Regression Models week1

Reboot: ip <- installed.packages() pkgs.to.remove <- ip[!(ip[,"Priority"] %in% c("base", "recommended")), 1] sapply(pkgs.to.remove, remove.packages)

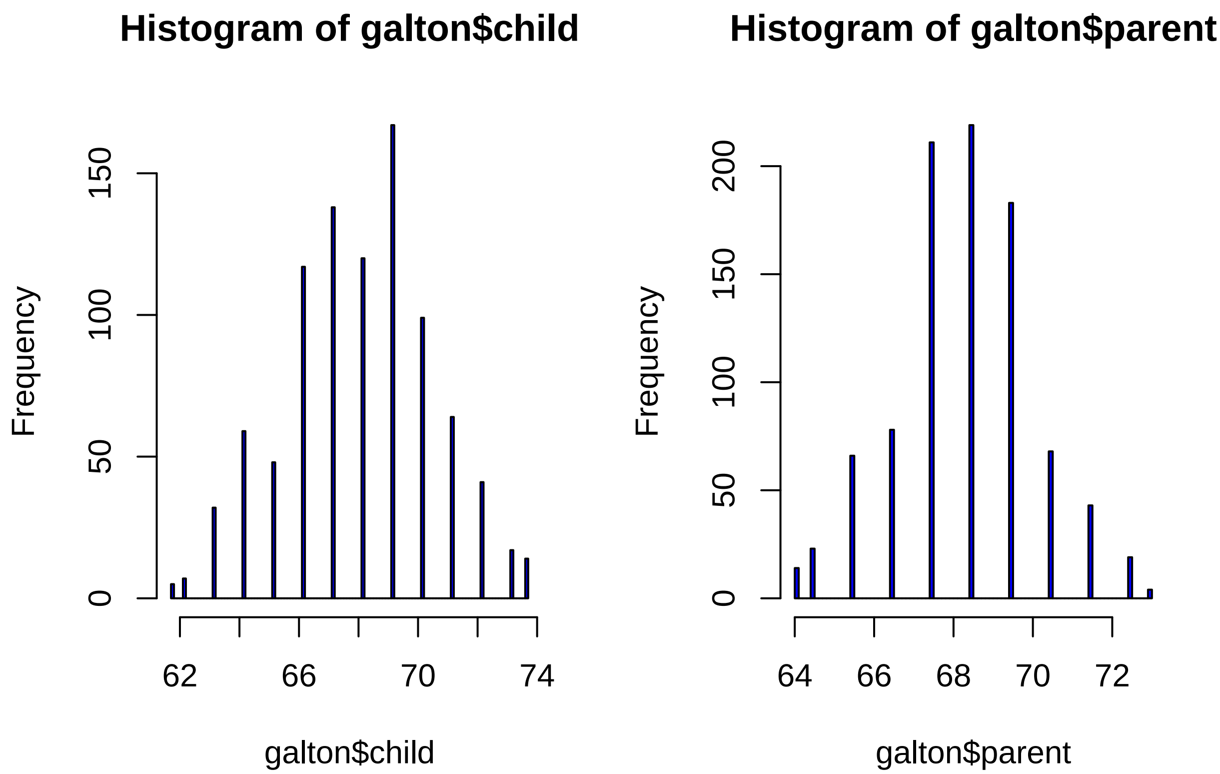
Regression

Galton, cousin of Darwin invented the idea of regression.

```r
library(UsingR)
```

```
## Loading required package: MASS
## Loading required package: HistData
## Loading required package: Hmisc
## Loading required package: grid
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
##
##
## Attaching package: 'UsingR'
##
## The following object is masked from 'package:ggplot2':
##
##     movies
##
## The following object is masked from 'package:survival':
##
##     cancer
```

```r
data(galton)
par(mfrow = c(1,2))
hist(galton$child, col = "blue", breaks = 100)
hist(galton$parent, col = "blue", breaks = 100)
```



Looks fairly similar, without the pairing.

Least square

The least square is defined as

$$\mu \to min \sum_{i=1}^{n} (Y_i - \mu)^2$$

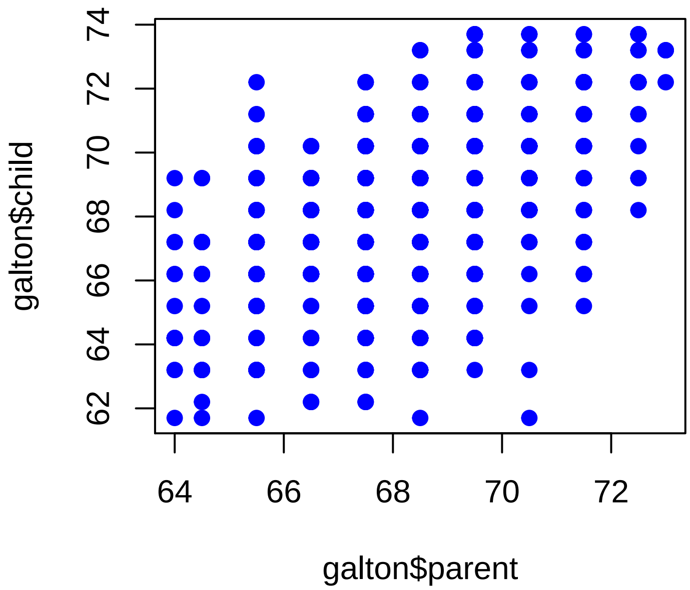
which is obviously

$$\overline{X} = \mu$$

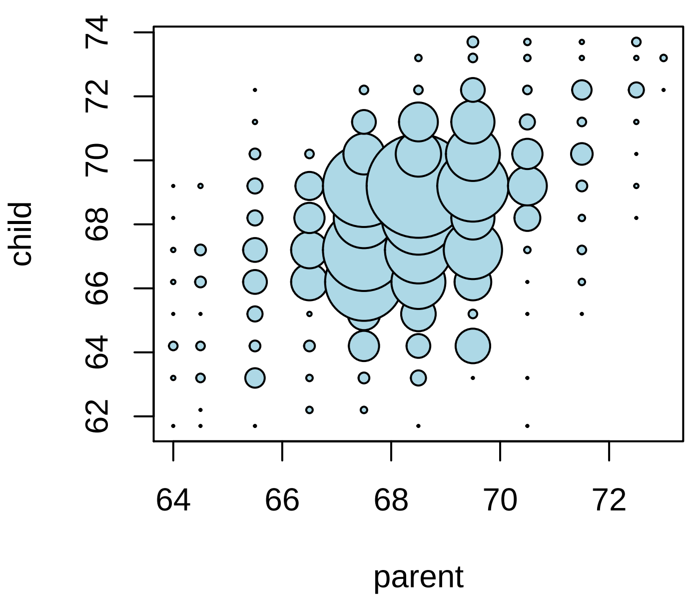
```
par(mfrow = c(1,1))
library(manipulate)
myHist <- function(mu){
  hist(galton$child, col = "blue", breaks = 100)
  lines(c(mu,mu), c(0,150), col = "red", lwd = "5")
  mse <- mean((galton$child - mu)^2)
  text(63, 150, paste("mu = ", mu))
  text(63, 140, paste("MSE = ", round(mse, 2)))
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

The least squares estimate is the empiracal mean

```
plot(galton$parent, galton$child, pch =19, col = "blue")
```




```
freqData <- as.data.frame(table(galton$child, galton$parent))
names(freqData) <- c("child", "parent", "freq")
plot(as.numeric(as.vector(freqData$parent)),
     as.numeric(as.vector(freqData$child)),
     pch = 21, col = "black", bg = "lightblue",
     cex = .15 * freqData$freq,
     xlab = "parent", ylab = "child")
```



Regression through the origin

```
myPlot <- function(beta){
  y <- galton$child - mean(galton$child)
  x <- galton$parent - mean(galton$parent)
```

```r
  freqData <- as.data.frame(table(x, y))
  names(freqData) <- c("child", "parent", "freq")
  plot(
    as.numeric(as.vector(freqData$parent)),
    as.numeric(as.vector(freqData$child)),
    pch = 21, col = "black", bg = "lightblue",
    cex = .15 * freqData$freq,
    xlab = "parent",
    ylab = "child"
    )
  abline(0, beta, lwd = 3)
  points(0, 0, cex = 2, pch = 19)
  mse <- mean( (y - beta * x)^2 )
  title(paste("beta = ", beta, "mse = ", round(mse, 3)))
}
manipulate(myPlot(beta), beta = slider(0.6, 1.2, step = 0.02))
```

R can do this

```r
lm(I(child - mean(child)) ~ I(parent - mean(parent)) -1, data = galton)
```

```
##
## Call:
## lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) -
##     1, data = galton)
##
## Coefficients:
## I(parent - mean(parent))
##                   0.6463
```

Basic notations

Sample variance, Sample covariance, sample corelation. also called emprical.

Linear least squares

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

It turns out to be

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = Cor(X, Y)\frac{Sd(X)}{Sd(Y)}$$