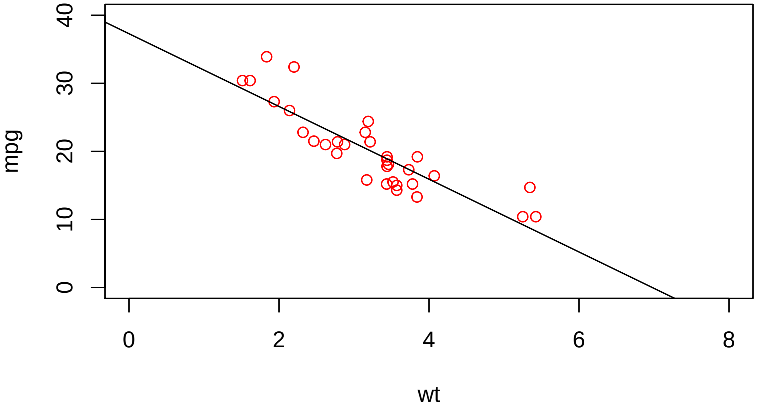
R language and data analysis: Linear regression

Qiang Shen

Jan 2, 2018

Bivariate linear regression

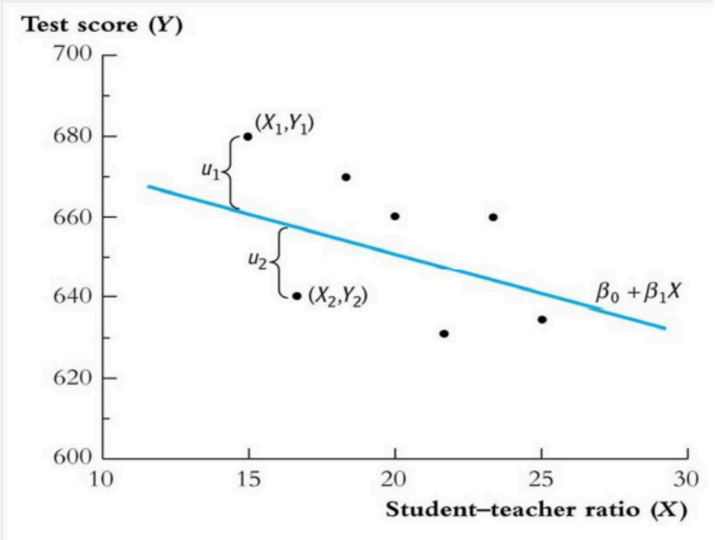
$$Y = \beta_0 + \beta_1 X_1 + \mu$$



Ordinary Least Square (OLS)

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

Ordinary Least Square (OLS)



Ordinary Least Square (OLS)

- Numerical solution.

```r
dat<-mtcars[,c('wt','mpg')]
min.RSS<-function(data, par) {
  with(data, sum((mpg-(par[1] + par[2]*wt))^2))
}
result<-optim(par = c(0, 0), min.RSS, data = dat)
result$par
```

```
[1] 37.275657 -5.342921
```

Maximum Likelihood Estimation (MLE)

$$LF(\beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma^n(\sqrt{2\pi})^n} e^{\{-\frac{1}{2}\sum\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma}\}}$$

Maximum Likelihood Estimation (MLE)

$$l = ln(LF(\beta_0, \beta_1, \sigma^2)) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)]^2$$

Maximum Likelihood Estimation (MLE)

```
library(maxLik)
dat<-mtcars[,c('wt','mpg')]
wt<-dat$wt;mpg<-dat$mpg
loglik=function (para){
  N=length(wt)
  e=mpg-para[1]-para[2]*wt
  ll=-0.5*N*log(2*pi)-0.5*N*log(para[3]^2)-0.5*sum(e^2/para
  return(ll)
}
mle1=maxLik(loglik,start=c(0.1,1,1))
coef(mle1)
```

```
[1] 37.285128 -5.344472  2.949162
```

Maximum Likelihood Estimation (MLE)

```r
rm(list=ls())
library(maxLik)
dat<-mtcars[,c('wt','mpg')]
wt<-dat$wt;mpg<-dat$mpg
loglik=function (pars){
  avg = pars[1]+pars[2]*wt
  ll=sum(dnorm(mpg-avg,0,pars[3],log=T))
  return(ll)
}
mle1=maxLik(loglik,start=c(0.1,1,1))
coef(mle1)
```

```
[1] 37.285133 -5.344474  2.949164
```

Ordinary Least Square (OLS)

- analytical solution

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Ordinary Least Square (OLS)

- analytical solution

```r
x=mtcars$wt;y=mtcars$mpg
meanx=mean(x)
meany=mean(y)
beta1<-sum((x-meanx)*(y-meany))/sum((x-meanx)^2)
beta0<-meany-beta1*meanx
c(beta0,beta1)
```

```
[1] 37.285126 -5.344472
```

Ordinary Least Square (OLS)

- correlation vs. b1

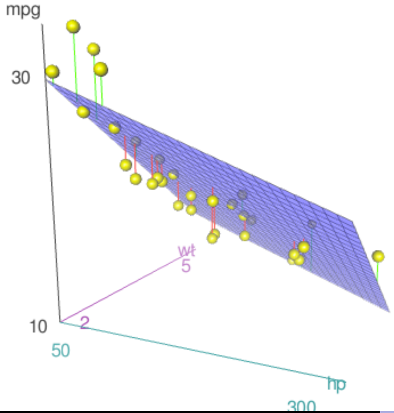
$$b_1 = r\frac{S_y}{S_x}$$

```r
x=mtcars$wt;y=mtcars$mpg
sy<-sqrt(sum((y-meany)^2))
sx<-sqrt(sum((x-meanx)^2))
cor(x,y)*sy/sx
```

```
[1] -5.344472
```

mutiple regression

```r
library(car)
library(rgl)
with(mtcars,scatter3d(x = wt, y = mpg , z = hp))
```



Coefficient with matrix

$$\mathbf{b} = (X'X)^{-1}X'Y$$

Coefficient with matrix

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 & 5 \\ 1 & 3 & 6 & 3 \\ 1 & 7 & 9 & 2 \\ 1 & 6 & 8 & 7 \\ 1 & 2 & 5 & 9 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 2 \\ 3 \\ 5 \\ 6 \\ 9 \end{bmatrix}$$

Coefficient with matrix

```r
x=mtcars$wt;y=mtcars$mpg
xmat<-cbind(1,x)
solve(t(xmat)%*%xmat) %*% t(xmat) %*% y
```

```
       [,1]
  37.285126
x -5.344472
```

```r
solve(crossprod(xmat)) %*% t(xmat) %*% y
```

```
       [,1]
  37.285126
x -5.344472
```

Coefficient with matrix

```
x=mtcars[,c('wt','hp')];y=mtcars$mpg
xmat<-as.matrix(cbind(1,x))
solve(t(xmat)%*%xmat) %*% t(xmat) %*% y
```

```
          [,1]
1  37.22727012
wt -3.87783074
hp -0.03177295
```

```
solve(crossprod(xmat)) %*% t(xmat) %*% y
```

```
          [,1]
1  37.22727012
wt -3.87783074
hp -0.03177295
```

Linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \mu$$

- Numerical solution: optimization
- Analytical solution
- Matrix

Coefficient with `lm`

```r
with(mtcars,lm(mpg~wt+hp))
```

```

Call:
lm(formula = mpg ~ wt + hp)

Coefficients:
(Intercept)           wt           hp
   37.22727     -3.87783     -0.03177
```

character vs. formula

character with or without quotation

```
equation<-mpg ~ wt + hp
class(equation)
lm(equation,mtcars)

equation<-'mpg ~ wt + hp' ## character
as.formula(equation)
PCs<-paste('PC',1:10,sep="",collapse=" +")
as.formula(paste('y ~ x +',PCs))
```

character vs. data

```r
data<-'mtcars'
equation<-'mpg ~ wt + hp' ## character
# lm(equation,data)
lm(as.formula(equation),get(data))

do.call("lm", list(as.formula(equation),as.name(data)))
# coef(summary(models))["wt","Pr(>|t|)"]
```