Solution Lab 21

May 25, 2016

Multiple Linear Regression

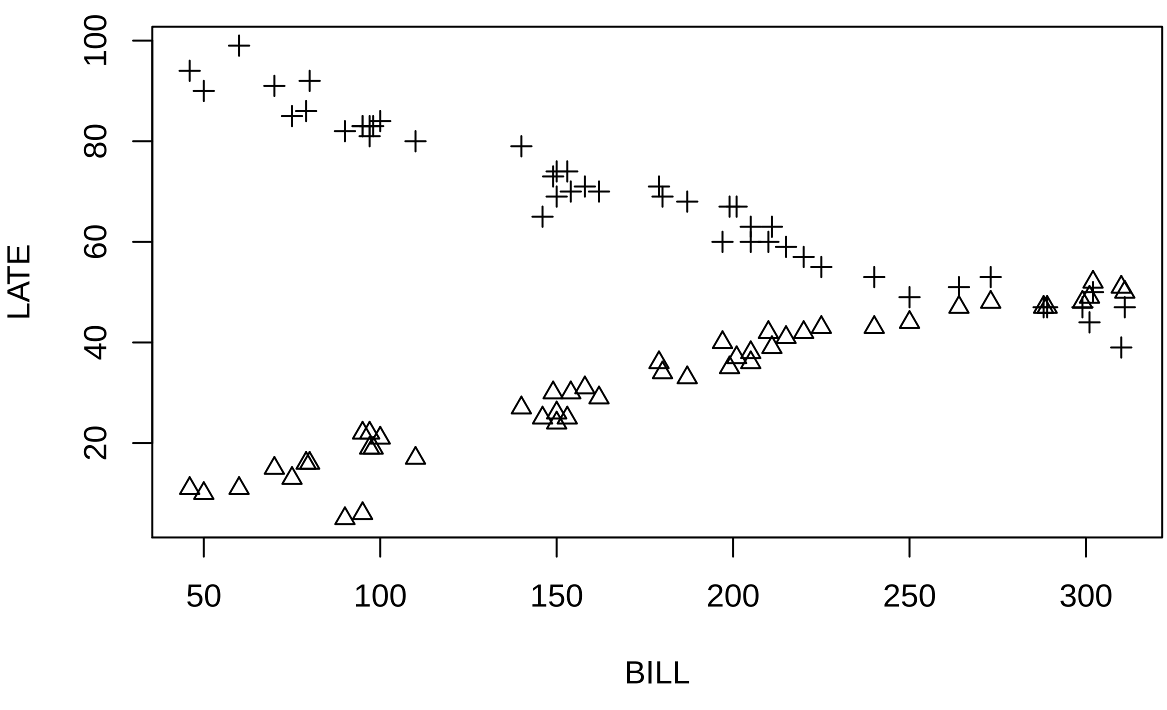
The three exercises of Chapter 5 have closely the same arguments. Thus, we use exercise 1 as a reference model and do not justify each milestones for the others.

Exercise 1

We introduce a **dummy** variable in the dataset. Set TYPE = 0 if the account is RESIDENTIAL or 1 if it is COMMERCIAL.

```
t <- read.table('overdue.txt',header=TRUE)
t <- cbind(t,TYPE=c(rep(0,48),rep(1,48)))
```

We represent each data account (BILL,LATE) with triangles if it belongs to TYPE = 1 or with crosses otherwise.



It is clear that we cannot ignore the variable TYPE since it clearly influences the linear relation. Nevertheless, the arrangement of the points for both TYPE groups suggests that there is a linear relation when considered separately. Thus, we consider the following model

$$LATE \sim BILL + TYPE + BILL \times TYPE.$$

```
lm(t$LATE~t$BILL+t$TYPE+t$BILL*t$TYPE)
```

```
##
## Call:
## lm(formula = t$LATE ~ t$BILL + t$TYPE + t$BILL * t$TYPE)
##
## Coefficients:
##   (Intercept)         t$BILL         t$TYPE  t$BILL:t$TYPE
##        2.2096         0.1657        99.5486        -0.3566
```

We have a model with *unrelated regression lines* such that

$$LATE = c_1 + c_2 BILL + c_3 TYPE + c_4 BILL \times TYPE,$$

which is equivalent to

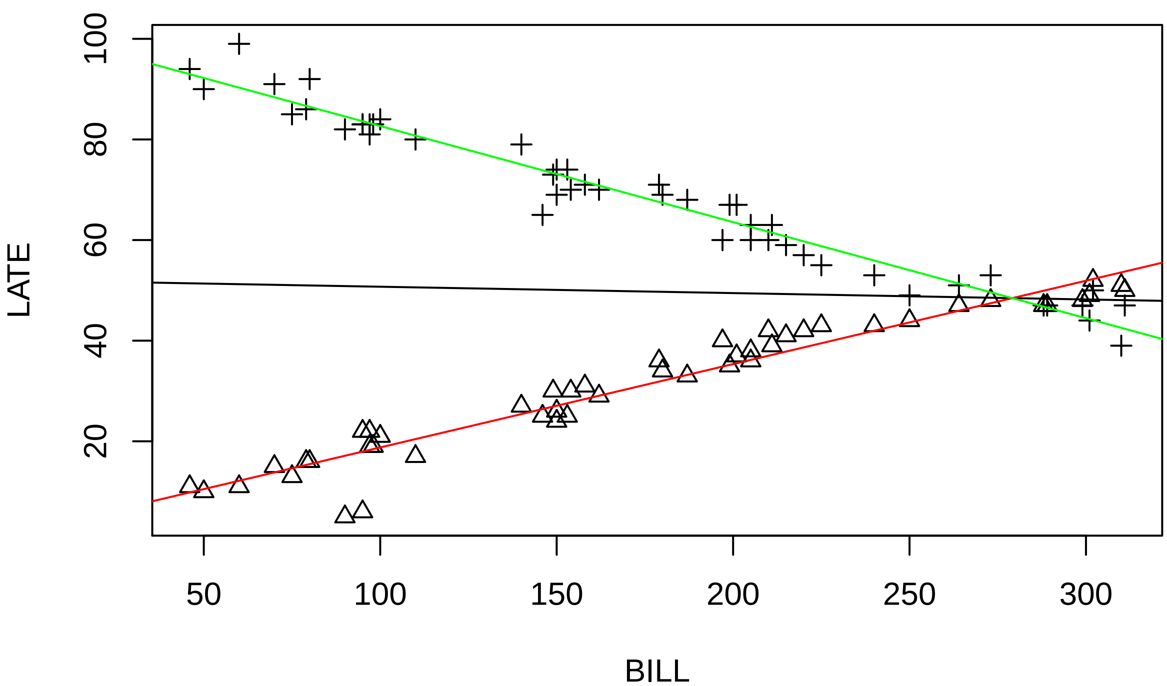
$$LATE = \beta_{0,0} + \beta_{1,0} BILL, \text{ if } TYPE = 0, \quad (1)$$
$$LATE = \beta_{0,1} + \beta_{1,1} BILL, \text{ if } TYPE = 1, \quad (2)$$

with $\beta_{0,0} = c_1$, $\beta_{1,0} = c_2$, $\beta_{0,1} = c_1 + c_3$ and $\beta_{1,1} = c_2 + c_4$. In other words, we have two different linear regression models given the value TYPE.

```
L <- lm(t$LATE~t$BILL+t$TYPE+t$BILL*t$TYPE)
l <- lm(t$LATE~t$BILL)
```

```
beta0_0 <- L$coefficients[1]
beta1_0 <- L$coefficients[2]
beta0_1 <- L$coefficients[1]+L$coefficients[3]
beta1_1 <- L$coefficients[2]+L$coefficients[4]
```

We represente the regression line for the linear model $LATE \sim BILL$, in red for model (1) and in green for model (2).



Remark that we retrieve the same estimated coefficients if we study the model separately, i.e.

```
lm(t$LATE[t$TYPE==0]~t$BILL[t$TYPE==0])
```

```
##
## Call:
## lm(formula = t$LATE[t$TYPE == 0] ~ t$BILL[t$TYPE == 0])
##
## Coefficients:
##         (Intercept)  t$BILL[t$TYPE == 0]
##              2.2096               0.1657
```

```
c(beta0_0,beta1_0)
```

```
## (Intercept)      t$BILL
##    2.209624    0.165683
```

```
lm(t$LATE[t$TYPE==1]~t$BILL[t$TYPE==1])
```

```
##
## Call:
## lm(formula = t$LATE[t$TYPE == 1] ~ t$BILL[t$TYPE == 1])
##
## Coefficients:
##         (Intercept)  t$BILL[t$TYPE == 1]
##             101.758               -0.191
```

```
c(beta0_1,beta1_1 )
```

```
## (Intercept)      t$BILL
## 101.7581844  -0.1909615
```

Exercise 2

We compute the models (a), (b) and (c) like in the previous exercise.

```
t <- read.csv('HoustonChronicle.csv')
Y <- t$X.Repeating.1st.Grade
X <- t$X.Low.income.students
X2 <- t$Year

la <- lm(Y~X)
lb <- lm(Y~X+X2)
lc <- lm(Y~X+X2+X*X2)
```

We study the relavance of the *full* model against the *reduced* models providing a partial-F test. One can use the R-built function `anova`.

```r
anova(la,lc)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X
## Model 2: Y ~ X + X2 + X * X2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    120 1751.9
## 2    118 1744.4  2     7.512 0.2541 0.7761
```

```r
anova(lb,lc)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X + X2
## Model 2: Y ~ X + X2 + X * X2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    119 1747.8
## 2    118 1744.4  1    3.4351 0.2324 0.6307
```

Since the p-values = 0.7761 and 0.6307 are pretty high, we cannot reject both reduced models. In order to chose between them, we look further to the results of `lm`

```r
summary(la)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9845 -2.5072 -0.4184  1.8505 11.1067
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.91419    0.83836   3.476 0.000709 ***
## X            0.07550    0.01823   4.141 6.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.821 on 120 degrees of freedom
## Multiple R-squared:  0.125,  Adjusted R-squared:  0.1177
## F-statistic: 17.14 on 1 and 120 DF,  p-value: 6.472e-05
```

```r
summary(lb)
```

```
##
## Call:
## lm(formula = Y ~ X + X2)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -8.6768 -2.5451 -0.4769  1.6624 11.3469
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -73.54333  145.12258  -0.507 0.613256
## X             0.07248    0.01917   3.782 0.000245 ***
## X2            0.03831    0.07272   0.527 0.599274
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.832 on 119 degrees of freedom
## Multiple R-squared:  0.127,  Adjusted R-squared:  0.1124
## F-statistic: 8.659 on 2 and 119 DF,  p-value: 0.0003083
```

The T-test p-value = 0.599274 for the coefficient of the variable X2 in model (b) is also high. It means that we can not decently reject the hypothesis that the coefficient for the variable Year is null. Thus, the simple linear model (a) seems to be the best choice.

Exercice 3

a)

```r
d <- read.table('Latour.txt',header=TRUE)

lmreduced <- lm(d$Quality~d$EndofHarvest +d$Rain)
lmfull <- lm(d$Quality~d$EndofHarvest +d$Rain+d$EndofHarvest*d$Rain)
anova(lmreduced,lmfull)
```

```
## Analysis of Variance Table
##
## Model 1: d$Quality ~ d$EndofHarvest + d$Rain
## Model 2: d$Quality ~ d$EndofHarvest + d$Rain + d$EndofHarvest * d$Rain
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     41 26.945
## 2     40 22.971  1    3.9749 6.9218 0.01203 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the partial-F test with p-value = 0.01203, we show that the coefficient of the interaction term in model (5.10) is statistically significant.

b)

We use the linear relationship with Quality from EndofHarvest according to Rain = 0 or 1, i.e.

$$Quality = Quality(EndofHarvest).$$

Thus, in both cases we can invert it in order to have a linear relationship with EndofHarvest from Quality according to Rain = 0 or 1, i.e.

$$EndofHarvest = EndofHarvest(Quality).$$

Since we want a 1 point Quality decreases and the relationships are linear, it is sufficient to evaluate EndofHarvest(0)-EndofHarvest(1) in both cases.

```r
beta0_0 <- lmfull$coefficients[1]
beta1_0 <- lmfull$coefficients[2]
beta0_1 <- lmfull$coefficients[1]+lmfull$coefficients[3]
beta1_1 <- lmfull$coefficients[2]+lmfull$coefficients[4]

# i)
(0-beta0_0)/beta1_0-(1-beta0_0)/beta1_0
```

```
## (Intercept)
##    31.80103
```

```r
# ii)
(0-beta0_1)/beta1_1-(1-beta0_1)/beta1_1
```

```
## (Intercept)
##    8.727273
```

Density estimation

Exercise 1

See example 10.1 in textbook

```r
n <- 100
X <- rlnorm(n)

# Sturges' rule
nclass <- ceiling(1+log2(n))    # One can use the in-built function nclass.Sturges
cwidth <- diff(range(X)/nclass)
breaks <- min(X)+cwidth*0:nclass
hist.sturges <- hist(X,breaks = breaks, plot = FALSE)
```

For Doane's rule, we need to compute the sample skeness coefficient

$$\sqrt{b_1} := \frac{1/n \sum_{i=1}^{n}(X_i - \bar{X})^3}{\left[1/n \sum_{i=1}^{n}(X_i - \bar{X})^2\right]^{3/2}}$$

```r
# Doane's rule
sqrtb1 <- mean((X-mean(X))^3)/(mean((X-mean(X))^2))^(3/2)
sigmab1 <- sqrt(6*(n-2)/((n+1)*(n+3)))
Ke <- log2(1+abs(sqrtb1)/sigmab1)
nclass <- ceiling(1+log2(100)+Ke)
cwidth <- diff(range(X)/nclass)
breaks <- min(X)+cwidth*0:nclass
hist.doane <- hist(X,breaks = breaks, plot = FALSE)
```

We set the different breaks and counts for both methods.

```r
hist.sturges$breaks
```

```
## [1] 0.0373773 1.2550009 2.4726244 3.6902480 4.9078716 6.1254951 7.3431187
## [8] 8.5607423 9.7783659
```

```r
hist.sturges$counts
```

```
## [1] 59 23  7  6  2  1  0  2
```

```r
hist.doane$breaks
```

```
##  [1] 0.0373773 0.8491263 1.6608754 2.4726244 3.2843735 4.0961225 4.9078716
##  [8] 5.7196206 6.5313697 7.3431187 8.1548678 8.9666168 9.7783659
```

```r
hist.doane$counts
```

```
##  [1] 40 30 12  6  5  2  2  1  0  0  0  2
```

```r
deciles <- qlnorm(1:9/10)
dlnorm(deciles)
```

```
## [1] 0.63218439 0.64954676 0.58740781 0.49773682 0.39894228 0.29987846
## [7] 0.20580277 0.12066672 0.04871943
```

```r
# We find what cell deciles_i belongs to
nbreaks <- sapply(deciles,function(x) which.min(hist.sturges$breaks<x)-1)
# which.min finds the indice of the first FALSE in a vector of logicals, see also
# which and which.max

hist.sturges$density[nbreaks]
```

```
## [1] 0.48455041 0.48455041 0.48455041 0.48455041 0.48455041 0.18889253
## [7] 0.18889253 0.18889253 0.05748903
```

```r
mean(abs(hist.sturges$density[nbreaks]-dlnorm(deciles)))
```

```
## [1] 0.07990821
```

```r
nbreaks <- sapply(deciles,function(x) which.min(hist.doane$breaks<x)-1)
hist.doane$density[nbreaks]
```

```
## [1] 0.49276313 0.49276313 0.49276313 0.49276313 0.36957235 0.36957235
## [7] 0.14782894 0.14782894 0.06159539
```

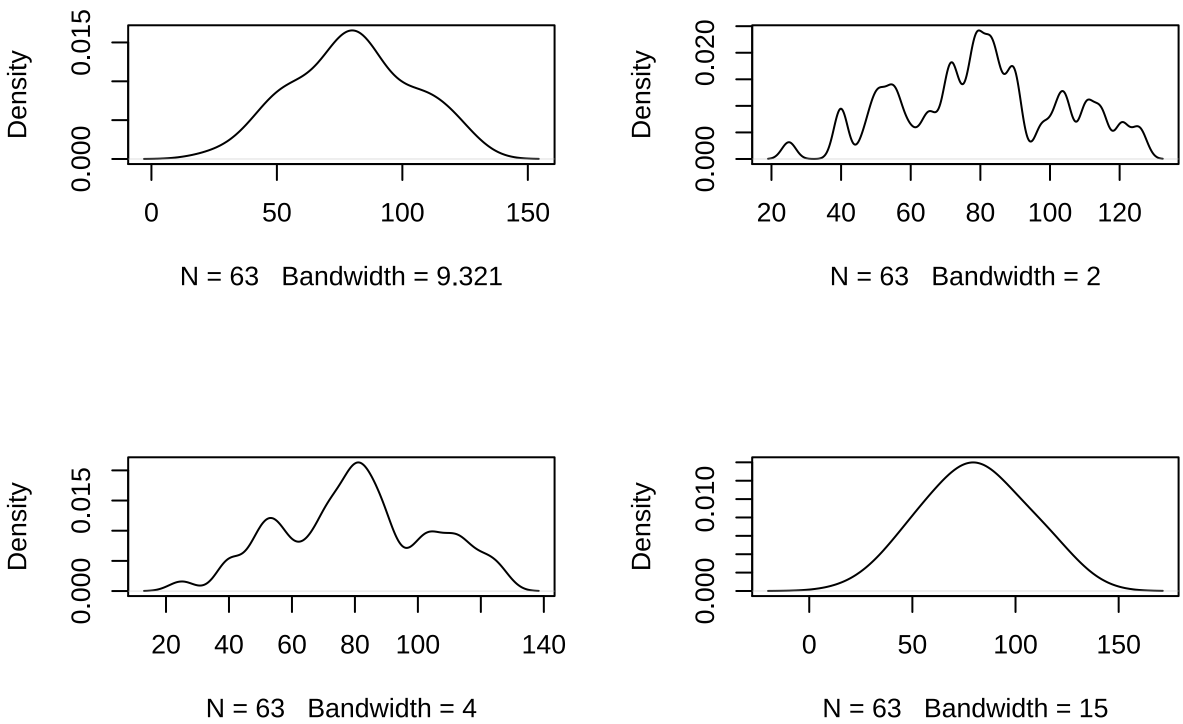
```r
mean(abs(hist.doane$density[nbreaks]-dlnorm(deciles)))
```

```
## [1] 0.06587768
```

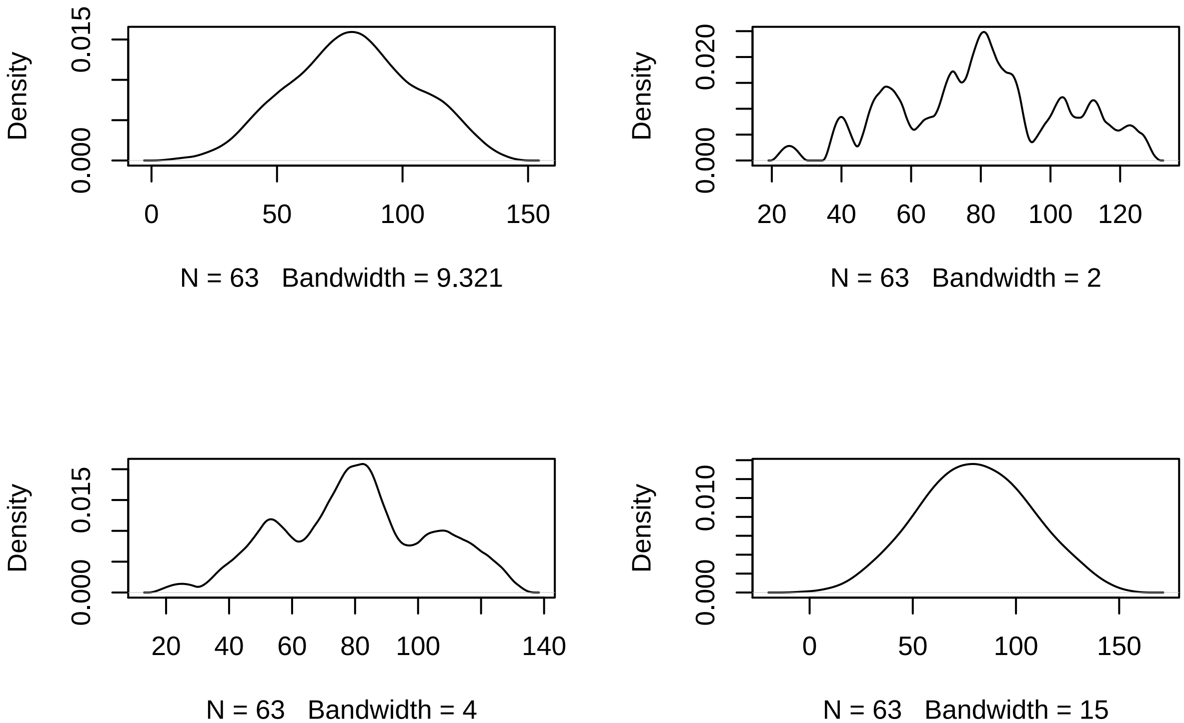
The mean error for Doane's rule is slightly lower than the one for Struges' rule and thus we shall prefer to use Doane's bin selection.

Exercise 8

```
library(gss)
data("buffalo")
par(mfrow=c(2,2))
plot(density(buffalo,kernel = 'gaussian'),main='')
plot(density(buffalo,kernel = 'gaussian',bw = 2),main='')
plot(density(buffalo,kernel = 'gaussian',bw = 4),main='')
plot(density(buffalo,kernel = 'gaussian',bw = 15),main='')
```




```
par(mfrow=c(2,2))
plot(density(buffalo,kernel = 'biweight'),main='')
plot(density(buffalo,kernel = 'biweight',bw = 2),main='')
plot(density(buffalo,kernel = 'biweight',bw = 4),main='')
plot(density(buffalo,kernel = 'biweight',bw = 15),main='')
```


Density
0.000
0.015
0
50
100
150
N = 63 Bandwidth = 9.321
Density
0.000
0.020
20
40
60
80
100
120
N = 63 Bandwidth = 2
Density
0.000
0.015
20
40
60
80
100
140
N = 63 Bandwidth = 4
Density
0.000
0.010
0
50
100
150
N = 63 Bandwidth = 15