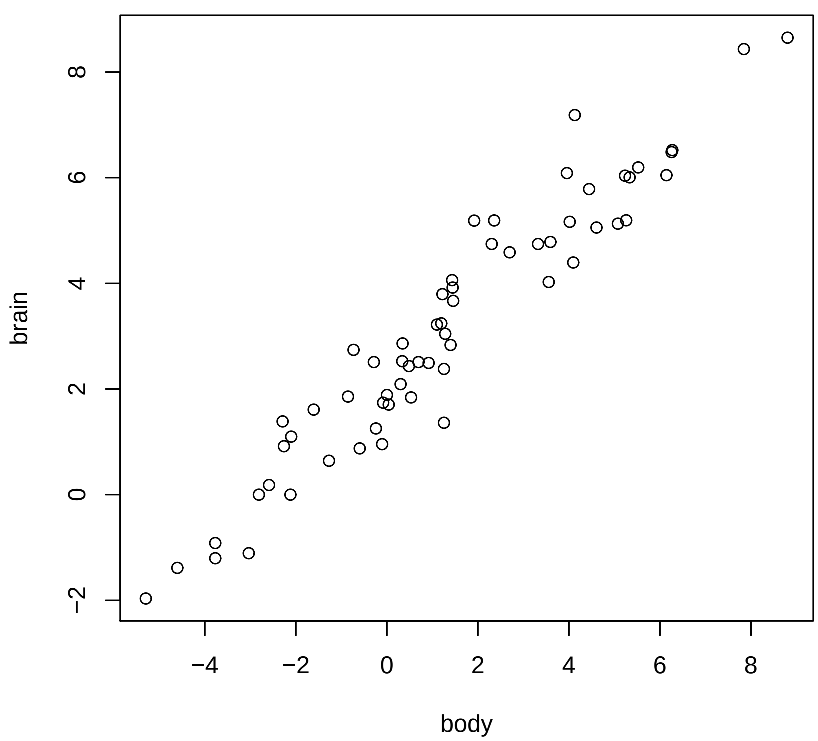
Regression

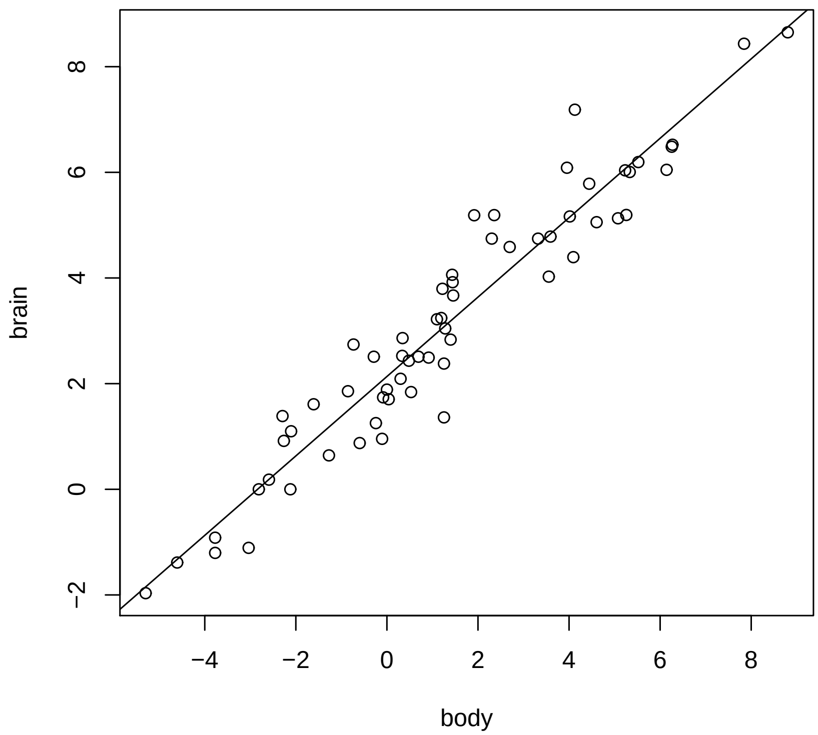
Regression analysis examines the effect of one or more *quantitative* independent variables upon another quantitative response variable.
Examples include relationships between

- leg-length and running speed
- inflorescence size and fitness
- temperature and enzymatic rate
- Plant size and fecundity
- body and brain size

Brain and body size



Brain and body size 2

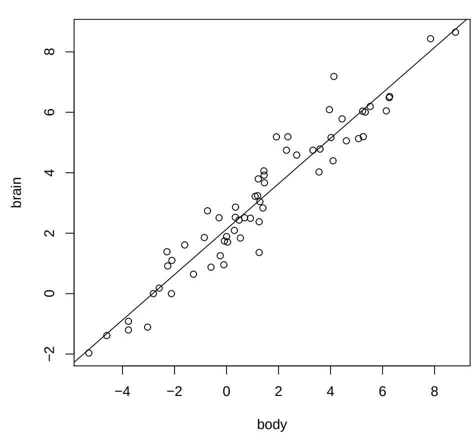


Regression Model

The model for regression looks like this:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

This should look a bit like $Y = b + mx$ with an error term.



In this relationship, β_0 = -2.135 and β_1 = 0.752.

Regression coefficients

It is important when considering regression coefficients (especially for $B_{>0}$) to examine both

- Sign (direction of effect)
- Magnitude (importance of effect)

A sizeable portion of learning about regression comes in the form of learning about how to decide if $\beta \neq 0$
Once the coefficients are in hand, it is possible to use the relationship to *predict*.

Linear Regression assumptions

Important assumptions of regression are:

- quantitative variables used as independent and dependent variables (not universal you will see that binary variables are also used)
- independent variables are measured without error
- ϵ is normally distributed with mean 0 and variance σ^2
- ϵ values for different observations are independent
- Linear relationship actually exists between X and Y.

Estimating coefficients

The equation described before is the "population" or parametric regression equation
The estimated equation looks like this

$$Y_i = b_0 + b_1 x_i + e_i$$

The basic idea in fitting the coefficients is to minimize the squared residual variation (squared Y-distance from the regression line).

Estimating coefficients 2

An analytical solution to this *least squares* estimator.
The slope of the regression line can be calculated by dividing the product of deviations from the means of X and Y by the sum of the square deviances in X.

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X - \bar{X})^2}$$

Estimating coefficients 3

Estimating the intercept depends upon the fact that the regression line *must pass through* the point $(\bar{X}, \bar{Y})$.
Once you know $\bar{X}$ and $\bar{Y}$ and b_1, it is possible to estimate b_0 (no error is assumed for this calculation)
Here is the equation for a line

$$\bar{Y} = b_0 + b_1 \bar{X}$$

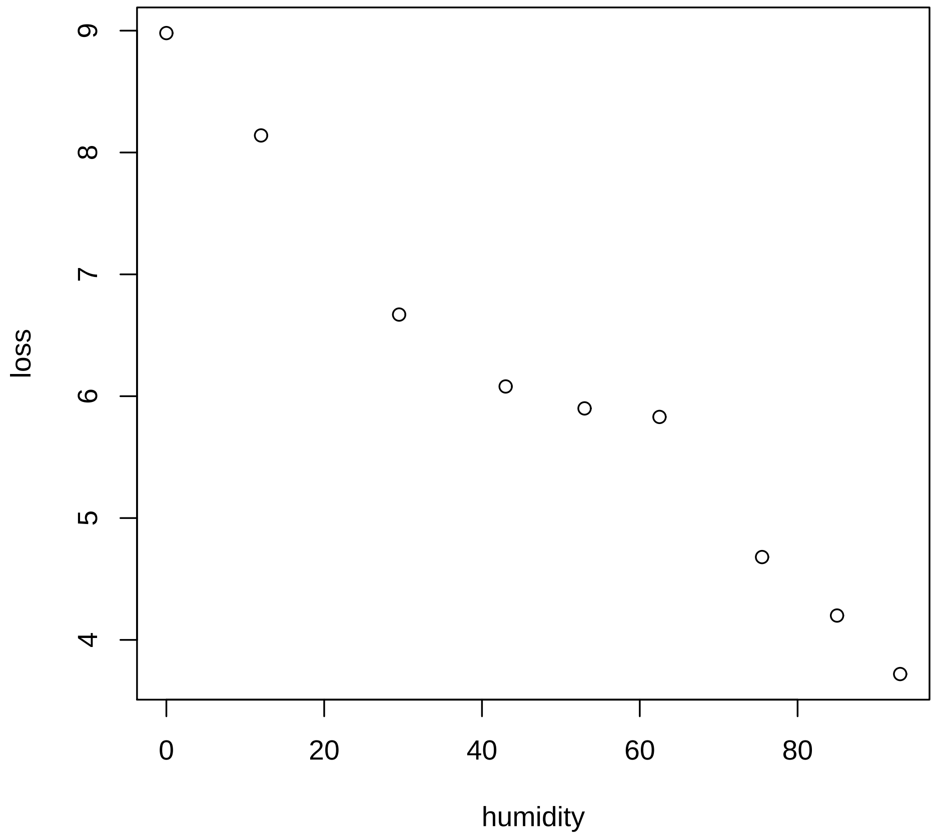
Here is the equation rearranged

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Example

Weight loss in *Tribolium* at different humidities (weight loss is an average of 25 animals).

```
> 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)
```



Example, cont

The slope:

```
> xdev <- humidity - mean(humidity)
> ydev <- loss - mean(loss)
> b1 <- sum((xdev) * (ydev))/sum(xdev^2)
> b1

[1] -0.05322215
```

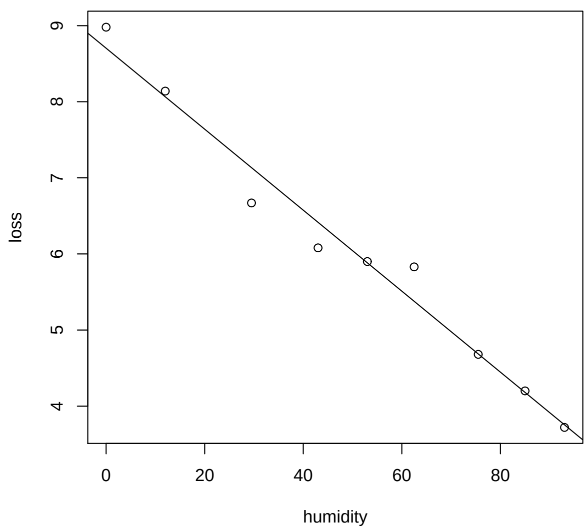
The intercept:

```
> b0 <- mean(loss) - mean(humidity) * b1
> b0

[1] 8.704027
```

Example cont

```
> plot(loss ~ humidity)
> abline(c(b0, b1))
```



Example cont

Linear relationships between variables in R can be calculated using the `lm()` function.

```
> regress.lm <- lm(loss ~ humidity)
> coef(regress.lm)

(Intercept)    humidity
 8.70402730 -0.05322215
```

Is b_1 different than zero?

If there is no relationship between X and Y, then b_1 should equal 0. So one could ask, how much deviation from 0 would be required before we can confidently say that there is a relationship between X and Y.
This is the same old question: how much variation in the data is explained by the model (regression line; signal) divided by the amount of variation unexplained by model (residuals; noise).

Inference on regression line

Estimated values of Y: After the regression line is estimated, one can use the measured values of X to develop estimates of Y assuming no noise. These estimates are denoted as $\hat{Y}$. The difference between $\hat{Y}$ and $\bar{Y}$ is the explained variation. The associated SS is $\Sigma_i(\hat{Y} - \bar{Y})^2$ (df = 1).
The difference between Y and $\hat{Y}$ is the unexplained variation. The associated SS is $\Sigma(Y - \hat{Y})^2$ The degrees of freedom are equal to $n - 2$.

Summary table

Source	df	SS	MS
Explained	1	$\Sigma(\hat{Y} - \bar{Y})^2$	SS/1
Unexplained	$n - 2$	$\Sigma(Y - \hat{Y})^2$	SS/$n - 2$

The appropriate F test is:

$$F = \frac{MS^2_{EX}}{MS^2_{UE}}$$

With degrees of freedom given above in the table.

Example cont (SS)

```
> y.hat <- b0 + b1 * humidity
> ss.explained <- sum((y.hat - mean(loss))^2)
> ss.unexplained <- sum((y.hat - loss)^2)
> ss.explained

[1] 23.51449

> ss.unexplained

[1] 0.6160628
```

Example cont (MS)

```
> ms.exp <- ss.explained/1
> ms.unexp <- ss.unexplained/7
> ms.exp

[1] 23.51449

> ms.unexp

[1] 0.08800897

> f <- ms.exp/ms.unexp
> f

[1] 267.1829

> 1 - pf(f, 1, 7)

[1] 7.816146e-07
```