mathbb{R}^{p \times p} && (17) \\
&= p && (18)
\end{aligned}$$

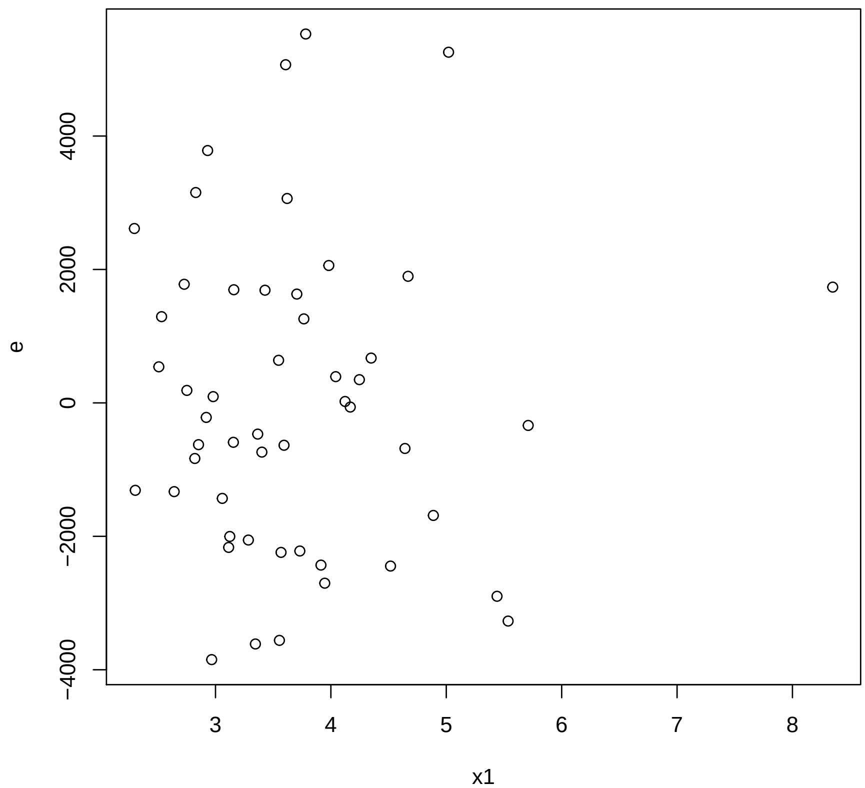
As shown before the trace of H_1 is 1 and this with the previous derivation proves (13).

Numerical part

```
require(MASS)

## Loading required package:  MASS
## Warning:  package 'MASS' was built under R version 3.3.1

bigx = cbind(matrix(1,length(x1)),x1)

n1 =length(x1)
H1 = matrix(1/n1,n1,n1)
sum(diag((diag(n1) - H1)))

## [1] 50

H = bigx %*%  ginv(t(bigx) %*% bigx) %*% t(bigx)
sum(diag(H - H1))

## [1] 1
```

h

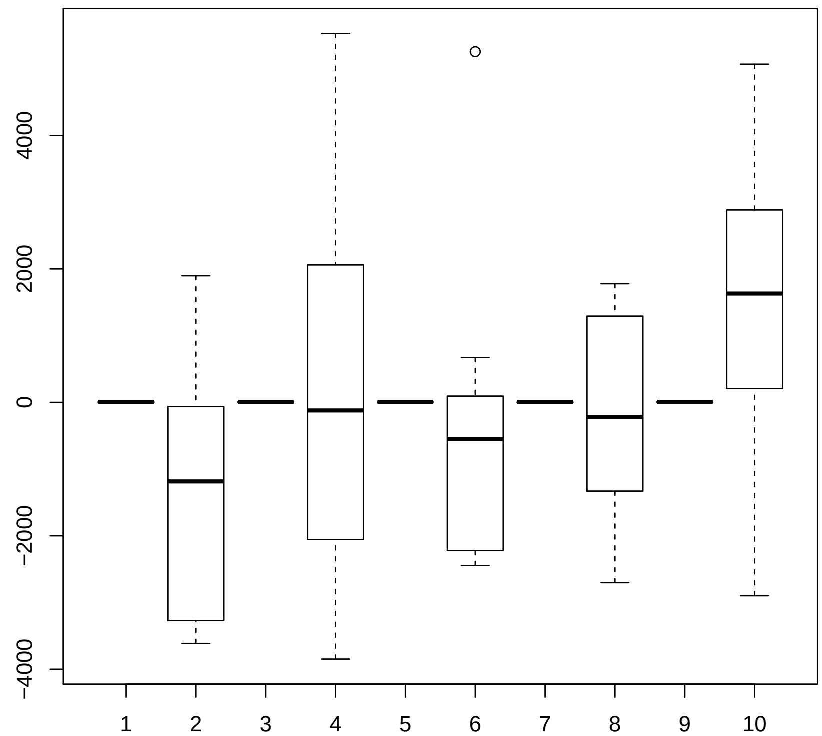
```
yhat = intercept + slope*x1
e = y - yhat
plot(x1,e)
```




```
mean(x1)

## [1] 3.696608
```

The residuals have zero mean.

```
boxplot(x1[1:10],e[1:10],x1[11:20],e[11:20],x1[21:30],
        e[21:30],x1[31:40],e[31:40],x1[41:41],e[41:51])
```



However a simple box plot shows quite clearly that they do not have constant variance.

```r
sum(e)#The sum of the residuals is zero, i.e. they are orthogonal to each other

## [1] -8.731149e-11

bigx = cbind(matrix(1,length(x1)),x1)
t(bigx)%*%e#the residuals are orthogonal to the regressors
```

	0
x1	0

```r
t(yhat)%*%e#the residuals are orthogonal to the fitted values
```

-2.4e-06

i

```
prediction = intercept + slope*4.8
print(prediction)

## [1] 28005.78
```

j

Let $x_0 := x_1^{new}$

$$\widehat{\mathrm{E}(y|x_0)} = \hat{\mu}_{u|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{19}$$
$$\tag{20}$$

The variance on a prediction is the variance of its estimator which is 19, so we compute as follows:

$$\mathrm{Var}(\hat{\mu}_{y|x_0}) = \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \tag{21}$$
$$= \mathrm{Var}[\bar{y} + \hat{\beta}_1(x_0 - \bar{x})] \tag{22}$$
$$= \frac{\sigma^2}{n} + \frac{\sigma^2(x_0 - \bar{x})^2}{S_{xx}} \tag{23}$$
$$= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \tag{24}$$

Now to estimate this value we use :

$$\hat{\sigma}^2 = MS_{Res}$$

```
x1new = 4800/1000
MSRes*(1/n1  + (x1new - xbar)^2/Sxx)

## [1] 13.78083
```