Solution lab 20, Chapter 2 : Linear regression.

May 19, 2016

Exercise 1

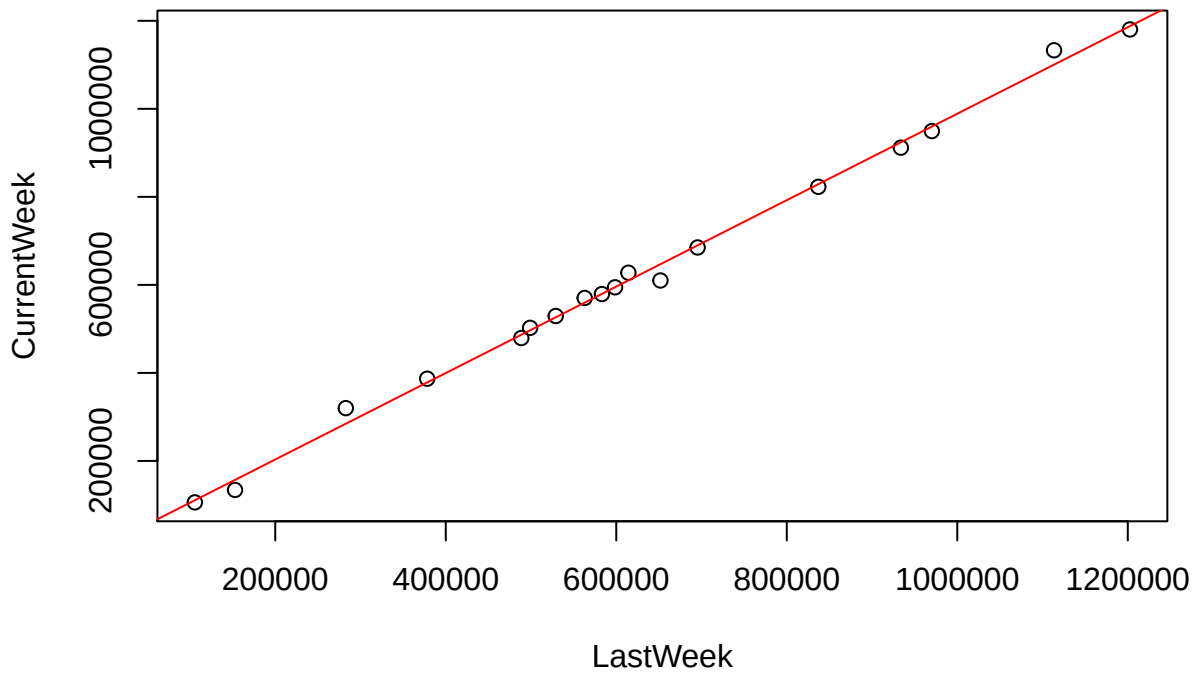
```
M <- read.csv('playbill.csv')
x <- M$LastWeek
y <- M$CurrentWeek

lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##   6804.8860       0.9821
```

```
l <- lm(y~x)

plot(x,y,xlab='LastWeek',ylab='CurrentWeek')
abline(a=l$coefficients[1],b=l$coefficients[2],col='red')
```



a)

```
n <- length(x)

SXX <- sum((x-mean(x))^2)
S <- sqrt(sum(l$residuals^2)/(n-2))

se1 <- S/sqrt(SXX)

ci1 <- l$coefficients[2]-qt(0.975,n-2)*se1
ci2 <- l$coefficients[2]+qt(0.975,n-2)*se1
```

The 95 % confidence interval is $C := [ci1, ci2] = [0.95, 1.01]$. Of course, it leads that $\beta_1 = 1$ is a plausible value since it lies in C and the confidence level is quite high.

b)

```
se0 <- S*sqrt(1/n+mean(x)^2/SXX)

T <- (l$coefficients[1]-10000)/se0
pval <- 2*min(1-pt(T,n-2),pt(T,n-2))
```

The p-value is approximatly equals 0.75. This is a very high p-value and thus we cannot decently reject the null hypothesis.

c)

```
y400000 <- l$coefficients[1]+l$coefficients[2]*400000

ci1 <- y400000-qt(0.975,n-2)*S*sqrt(1+1/n+(400000-mean(x))^2/SXX)
ci2 <- y400000+qt(0.975,n-2)*S*sqrt(1+1/n+(400000-mean(x))^2/SXX)
```

The estimate for the gross box office for a 400,000$ production the last week is $\widehat{y} = 399637.5$. We also have a 95% prediction interval $P = [359832.8, 439442.2]$. Thus, a 450,000$ production for the gross box office in the current week doesn't seem to be a plausible value since it doesn't lie in P which is a prediction interval with a quite high level.

d)

According to the estimation, the gross box office results mainly increases from one week to the following. Indeed, the slope estimate is close to 1 but the intercept estimate is relatively far from 0. Nevertheless, if the promotors of Broadway assumes that a difference of 6800$ is negligible, then we could say that this rule is probably true.

Exercise 2

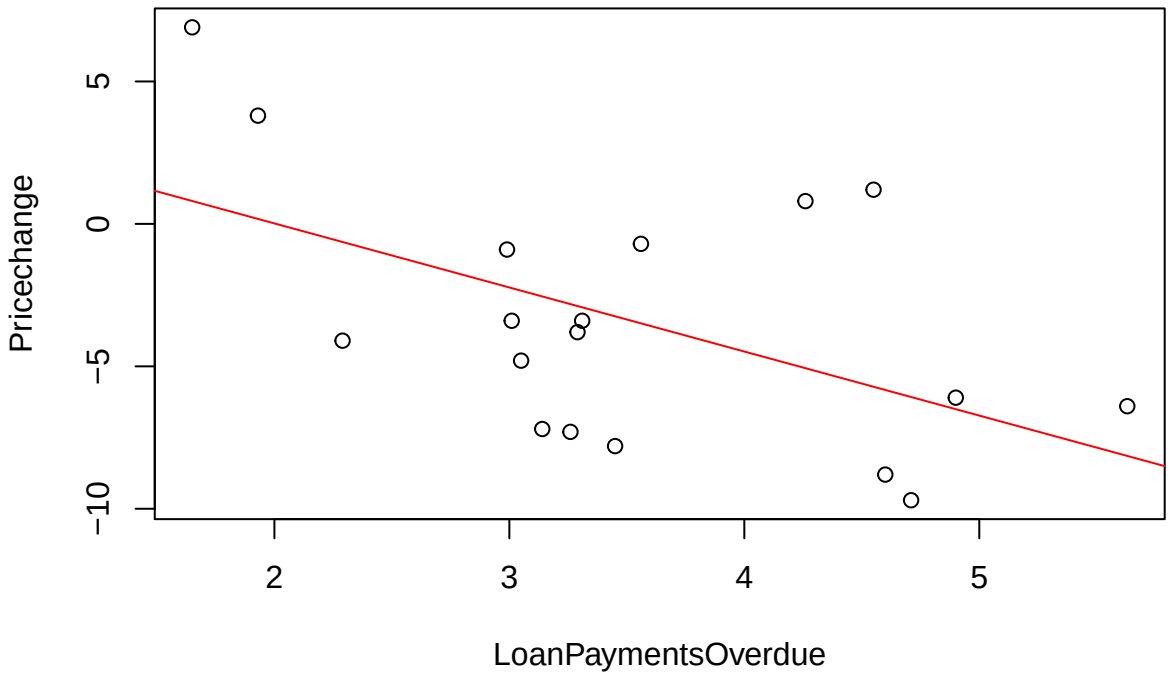
```
M <- read.table('indicators.txt',header=TRUE)
x <- M$LoanPaymentsOverdue
y <- M$PriceChange

lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##       4.514       -2.249
```

```
l <- lm(y~x)

plot(x,y,xlab='LoanPaymentsOverdue',ylab='Pricechange')
abline(a=l$coefficients[1],b=l$coefficients[2],col='red')
```



a)

```
n <- length(x)

SXX <- sum((x-mean(x))^2)
S <- sqrt(sum(l$residuals^2)/(n-2))
```

```
se1 <- S/sqrt(SXX)

ci1 <- l$coefficients[2]-qt(0.975,n-2)*se1
ci2 <- l$coefficients[2]+qt(0.975,n-2)*se1
```

The 95 % confidence interval is $C := [ci1, ci2] = [-4.16, -0.33]$. Of course, it leads that β_1 is most probably negative since C lies in $\mathbb{R}_-$ with high confidence level.

b)

```
y4 <- l$coefficients[1]+l$coefficients[2]*4

ci1 <- y4-qt(0.975,n-2)*S*sqrt(1/n+(4-mean(x))^2/SXX)
ci2 <- y4+qt(0.975,n-2)*S*sqrt(1/n+(4-mean(x))^2/SXX)
```

We estimate $\mathbb{E}[Y|X = 4] = -4.479585$. We also have a 95% confidence interval $I = [-6.64, -2.31]$. Thus, 0% is probably not a feasible value for $\mathbb{E}[Y|X = 4]$ since with high confidence, it is less than -2.31.

Exercise 3

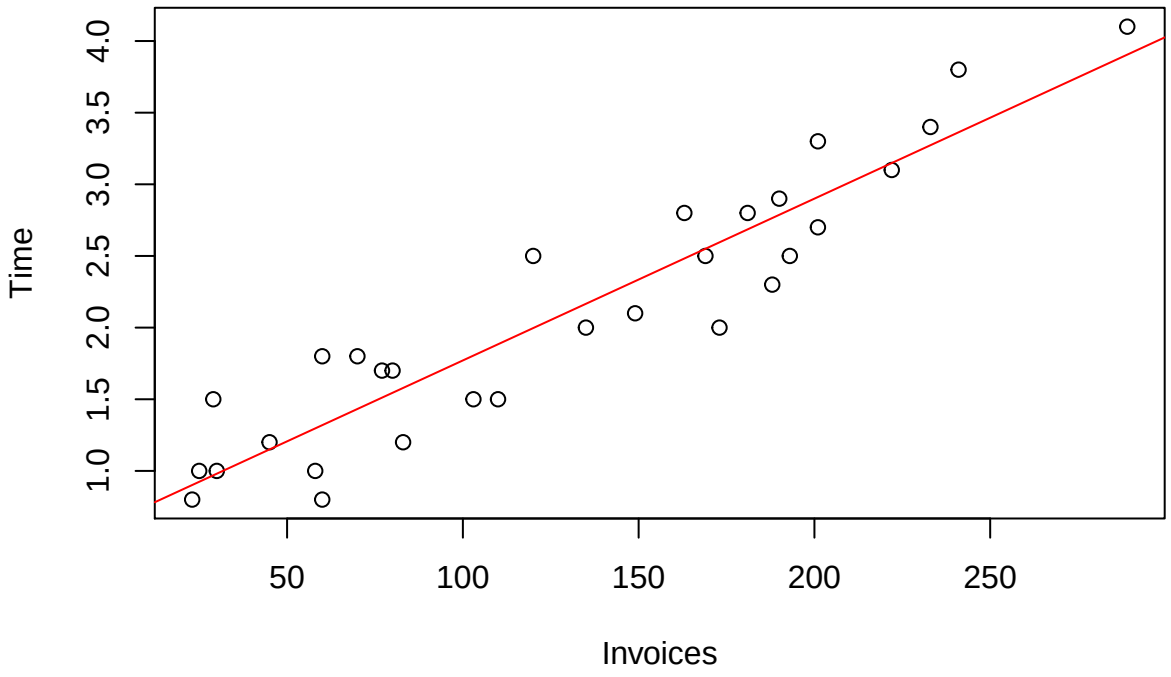
```
M <- read.table('invoices.txt',header=TRUE)
x <- M$Invoices
y <- M$Time

lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     0.64171      0.01129
```

```
l <- lm(y~x)

plot(x,y,xlab='Invoices',ylab='Time')
abline(l,col='red')
```



a)

```
confint(l,level=0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 0.391249620 0.89217014
## x           0.009615224 0.01296806
```

The 95 % confidence interval is $C := [ci1, ci2] = [0.39, 0.89]$.

b)

```
SXX <- sum((x-mean(x))^2)
se1 <- S/sqrt(SXX)

T <- (l$coefficients[2]-0.01)/se1
pval <- 2*min(1-pt(T,n-2),pt(T,n-2))
```

The p-value is approximatly equals 0.12. This is an high p-value and thus we cannot reject the null hypothesis.

c)

```
y130 <- l$coefficients[1]+l$coefficients[2]*130

ci1 <- y130-qt(0.975,n-2)*S*sqrt(1+1/n+(130-mean(x))^2/SXX)
ci2 <- y130+qt(0.975,n-2)*S*sqrt(1+1/n+(130-mean(x))^2/SXX)
```

We estimate $\widehat{y} = 2.109624$. We also have a 95% prediction interval $P = [1.42, 2.19]$.

Exercise 5

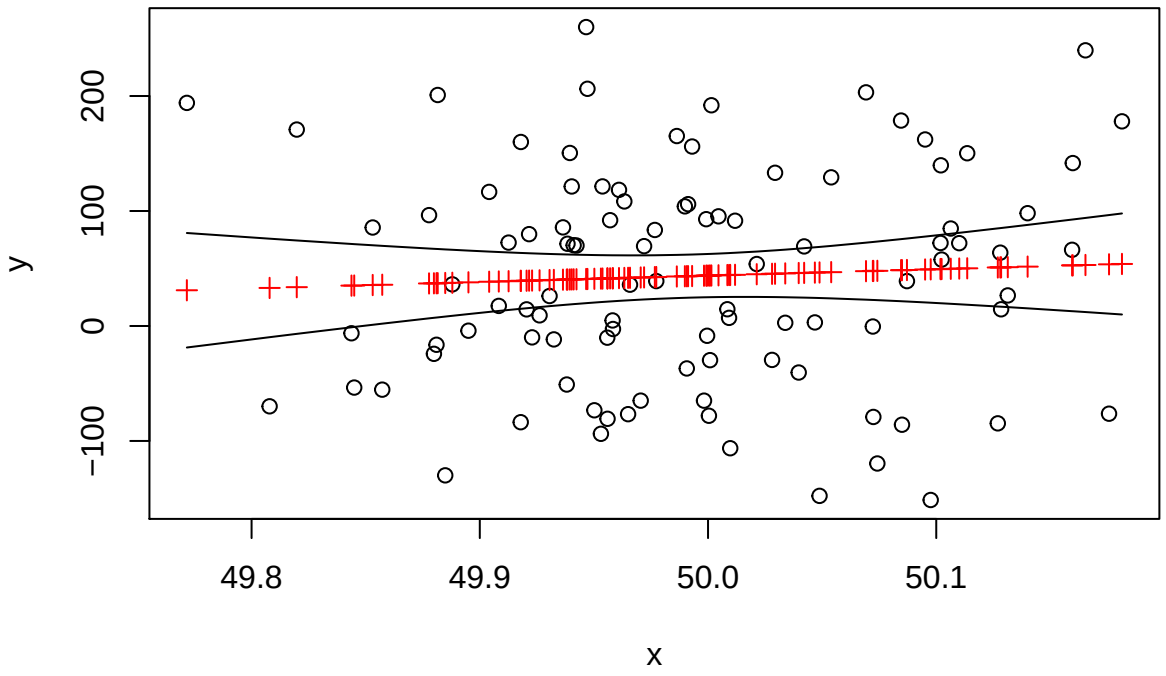
The right answer is statement (d). Indeed RSS recovers the global differency between the y_i's and $\widehat{y}_i$'s and the points in the first model are much more gathered around their least squares line than the second model. On the contrary, the SSreg relates the global difference between the $\widehat{y}_i$'s and $\bar{y}$. Thus, it is the second model that will assume a lower SSreg. Indeed, it's least squares line slope is lower in absolute value and it's intercept seems to be close to $\bar{y}$.

Exercise 7

```
x <- rnorm(100,mean = 50,sd = 0.1)
x <- sort(x)
y <- x+rnorm(100,sd=100)

l <- lm(y~x)

conf <- predict(l,interval='confidence',level=0.95)
plot(x,y)
points(x,l$fitted.values,col='red',pch=3)
lines(x,conf[,2])
lines(x,conf[,3])
```



The 95% confident intervals (black lines) are intended to represente a range where the fitted value (red crosses) belongs with 95% confidence. Now, because those ranges are aimed for the regression line and not the data points, it is possible that most of the observations falls outside, in particular when σ is large. In this figure, only 15% of the data points belongs to confident intervals.