Simple linear regression

Johan A. Elkink

University College Dublin

16 February 2012

Outline

Example

File `oecd_1960.sav` contains data on OECD countries in 1960, with variables:

COUNTRY	Country name
PCINC	Income per capita
AGR	Percentage employed in agriculture
IND	Percentage employed in industry
SER	Percentage employed in services

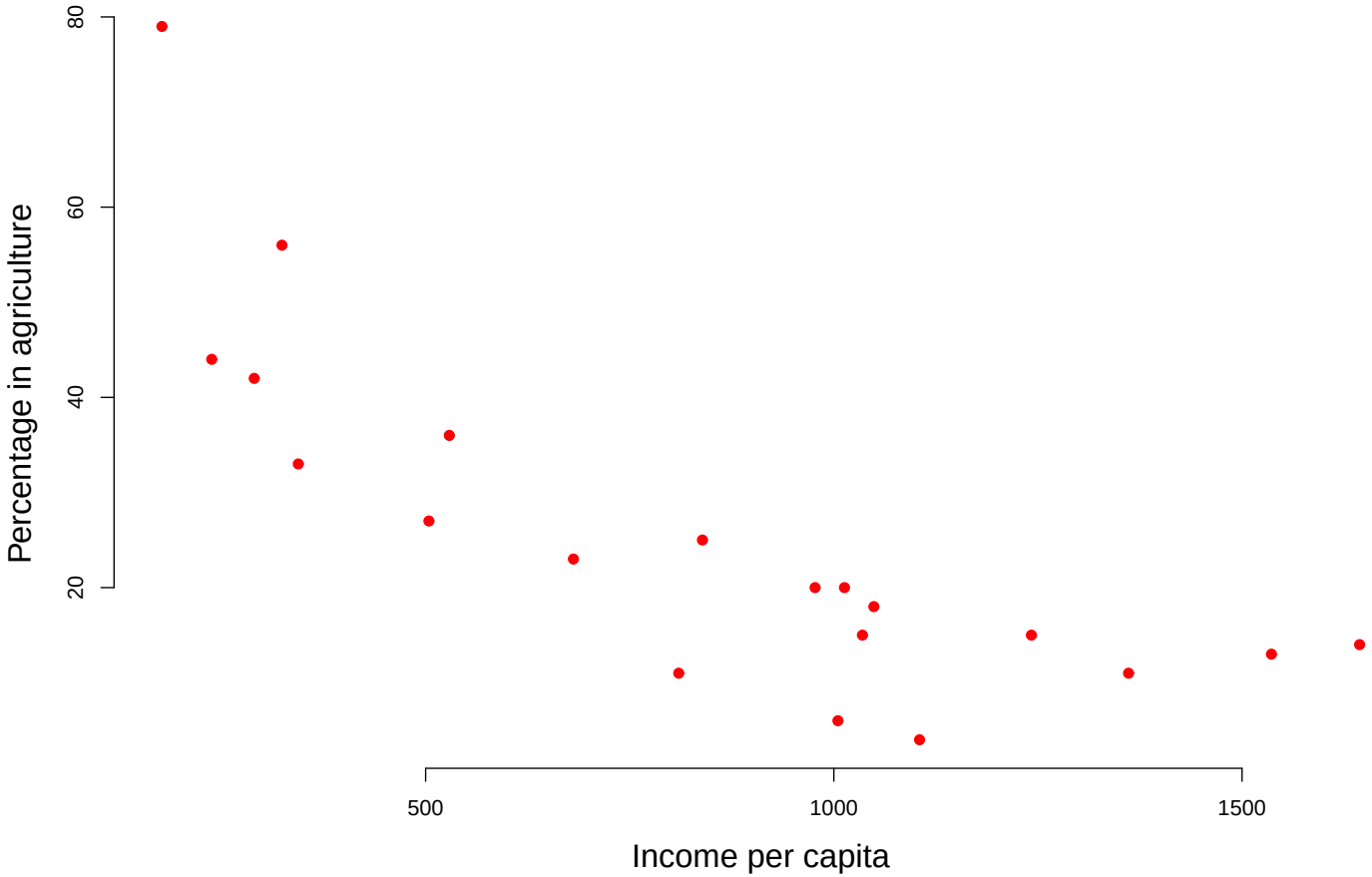
Example

File `oecd_1960.sav` contains data on OECD countries in 1960, with variables:

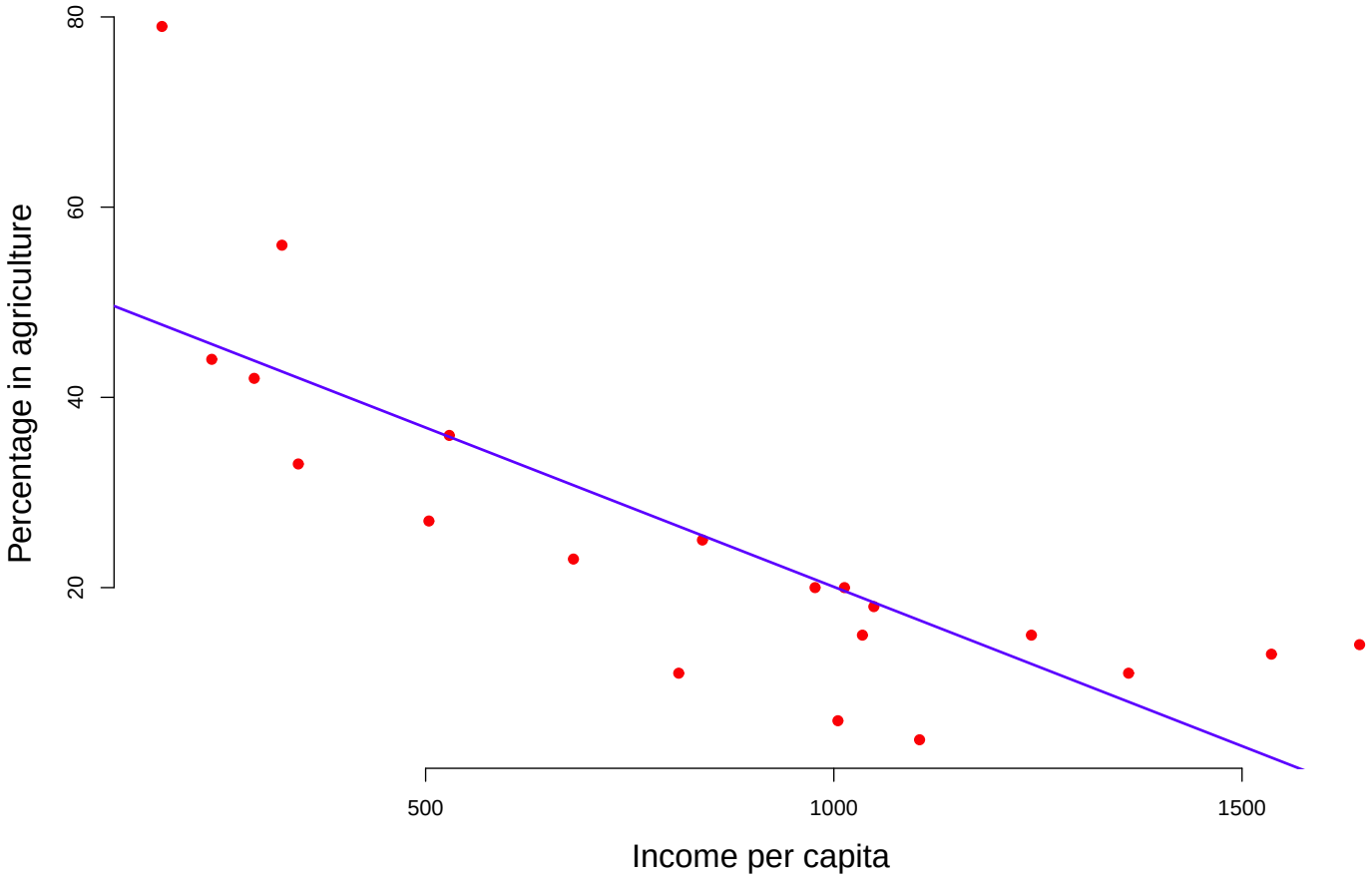
COUNTRY	Country name
PCINC	Income per capita
AGR	Percentage employed in agriculture
IND	Percentage employed in industry
SER	Percentage employed in services

We will look at the relation between per capita income and employment in agriculture. What do you expect to see?

Linear regression



Linear regression



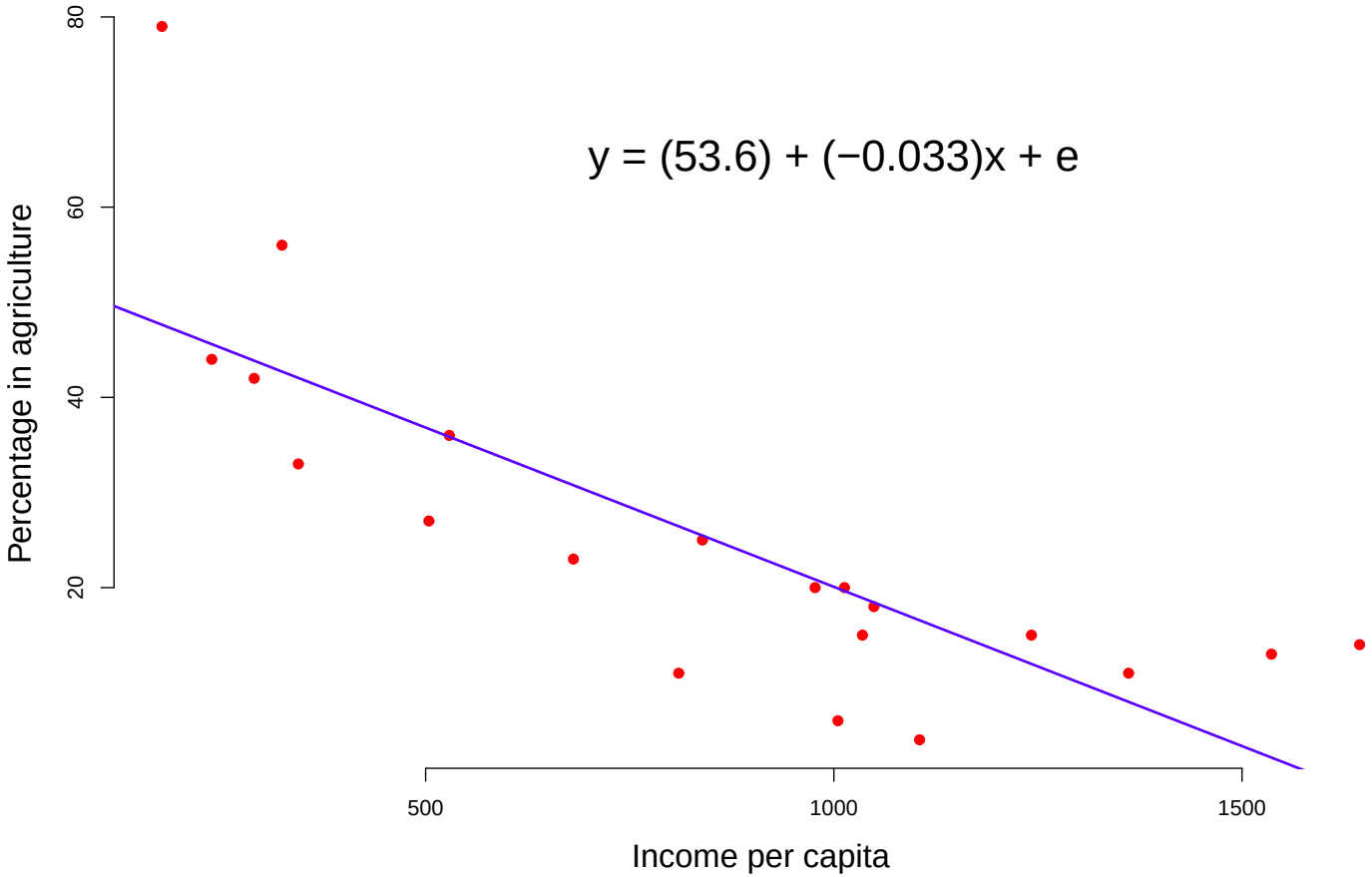
Example

Which is the dependent and which the independent variable in this example?

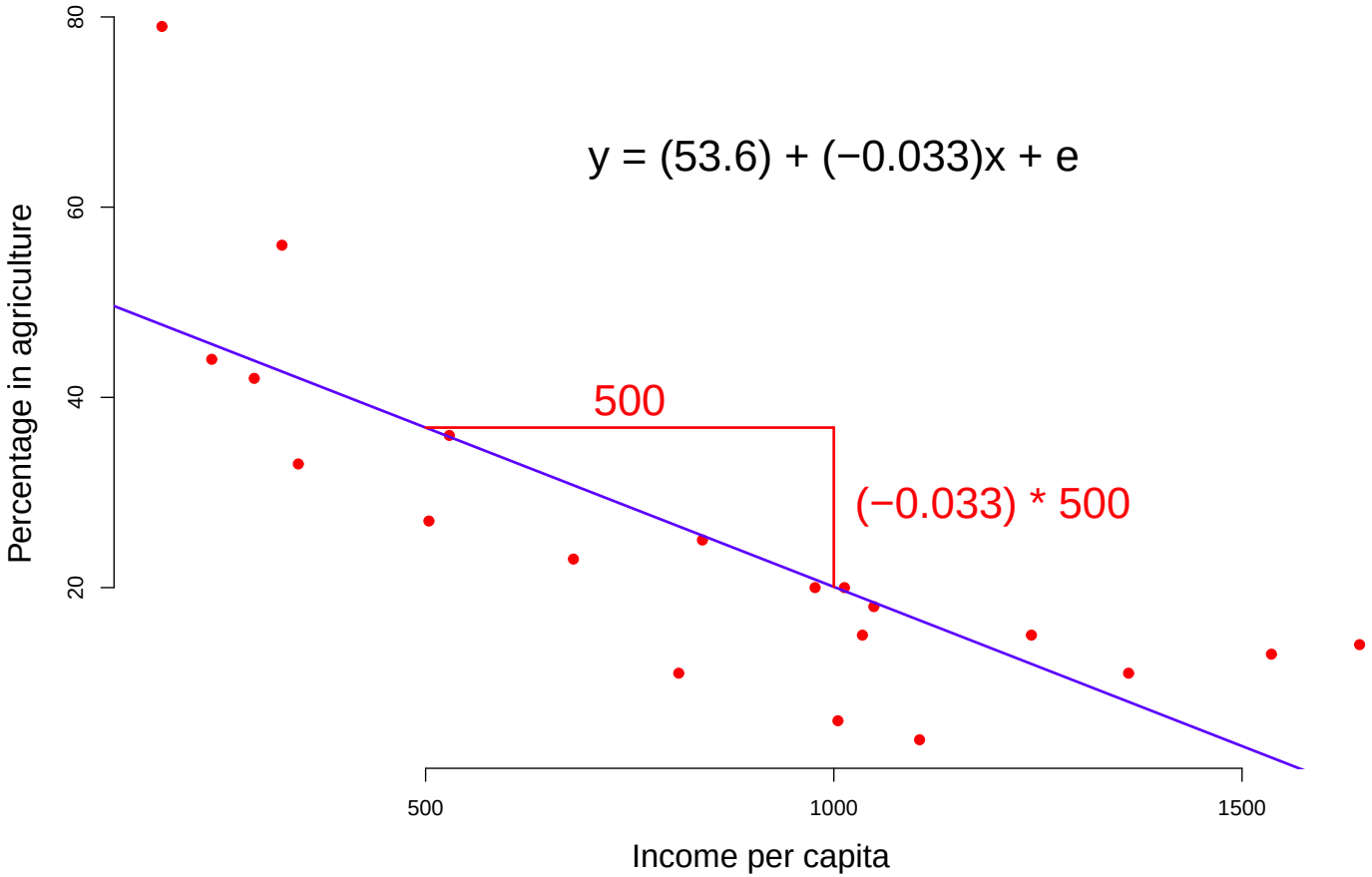
Exercise

- Open `oecd_1960.sav`
- Create a scatter plot for percentage in agriculture by per capita income
- Add a regression line to the plot
- Run a linear regression with those two variables

Linear regression



Linear regression



Example: interpretation

$$y_i = 53.6 - 0.033x_i + \varepsilon_i$$

Example: interpretation

$$y_i = 53.6 - 0.033x_i + \varepsilon_i$$

- There is a negative relation between income per capita and percentage employed in agriculture.

Example: interpretation

$$y_i = 53.6 - 0.033x_i + \varepsilon_i$$

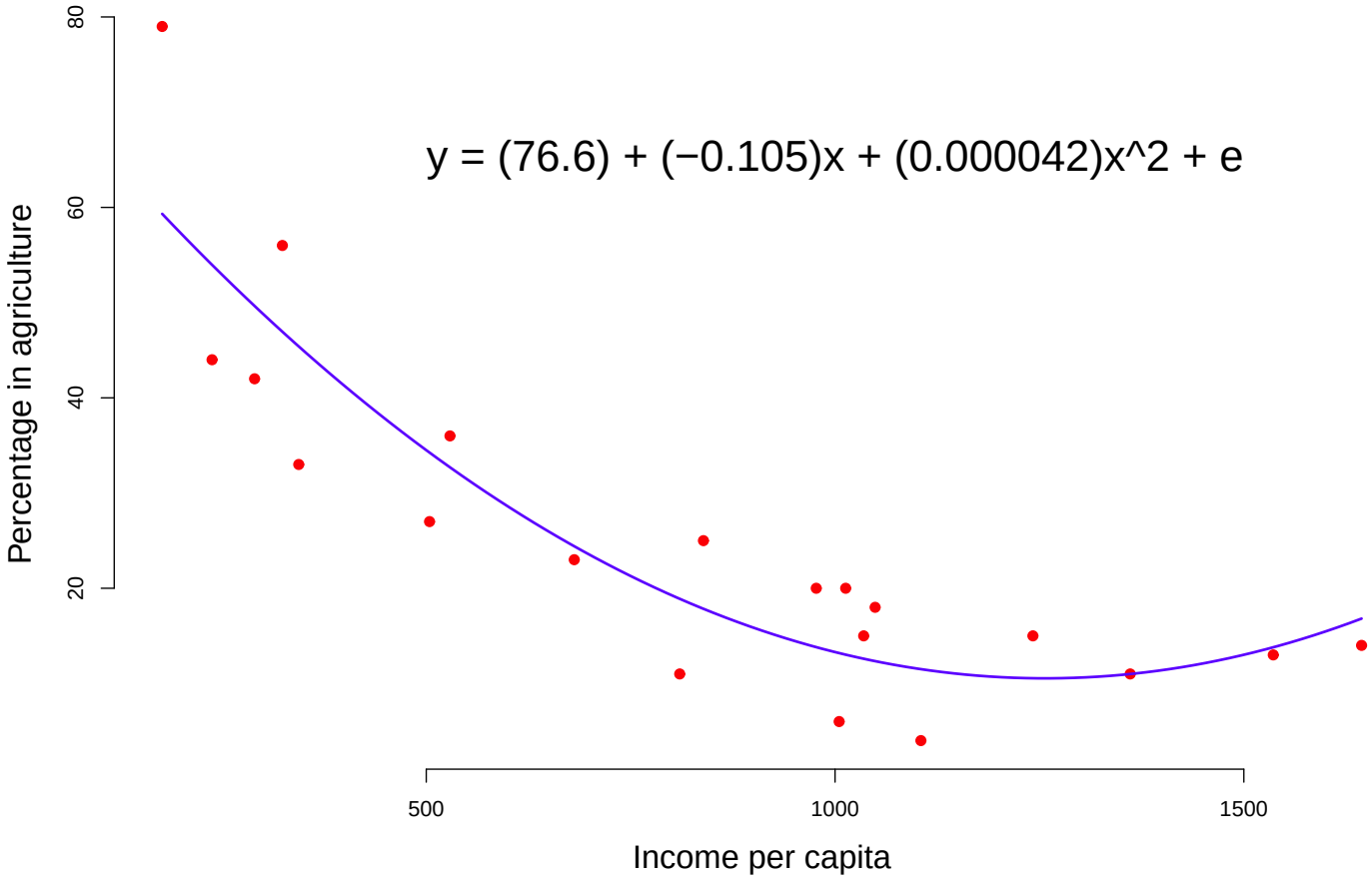
- There is a negative relation between income per capita and percentage employed in agriculture.
- For every increase in income per capita by 1,000 dollar, the percentage employed in agriculture decreases by 33 percent points ($-0.033 \cdot 1000 = -33$).

Example: interpretation

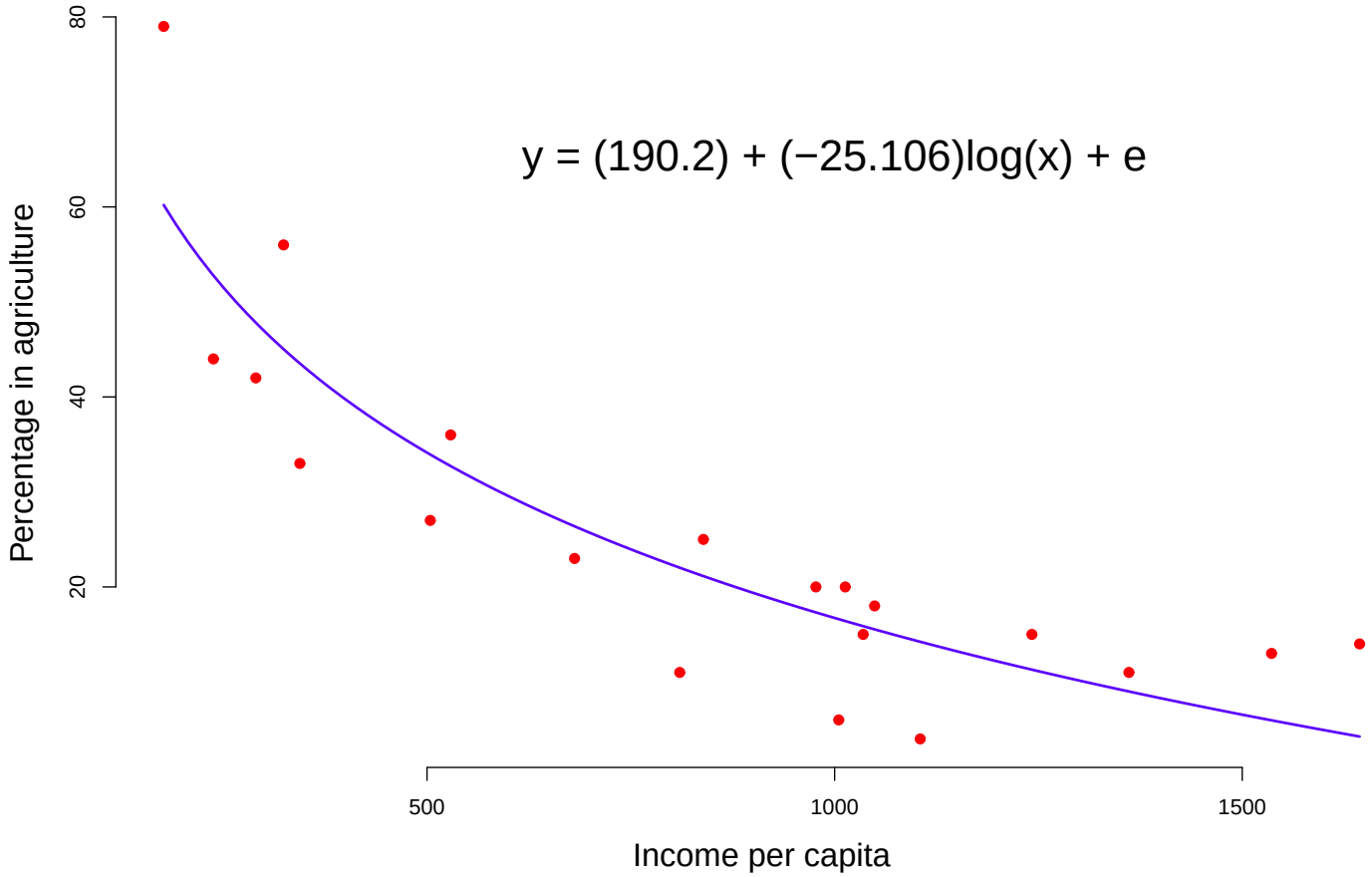
$$y_i = 53.6 - 0.033x_i + \varepsilon_i$$

- There is a negative relation between income per capita and percentage employed in agriculture.
- For every increase in income per capita by 1,000 dollar, the percentage employed in agriculture decreases by 33 percent points ($-0.033 \cdot 1000 = -33$).
- For a (hypothetical) country where income per capita is 0 dollar, 53.6% would work in agriculture.

Linear regression (squared)



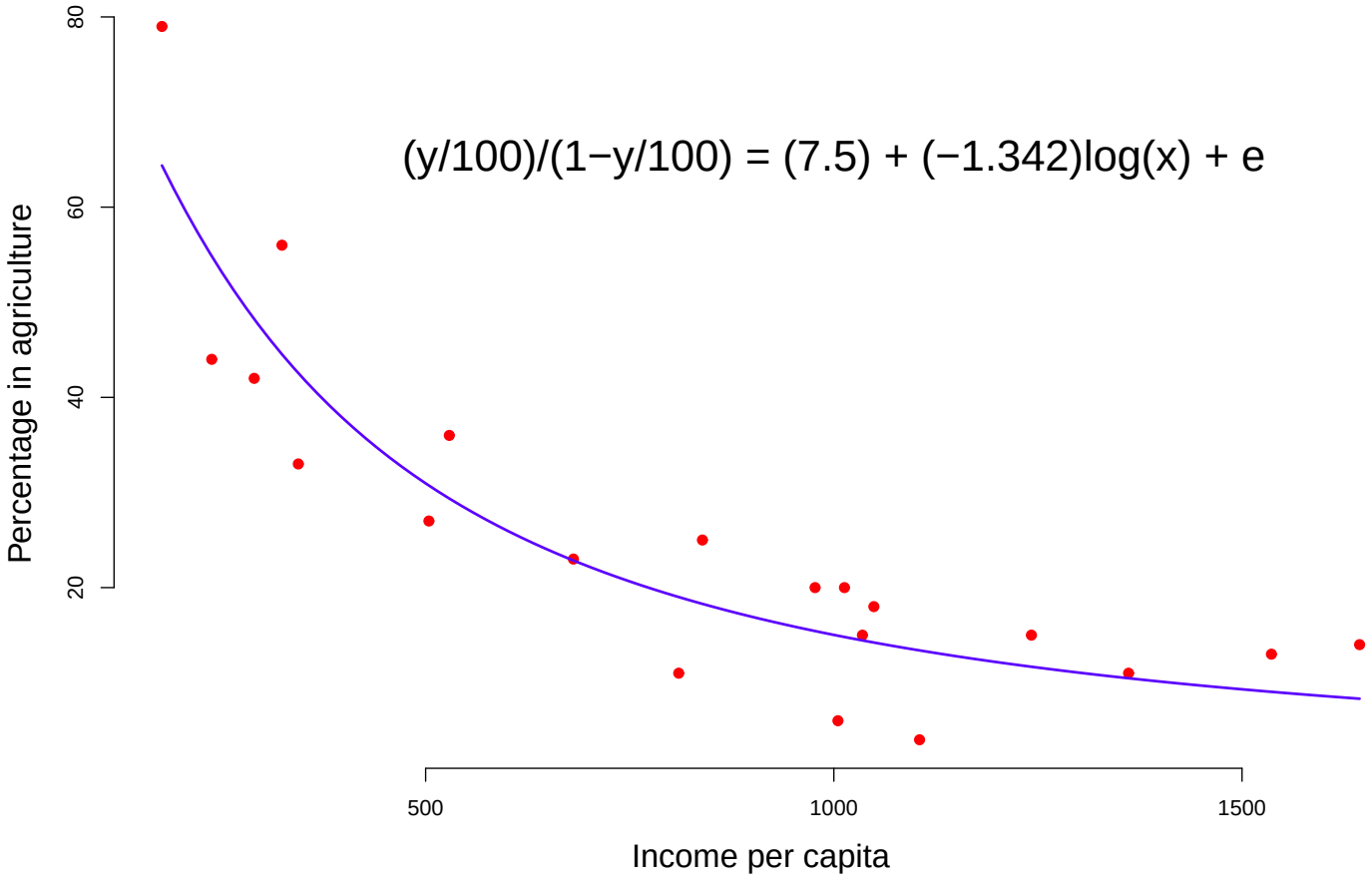
Linear regression (log)



Exercise

- Run a linear regression with $log(x)$ instead of x.
- Run a linear regression with x^2 and x instead of just x.

Linear regression (logit)



Outline

Linear model

The regression equation here is

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

whereby **y** is the dependent variable, **x** the independent variable, i an indicator of the case (country), b_0 and b_1 the model parameters, and ε the error term.

Residuals

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

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

Residuals

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

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

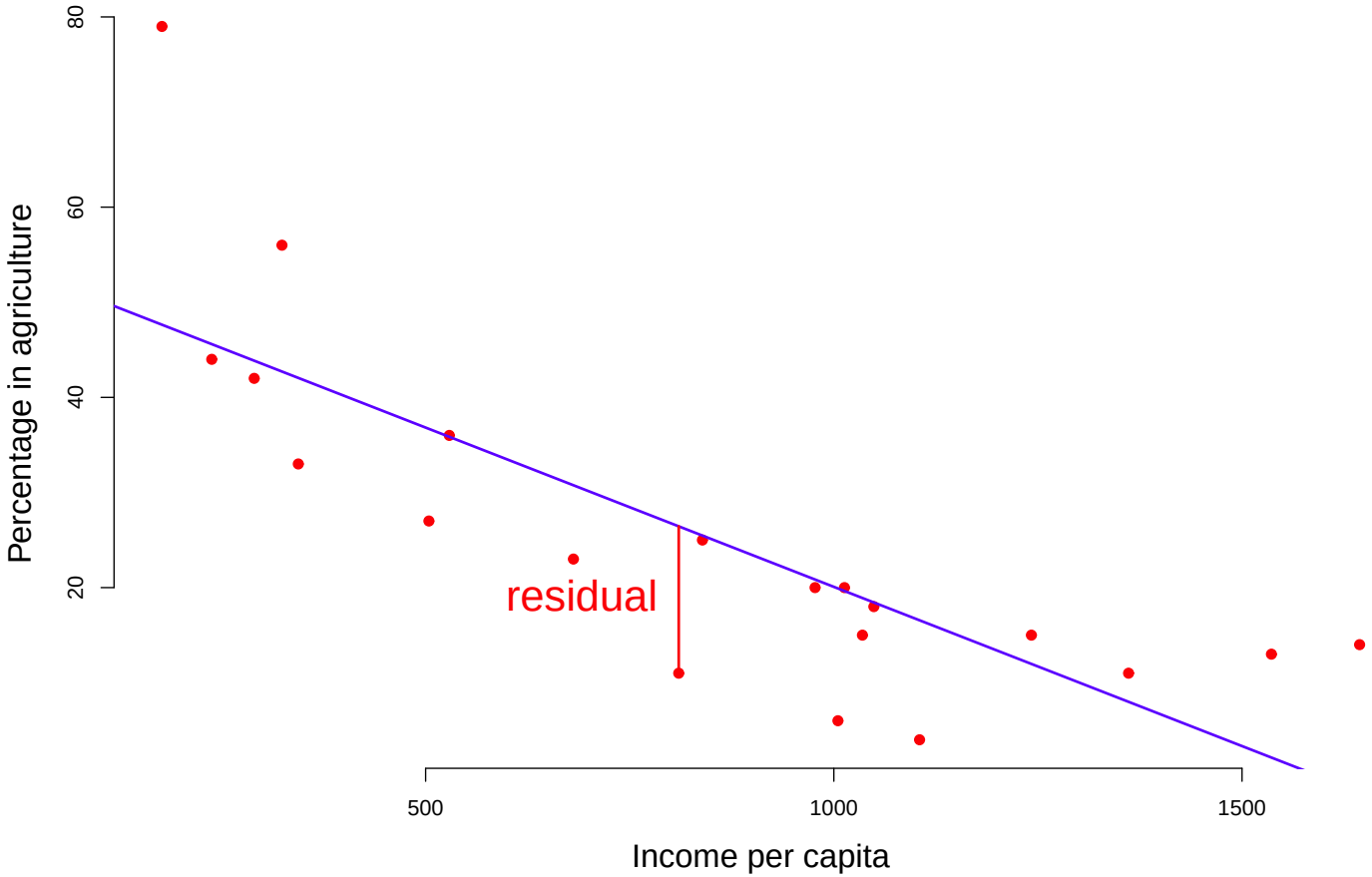
The extend to which the real value differs from the predicted value is:

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

Residuals

By this formulation, the residuals ($\mathbf{e}$) 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

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.

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.

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:

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

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 **y** explained by this model.

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 **y** explained by this model.

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

Outline

Exercise

Repeat for both industry (IND) and services (SER):

1. Plot percentage in sector against income per capita.
2. Regress percentage in sector on income per capita.
3. Interpret the regression results.
4. Evaluate the model fit.

Exercise

Open `bes_class_data.sav` and investigate the relation between:

- `lr` and `trustprl`
- `eumember` and `lr`