Simple linear regression

Johan A. Elkink
School of Politics & International Relations
University College Dublin

28 September 2015

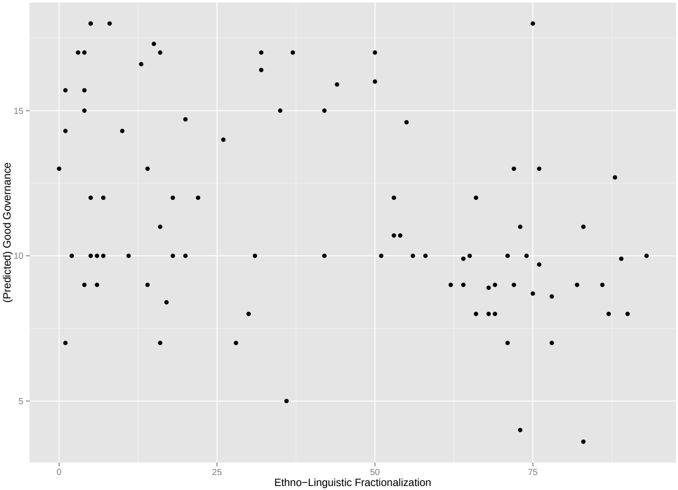
Outline

Example

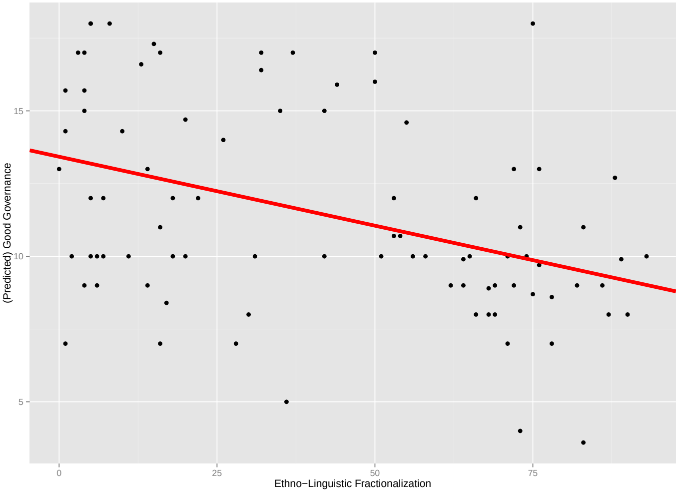
For this example we will make use of replication data from Ross (2004), taking only data from 1997, where we take as the dependent variable the level of good governance in a country and as explanatory variable the level of ethno-linguistic fractionalization (ELF).

Basic interpretation of ELF: If I take two random individuals in a country, how likely is it that they are from two different ethno-linguistic groups?

Linear regression



Linear regression



Linear regression

	Dependent variable:
	goodgovt
elf	-0.047^{***}
	(0.012)
Constant	13.423^{***}
	(0.597)
Observations	87
R^2	0.162
Adjusted R^2	0.153
Residual Std. Error	3.247 (df = 85)
F Statistic	16.484^{***} (df = 1; 85)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Example: interpretation

$$goodgovt_i = 13.42 - 0.05elf_i + \varepsilon_i$$

Example: interpretation

$$goodgovt_i = 13.42 - 0.05elf_i + \varepsilon_i$$

- There is a negative relation between ethno-linguistic fractionalization and good governance.

Example: interpretation

$$goodgovt_i = 13.42 - 0.05 elf_i + \varepsilon_i$$

- There is a negative relation between ethno-linguistic fractionalization and good governance.
- For every increase in ELF by 1, the good governance measure decreases by 0.05.

Example: interpretation

$$goodgovt_i = 13.42 - 0.05elf_i + \varepsilon_i$$

- There is a negative relation between ethno-linguistic fractionalization and good governance.
- For every increase in ELF by 1, the good governance measure decreases by 0.05.
- For a (hypothetical) country where the ELF is 50, the level of good governance would be $13.42 - 0.05 \times 50 = 11.07$.

Outline

Linear model

The regression equation here is

$$y_i = b_0 + b_1 x_i + \varepsilon_i,$$

whereby $\mathbf{y}$ is the dependent variable, $\mathbf{x}$ the independent variable, i an indicator of the case (country), b_0 and b_1 the model parameters, and $\boldsymbol{\varepsilon}$ the error term.

Residuals

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

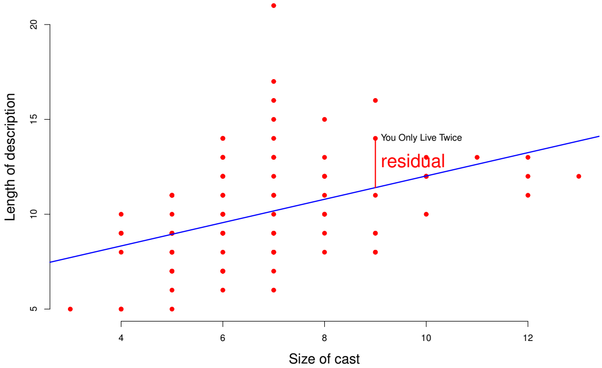
The linear prediction given the parameters would be $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$.

The extend to which the real value differs from the predicted value is:

$$y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 x_i = e_i.$$

By this formulation, the **residuals** ($\boldsymbol{e}$) are the vertical distance between a point and the regression line (i.e. not the shortest distance between the point and the line).

Linear regression (residuals)



Ordinary Least Squares

To estimate the regression line, we need to estimate the parameters b_0 and b_1.

Ordinary Least Squares (OLS) is the most common method to do so. With OLS, we estimate the parameters such that the **sum of squared residuals** are minimized.

(This is the same as minimizing the variance of the residuals.)

Outline

Model fit

Once we have estimated a line, we might ask how well this line summarizes the relationship between those two variables.

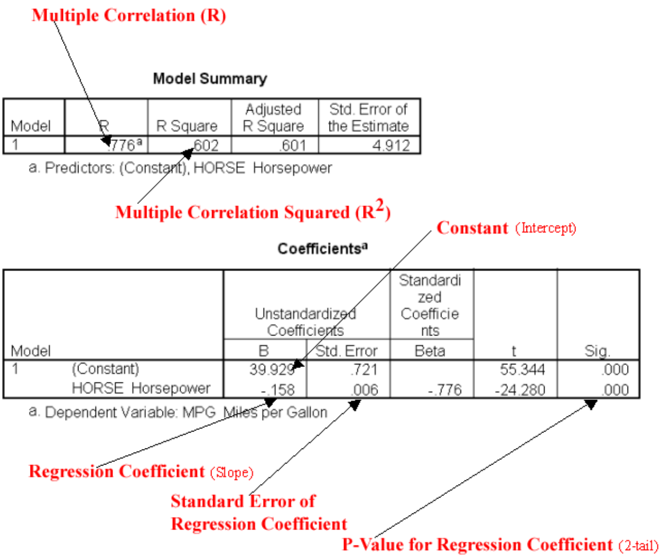
A common measure is R^2:

$$R^2 = 1 - \frac{\text{residual sum of squares}}{\text{total sum of squares}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

This can be interpreted as the proportion of the variation in $\mathbf{y}$ explained by this model.

Note the relation with correlation coefficient Peason's r: $r = \sqrt{R^2}$.

SPSS regression output


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.776[a]	.602	.601	4.912

a. Predictors: (Constant), HORSE Horsepower

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	39.929	.721		55.344	.000
	HORSE Horsepower	-.158	.006	-.776	-24.280	.000

a. Dependent Variable: MPG Miles per Gallon

Ross, Michael. 2004. “Does taxation lead to representation?” *British Journal of Political Science* 34:229–249.