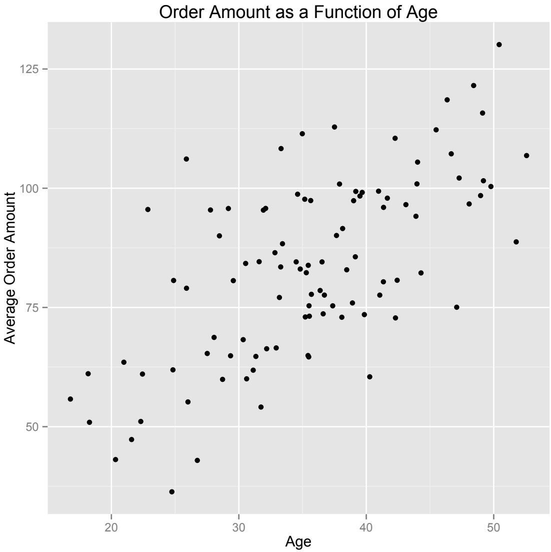
Regression

Scott Hoover

September 3, 2013

- The basic idea of regression analysis is to fit a line or curve to data.
- Regression is primarily used to (i) quantify relationships between variables and to (ii) make predictions (arguably the more interesting of the two).

Suppose we were interested in two variables: the age of a customer and their average order amount. Visualizing the data, we suspect there's a positive relationship between the two variables.


Order Amount as a Function of Age
Average Order Amount
125
100
75
50
20
30
40
50
Age

If we were to guess the relationship (*i.e.*, draw a line), perhaps it would look something like this:

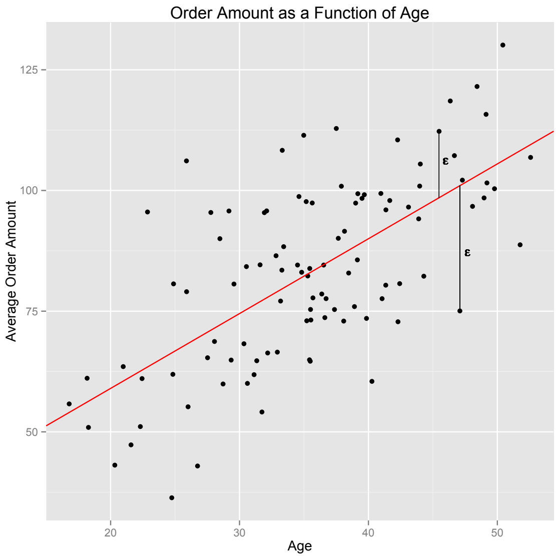

Order Amount as a Function of Age
Average Order Amount
Age
125
100
75
50
20
30
40
50

More concretely, we suspect that the mathematical relationship looks something like this:

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

where α is the intercept, β is the slope, x_i is the age of person i, y_i is the average order amount of person i, and ε is an error term that captures the disturbance from the other variables we cannot observe. $i = 1, 2, ..., n$ is an index; in other words, there are n rows in our table that consists of two columns.

ε is the only thing that is different from the basic equation for a line. It is included because the points do not fall perfectly on a straight line; there is some variation to the data. Visually, we can think of ε_i in the following way:


Order Amount as a Function of Age
Average Order Amount
Age
125
100
75
50
20
30
40
50
ε
ε

There is a basic and efficient algorithm to get a line of best fit:

$$SSE = \min_{\{\alpha,\beta\}} \sum_{i=1}^{n} \varepsilon_i^2 \tag{1}$$

If $\varepsilon_i = y_i - \alpha - \beta x_i$ then (1) can be re-written as

$$SSE = \min_{\{\alpha,\beta\}} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{2}$$

Calculus!

$$\frac{\partial SSE}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left[\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right]$$

$$\frac{\partial SSE}{\partial \alpha} = \sum_{i=1}^{n}\left[\frac{\partial}{\partial \alpha}(y_i - \alpha - \beta x_i)^2\right]$$

$$\frac{\partial SSE}{\partial \alpha} = \sum_{i=1}^{n}\left[-2(y_i - \alpha - \beta x_i)\right]$$

$$\frac{\partial SSE}{\partial \alpha} = -2\sum_{i=1}^{n}\left[(y_i - \alpha - \beta x_i)\right]$$

To minimize this function, we set it equal to zero and solve for α

$$0 = -2\sum_{i=1}^{n}\left[(y_i - \alpha - \beta x_i)\right]$$

$$0 = \sum_{i=1}^{n}\left[(y_i - \alpha - \beta x_i)\right]$$

$$0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\alpha - \sum_{i=1}^{n}\beta x_i$$

$$\sum_{i=1}^{n}\alpha = \sum_{i=1}^{n} y_i - \beta\sum_{i=1}^{n} x_i$$

$$n\alpha = \sum_{i=1}^{n} y_i - \beta\sum_{i=1}^{n} x_i$$

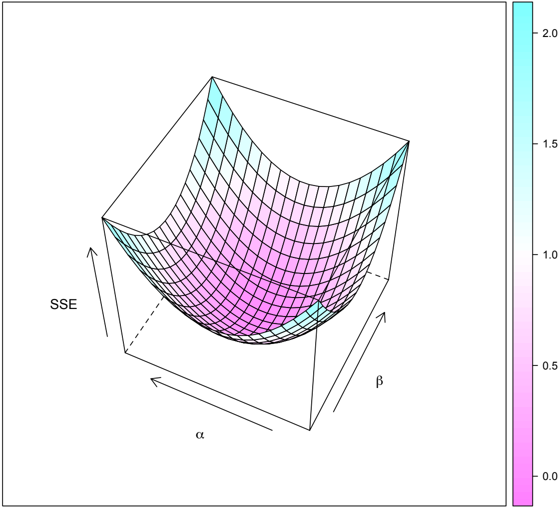
$$\alpha = \left[\frac{\sum_{i=1}^{n} y_i}{n}\right] - \beta\left[\frac{\sum_{i=1}^{n} x_i}{n}\right]$$

$$\alpha = \bar{y} - \beta\bar{x}$$

The above procedure for β is a bit more involved, but results in the following:

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

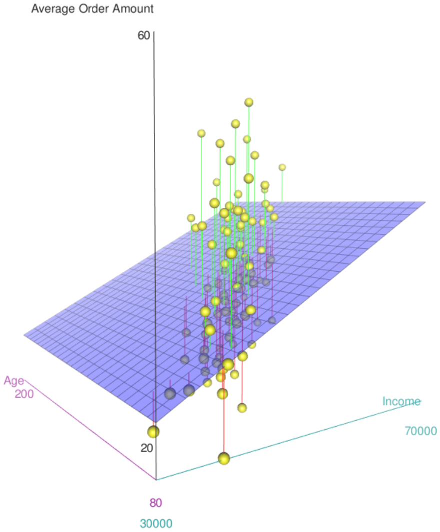
Visualization of the objective function



Our example is limited to two variables; however, in more realistic situations we'd be looking at a number of explanatory variables.

For the multivariate case, matrix notation simplifies the math considerably:

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


Average Order Amount
60
20
Age
200
80
Income
70000
30000

```
Call:
lm(formula = y ~ x1, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-30.147 -11.640  -1.213   9.059  37.867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.2089     6.4357   4.383 2.94e-05 ***
age           1.5473     0.1763   8.777 5.42e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.41 on 98 degrees of freedom
Multiple R-squared:  0.4401,Adjusted R-squared:  0.4344
F-statistic: 77.04 on 1 and 98 DF,  p-value: 5.421e-14
```