Regression models

Peter Hoff
STAT 423
Applied Regression and Analysis of Variance
University of Washington
Data on heights of 1375 mother-daughter pairs.

\[ x_i = \text{height of mother } i, \ y_i = \text{height of daughter } i \]
Height data

Data on heights of 1375 mother-daughter pairs.

\[ x_i = \text{height of mother } i, \ y_i = \text{height of daughter } i \]
Height data

Data on heights of 1375 mother-daughter pairs.

\[ x_i = \text{height of mother } i, \quad y_i = \text{height of daughter } i \]
mean(y[x==60])

## [1] 62.46053

sd(y[x==60])

## [1] 2.391712
Empirical conditional distribution

```r
mean(y[x==61])
## [1] 63.0407
sd(y[x==61])
## [1] 2.235695
```
mean(y[x==62])

## [1] 63.51012

sd(y[x==62])

## [1] 2.3734
Empirical conditional distribution

```
mean(y[x==63])
## [1] 64.24121
sd(y[x==63])
## [1] 2.385163
```
Empirical conditional distribution

\[
\text{mean}(y[x==64])
\]

## [1] 64.29798

\[
\text{sd}(y[x==64])
\]

## [1] 2.129583
Empirical conditional distribution

\[
\text{mean}(y[x==65])
\]

## [1] 65.25893

\[
\text{sd}(y[x==65])
\]

## [1] 2.415213
Empirical mean model and variance model

```r
mux <- tapply(y, x, "mean") ; sdx <- tapply(y, x, "sd")
mux[1:5]
##       55      56      57      58      59
## 57.00000 59.66667 60.83333 61.32500 62.03659

sdx[1:5]
##       55      56      57      58      59
##      NA 1.527525 2.229482 2.223274 2.168490
```
Residual sums of squares

```r
bols<-lm(y~x)$coef

bols
## (Intercept) x
## 31.2082961 0.5209411

yfit_ls<-bols[1]+bols[2]*x

mean( (y-yfit_ls)^2 )
## [1] 5.331938

yfit_np<-mux[ match(x,ux) ]

mean( (y-yfit_np)^2 )
## [1] 5.266022
```
Cross validation

```r
yprd_ls<-yprd_np<-NULL
for(i in 1:n)
{
  xmi<-x[-i]
  ymi<-y[-i]
  bmi<-lm(ymi~xmi)$coef
  yprd_ls<-c(yprd_ls, bmi[1] + bmi[2]*x[i] )
  yprd_np<-c(yprd_np,mean(ymi[xmi==x[i]]) )
}

mean( (y-yprd_ls)^2 )
## [1] 5.347388

mean( (y-yprd_np)^2 )
## [1] NaN

mean( (y-yprd_np)^2, na.rm=TRUE )
## [1] 5.394983
```