Simple Linear Regression

ST552 Lecture 2

Charlotte Wickham

January 6, 2015

Review

High level review

Since simple linear regression is a special case of multiple linear regression, we'll leave the "whys?" to when we cover multiple linear regression.

- the simple linear regression model → $y = ax + b + \text{error}$, $X\beta + \varepsilon$
- interpretation
- assumptions
- how the estimates are found
- properties of the estimates
- F-tests

The simple linear regression model

n observations are collected in pairs, $(x_i, y_i), i = 1, \ldots, n$ where the y_i are generated according to the model,

straight line

noise or error term random

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

response variable — intercept — slope

What is random?

assumptions

where ϵ_i are independent and identically distributed with expected value zero, and variance σ^2.

For inference, we often also assume the ϵ_i are Normally distributed,

$$\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$$

β_0, β_1 fixed parameters

$\hat{\beta}_0, \hat{\beta}_1$ random variables

x_i fixed

The assumptions in words

- Linearity: the mean response is a straight line function of the explanatory variable
- Constant spread: the standard deviation around the mean response is the same at all values of the explanatory variable
- Normality: the deviations from the mean response, the errors, are Normally distributed.
- Independence: the deviations from the mean response are independent.

Interpretation of the parameters

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$E[y_i] = \beta_0 + \beta_1 x_i \qquad E(\varepsilon_i) = 0$$

Intercept, β_0,
When the explanatory variable is zero, the mean response is β_0.

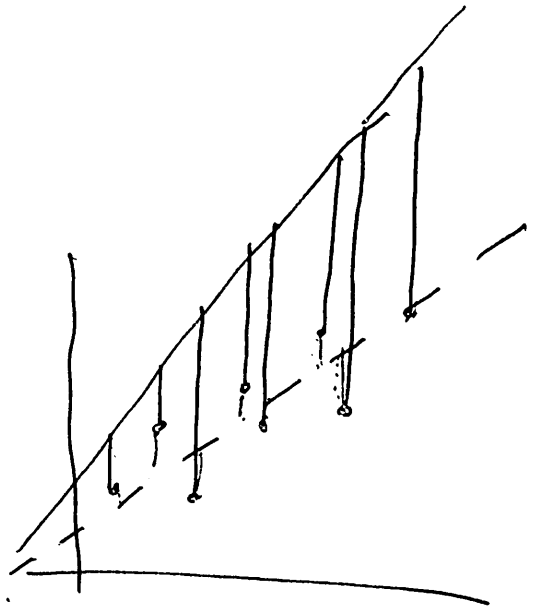
$x_i = 0$

Slope, β_1,
An increase in the explanatory variable of one unit is associated with a change in mean response of β_1.

(Careful with causal language... is it justified?)

But we don't know β_0 and β_1...

The least squares estimates



Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

$$\sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2$$

fitted values

is minimized.

Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Residuals: $e_i = y_i - \hat{y}_i$

We don't require any properties of random variables to derive these estimates.

There are formulae for $\hat{\beta}_0$ and $\hat{\beta}_1$, how do you derive them?

Properties of the least squares estimates

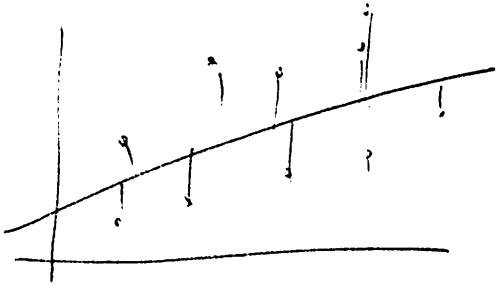
$\hat{\beta}_0$ $\hat{\beta}_1$

$E(\epsilon_i) = 0$ $E(\hat{\beta}_0) = \beta_0$

Using the moment assumptions of ϵ_i, the least squares estimates can be shown to be unbiased. You can derive their variances, $\text{Var}(\hat{\beta}_0)$, $\text{Var}(\hat{\beta}_1)$, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$, but they depend on the unknown σ.

An unbiased estimate of σ is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$



Intuition: $\frac{1}{n}\sum_{i=1}^{n} e_i^2$ seems a reasonable place to start to estimate the variance of the errors, but this tends to underestimate the variance because we picked our estimates to make the sum of squared errors as small as possible.