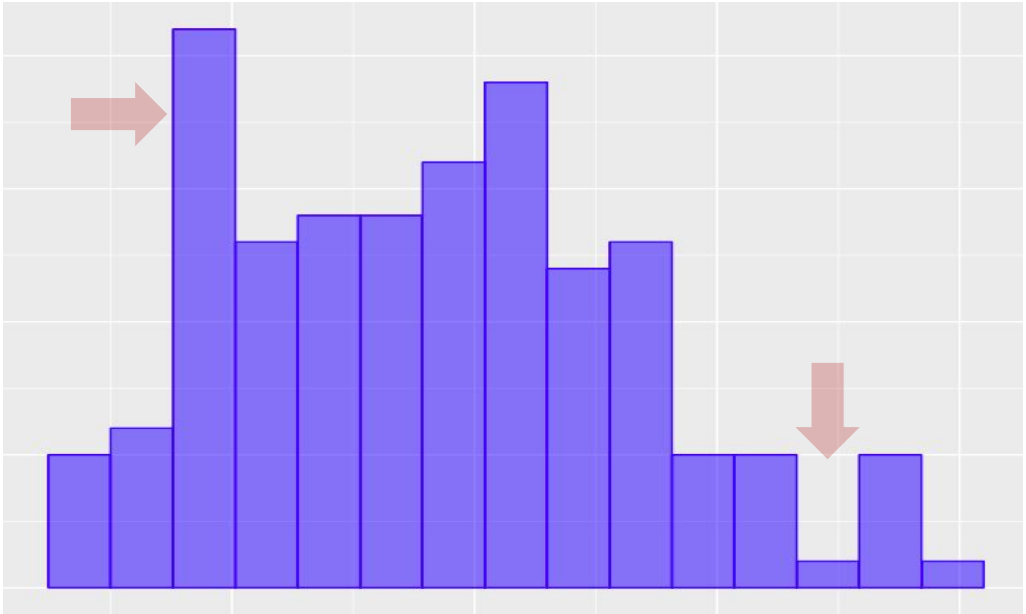
Linear Regression

Statistics in Five Minutes (It's Short We Promise)

Probability distribution - "How frequently do we expect to see certain values?"

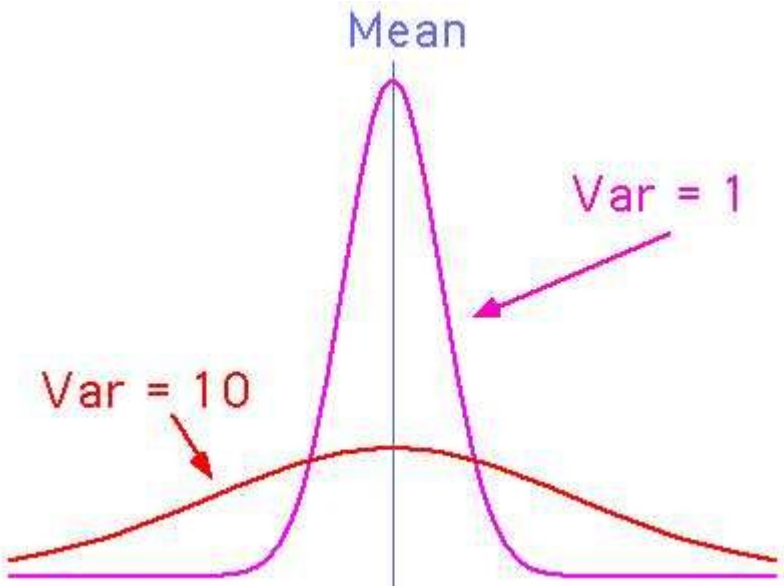


Statistics in Five Minutes

Gaussian (normal distribution) - “Bell curve” with lots of data concentrated around the mean.

Mean (expected value) - the average value of the data

Variance - the amount of spread or deviation in the data

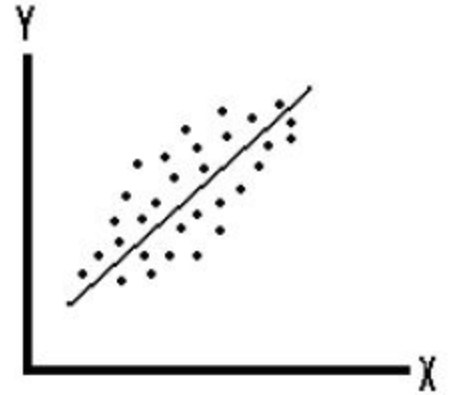


http://nitro.biosci.arizona.edu/courses/EEB600A-2003/lectures/lecture27/figures/var.jpeg

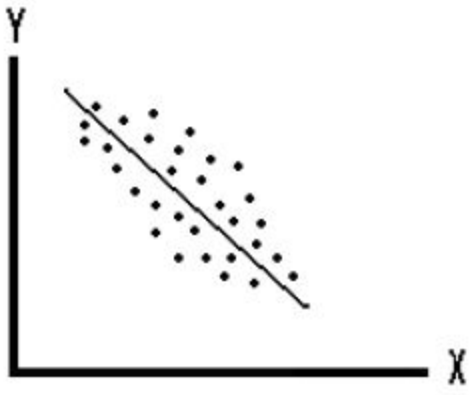
Statistics in Five Minutes

Correlation - a number between -1 and 1 describing the degree of linear relationship between two variables

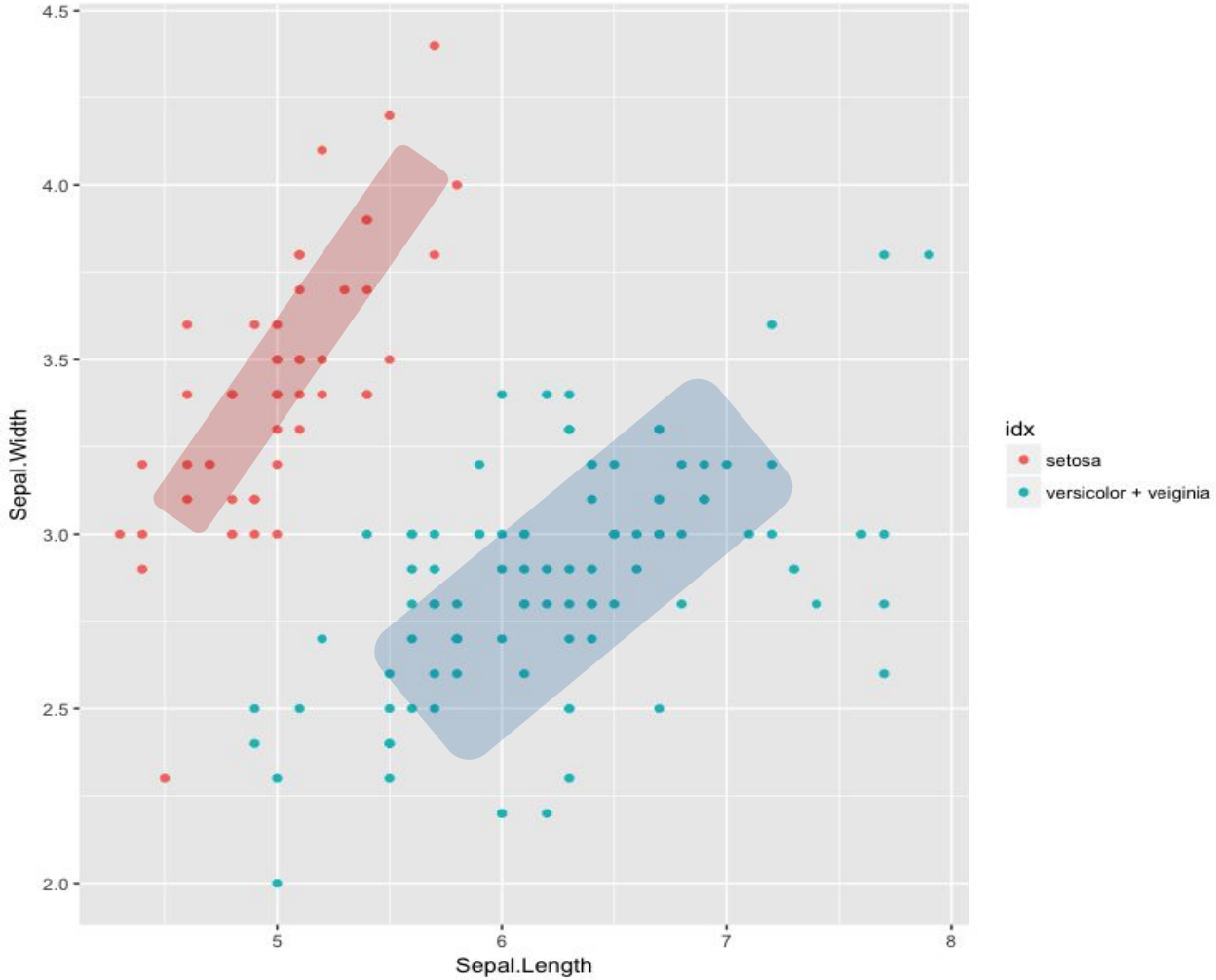
1: positive linear relationship 0: no relationship -1: negative linear relationship



Positive Correlation

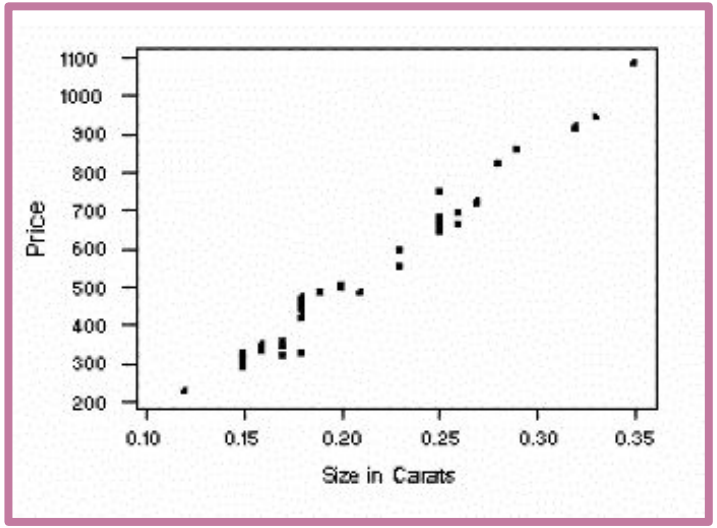


Negative Correlation


4.5
4.0
3.5
3.0
2.5
2.0
Sepal.Width
5
6
7
8
Sepal.Length
idx
setosa
versicolor + veiginia

Regression

We want to find a **hypothesis** that explains the behavior of *y*.



$$y = B_0 + B_1 x_1 + \ldots + B_p x_p + \varepsilon$$

Source

Linear Regression

$$y = B_0 + B_1 x_1 + \ldots + B_p x_p + \varepsilon$$

y is the **dependent variable** - what we want to predict

x_i's are n **independent variables** that influence y.

ε is inherent "noise" for each y.

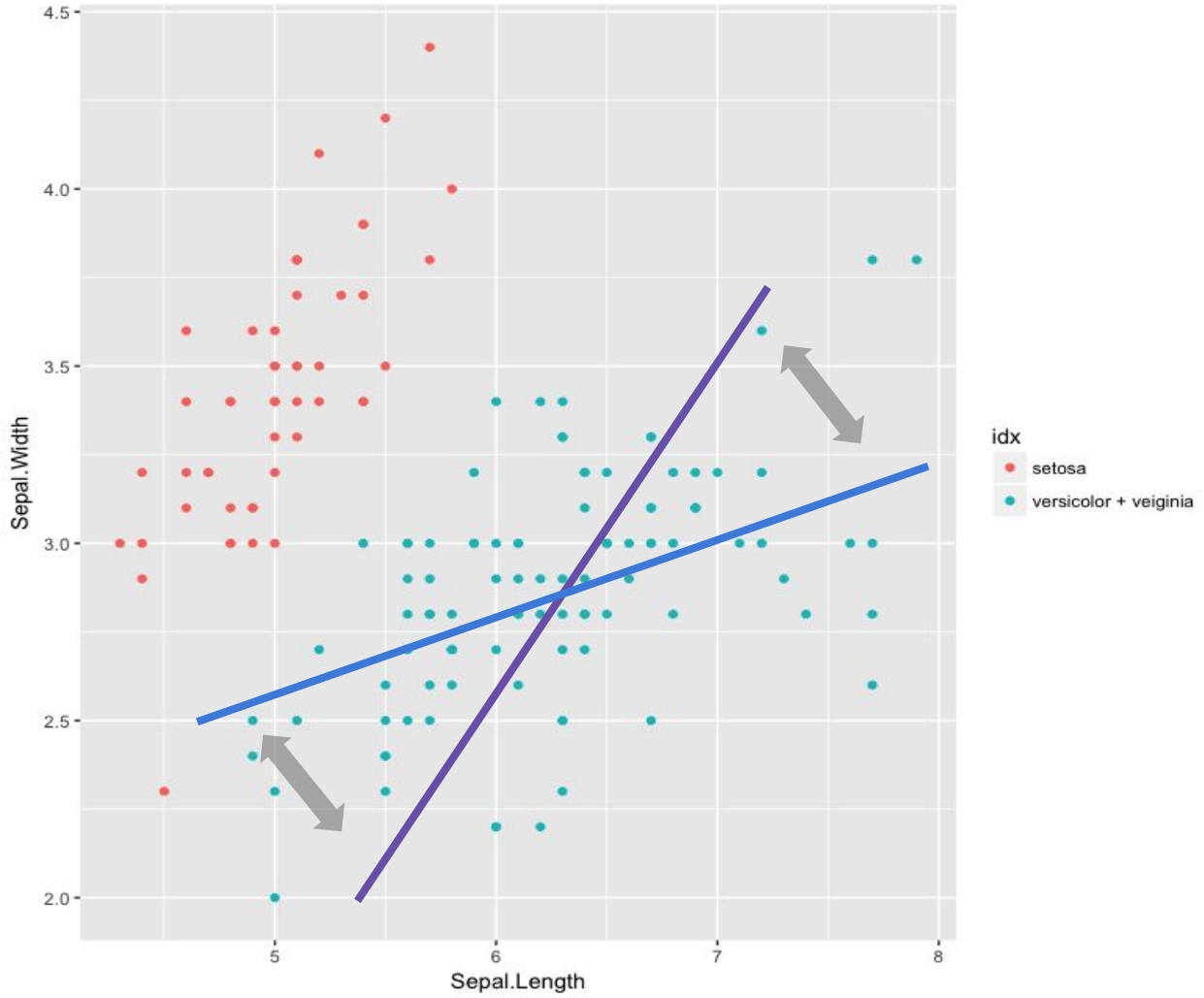
We want to find the **coefficients** (B_i's) that most accurately relate the x_i's to y.

Can think of data as n x p matrix (n data points, p variables).

Question:

What are some examples of quantities that are linearly related?

Error


4.5
4.0
3.5
3.0
2.5
2.0
Sepal.Width
5
6
7
8
Sepal.Length
idx
setosa
versicolor + veiginia

What do the coefficients mean?

Coefficient B_i for x_i- The amount by which y changes if we change x_i by one unit, keeping all other x's constant.

NOTE: The size of the coefficient says nothing about how significant a variable is! Variables large in magnitude can have very small coefficients but still be important.

Least Squares Error

We define our error as follows:

$$\sum_{i=0}^{n}(y_i - (B_0 + B_1x_1 + ... + B_px_p))^2$$

theoretical

observed

We call this **Least Squares Error**. Sum of squared *vertical* distance between observed and theoretical values.

Our Mission

Find the set of $B_0, B_1, \ldots, B_p$ that minimizes the least-squares error over the whole data set.

Source

P-values

P-value - probability that we chose samples that “happened” to have a linear relationship when there really is none.

Results are **statistically significant** if the p-value is less than or equal to 0.05.

The smaller the p-value, the more significant our results are!

Model "Goodness of Fit"

Common metric is called **R^2**.

- We compare our model to a **benchmark model**
  - Predict the mean y value, no matter what the x_i's are
- SST = least-squares for benchmark
- SSE = least-squares error for our model
- $R^2 = 1 - SSE/SST$

Source

Adjusted R^2

Pitfall of R^2: It will always increase as you increase the number of variables.

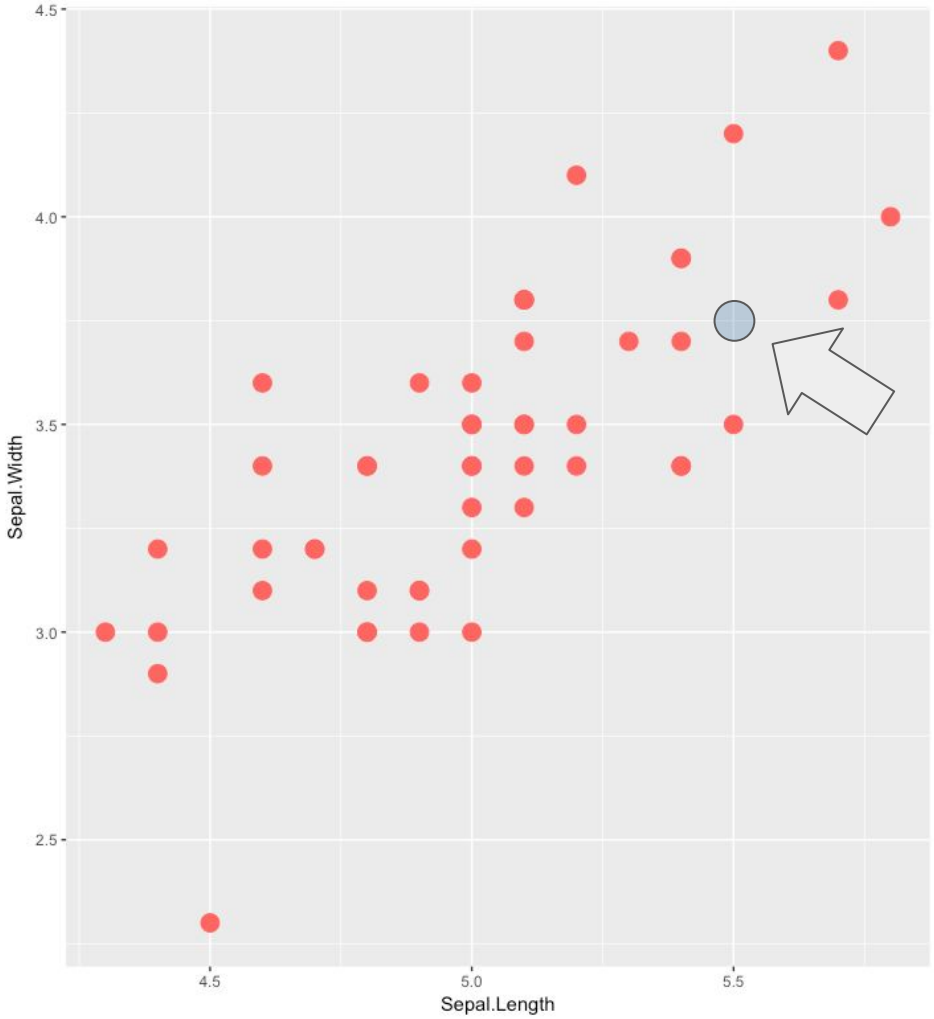
Adjusted R^2 penalizes large numbers of parameters. You should look at the *adjusted* value whenever you examine your R output.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Source

Predicting Values

So far, everything we've done has been **in-sample**. But we need to use the `predict` function to predict y values for our dataset.


4.5
4.0
3.5
3.0
2.5
Sepal.Width
4.5
5.0
5.5
Sepal.Length

```
> Model <- lm(Sepal.Width ~ Sepal.Length,
iris)
> predict(Model, data.frame(Sepal.Length =
c(5.5)))

3.07858
```


```
> Model <- lm(Sepal.Width ~ Sepal.Length,
iris[which(iris$Species == “setosa”),])
> predict(Model, data.frame(Sepal.Length =
c(5.5)))

3.822473
```

```
> summary(Model)
```

```
Call:
lm(formula = Sepal.Width ~ Sepal.Length, data = iris[which(iris$Species ==
    "setosa"), ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.72394 -0.18273 -0.00306  0.15738  0.51709

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.5694     0.5217  -1.091    0.281
Sepal.Length   0.7985     0.1040   7.681 6.71e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2565 on 48 degrees of freedom
Multiple R-squared:  0.5514,    Adjusted R-squared:  0.542
F-statistic: 58.99 on 1 and 48 DF,  p-value: 6.71e-10
```

Regression is powerful.

Source

When to Use Linear Regression

Use `lm` if...

- The data seems sufficiently linear
- You're constrained by a very strict timeline
    - Linear regression is computationally efficient
    - Helpful when there's a constant stream of new data to be processed. This is called **online learning**.

Coming Up

Your problem set: None this week

Next week: Classification using logistic regression and decision trees

See you then!