Linear Modelling: Multiple Regression

7th of February 2023

R. Nicholls / D.-L. Couturier / C.S.R. Chilamakuri

Linear models:

Simple/single regression: $$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

Multiple regression: $$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \cdots + \beta_n \boldsymbol{x_n} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Linear models:

Simple/single regression: $$\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon}$$

Multiple regression: $$\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \cdots + \beta_n\boldsymbol{x_n} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Parameter estimation:

Minimise sum of squares of residuals: $$\sum_i {\varepsilon_i}^2 \to \min$$

$$SS_{\mathrm{error}} = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) \to \min$$

Solution: $$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Compare with the simple case: $$\hat{\beta} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y})}{\mathrm{var}(\boldsymbol{x})}$$

Linear models:

Simple/single regression: $\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon}$

Multiple regression: $\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \cdots + \beta_n\boldsymbol{x_n} + \boldsymbol{\varepsilon}$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Assumptions:

1. Model is linear in parameters.
2. Gaussian error model. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
3. Additive error model.
4. Independence of errors. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$
5. Homoscedasticity. $\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{x}) = \sigma^2\mathbf{I}$

   and…
6. Lack of multicollinearity in the predictors (no highly correlated variables).

Linear models:

Simple/single regression: $\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$

Multiple regression: $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \cdots + \beta_n \boldsymbol{x_n} + \boldsymbol{\varepsilon}$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Assumptions:

1. Model is linear in parameters.

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \boldsymbol{\varepsilon}$$

Linear models:

Simple/single regression: $\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$

Multiple regression: $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \cdots + \beta_n \boldsymbol{x_n} + \boldsymbol{\varepsilon}$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Assumptions:

1. Model is linear in parameters.

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2^2} + \boldsymbol{\varepsilon}$$

Linear models:

Simple/single regression: $$\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon}$$

Multiple regression: $$\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \cdots + \beta_n\boldsymbol{x_n} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Assumptions:

1. Model is linear in parameters.

$$\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2^2} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_1^2} + \boldsymbol{\varepsilon}$$

Linear models:

Simple/single regression: $$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

Multiple regression: $$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \cdots + \beta_n \boldsymbol{x_n} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Assumptions:

1. Model is linear in parameters.

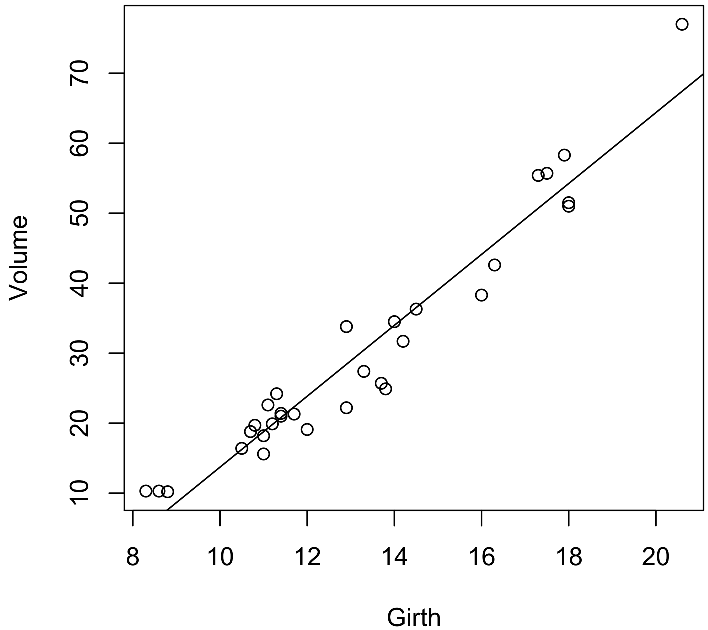
$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2^2} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_1^2} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_1^2 \boldsymbol{x_1^2} + \boldsymbol{\varepsilon}$$

Example – Predict Black Cherry Tree Timber Volume

Example: *Predicting timber volume of cherry trees*

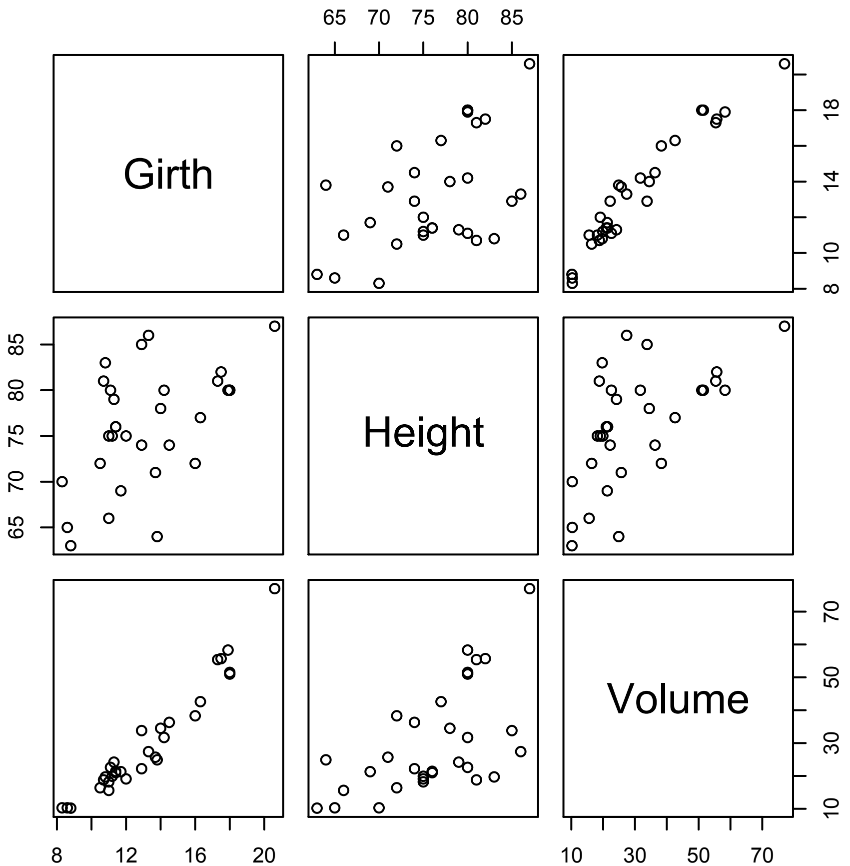
$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

```
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
```



Response: y = Volume
Predictor: x = Girth

Example – Predict Black Cherry Tree Timber Volume

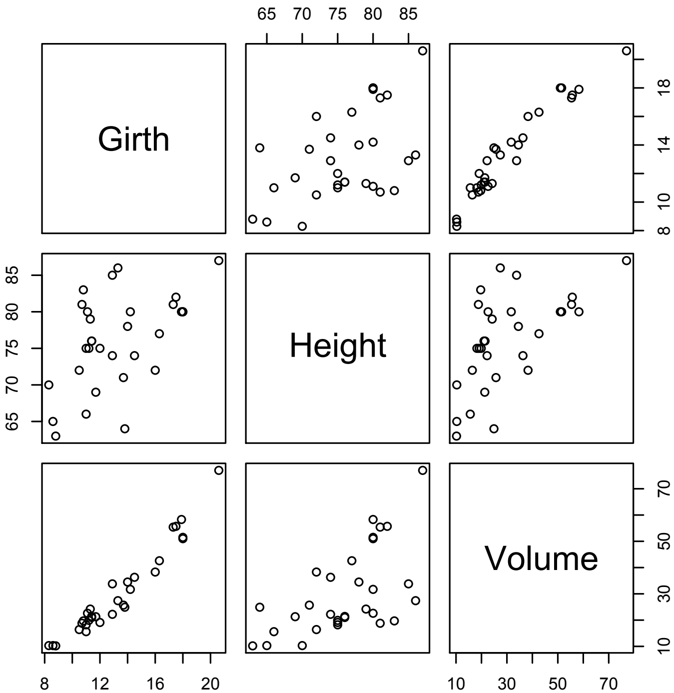


Example – Predict Black Cherry Tree Timber Volume

Simple Regression

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Response: y = Volume
Predictor: x = Girth




```
            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
```

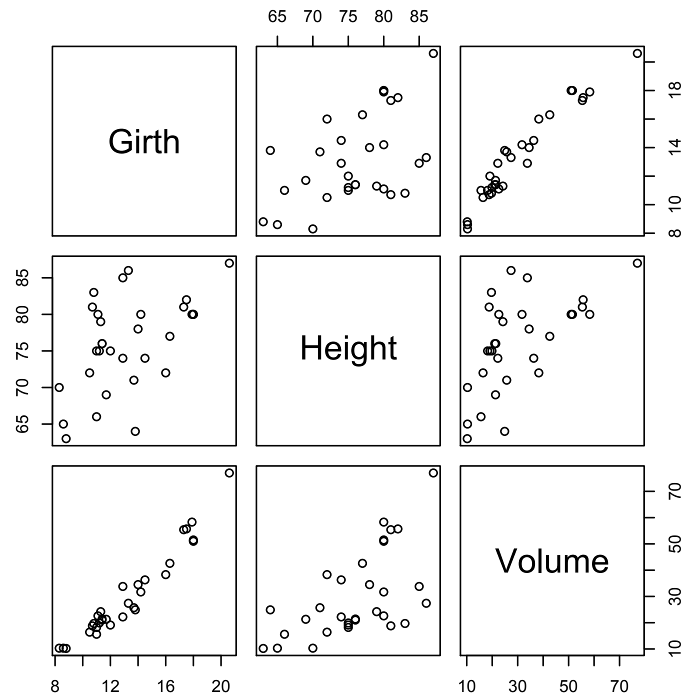
Example – Predict Black Cherry Tree Timber Volume

Multiple Regression – main effects only

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \boldsymbol{\varepsilon}$$

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height

```
            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
```



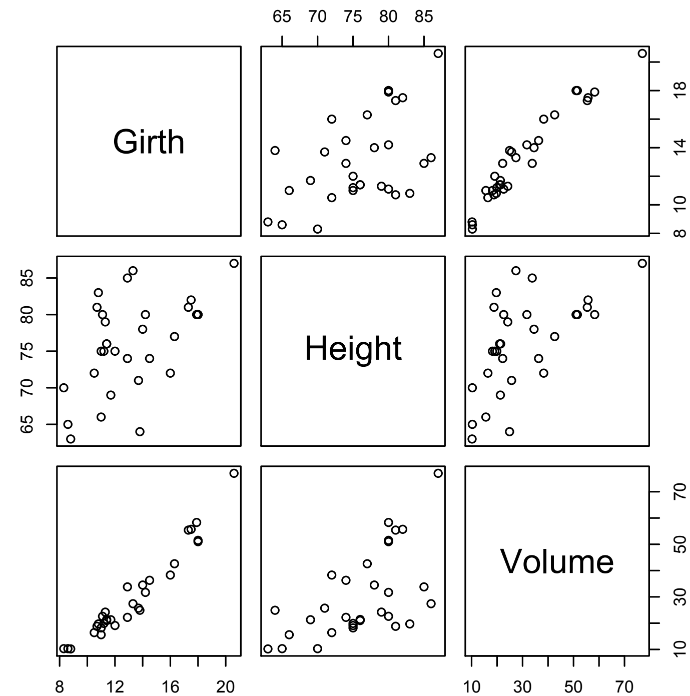
R^2 is improved
Height term is significant
But less significant than Girth

Example – Predict Black Cherry Tree Timber Volume

Multiple Regression – including interaction

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \beta_3 \boldsymbol{x_1 x_2} + \boldsymbol{\varepsilon}$$

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height




```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.39632   23.83575   2.911  0.00713 **
Girth        -5.85585    1.92134  -3.048  0.00511 **
Height       -1.29708    0.30984  -4.186  0.00027 ***
Girth:Height  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
```

R^2 is improved
All terms are significant
Height term is more significant(!)

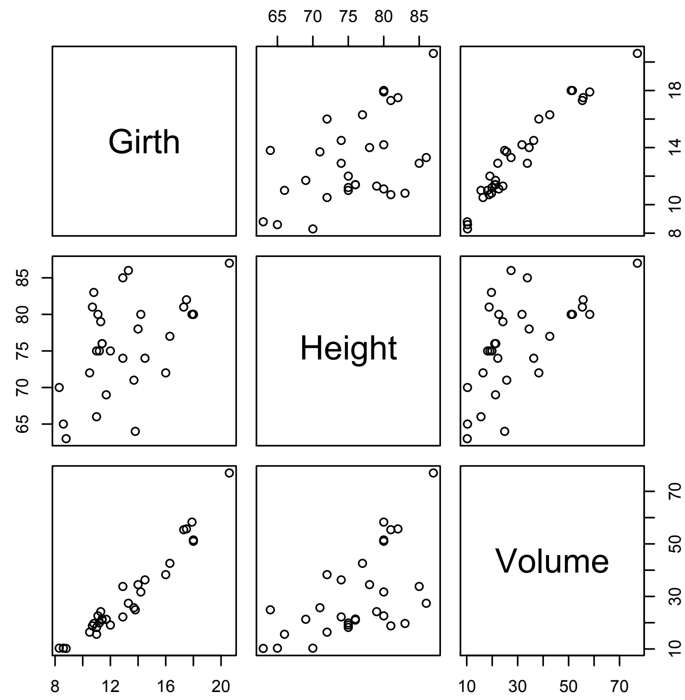
Example – Predict Black Cherry Tree Timber Volume

Log-transform response and predictors? No interaction

$$\log(\boldsymbol{y}) = \beta_0 + \beta_1 \log(\boldsymbol{x_1}) + \beta_2 \log(\boldsymbol{x_2}) + \boldsymbol{\varepsilon}$$

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
log(Girth)   1.98265    0.07501  26.432  < 2e-16 ***
log(Height)  1.11712    0.20444   5.464 7.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared:  0.9777,	Adjusted R-squared:  0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

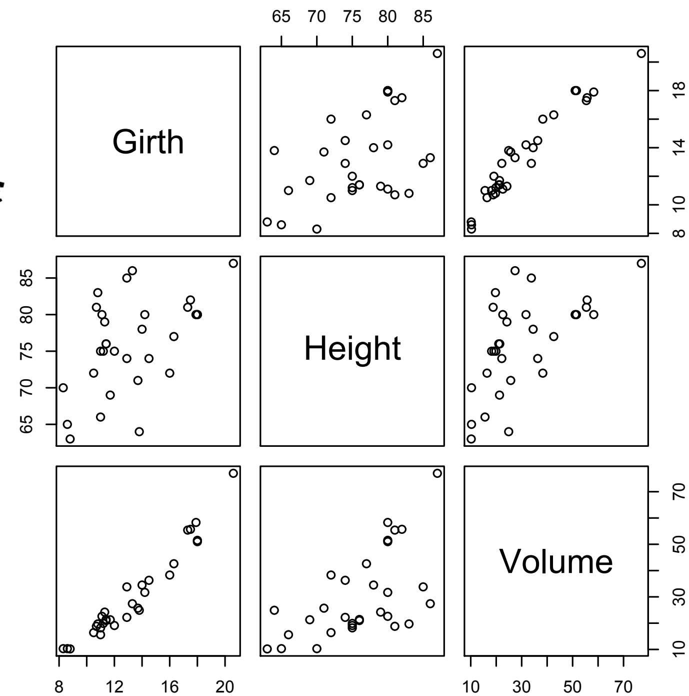


R^2 is improved
Fewer parameters
All terms are significant
Residual standard error!!!

Example – Predict Black Cherry Tree Timber Volume

Log-transform response and predictors? With interaction

$$\log(\boldsymbol{y}) = \beta_0 + \beta_1 \log(\boldsymbol{x_1}) + \beta_2 \log(\boldsymbol{x_2}) + \beta_3 \log(\boldsymbol{x_1}) \log(\boldsymbol{x_2}) + \boldsymbol{\varepsilon}$$



Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height

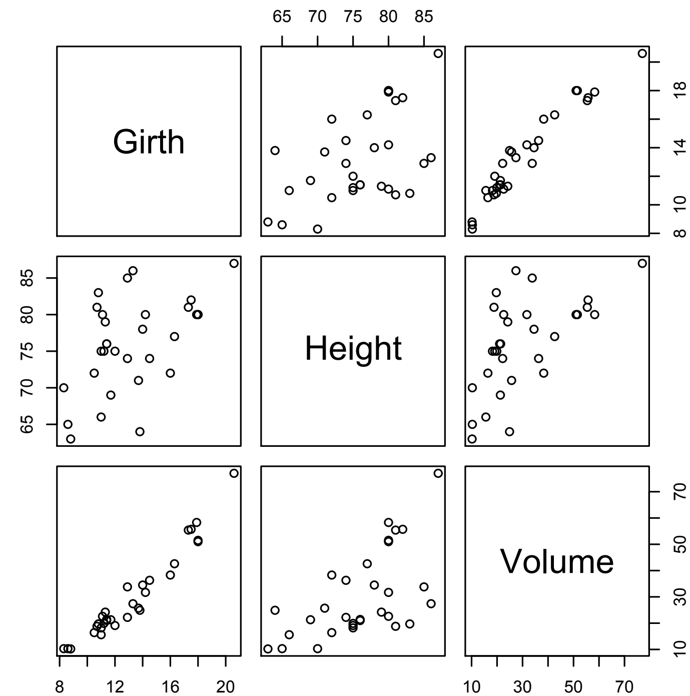
```
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            -3.6869     7.6996  -0.479    0.636
log(Girth)              0.7942     3.0910   0.257    0.799
log(Height)             0.4377     1.7788   0.246    0.808
log(Girth):log(Height)  0.2740     0.7124   0.385    0.704

Residual standard error: 0.08265 on 27 degrees of freedom
Multiple R-squared:  0.9778,   Adjusted R-squared:  0.9753
F-statistic: 396.4 on 3 and 27 DF,  p-value: < 2.2e-16
```

R^2 marginally improved
No terms are significant!!!

Example – Predict Black Cherry Tree Timber Volume

Favourite model so far:

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height



$$\log(\boldsymbol{y}) = \beta_0 + \beta_1 \log(\boldsymbol{x_1}) + \beta_2 \log(\boldsymbol{x_2}) + \boldsymbol{\varepsilon}$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
log(Girth)   1.98265    0.07501  26.432  < 2e-16 ***
log(Height)  1.11712    0.20444   5.464 7.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared:  0.9777,  Adjusted R-squared:  0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

$$\boldsymbol{y} = e^{\beta_0} \boldsymbol{x_1}^{\beta_1} \boldsymbol{x_2}^{\beta_2} e^{\boldsymbol{\varepsilon}}$$

Volume ∝ Girth^2 x Height

Confidence Intervals:

	2.5 %	97.5 %
(Intercept)	-8.269912	-4.993322
log(Girth)	1.828998	2.136302
log(Height)	0.698353	1.535894

$$\hat{\beta}_1 \approx 2$$
$$\hat{\beta}_2 \approx 1$$

$$\log(\boldsymbol{y}) = \beta_0 + 2\log(\boldsymbol{x_1}) + \log(\boldsymbol{x_2}) + \boldsymbol{\varepsilon}$$

Example – Predict Black Cherry Tree Timber Volume

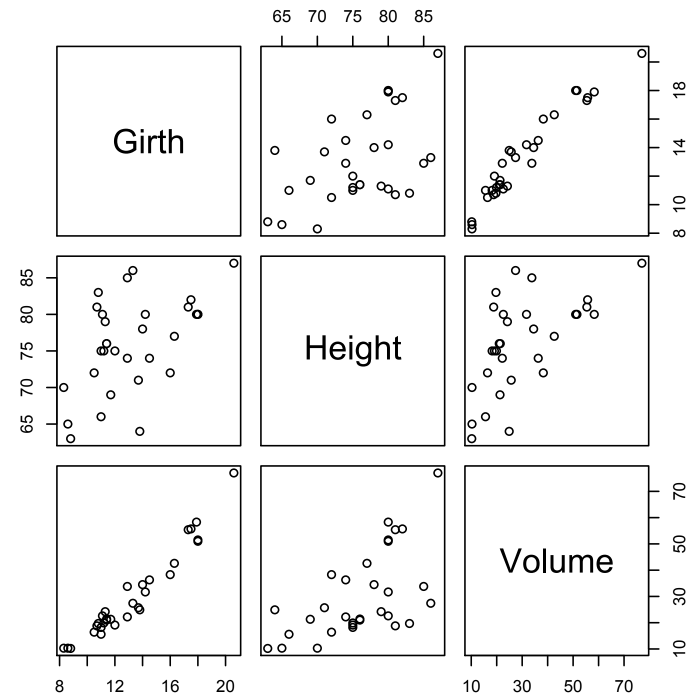
Fix parameters to mechanistically sensible values

$$\log(\boldsymbol{y}) = \beta_0 + 2\log(\boldsymbol{x_1}) + \log(\boldsymbol{x_2}) + \boldsymbol{\varepsilon}$$

$$\log\left(\frac{\boldsymbol{y}}{\boldsymbol{x_1^2 x_2}}\right) = \beta_0 + \boldsymbol{\varepsilon}$$

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.16917    0.01421  -434.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0791 on 30 degrees of freedom
```



No R^2
Intercept is significant
Again, can’t compare RSE…

Example – Predict Black Cherry Tree Timber Volume

Fix parameters to mechanistically sensible values

$$\log\left(\frac{\boldsymbol{y}}{\boldsymbol{x}_1^2\boldsymbol{x}_2}\right) = \beta_0 + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \beta_1\boldsymbol{x}_1^2\boldsymbol{x}_2 e^{\boldsymbol{\varepsilon}}$$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.16917    0.01421  -434.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0791 on 30 degrees of freedom
```

Why not instead fix the intercept, and estimate the coefficient of $x_1^2x_2$???

$$\boldsymbol{y} = \beta_1\boldsymbol{x}_1^2\boldsymbol{x}_2 + \boldsymbol{\varepsilon}$$

```
                  Estimate Std. Error t value Pr(>|t|)
I(Girth^2):Height 2.108e-03  2.722e-05   77.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.455 on 30 degrees of freedom
Multiple R-squared:  0.995,	Adjusted R-squared:  0.9949
F-statistic:  5996 on 1 and 30 DF,  p-value: < 2.2e-16
```

Produces R^2

But… R^2 incomparable when intercept is fixed.

Again, can’t compare RSE…

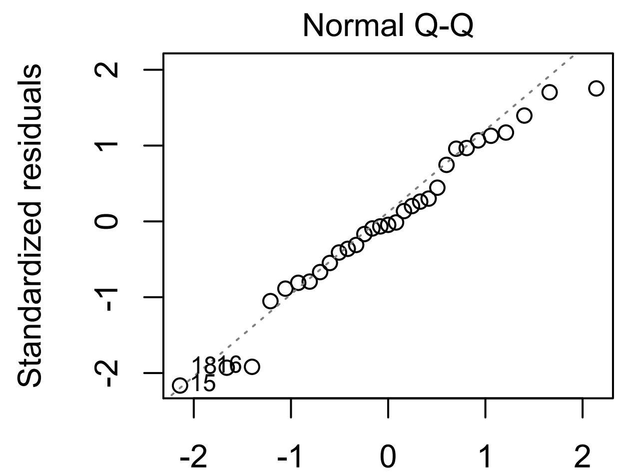
Hang on…   exp(-6.16917) = 2.092e-03 …?!

Example – Predict Black Cherry Tree Timber Volume

Multiplicative error model:

$$\log\left(\frac{\boldsymbol{y}}{\boldsymbol{x}_1^2\boldsymbol{x}_2}\right) = \beta_0 + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \beta_1\boldsymbol{x}_1^2\boldsymbol{x}_2 e^{\boldsymbol{\varepsilon}}$$

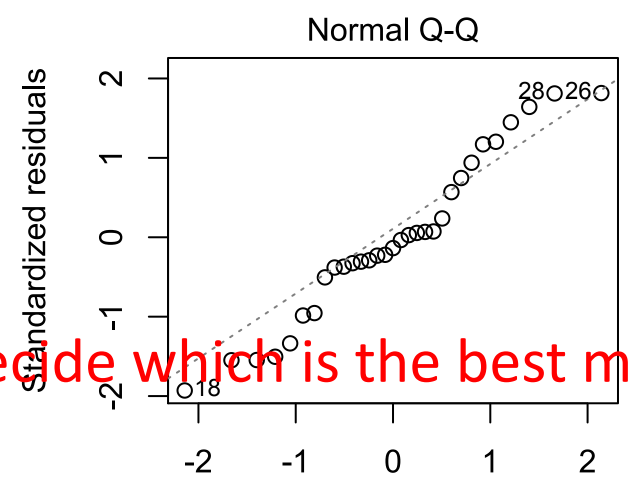


Shapiro-Wilk test

p-value: 0.5225

Additive error model:

$$\boldsymbol{y} = \beta_1\boldsymbol{x}_1^2\boldsymbol{x}_2 + \boldsymbol{\varepsilon}$$



p-value: 0.2655

How can we decide which is the best model?

Model Selection: Choosing the best model

Occam's Razor:
Among competing hypotheses, the one with the fewest assumptions should be selected

Parsimonious modelling:
Only choose a more complex model if the benefits are sufficiently substantial

We want:
1. The model that fits the data the best
2. Not to suffer from excessive overfitting

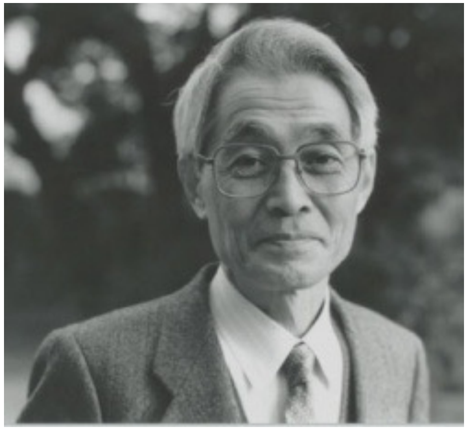
Objective solution: use "information criteria"

- Akaike information criterion – AIC (1974)
  - Measures a trade-off between model goodness-of-fit and complexity (i.e. number of parameters)
  - Used for comparing models – relative only

$$AIC = 2k - 2\log(L)$$

k : number of parameters
L : maximum of the likelihood function.

  - **Lower** AIC indicates higher quality model



Hirotugu Akaike

Gideon Schwarz

- Bayesian information criterion – BIC (1978)

$$BIC = \log(n)k - 2\log(L)$$

Example – Predict Black Cherry Tree Timber Volume

Choosing the best model…

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height

Model		R^2	AIC
$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon}$		0.9353	181.6
$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \boldsymbol{\varepsilon}$		0.9480	176.9
$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \beta_3 \boldsymbol{x_1 x_2} + \boldsymbol{\varepsilon}$		0.9756	155.5
$\log(\boldsymbol{y}) = \beta_0 + \beta_1 \log(\boldsymbol{x_1}) + \beta_2 \log(\boldsymbol{x_2}) + \boldsymbol{\varepsilon}$		0.9777	-62.71
$\log(\boldsymbol{y}) = \beta_0 + \beta_1 \log(\boldsymbol{x_1}) + \beta_2 \log(\boldsymbol{x_2}) + \beta_3 \log(\boldsymbol{x_1}) \log(\boldsymbol{x_2}) + \boldsymbol{\varepsilon}$		0.9778	-60.88
$\log\left(\frac{\boldsymbol{y}}{\boldsymbol{x_1^2 x_2}}\right) = \beta_0 + \boldsymbol{\varepsilon}$	$\boldsymbol{y} = \beta_1 \boldsymbol{x_1^2 x_2} e^{\boldsymbol{\varepsilon}}$	NA	-66.34
$\boldsymbol{y} = \beta_1 \boldsymbol{x_1^2 x_2} + \boldsymbol{\varepsilon}$		0.9950	146.6

So how could we compare the two best models (4) and (6)?

Example – Predict Black Cherry Tree Timber Volume

What if we *really* wanted to try to estimate parameters for this model?

$$\boldsymbol{y} = \beta_0 \boldsymbol{x}_{\mathbf{1}}^{\beta_1} \boldsymbol{x}_{\mathbf{2}}^{\beta_2} + \boldsymbol{\varepsilon}$$

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height

Can't solve using the standard linear regression approach.

Example – Predict Black Cherry Tree Timber Volume

What if we *really* wanted to try to estimate parameters for this model?

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} + \varepsilon$$

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height

Can't solve using the standard linear regression approach.

Instead, use a library that can estimate parameters for non-linear models, e.g. "nls" in R.

```
Parameters:
            Estimate   Std. Error t value Pr(>|t|)
beta0 0.001449   0.001367   1.060 0.298264
beta1 1.996921   0.082077  24.330  < 2e-16 ***
beta2 1.087647   0.242159   4.491 0.000111 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Residual standard error: 2.533 on 28 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 8.255e-07
```

AIC = 150.4

```
Parameters:
            Estimate   Std. Error t value Pr(>|t|)
beta1  2.27405    0.12967   17.54  < 2e-16 ***
beta2 -0.58432    0.08242   -7.09 8.44e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 4.216 on 29 degrees of freedom

Number of iterations to convergence: 10
Achieved convergence tolerance: 8.673e-06
```

AIC = 181.1

Poor parameter interpretation

Conclusion: the simpler model with only β_0 is better (AIC: 146.6)
And we prefer the multiplicative log-Normal error model

Example – Predict Black Cherry Tree Timber Volume

What if we *really* wanted to try to estimate parameters for this model?

$$\boldsymbol{y} = \beta_0 \boldsymbol{x}_{\boldsymbol{1}}^{\beta_1} \boldsymbol{x}_{\boldsymbol{2}}^{\beta_2} + \boldsymbol{\varepsilon}$$

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height

Can't solve using the standard linear regression approach.

Instead, use a library that can estimate parameters for non-linear models, e.g. "nls" in R.

Con's:

- May require initial parameter estimates
- May not find globally optimal solution – depends on initial parameter estimates
- May not converge at all
- Slower – iterative approach
- Becomes slower and less reliable as the function becomes more complex

Pro's:

- Allows dealing with a wider class of model functional forms

Model Selection: Choosing the best model

Sometimes selecting the best model can be difficult (time consuming & subjective)

Especially when there are a huge number of independent variables

Stepwise Regression – automatically selects “the best” model:

- Start from a given model
- Add or remove terms one at a time
- Score model (AIC)
- Repeat until optimal solution is found

Two options:

- Forward selection – start from simple model and add terms one at a time
- Backward elimination – start from a complex model and remove terms one at a time

Warning:

These strategies can lead to different models being selected

Neither strategy guarantees the optimal solution, but they are quick

Stepwise Regression:

Example: *Swiss fertility and socioeconomic indicators*

Regress Fertility against all available indicators:

```
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *  
Examination      -0.25801    0.25388  -1.016  0.31546    
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 ** 
Infant.Mortality  1.07705    0.38172   2.822  0.00734 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,  Adjusted R-squared:  0.671 
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

Stepwise Regression:

Example: *Swiss fertility and socioeconomic indicators*

Regress Fertility against all available indicators:

```
Start:  AIC=190.69
Fertility ~ Agriculture + Examination +
Education + Catholic + Infant.Mortality

                   Df Sum of Sq    RSS    AIC
- Examination       1     53.03 2158.1 189.86
<none>                          2105.0 190.69
- Agriculture       1    307.72 2412.8 195.10
- Infant.Mortality  1    408.75 2513.8 197.03
- Catholic          1    447.71 2552.8 197.75
- Education         1   1162.56 3267.6 209.36
```

```
Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic
+ Infant.Mortality

                   Df Sum of Sq    RSS    AIC
<none>                          2158.1 189.86
- Agriculture       1    264.18 2422.2 193.29
- Infant.Mortality  1    409.81 2567.9 196.03
- Catholic          1    956.57 3114.6 205.10
- Education         1   2249.97 4408.0 221.43
```

Stepwise Regression:

Example: *Swiss fertility and socioeconomic indicators*

Regress Fertility against all available indicators:

```
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Agriculture      -0.15462    0.06819  -2.267  0.02857 *  
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared:  0.6993,  Adjusted R-squared:  0.6707 
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

- Compared to before stepwise regression, R^2 is lower, and RSE is higher
- AIC favoured the model with fewer parameters.

Final Message:

Linear regression is well-suited to dealing with continuous data…

However it is also suited to:

- Discrete data (e.g. Poisson, Binomial)
- Categorical data (indicator variables, factors)
- Binary data (e.g. Bernoulli)

We have already seen a linear model be used to estimate the mean…

Consider similarities to other techniques:

One-sample Student's t-test: $$Y_i = \mu + \varepsilon_i$$

Two independent sample t-test:
One-way ANOVA: $$Y_{i(g)} = \mu + \delta_g + \varepsilon_{i(g)}$$

Two-way ANOVA: $$Y_{i(gk)} = \mu + \delta_g + \delta_k + \delta_{gk} + \varepsilon_{i(gk)}$$

These are all linear models! The only difference is in the questions we ask…
Linear modelling is extremely flexible.

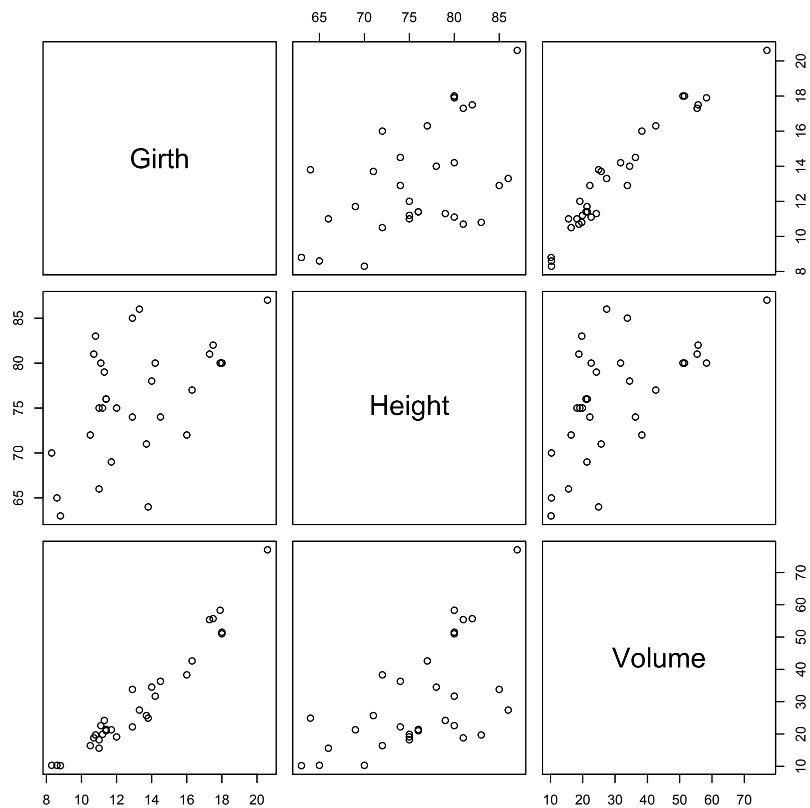
What to use and when:

	Multiple regressors	Non-Gaussian error model	Non-linear data	Autocorrellated data
Simple regression				
Multiple regression	✓			
Generalised linear model	✓	✓		
Non-linear model	✓	✓	✓	
Time series analysis				✓

Multiple Regression in R:

R functions:

plot(x,y)



Multiple Regression in R:

R functions:

plot(x,y)

m1 <- lm(y~x)
summary(m1)

confint(m1)

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

Residuals:
     Min        1Q    Median       3Q       Max
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
log(Girth)   1.98265    0.07501  26.432  < 2e-16 ***
log(Height)  1.11712    0.20444   5.464 7.81e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared:  0.9777,  Adjusted R-squared:  0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

```
                2.5 %    97.5 %
(Intercept) -8.269912 -4.993322
log(Girth)   1.828998  2.136302
log(Height)  0.698353  1.535894
```

Multiple Regression in R:

R functions:

```
plot(x,y)

m1 <- lm(y~x)
summary(m1)

confint(m1)

shapiro.test(residuals(m1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(m1)
## W = 0.97013, p-value = 0.5225
```

Multiple Regression in R:

R functions:

plot(x,y)

m1 <- lm(y~x)
summary(m1)

confint(m1)

shapiro.test(residuals(m1))

AIC(m1)

stepAIC(m1)

```
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                  Df Sum of Sq    RSS    AIC
- Examination      1     53.03 2158.1 189.86
<none>                         2105.0 190.69
- Agriculture      1    307.72 2412.8 195.10
- Infant.Mortality 1    408.75 2513.8 197.03
- Catholic         1    447.71 2552.8 197.75
- Education        1   1162.56 3267.6 209.36

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic +
Infant.Mortality

                  Df Sum of Sq    RSS    AIC
<none>                         2158.1 189.86
- Agriculture      1    264.18 2422.2 193.29
- Infant.Mortality 1    409.81 2567.9 196.03
- Catholic         1    956.57 3114.6 205.10
- Education        1   2249.97 4408.0 221.43

Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic +
    Infant.Mortality, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6765  -6.0522   0.7514   3.1664  16.1422

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Agriculture      -0.15462    0.06819  -2.267  0.02857 *
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared:  0.6993,  Adjusted R-squared:  0.6707
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

Multiple Regression in R:

R functions:

```
plot(x,y)

m1 <- lm(y~x)
summary(m1)

confint(m1)

shapiro.test(residuals(m1))

AIC(m1)

stepAIC(m1)

nls(volume~beta0*girth^beta1*height^beta2, start=list(beta0=1,beta1=2,beta2=1))
```