Overview of simple linear regression

(STATS310.3: Simple linear regression)

Gunnar Stefansson

June 9, 2012

Background

- This lecture gives an overview of simple linear regression (SLR) at an advanced level
- See other tutorials for more detail
- This tutorial will eventually become less theoretical (more applied)

Typical SLR overview, as an intro to mulreg:
Week 1: Introduction; R; t-tests; P-values; SLR; matrices in passing
Week 2: Data sets and files; case study intro, reading into R; SLR in R; simple scatter plots with regression line; interpreting r and R^2.
Week 3: Case study data sets: plots

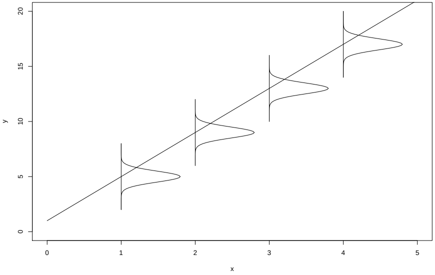
Informal regression

Have data as (x,y)-pairs
Scatterplot indicates relationship
Want to "fit a line" through the data
Evaluate the fit

Formal regression

Fixed numbers, x_i
Random variables: $Y_i \sim n(\alpha + \beta x_i, \sigma^2)$
or: $Y_i = \alpha + \beta x_i + \epsilon_i$
$\epsilon_i \sim n(0, \sigma^2)$ independent and identically distributed (i.i.d.)
The data:

$$y_i = \alpha + \beta x_i + e_i$$



Estimation methods

Least squares estimation technique minimizes: $S = \sum(y_i - (\alpha + \beta x_i))^2$
Maximum likelihood assumes a probability distribution for the data and maximizes the corresponding likelihood function.

The point estimates of a and b

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

These are the least squares estimates of the coefficient of a regression line through the data points (x, y).
It is implicitly assumed that the only errors are in the y-measurements.

The estimator and the estimate

The number b should be viewed as the outcome of the random variable,

$$\hat{\beta} = \frac{\sum (x - \bar{x}) Y}{\sum (x - \bar{x})^2}$$

(note the rewrite from earlier formula b).
i.e. $\hat{\beta}$ is a linear combination of $Y_1, \ldots, Y_n$, commonly assumed to be normally distributed.

Assumptions

Common assumption: Gaussian
Leads to same numerical estimates as OLS
But can also use OLS without explicitly stating a Gaussian assumption
Need to be careful in what results hold with and without normality!

On expected values and variances

Estimating dispersion

A point estimate of σ^2, the variance of the y-measurements, is obtained with

$$s^2 = \frac{\sum_i (y_i - (a + bx_i))^2}{n-2}$$

The predicted value of y at a given x is often denoted by $\hat{y} = a + bx$ and therefore

$$s^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}$$

Commonly $\hat{\sigma}^2$ is used in place of s^2.

Correlation and explained variation

Recall the the correlation coeffficient r is always between -1 and 1.
Write $SSE = \sum(y-\hat{y})^2$ (sum of squared errors, i.e. error after regression), and $SSTOT = \sum(y-\bar{y})^2$ (total sum of squares, i.e. before regression)
Definition: The explained variation is

$$R^2 = 1 - \frac{SSE}{SSTOT}$$

Note:

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = \ldots = r^2$$

Output from regression software

Overview and vocabulary

Vocabulary:

- Regression
- Standard error
- Regression analysis
- Least squares estimation
- ...