Linear Modeling in R

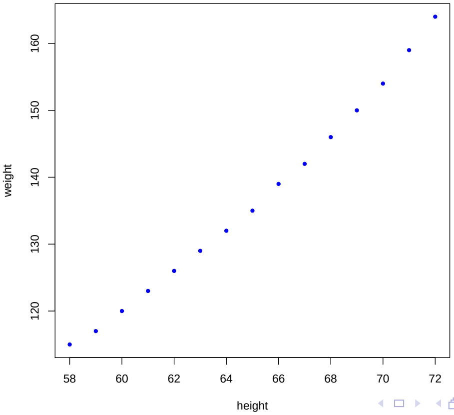
©2015 – Stott Parker

UCLA

April 20, 2015

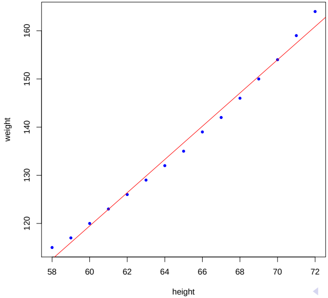
Massive Dataset

```
data(women)      # R demo dataset with only two columns
str(women)       # get quick summary of dataset structure
attach(women)    # make the namespace of "women" accessible

plot(women, pch=20, col="blue") # plot the dataset
```



Linear Models in R

```
data(women)     # R demo dataset with only two columns
str(women)      # get quick summary of dataset structure
attach(women)   # make the namespace of "women" accessible

plot(women, pch=20, col="blue") # plot the dataset

linear.model <- lm( weight ~ height )  # linear model: weight vs height
abline(linear.model, col="red") # superimpose a red line (the fit)

summary(linear.model)
```



Linear Models in R

```
> data(women)     # R demo dataset with only two columns
> str(women)      # get quick summary of dataset structure
'data.frame': 15 obs. of  2 variables:
 - height: num  58 59 60 61 62 63 64 65 66 67 ...
 - weight: num  115 117 120 123 126 129 132 135 139 142 ...
> attach(women)  # make the namespace of "women" accessible
>
> plot(women, pch=20, col="blue") # plot the dataset
>
> linear.model <- lm( weight ~ height )  # linear model: weight vs heig
> abline(linear.model, col="red") # superimpose a red line (the fit)
>
> linear.model

Call:
lm(formula = weight ~ height)

Coefficients:
(Intercept)       height
     -87.52         3.45
```

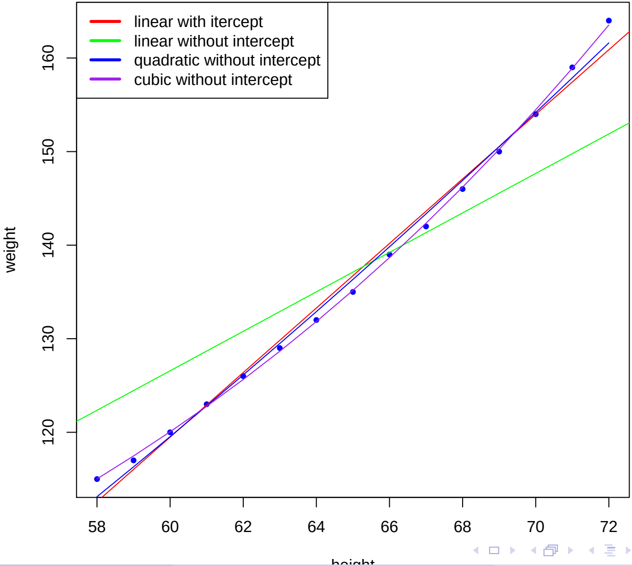
Linear Models in R

```
> linear.model = lm(weight ~ height)
> summary(linear.model)

Call:
lm(formula = weight ~ height)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7333 -1.1333 -0.3833  0.7417  3.1167

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-squared:  0.991, Adjusted R-squared:  0.9903
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

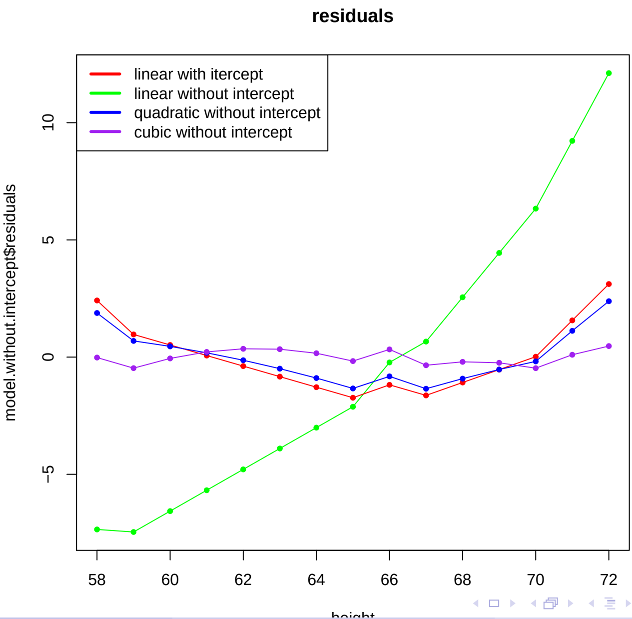
Quadratic and Cubic Models

```
plot(women, pch=20, col="blue") # plot the dataset
abline(linear.model, col="red") # superimpose a red line (the fit)

model = lm( weight ~ height - 1 )
abline(model, col="green")
summary(model)

quadratic.model = lm( weight ~ height + I(height^2) - 1 )
summary(quadratic.model)

cubic.model = lm( weight ~ height + I(height^2)  + I(height^3) - 1 )
summary(cubic.model)

input.points = data.frame( height )

quadratic.predictions = predict( quadratic.model, input.points )
lines(height, quadratic.predictions, col="blue")

cubic.predictions = predict( cubic.model, input.points )
lines(height, cubic.predictions, col="purple")
```

Results: Quadratic and Cubic Models



Residuals of all Models



Linear Models in R

$Y \sim X \mid G$ (Y is modeled as X (with models grouped by G).)

- $MPG \sim Weight + Horsepower$

- $Fuel \sim Weight + Horsepower \mid Country$
- $Fuel \sim Weight + Horsepower + Weight : Horsepower \mid Country$
- $Fuel \sim Weight$ * $Horsepower \mid Country$

- $Fuel \sim Weight + Horsepower \mid$ *cut(Power,2)*
- $Fuel \sim Weight + Horsepower \mid Make$ * $Model$
- *1/MPG* $\sim$ *poly(Weight,3)* $+$ *sqrt(Power)* $\mid Make$ * $Model$ * $Country$

General Linear Models are expressible this way.
Expressions can be nested also, permitting hierarchical structure.