Regression continued

What happens if data do not fit regression assumptions

- Do nothing
- Transform
- Try a method that does not depend on the violated assumption

This section will focus on the third approach

Model I versus Model II

We have been talking about model I regressions. When the independent variables are not known without error, it is known as a Model II regression. It is important to note the distinction, though that's all we'll say about it. Sokal et al. (1991) outlines three approaches in chapter 15. Gotelli and Ellison avoid the topic.

Non-parametric regression

Just as there are so-called non-parametric alternatives to ANOVA there are nonparametric approaches to regression.
Gotelli and Ellison emphasize robust regression. Sokal et al. (1991) highlight Kendall's robust line-fit method. It is important to note that most nonparametric regression approaches estimate parameters, but depend upon less restrictive assumption

Kendall's method

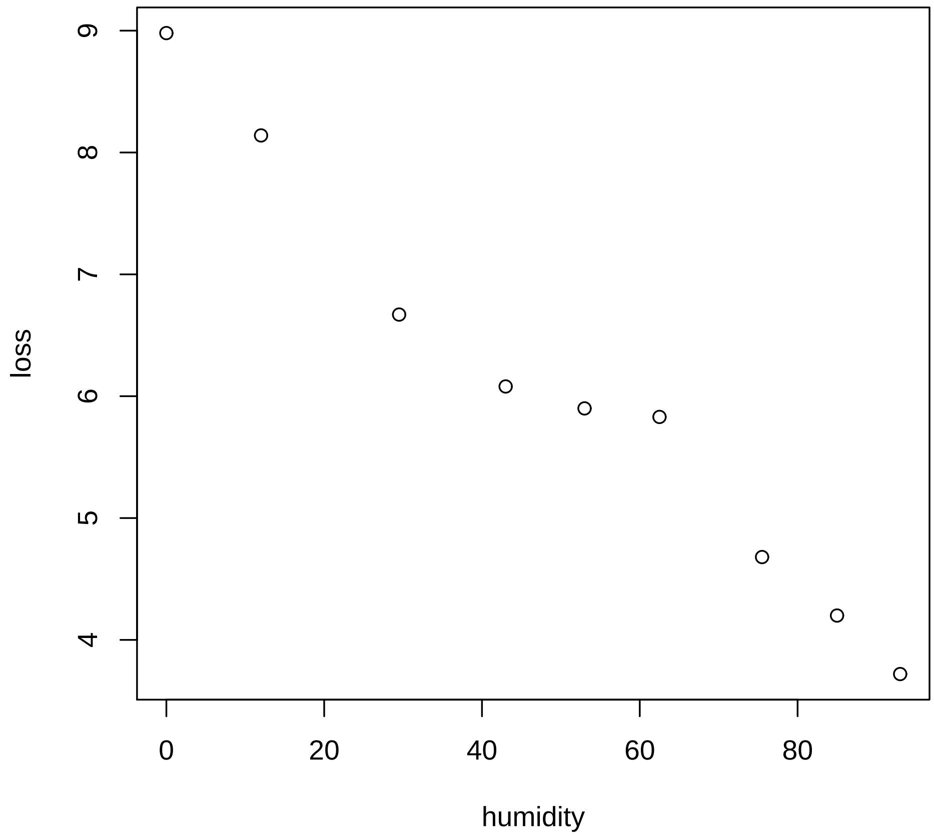
In Kendall's method, you calculate the slope between each pair of points:

$$S_{ji} = \frac{Y_j - Y_i}{X_j - X_i}$$

Where $j > i$ In other words, don't calculate slopes twice.
The slope is the median of the S_{ij} values

Calculation of Kendall's method: Data

```
> humidity <- c(0, 12, 29.5, 43, 53, 62.5, 75.5, 85, 93)
> loss <- c(8.98, 8.14, 6.67, 6.08, 5.9, 5.83, 4.68, 4.2, 3.72)
> plot(loss ~ humidity)
```



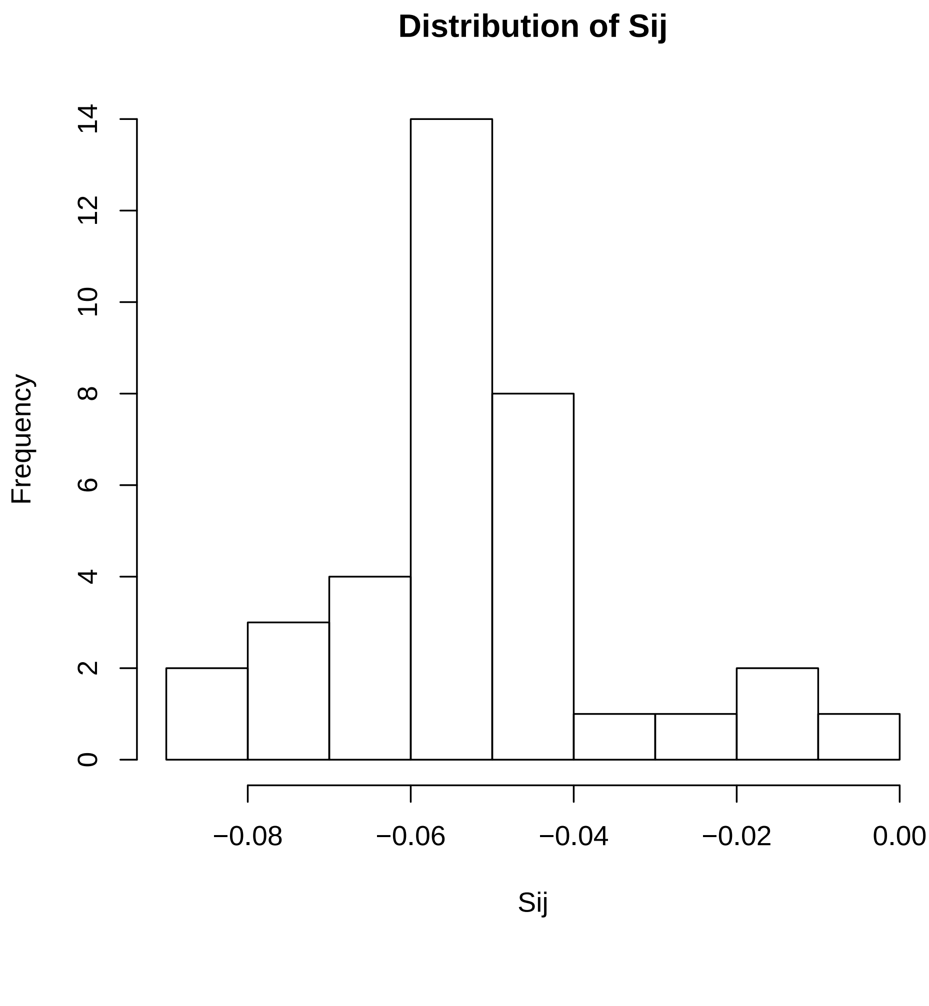
Calculations of Kendall's method

```
> Svec <- NULL
> for (i in 1:length(humidity)) {
+     for (j in 1:length(humidity)) {
+         if (j > i) {
+             Svec <- c(Svec, (loss[j] - loss[i])/(humidity[j] -
+                 humidity[i]))
+         }
+     }
+ }
> Svec

 [1] -0.070000000 -0.078305085 -0.067441860 -0.058113208 -0.050400000
 [6] -0.056953642 -0.056235294 -0.056559140 -0.084000000 -0.066451613
```

```
[11] -0.054634146 -0.045742574 -0.054488189 -0.053972603 -0.054567901
[16] -0.043703704 -0.032765957 -0.025454545 -0.043260870 -0.044504505
[21] -0.046456693 -0.018000000 -0.012820513 -0.043076923 -0.044761905
[26] -0.047200000 -0.007368421 -0.054222222 -0.053125000 -0.054500000
[31] -0.088461538 -0.072444444 -0.069180328 -0.050526316 -0.054857143
[36] -0.060000000
```

```
> hist(Svec, main = "Distribution of Sij", xlab = "Sij")
> b1 <- median(Svec)
> b1
```

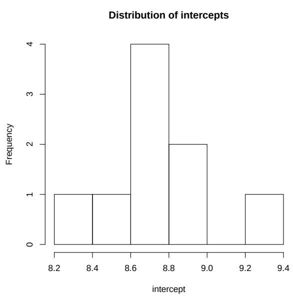
```
[1] -0.05435521
```



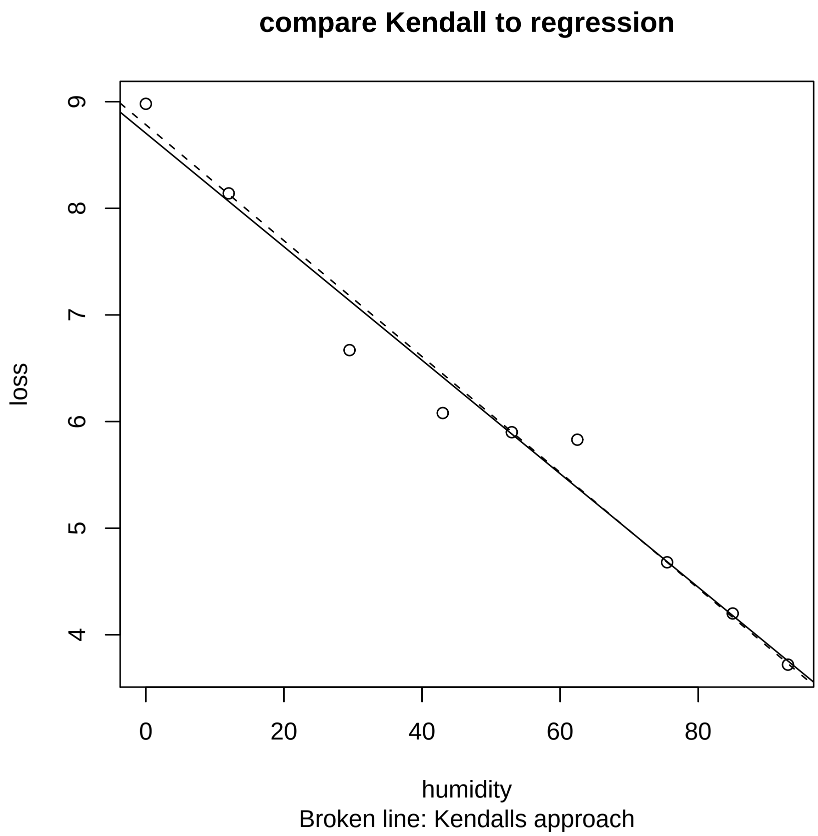
Kendall's method (intercept)

The Y intercept term (β_0 in linear regression) can be calculated by taking the median of the results of $Y_i - b_1 X_i$ for all X, Y pairs.

```
> intvec <- loss - b1 * humidity
> b0 <- median(intvec)
> b0

[1] 8.783818

> hist(intvec, main = "Distribution of intercepts", xlab = "intercept")
```



Compare with linear regression



Function for Kendall's robust regression

```
> kendall.robust <- function(X, Y) {
+     Svec <- NULL
+     for (i in 1:length(X)) {
+         for (j in 1:length(X)) {
+             if (j > i) {
+                 Svec <- c(Svec, (Y[j] - Y[i])/(X[j] - X[i]))
+             }
+         }
+     }
+     b1 <- median(Svec)
+     b0 <- median(Y - b1 * X)
+     c(b0, b1)
+ }
> kendall.robust(humidity, loss)

[1]  8.78381802 -0.05435521
```

Is slope different than zero?

Most nonparametric tests of regression (and association) are based upon ranks. Sokal et al. (1991) highlight Kendall's τ. The approach for calculating this quantity by hand is given in box 15.7 of Sokal et al. (1991)
There is a function in R that calculates Kendall's tau and estimates its significance in testing whether there is no relationship between X and Y.

Actual test

```
> cor.test(loss, humidity, method = "kendall")

        Kendall's rank correlation tau

data:  loss and humidity
T = 0, p-value = 5.511e-06
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
 -1

> cor.test(loss, humidity, method = "pearson")

        Pearson's product-moment correlation

data:  loss and humidity
t = -16.3457, df = 7, p-value = 7.816e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9973935 -0.9379224
sample estimates:
       cor
-0.9871523
```

Regression as a modeling tool

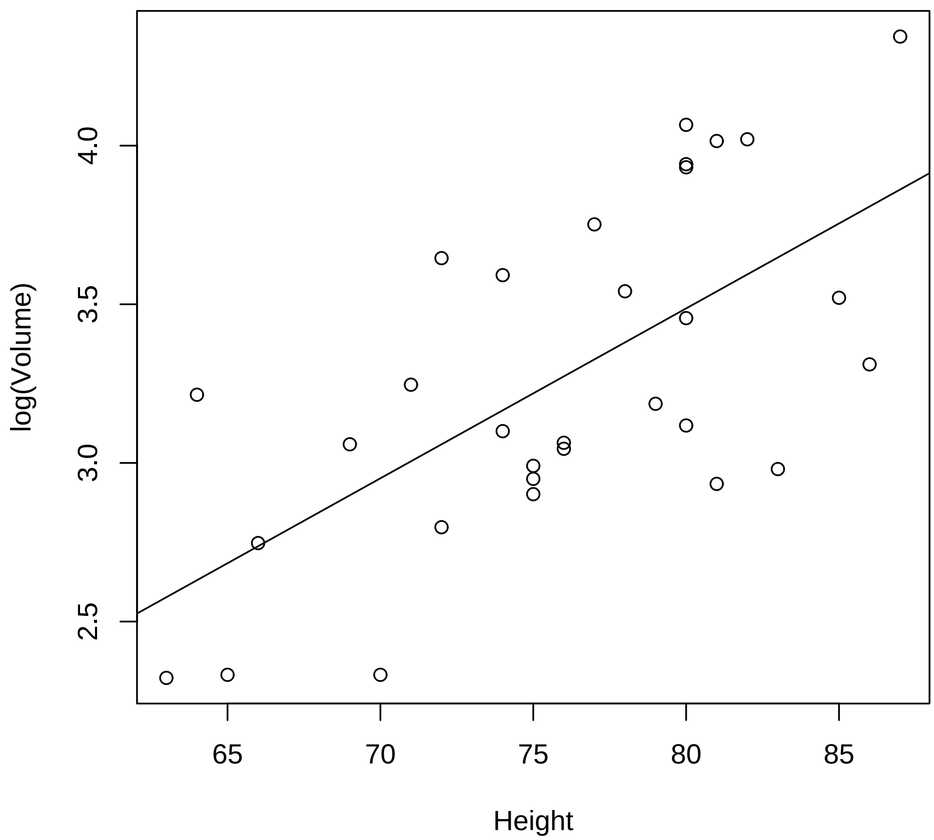
So far we have talked about regression as a mechanism to establish a causal relationship (accompanied by careful exp. design) between an independent and dependent variable.
Regression can also be though of as a modeling tool, or a mechanism to explain the variation observed in the dependent variable.
If regression does not explain the relationship between two variables, it does a poor job of explaining a variance in the dependent variable.
If a regression has a significant slope and a large R^2, it *does* explain variation

Multiple regression

When regression is used as a modeling tool, it frequently uses multiple independent variables to explain the variation in a dependent variable.
This could be done by estimating the regression equation of X_1 on Y then examining the effects of X_2 on the residuals of the first equation, and so on.

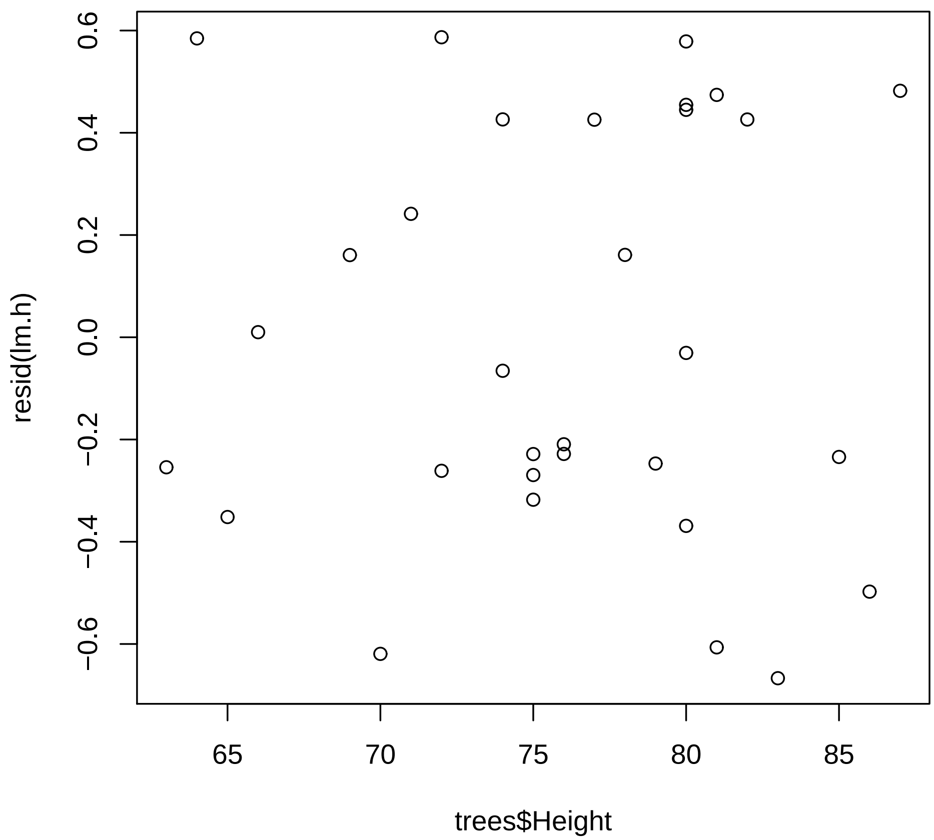
trees

```
> data(trees)
> lm.h <- lm(log(Volume) ~ Height, data = trees)
> plot(log(Volume) ~ Height, data = trees)
> abline(coef(lm.h))
```

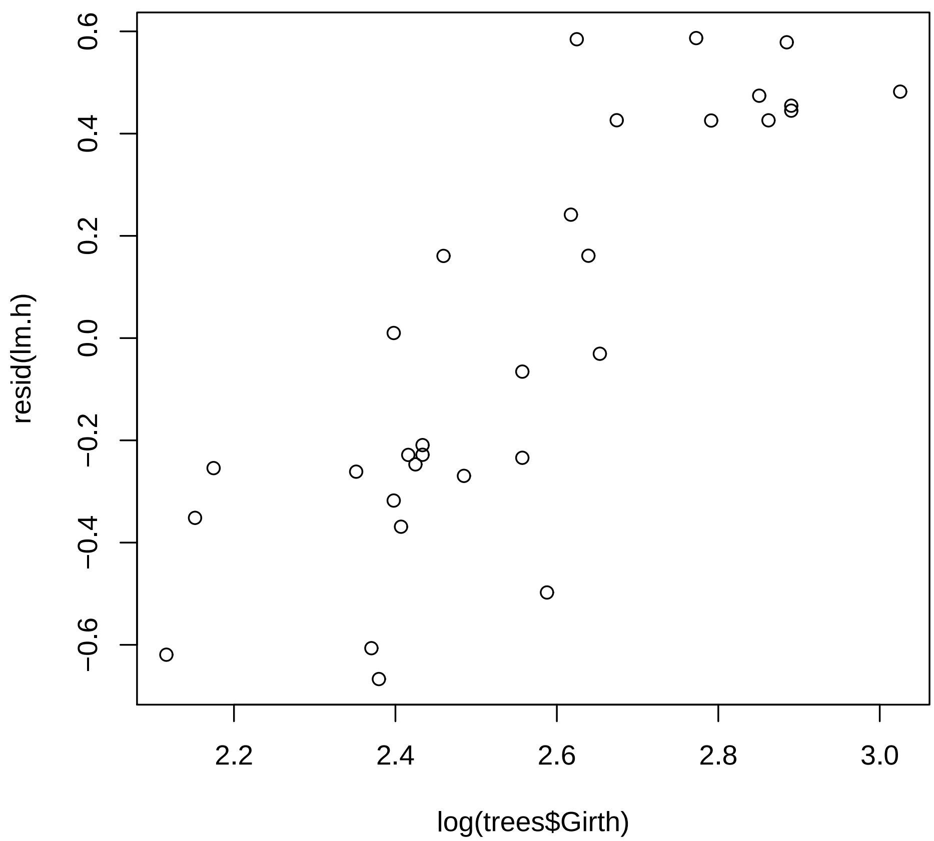


Examine residuals

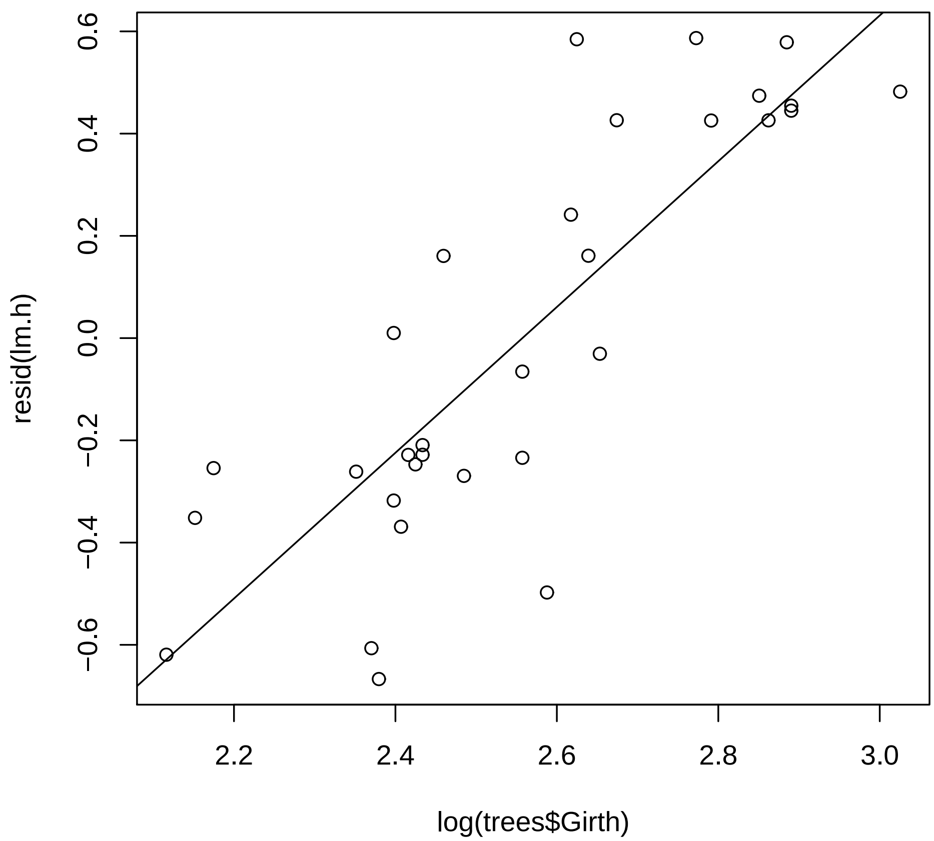
```
> plot(resid(lm.h) ~ trees$Height)
```



Residuals with Girth



Residuals with Girth2



Coefficients

```
> summary(lm.h)

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.66691 -0.26539 -0.06555  0.42608  0.58689

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.79652    0.89053  -0.894    0.378
Height       0.05354    0.01168   4.585 8.03e-05 ***
```

```
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.4076 on 29 degrees of freedom
Multiple R-Squared: 0.4203,       Adjusted R-squared: 0.4003
F-statistic: 21.02 on 1 and 29 DF,  p-value: 8.026e-05
```

Coefficients 2

```
> summary(lm.gr)

Call:
lm(formula = resid(lm.h) ~ log(trees$Girth))

Residuals:
     Min       1Q   Median       3Q      Max
-0.54100 -0.12249 -0.02016  0.13382  0.48864

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3.6462     0.4545  -8.022 7.57e-09 ***
log(trees$Girth)   1.4258     0.1770   8.055 6.98e-09 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.2265 on 29 degrees of freedom
Multiple R-Squared: 0.6911,       Adjusted R-squared: 0.6804
F-statistic: 64.88 on 1 and 29 DF,  p-value: 6.977e-09
```

Better approach

A more elegant approach takes all X variables at once and looks at their effect on Y. This is known as *multiple regression* .

Linear model

The linear model for multiple regression looks like this:

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \cdots + \beta_n X_{n_i} + \epsilon_i$$

The X_{j_i}'s represent observations on independent variables for each dependent variable Y_i. ϵ has its usual meaning.
You could use an approach similar to the one we used for simple regression to estimate these coefficients. As the number of independent variables increases, this gets really hairy, though

Multiple regression: elegance

Matrices provide an elegant mechanism to summarize and solve the regression equations.

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

Where $\mathbf{Y}$, and $\mathbf{e}$ are vectors of length n, $\mathbf{X}$ is a matrix with n rows and $k+1$ columns, and β is a vector with length $k+1$.
(slight digression on matrices)
The coefficients are estimated by this equation:

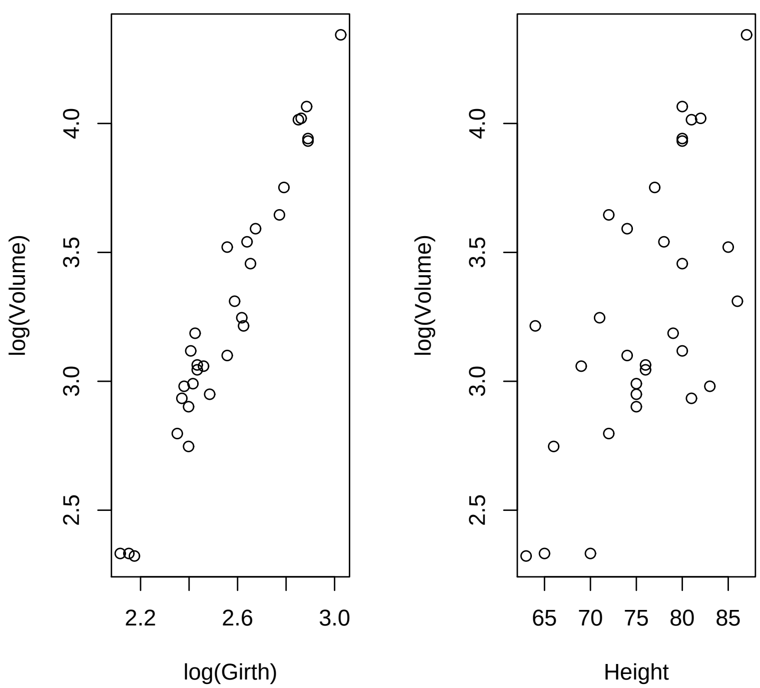
$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

This happens behind the scenes in `lm()` in R

Why use multiple regression: examples

- rat mass depends upon both size and diet
- plant yield depends upon both [K] and [N]
- severity of disease depends upon age, weight, diet, etc...

Example with two predictors



```
> data(trees)
> summary(lm(log(Volume) ~ log(Girth), data = trees))

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

Residuals:
      Min        1Q    Median        3Q       Max
-0.205999 -0.068702  0.001011  0.072585  0.247963

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.35332    0.23066  -10.20 4.18e-11 ***
log(Girth)   2.19997    0.08983   24.49  < 2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.115 on 29 degrees of freedom
Multiple R-Squared: 0.9539,      Adjusted R-squared: 0.9523
F-statistic: 599.7 on 1 and 29 DF,  p-value: < 2.2e-16
```



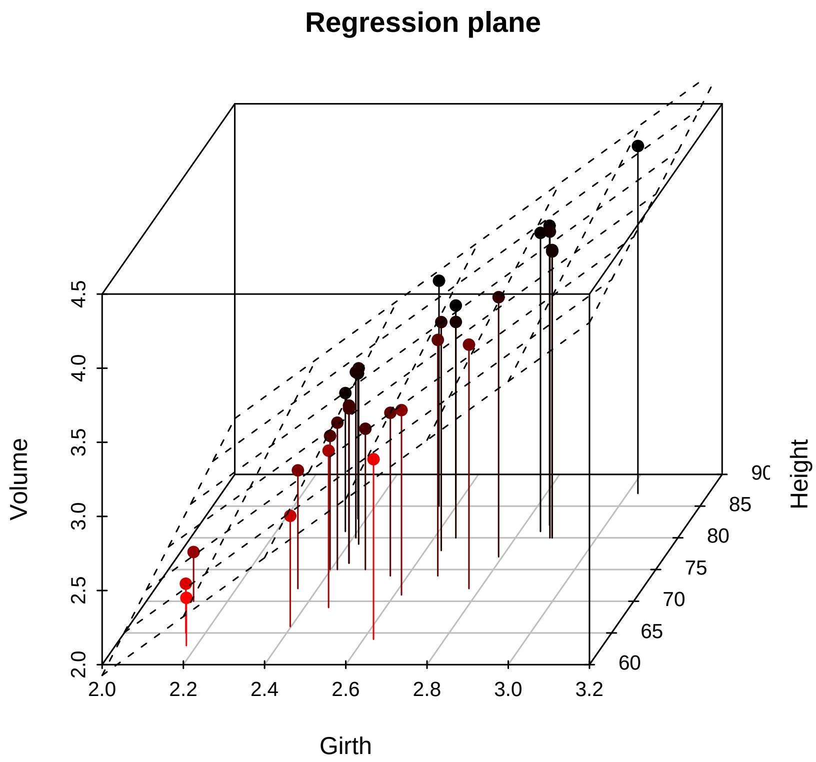
```
> summary(lm(log(Volume) ~ Height, data = trees))

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.66691 -0.26539 -0.06555  0.42608  0.58689

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.79652    0.89053  -0.894    0.378
Height       0.05354    0.01168   4.585 8.03e-05 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.4076 on 29 degrees of freedom
```

```
Multiple R-Squared: 0.4203,        Adjusted R-squared: 0.4003
F-statistic: 21.02 on 1 and 29 DF,  p-value: 8.026e-05
```

Multiple regression

```
> data(trees)
> summary(lm(log(Volume) ~ log(Girth) + Height, data = trees))

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

Residuals:
      Min        1Q    Median        3Q       Max
-0.172440 -0.048026  0.003274  0.064428  0.131489

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.938289   0.195950 -14.995 6.58e-15 ***
log(Girth)   1.983227   0.075215  26.367  < 2e-16 ***
Height       0.014990   0.002758   5.435 8.45e-06 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.08161 on 28 degrees of freedom
Multiple R-Squared: 0.9776,        Adjusted R-squared: 0.976
F-statistic: 609.8 on 2 and 28 DF,  p-value: < 2.2e-16
```

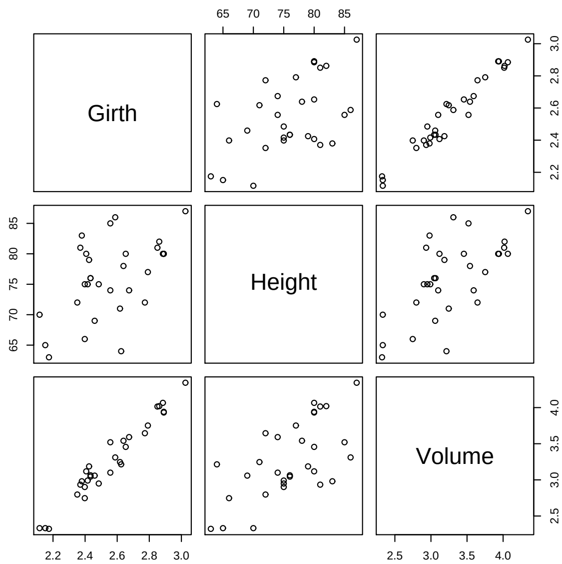
Regression plane



Why doesn't Height help much?

Two reasons

- Height does not explain a lot of the variation in Volume
- Height and Girth are colinear



Overfitting

If you are interested in just coming up with a predictive function, one strategy is to include all possible predictors. A better strategy is to include just those predictors that are important by either adding terms in one at a time or removing terms one at a time until a good fit is achieved that:

- Has as many significant terms as possible and
- Has a high adjusted R^2, or preferably AIC.

If you are interested in causal relationships however, including terms willy nilly may result in spurious assessment of causation.

References

Sokal, R. R., Oden, N. L. and Wilson, C. (1991). Genetic evidence for the spread of agriculture in Europe by demic diffusion, *Nature* **351**: 143–145.