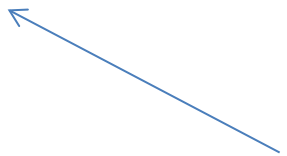
# NRES_798_7_201501

Variance partitioning in regression and ANOVA models

# Linear models

$$Y \sim Normal(b_0 + b_1 X, \sigma^2)$$

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$



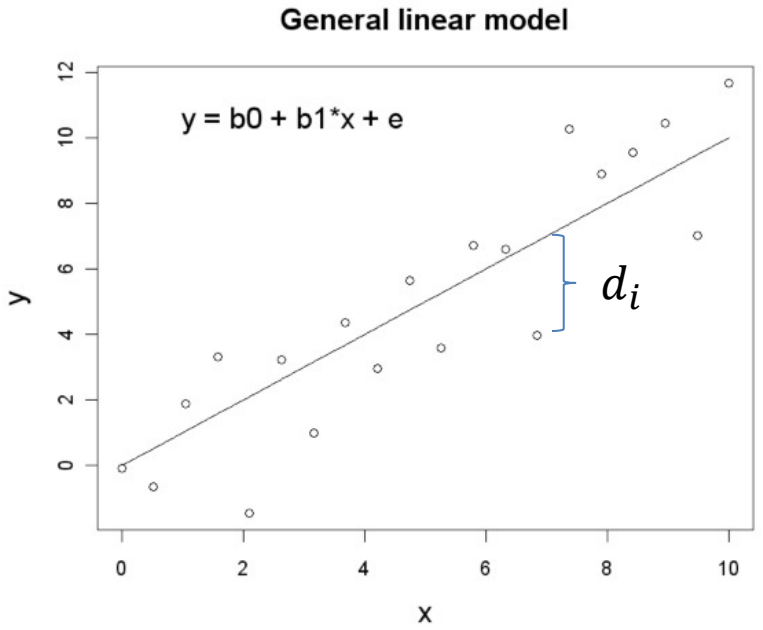
$$Normal(0, \sigma^2)$$

# Estimating of parameters

- We want to find the equation of the line that "best" fits the data. It means finding $b_0$ and $b_1$ such that the fitted values of $y_i$ given by

$$\hat{Y}_i = b_0 + b_1 X_i$$

- Are as close as possible to the observed values of $y_i$




$$d_i = y_i - \hat{y}_i$$

Fit $b_0$ and $b_1$ such that

$$\sum {d_i}^2$$

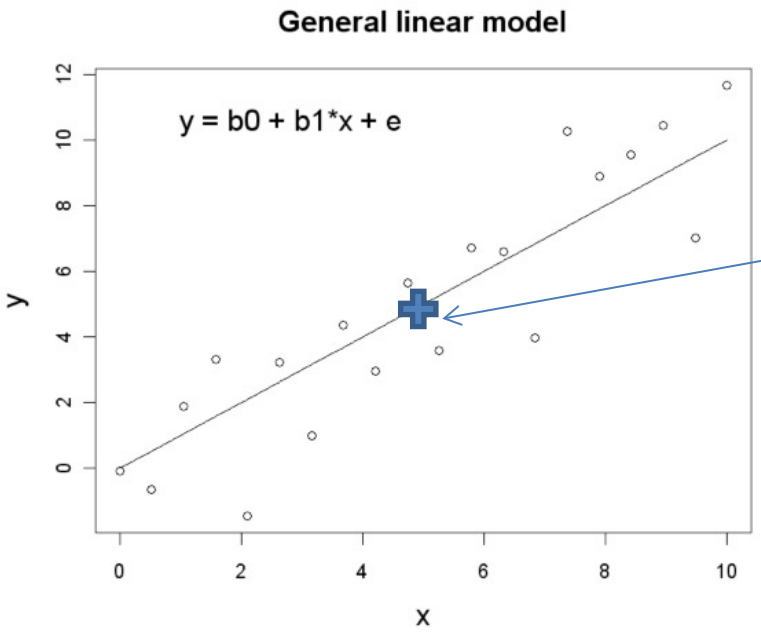
is minimized

# Minimizing residual sum of squares

$$\sum {d_i}^2$$

$$\sum {d_i}^2 = \sum (y_i - b_0 - b_1 X_i)^2$$

Least square estimate of $b_1$




Midpoint of the data, "eye" $(\bar{X}, \bar{Y})$

$\bar{X}$ estimate of the mean along the X axis

$\bar{Y}$ estimate of the man along the Y axis

# Least-squares estimates

$$\sum {d_i}^2 = \sum (y_i - b_0 - \ b_1 X_i)^2$$

- To minimize the equation we need

Predicted regression value

Residual Sum of Squares $$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

"eye" of the data

Sum of Squares of a variable $$ss_x = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Sample variance of a variable $${s_x}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Sum of the cross products $$ss_{xy} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Sample covariance $$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

# Least-squares Parameter Estimates: Slope and intercept

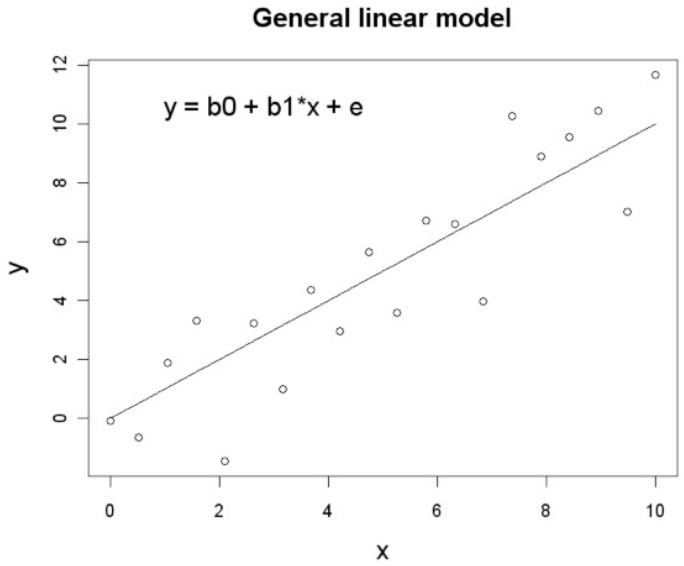
$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

$$\widehat{\beta_1} = \frac{s_{xy}}{{s_x}^2} = \frac{SS_{xy}}{SS_x}$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$${s_x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

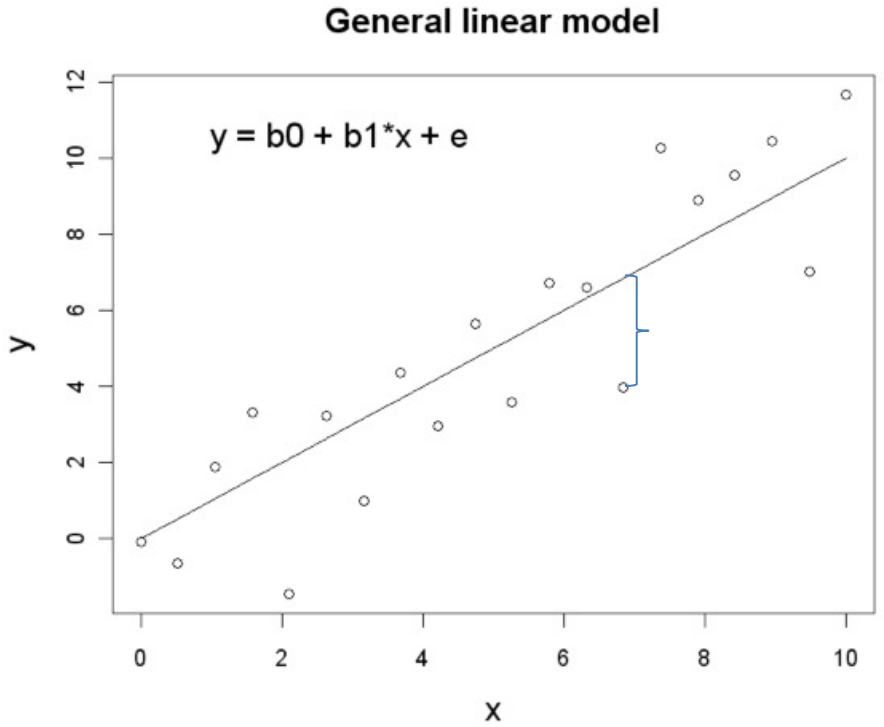
$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$



# Least-squares Parameter Estimates: error term (variance)

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

$$\varepsilon_i \sim Normal(0, \sigma^2)$$




$$\widehat{\sigma^2} = \frac{RSS}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2}{n-2} = \frac{\sum(y_i - b_0 - b_1 X_i)^2}{n-2}$$

Standard error of regression

# Variance components

- Variance partitioning:
  - Determining how much of the observed variance can be attributed to different factors.
  - Partition a sum of squares into different components (sources, factors)
  - The relative importance of a factor being quantified as its relative SS contribution to the overall sum of squares

# Variance components

- Components of variation
  - Pure or random error (random sampling from a random variable (normal distribution))
    - RSS
  - Systematic variation related to a variable(s)
    - $SS_{reg}$

$$SS_y = SS_{reg} + RSS$$

Total variation in Y

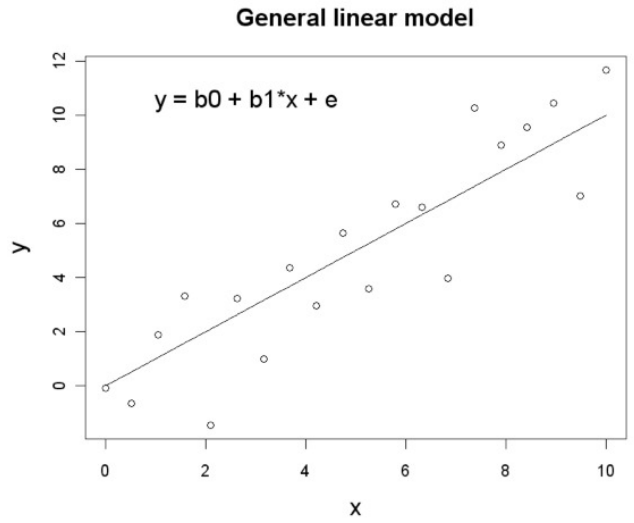
Variation due to regression model

Random error

$$SS_{reg} = SS_y - RSS$$

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

$$TSS = ESS + RSS$$



# Variance components

- Coefficient of determination
  - Relative importance of regression vs. residual variation
  - $r^2$, explanatory value of the regression model

$$r^2 = \frac{SS_{reg}}{SS_y} = \frac{SS_{reg}}{SS_{reg} + RSS}$$

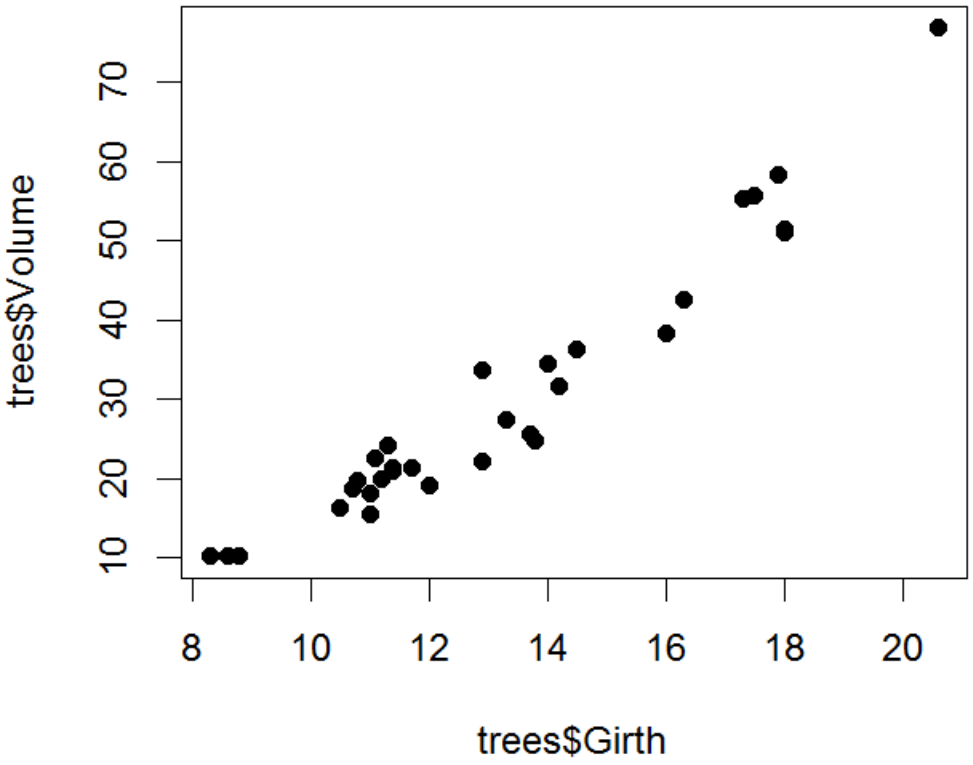
  - Scaled to 100: Percent of variation in Y that is explained by the model

- Product-moment correlation coefficient
  - r
  - Sign indicates directional relationship between X and Y (slope)

# Linear regression in R

```
> data(trees)
> summary(trees)
```

| Girth | Height | Volume |
|---|---|---|
| Min. : 8.30 | Min. :63 | Min. :10.20 |
| 1st Qu.:11.05 | 1st Qu.:72 | 1st Qu.:19.40 |
| Median :12.90 | Median :76 | Median :24.20 |
| Mean :13.25 | Mean :76 | Mean :30.17 |
| 3rd Qu.:15.25 | 3rd Qu.:80 | 3rd Qu.:37.30 |
| Max. :20.60 | Max. :87 | Max. :77.00 |

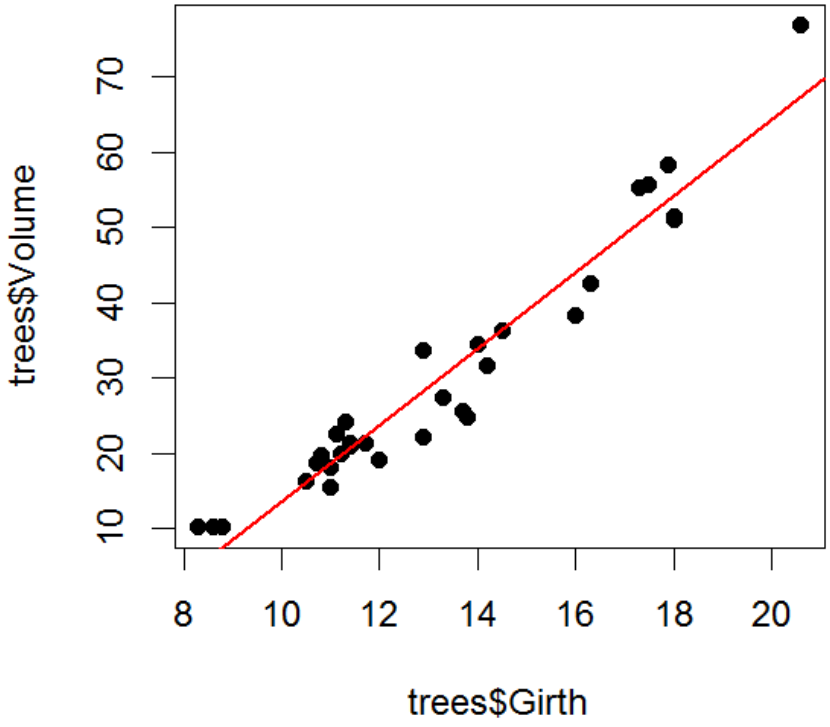


# Linear regression in R

# R General linear model (lm) framework

# lm.reg <- lm(y ~ x)

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$




lm.reg1 <- lm(Volume ~ Girth, data = trees)

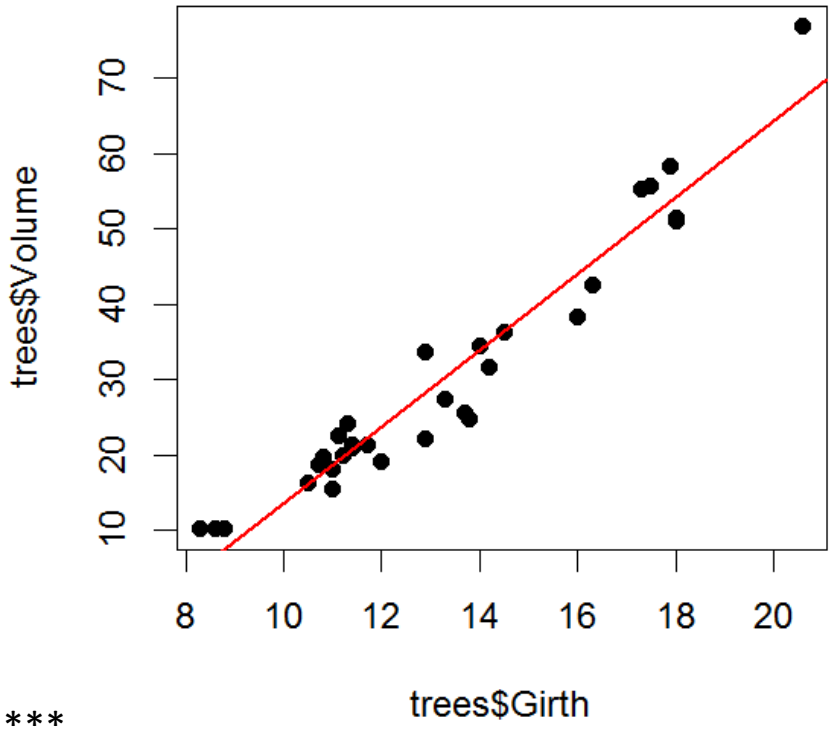
# Linear regression in R

```
lm.reg1 <- lm(Volume ~ Girth, data = trees)
```

```
> is.list(lm.reg1)
[1] TRUE
```

```
> names(lm.reg1)
 [1] "coefficients"  "residuals""effects"  "rank"  "fitted.values" "assign""qr"  "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

# Linear regression in R




```
> summary(lm.reg1)

Call:
lm(formula = Volume ~ Girth, data = trees)

Residuals:
   Min     1Q Median     3Q    Max
-8.065 -3.107  0.152  3.495  9.587

Coefficients:
             Estimate Std. Error t value     Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353,    Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

# Linear regression in R

```
> anova(lm.reg1)
Analysis of Variance Table

Response: Volume
          Df  Sum Sq Mean Sq F value    Pr(>F)
Girth      1 7581.8  7581.8  419.36 < 2.2e-16 ***
Residuals 29  524.3    18.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Linear regression in R

lm(Volume ~ Girth)

```
> anova(lm.reg1)
Analysis of Variance Table

Response: Volume
          Df  Sum Sq Mean Sq F value    Pr(>F)
Girth      1 7581.8  7581.8  419.36 < 2.2e-16 ***
Residuals 29  524.3    18.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## Variance Components

$$SS_y = SS_{reg} + RSS$$

```
SSy  <- deviance(lm(Volume ~ 1,data = trees))
RSS <- deviance(lm(Volume ~ Girth, data = trees))
SSreg <- SSy - RSS
```

```
> SSy     8106.084
> RSS     524.3025
> SSreg  7581.781
```

$$r^2 = \frac{SS_{reg}}{SS_y} = \frac{7581.8}{8106.1} = 0.9353$$

# Linear regression in R

## Variance Components

$$SS_y = SS_{reg} + RSS$$

```
SSy  <- deviance(lm(Volume ~ 1,data = trees))
RSS <- deviance(lm(Volume ~ Girth, data = trees))
SSreg <- SSy - RSS
```

```
> SSy     8106.084
> RSS     524.3025
> SSreg  7581.781
```

$$r^2 = \frac{SS_{reg}}{SS_y} = \frac{7581.8}{8106.1} = 0.9353$$

```
> summary(lm.reg1)

Call:
Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared: 0.9353, Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

# Multiple linear regression in R

```
# Multiple linear regression, no interaction
lm.reg2 <- lm(Volume ~ Girth + Height, data = trees)
Summary(lm.reg2)
```

```
Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
             Estimate   Std. Error  t value      Pr(>|t|)
(Intercept) -57.9877     8.6382     -6.713       2.75e-07 ***
Girth         4.7082     0.2643     17.816       < 2e-16 ***
Height        0.3393     0.1302      2.607       0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,          Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

# Multiple linear regression in R

```
> anova(lm.reg2)
Analysis of Variance Table

Response: Volume
           Df  Sum Sq Mean Sq F value    Pr(>F)
Girth       1 7581.8  7581.8  503.1503 < 2e-16 ***
Height      1  102.4   102.4    6.7943 0.01449 *
Residuals  28  421.9    15.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> anova(lm.reg1)
Analysis of Variance Table

Response: Volume
           Df  Sum Sq Mean Sq F value    Pr(>F)
Girth       1 7581.8  7581.8  419.36 < 2.2e-16 ***
Residuals  29  524.3    18.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Multiple linear regression in R

```
# Multiple linear regression, no interaction
lm.reg2 <- lm(Volume ~ Girth + Height, data = trees)
```

```
Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
             Estimate   Std. Error   Pr(>|t|)
(Intercept) -57.9877     8.6382      2.75e-07 ***
Girth         4.7082     0.2643      < 2e-16 ***
Height        0.3393     0.1302      0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

```
> summary(lm.reg1)

Call:
lm(formula = Volume ~ Girth, data = trees)

Residuals:
   Min     1Q Median     3Q    Max
-8.065 -3.107  0.152  3.495  9.587

Coefficients:
             Estimate   Std. Error   Pr(>|t|)
(Intercept) -36.9435     3.3651      7.62e-12 ***
Girth         5.0659     0.2474      < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

# Multiple linear regression in R

$$Y \sim Normal\left(b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2, \sigma^2\right)$$

```
# With interaction
Girthc <- trees$Girth - mean(trees$Girth)          # centre the Girth variable
Heightc <- trees$Height - mean(trees$Height)       # centre the Height variable
trees$GH.int <- Girthc*Heightc                      # create the interaction term

lm.reg3 <- lm(Volume ~ Girth + Height + GH.int, data = trees)

# Alternative formulation
lm.reg4 <- lm(Volume ~ Girth + Height + Girth:Height, data = trees)
```

---

Response: Volume

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Girth | 1 | 7581.8 | 7581.8 | 1033.469 | < 2.2e-16 *** |
| Height | 1 | 102.4 | 102.4 | 13.956 | 0.0008867 *** |
| Girth:Height | 1 | 223.8 | 223.8 | 30.512 | 7.484e-06 *** |
| Residuals | 27 | 198.1 | 7.3 | | |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Multiple linear regression in R

```
> summary(lm.reg3)

Call:
lm(formula = Volume ~ Girth + Height + GH.int, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5821 -1.0673  0.3026  1.5641  4.6649

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) -66.18415    6.20729   -10.662  3.52e-11 ***
Girth         4.37789    0.19384    22.585   < 2e-16 ***
Height        0.48687    0.09466     5.143  2.07e-05 ***
GH.int        0.13465    0.02438     5.524  7.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.709 on 27 degrees of freedom
Multiple R-squared:  0.9756,  Adjusted R-squared:  0.9728
F-statistic: 359.3 on 3 and 27 DF,  p-value: < 2.2e-16
```

# Robust linear regression

- Minimizes the influence of outliers through a modification of how variance is calculated

$$residual = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$$

$$Robust = \sum_{i=1}^{n} \left|\left(Y_i - \widehat{Y}_i\right)\right|$$