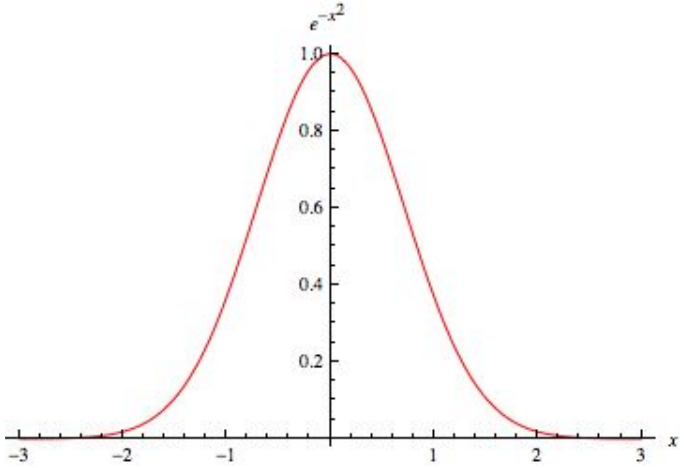
Linear Regression

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

Probability distribution - an assignment of probabilities to possible data points. Answers the question: “How frequently do we expect to see certain values?”

Statistics in Five Minutes

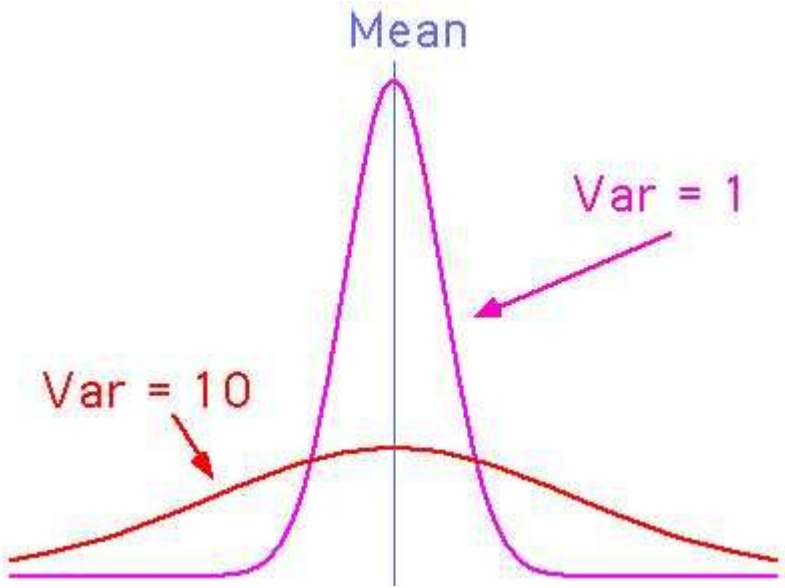
Gaussian (normal distribution) - “Bell curve” with lots of data concentrated around the mean. Examples include test scores, weight, height.



Statistics in Five Minutes

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

Variance - the amount of spread or deviation in the data

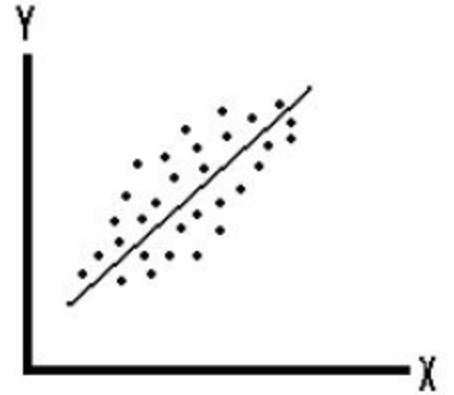


http://nitro.biosci.arizona.edu/courses/EEB600A-2003/lectures/lecture27/figures/var.jpeg

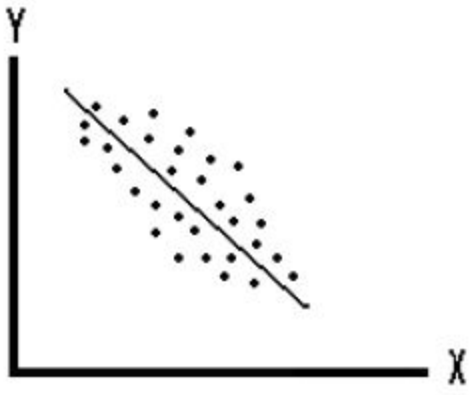
Statistics in Five Minutes

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

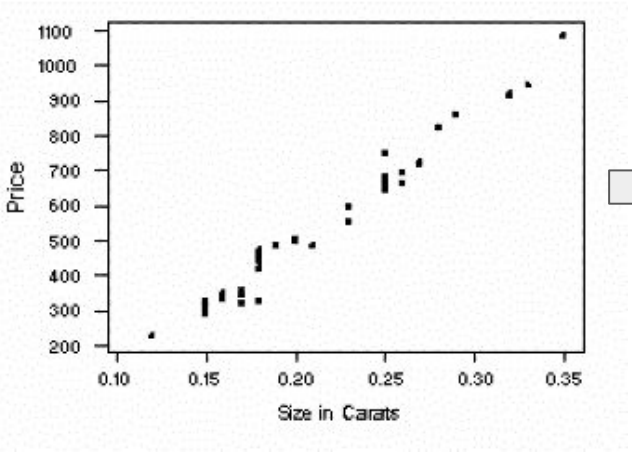


Negative Correlation

Regression

In real life we encounter **error** which causes deviations from what's expected.

Regression uses measured values to obtain a mathematical relationship between several quantities called a **hypothesis**. (What would things look like without error?)



$$y = mx + b$$

http://www.stat.yale.edu/Courses/1997-98/101/scatex.gif

{Extra, Inter}-polation

We usually use regression for two things:

- **Extrapolation** - estimate values outside the range of observation
- **Interpolation** - estimate values in “gaps” inside the range of observation

These are forms of **inductive reasoning**.

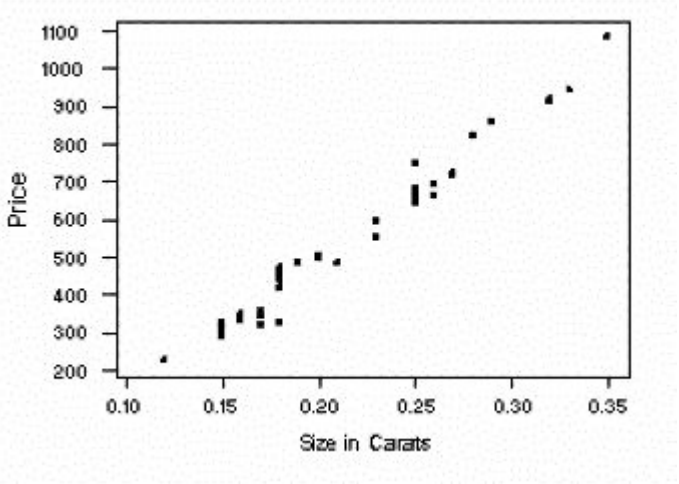
Linear Regression

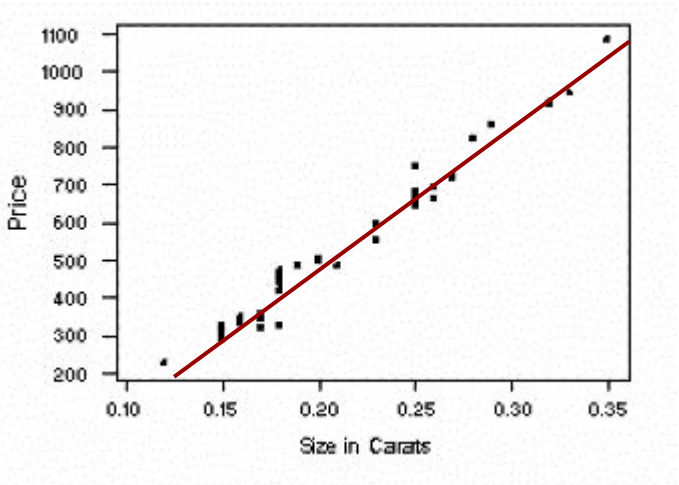
Assumes linear relationship between variables (other assumptions too - check notes)

General form:

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

This is a hyperplane in p-dimensional space.





What does it mean?

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

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

x_i's are n **independent variables** that influence y. ε is inherent "noise".

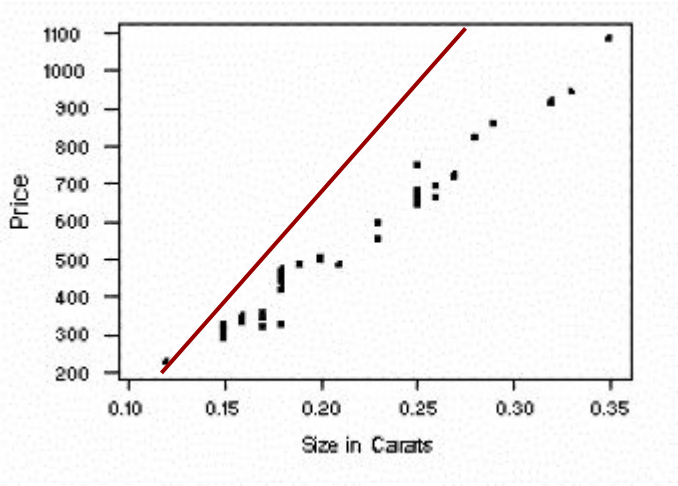
Our task: Find the **coefficients** (B_i's) that most accurately relate the x_i's to y.

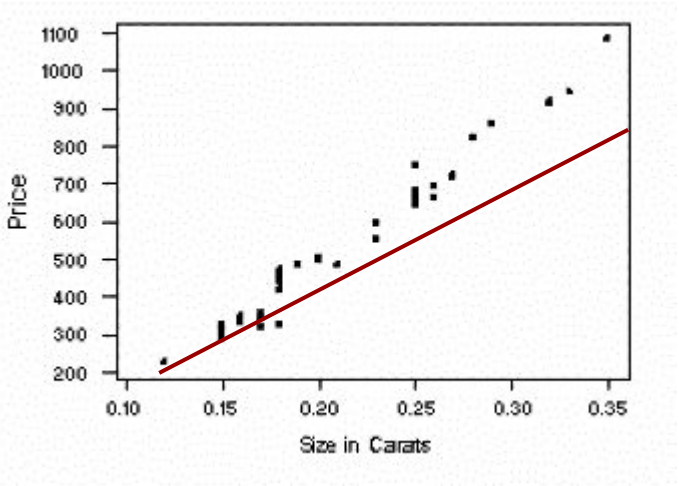
An example of **parametric learning**: we know the "shape" of the data (in our case linear), but want to find the exact parameters that define this shape.

Error

We want to reflect the real world as accurately as possible.

In other words, our model should minimize error (deviations between our theoretical model and observed data).





Question:

How do we represent error (the distance between our line and a data point) mathematically?

HINT: Recall the formula for geometric (Euclidean) distance.

Least Squares Error

We define our error as follows:

theoretical

$$\sum_i (y_i - (B_0 + B_1 x_1 + \ldots + B_n x_n))^2$$

observed

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

Our Goal

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

http://cdn.artofthetitle.com/assets/sm/upload/2i/6b/v9/9i/mission_impossible_t.jpg

Luckily, we have R...

We will use the `lm` built-in function to create a linear model for a dataset.

Demo time!

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.

Size of coefficient says nothing about how significant a variable is. (The variable can take on very large values, for example)

P-values

- **P-value** - probability that we chose samples that “happened” to have a relationship when there really is none.
- The smaller the p-value, the more significant our results are.
- Results are **statistically significant** if the p-value is less than or equal to 0.05.

Model "Goodness of Fit"

Common metric is called **R^2**

- We compare our model to a **benchmark model** (where we predict the mean y value, no matter what the x_i's are).
- Let SST be the least-squares value for the benchmark and SSE be the least-squares error for our model.
- Then our R^2 value is $1 - SSE/SST$.

Adjusted R^2

R^2 will always increase as you increase the number of variables

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

Adjusted R^2 does not follow this behavior, and actually decreases when number of parameters (p) becomes too high. You should look at this whenever you examine your R output.

http://www.haghish.com/statistics/stata-blog/stata-programming/images/adjr2.png

Predicting Values

So far, everything we've done has been **in-sample**. Now it's time to use our model on new data.

We'll use the `predict` function to predict some y values for our dataset given some new x_i values.

Demo time!

Regression is powerful.

http://cdn3-www.craveonline.com/assets/uploads/2015/07/Mission-Impossible-2.jpg

When to Use Linear Regression

- When linearity is a sufficiently good assumption of the data
- If constrained by a very strict timeline
  - Linear regression is computationally efficient
  - If there is a constant stream of new data to be processed - “online learning”

Coming Up

Your problem set: Make predictions using (you guessed it) linear regression

Next week: Classification using logistic regression and decision trees

See you then!