

# Chapter 3: Linear Regression and Graphical User Interface

## I. Statistics

### 3.1. Linear regression

#### 1. Estimation:

**Response:**  $Y_1, Y_2, \dots, Y_n$ ; **Covariates:**  $X_{i1}, X_{i2}, \dots, X_{i(p-1)}$ ,  $i = 1, \dots, n$ .

Observation 1	Observation 2	...	Observation $n$
$(Y_1, X_{11}, X_{12}, \dots, X_{1(p-1)})$	$(Y_2, X_{21}, X_{22}, \dots, X_{2(p-1)})$	...	$(Y_n, X_{n1}, X_{n2}, \dots, X_{n(p-1)})$

The standard linear regression is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i(p-1)} + \epsilon_i, \epsilon_i \sim N(\mathbf{0}, \sigma^2)$$

where  $\beta_0, \beta_1, \dots, \beta_{p-1}$  are unknown parameters. Then, the least squares estimator is

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} = (X^t X)^{-1} X Y,$$

where

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1(p-1)} \\ 1 & X_{21} & \cdots & X_{2(p-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{n(p-1)} \end{bmatrix}, Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}.$$

#### Example (Splus):

```
### lm command
help(air)
ozone=air[,1]
radi=air[,2]
temper=air[,3]
wind=air[,4]
ozonelm1=lm(ozone~radi)      # ozone = β₀ + β₁(radiation) + ε
ozonelm1
ozonelm1$coefficients
ozonelm2=lm(ozone~1+radi)    # ozone = β₁(radiation) + ε
ozonelm2
```

```

ozonelm3=lm(ozone~radi+temper+wind) # multiple linear regression model:
#  $ozone = \beta_0 + \beta_1(radiation) + \beta_2(temperature) + \beta_3(wind) + \epsilon$ 

ozonelm4=lm(ozone~-1+radi+temper+wind)
lm(ozone~radiation)
ozonelm5=lm(ozone~radiation,data=air)
ozonelm5
ozonelm1
attributes(air)
ozonedata=data.frame(myozone=ozone,myradi=radi,mytemper=temper,
                      mywind=wind)
attributes(ozonedata)
lm(myozone~mytemper,data=ozonedata)

### matrix manipulation
xm=cbind(1,radi,temper,wind)
b=solve(t(xm)%*%xm)%*%t(xm)%*%ozone #  $b = (X^t X)^{-1} X Y$ 
b
ozonelm3$coefficients
yhat4=xm%*%b #  $\hat{Y} = X b$ 
yhat4-ozonelm3$fitted.values
residual4=ozone-yhat4 #  $e = Y - \hat{Y}$ 
residual4-ozonelm3$residuals

```

### Note:

The following simulation illustrates the performance of the least squares estimator.

#### **Example (Splus):**

```

lsq=matrix(0,1000,2)
x=seq(0.025,1,by=0.025)
for(i in 1:1000)
{
  y=3+5*x+rnorm(40,sd=1)      # generated data
  lsq[i,]=lm(y~x)$coefficients # least squares estimate
}
apply(lsq,2,mean)

```

## 2. Hypothesis testing:

**Model 0:**  $Y = \varepsilon$

**Model 1:**  $Y = \beta_0 + \varepsilon$

**Model 2:**  $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

⋮

**Model p:**  $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$

Let  $\hat{Y}$ (model  $i$ ),  $i = 1, \dots, p$ , be the vector of the predicted values, for example,

$$\hat{Y}(\text{model } p) = Xb = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}.$$

**(a) Test  $H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ :**

$$F = \frac{\left[ \|\hat{Y}(\text{model } p) - \hat{Y}(\text{model } 1)\|^2 \right] / (p-1)}{s^2}, s^2 = \frac{\|Y - \hat{Y}(\text{model } p)\|^2}{n-p}$$

Large F value might imply the difference between model  $p$  and model 1 is large relative to the random variation. As  $H_0$  is true, F statistic is distributed as F distribution with degrees of freedom  $p-1$  and  $n-p$ , respectively.

### Example:

Let  $p = 2$  and  $H_0: \beta_1 = 0$ . Then

**Model 2:**  $Y = \beta_0 + \beta_1 X_1 + \varepsilon$  (the original (full) model)

**Model 1:**  $Y = \beta_0 + \varepsilon$  (the reduce model as  $H_0: \beta_1 = 0$  is true).

The test statistic is

$$F = \frac{\left[ \|\hat{Y}(\text{model } 2) - \hat{Y}(\text{model } 1)\|^2 \right] / (2-1)}{s^2}, s^2 = \frac{\|Y - \hat{Y}(\text{model } 2)\|^2}{n-2}$$

### Example (Splus):

```
summary(ozonelm1) # n = 111, p = 2, n - p = 109,
          # F = 23.62, p-value = 3.964 * 10^-6
help(lm.object)
help(summary.lm)
```

```

ozsummary1=summary(ozonelm1)
ozsummary1
ozsummary1$fstatistic

yhat1=rep(mean(ozone),111)      #  $\widehat{Y}(\text{model 1}) = \begin{bmatrix} \bar{Y} \\ \bar{Y} \\ \vdots \\ \bar{Y} \end{bmatrix}_{111 \times 1}$ 
yhat2=ozonelm1$fitted.values    #  $\widehat{Y}(\text{model 2})$ 

diffm2m1=sum((yhat2-yhat1)^2)   #  $\|Y - \widehat{Y}(\text{model 2})\|^2$ 
s2=sum((ozone-yhat2)^2)/109
#  $s^2 = \frac{\|Y - \widehat{Y}(\text{model 2})\|^2}{109}$ 

f=diffm2m1/s2                  # F statistic for testing  $H_0: \beta_1 = 0$ 
f-ozsummary1$fstatistic[1]
ozsummary1$sigma^2             #  $s^2$ 

```

**(b) Test general hypothesis:**

$H_0$ : some linear equations, for example,

$$H_0: \beta_0 - 2\beta_1 - 4\beta_2 = 0, \beta_1 + 5\beta_2 - 3\beta_3 = 0.$$

In general,

Model  $p$ :  $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$  (the original (full) model)

Model  $q$ :  $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1} X_{q-1} + \varepsilon$  (the model as  $H_0$  is true)

The test statistic is

$$F = \frac{\left[ \|\widehat{Y}(\text{model } p) - \widehat{Y}(\text{model } q)\|^2 / (p - q) \right]}{s^2}$$

**Example:**

Let  $p = 4$  and  $H_0: \beta_1 = \beta_3 = 0$ . Then

Model 4:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$  (the original (full) model)

Model 2:  $Y = \beta_0 + \beta_2 X_2 + \varepsilon$  (the reduce model as  $H_0: \beta_1 = \beta_3 = 0$  is true).

We need to find  $\widehat{Y}(\text{model 4})$  and  $\widehat{Y}(\text{model 2})$  by fitting the two models, then compute the F statistic.

**Example (Splus):**

```

yhat4=xm%*%b                         #  $\widehat{Y}(\text{model } 4)$ 
ozonelm6=lm(ozone~temper)             #  $\text{ozone} = \beta_0 + \beta_2(\text{temperature}) + \varepsilon$ 
yhat2=ozonelm6$fitted.values        #  $\widehat{Y}(\text{model } 2)$ 
diffm4m2=sum((yhat4-yhat2)^2)/(4-2)
# 
$$\frac{\|\widehat{Y}(\text{model } 4) - \widehat{Y}(\text{model } 2)\|^2}{4 - 2}$$

s2=sum((ozone-yhat4)^2)/107          #  $s^2$ 
f=diffm4m2/s2                        # F statistic for testing  $H_0: \beta_1 = \beta_3 = 0$ 
f

```

**3. Prediction:****(a) Prediction:**

The fitted equation is  $\widehat{Y} = b_0 + b_1 X_1 + \cdots + b_{p-1} X_{p-1}$ . Thus, the predicted value for an observation with covariates  $X_{(n+1)1}, X_{(n+1)2}, \dots, X_{(n+1)(p-1)}$  is

$$\widehat{Y}_{n+1} = b_0 + b_1 X_{(n+1)1} + \cdots + b_{p-1} X_{(n+1)(p-1)} = X_{n+1} \mathbf{b},$$

where  $X_{n+1} = [1 \quad X_{(n+1)1} \quad \cdots \quad X_{(n+1)(p-1)}]$ .

**Example (Splus):**

```

summary(air$radiation)
summary(air$temperature)
summary(air$wind)
newx=data.frame(radi=c(5,340),temper=c(55,100),wind=c(2,22))
predict(ozonelm3,newx)

```

```

x1=c(1,5,55,2)
newy=sum(ozonelm3$coefficient*x1)      #  $\widehat{Y}_{n+1} = X_{n+1} \mathbf{b}$ 
newy
predict(ozonelm3)                      #  $\widehat{Y}_1, \widehat{Y}_2, \dots, \widehat{Y}_n$ 

```

**(b) Confidence interval:**

$$E(\widehat{Y}_{n+1}) = \beta_0 + \beta_1 X_{(n+1)1} + \cdots + \beta_{p-1} X_{(n+1)(p-1)}$$

and

$$s.e.(\widehat{Y}_{n+1}) = \{s^2 [X_{n+1} (X^t X)^{-1} X_{n+1}^t]\}^{1/2}.$$

Then

$$\widehat{Y}_{n+1} \pm t_{n-p,\alpha/2} s.e.(\widehat{Y}_{n+1})$$

$$= [\widehat{Y}_{n+1} - t_{n-p,\alpha/2} s.e.(\widehat{Y}_{n+1}), \widehat{Y}_{n+1} + t_{n-p,\alpha/2} s.e.(\widehat{Y}_{n+1})]$$

is a  $100(1 - \alpha)\%$  confidence interval for  $E(\widehat{Y}_{n+1})$ .

**Example (Splus):**

```
x=cbind(1,rad,temper,wind)
error=qt(0.975,107)*sqrt(s2*(x1%*%solve(t(x)%*%x)%*%as.matrix(x1)))
newy-error # lower limit of 95% confidence interval
newy+error # upper limit of 95% confidence interval
```

#### 4. Model selection:

Mallows'  $C_p$  statistic:

$$C_p = \frac{RSS(model\ p)}{s^2} - (n - 2p),$$

where  $RSS(model\ p)$  is the residual sum of squares for model  $p$  and  $s^2$  is the mean residual sum of squares of the model with all covariates.

$AIC_p$  and  $SBC_p$  (or  $BIC_p$  criteria):

$$AIC_p = n \cdot \ln[RSS(model\ p)] - n \cdot \ln(n) + 2p$$

$$SBC_p = n \cdot \ln[RSS(model\ p)] - n \cdot \ln(n) + \ln(n) \cdot p.$$

**Example:**

Here is a set of data with the model  $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, i = 1, \dots, 5.$

$Y_i$	15	15	25	10	30
$X_{i1}$	-2	-1	0	1	2
$X_{i2}$	1	-1	0	-1	1

Find all possible  $AIC_p$ ,  $SBC_p$ , and  $C_p$  values and determine the best model based on these criteria.

**Example (Splus):**

```
y=c(15,15,25,10,30)
x0=rep(1,5)
x1=c(-2,-1,0,1,2)
x2=c(1,-1,0,-1,1)
n=length(y)
x=cbind(x0,x1,x2)
```

```

bhat=solve(t(x)%*%x)%*%t(x)%*%y
yhat=x%*%bhat
s2=sum((y-yhat)^2)/(5-3)

# x0
bhat0=solve(t(x0)%*%x0)%*%t(x0)%*%y
yhat0=x0%*%bhat0
rss0=sum((y-yhat0)^2)
cp0=(rss0/s2)-(5-2*1)
aic0=n*log(rss0)-n*log(n)+2*1
bic0=aic0-2*1+log(n)*1

# x0, x1
x01=cbind(x0,x1)
bhat01=solve(t(x01)%*%x01)%*%t(x01)%*%y
yhat01=x01%*%bhat01
rss01=sum((y-yhat01)^2)
cp01=rss01/s2-(5-2*2)
aic01=n*log(rss01)-n*log(n)+2*2
bic01=aic01-2*2+log(n)*2

# x0, x2
x02=cbind(x0,x2)
bhat02=solve(t(x02)%*%x02)%*%t(x02)%*%y
yhat02=x02%*%bhat02
rss02=sum((y-yhat02)^2)
cp02=rss02/s2-(5-2*2)
aic02=n*log(rss02)-n*log(n)+2*2
bic02=aic02-2*2+log(n)*2

# x0, x1, x2
x012=cbind(x0,x1,x2)
bhat012=solve(t(x012)%*%x012)%*%t(x012)%*%y
yhat012=x012%*%bhat012
rss012=sum((y-yhat012)^2)
cp012=rss012/s2-(5-2*3)
aic012=n*log(rss012)-n*log(n)+2*3

```

```
bic012=aic012-2*3+log(n)*3

table=cbind(c(cp0,cp01,cp02,cp012),c(aic0,aic01,aic02,aic012),
c(bic0,bic01,bic02,bic012))
table

### By command "step"
examplelm=lm(y~x1+x2)
step(examplelm,~x1+x2)
step(examplelm,~x1+x2,trace=F)
```

## Useful links:

R Commander: <http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/>