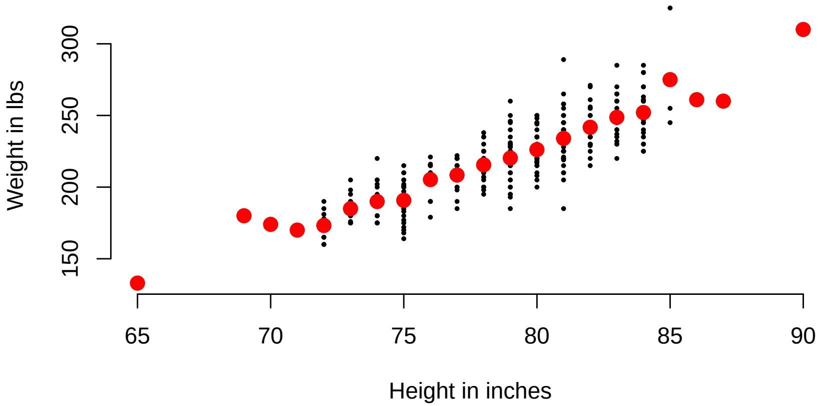
Business Statistics 41000

NBA height and weigth

Mladen Kolar

NBA height and weight: $E(Y \mid X = x)$

```
nba = read.csv("nba.csv")
EY <- aggregate(weight~height, nba, FUN='mean')
plot(weight~height,data=nba,pch=20,cex=0.5,bty='n',
     xlab='Height in inches',ylab='Weight in lbs')
points(EY$height,EY$weight,col='red',pch=20,cex=2)
```

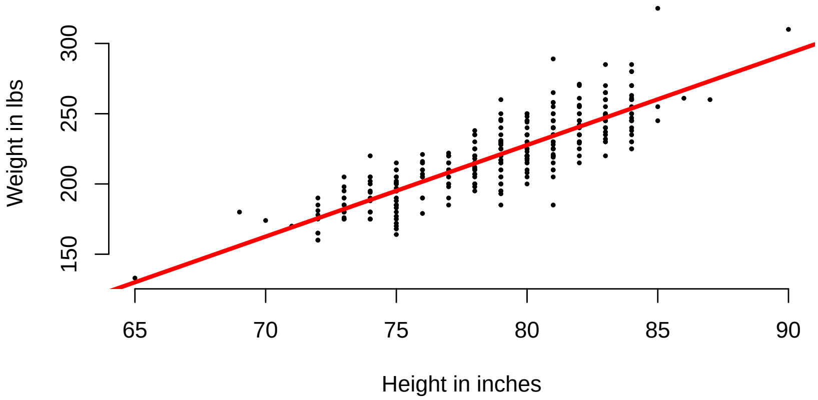


A few heights have only one observation. Is that problematic?

The least square fit line

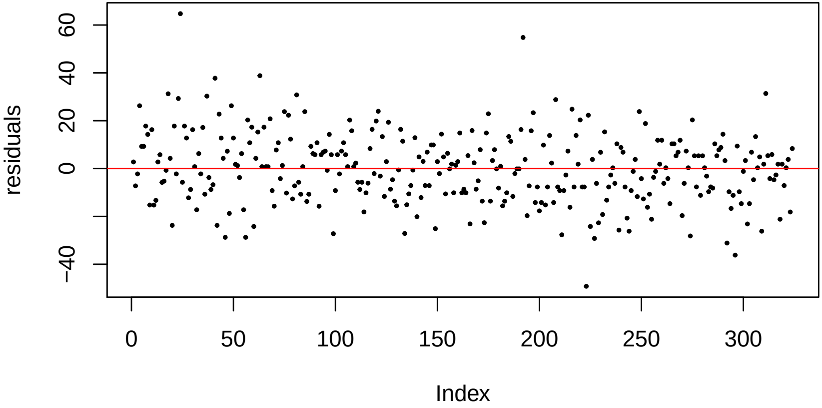
$$\widehat{\text{weight}} = -293.33 + 6.513 \cdot \text{height}$$

```r
plot(weight~height,data=nba,pch=20,cex=0.5,bty='n',
     xlab='Height in inches',ylab='Weight in lbs')
fit <- lm(weight~height,data=nba)
abline(fit,col='red',lwd=3)
```



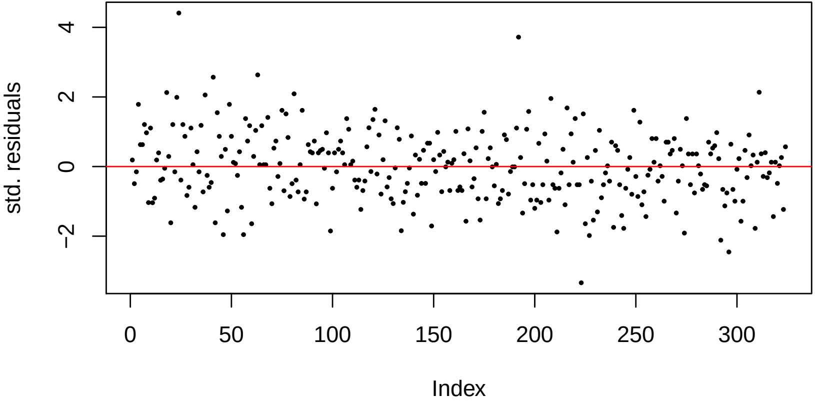
We can plot residuals

```r
plot(fit$residuals, pch=20, cex=0.5, ylab="residuals")
abline(h=0, col="red")
```



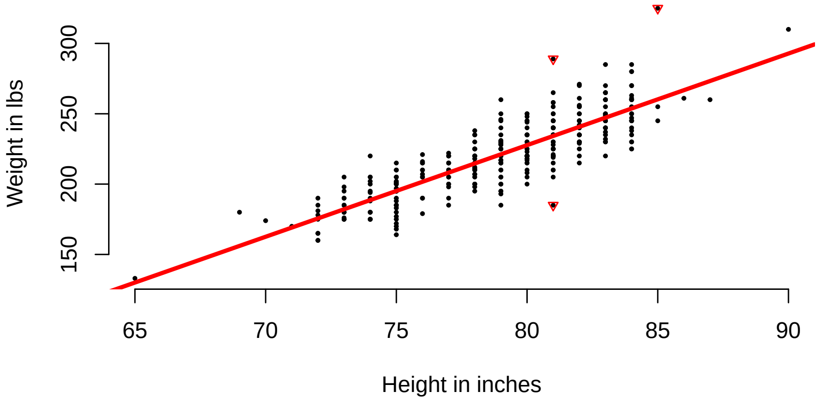
. . . or standardized residuals

```r
plot(rstandard(fit), pch=20, cex=0.5, ylab="std. residuals")
abline(h=0, col="red")
```



```
outlier = abs(rstandard(fit)) > 3
nba[outlier, ]
```

```
##                 name weight height team
## 24   oneal,shaquille    325     85  pho
## 192       davis,glen    289     81  bos
## 223     brewer,corey    185     81  min
```



```r
summary(fit)
```

```
##
## Call:
## lm(formula = weight ~ height, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.21  -9.70   0.33   9.26  64.74
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -293.329     17.825   -16.5   <2e-16 ***
## height         6.513      0.226    28.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.8 on 322 degrees of freedom
## Multiple R-squared:  0.721,  Adjusted R-squared:  0.72
## F-statistic:  833 on 1 and 322 DF,  p-value: <2e-16
```

The `anova()` command, which stands for “analysis of variance”.

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: weight
##            Df Sum Sq Mean Sq F value Pr(>F)
## height      1 181427  181427     833 <2e-16 ***
## Residuals 322  70164     218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let’s check that the sum-of-squares regression divided by the sum-of-squares total equals R^2 from the previous slide.

Forecasting interval

```r
new_height = data.frame(height=c(75, 81,85))
predict.lm(fit, newdata=new_height, interval="pred", level=0.95)
```

```
##      fit    lwr    upr
## 1 195.13 166.00 224.27
## 2 234.21 205.11 263.31
## 3 260.26 231.05 289.47
```

```r
predict.lm(fit, newdata=new_height, interval="pred", level=0.90)
```

```
##      fit    lwr    upr
## 1 195.13 170.70 219.56
## 2 234.21 209.81 258.61
## 3 260.26 235.77 284.76
```