t-tests and F-tests in regression

Johan A. Elkink

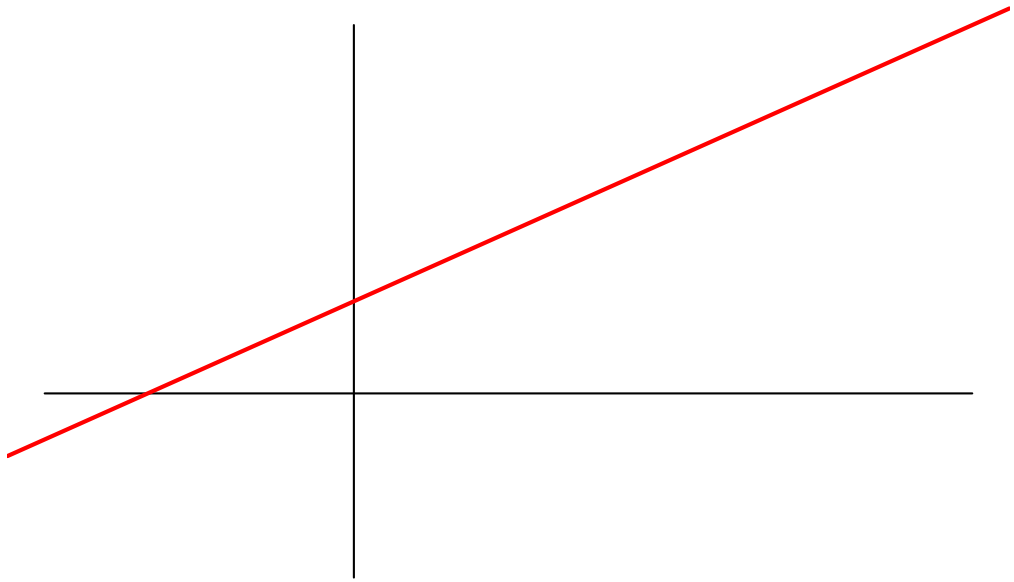
University College Dublin

5 April 2012

Outline

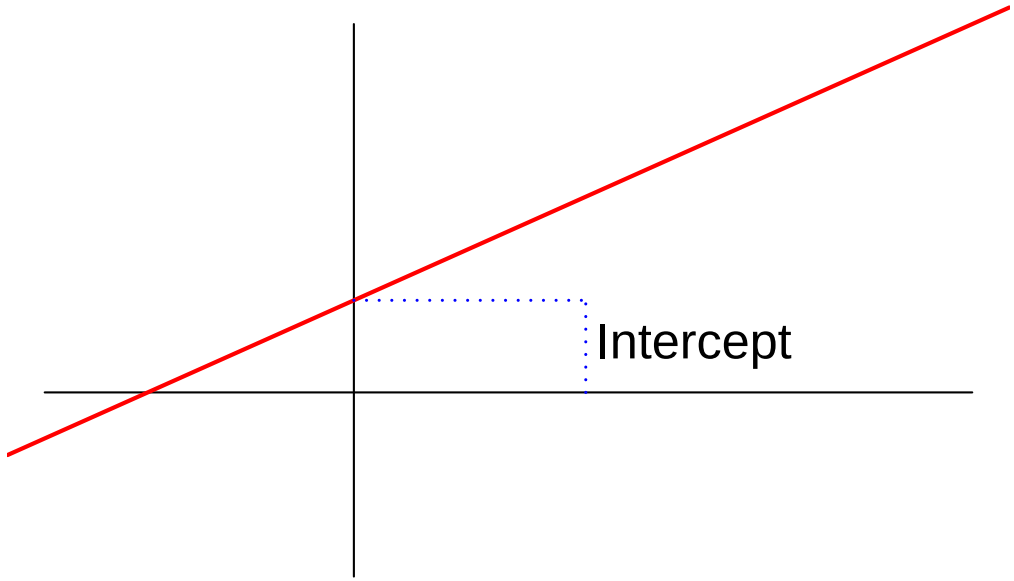
Outline

Linear equations



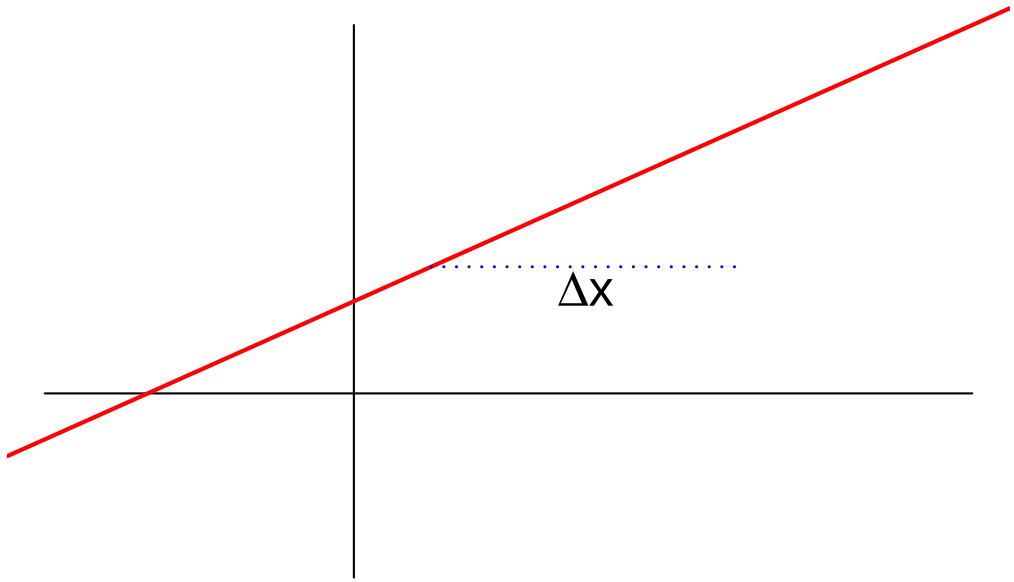
$$y = Intercept + Slope * x$$

Linear equations



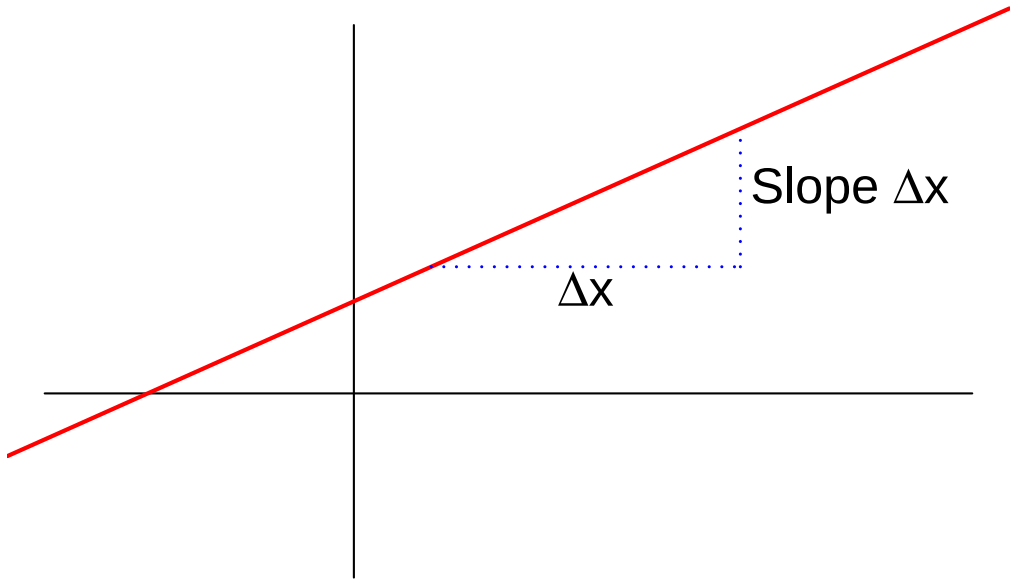
$$y = Intercept + Slope * x$$

Linear equations



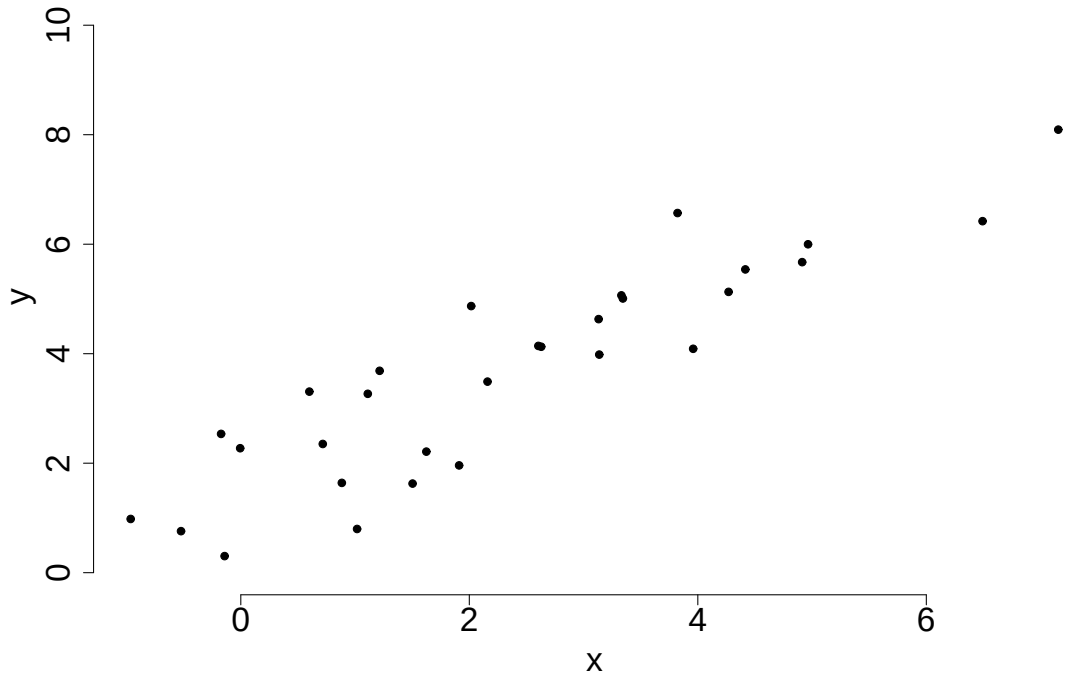
$$y = Intercept + Slope * x$$

Linear equations

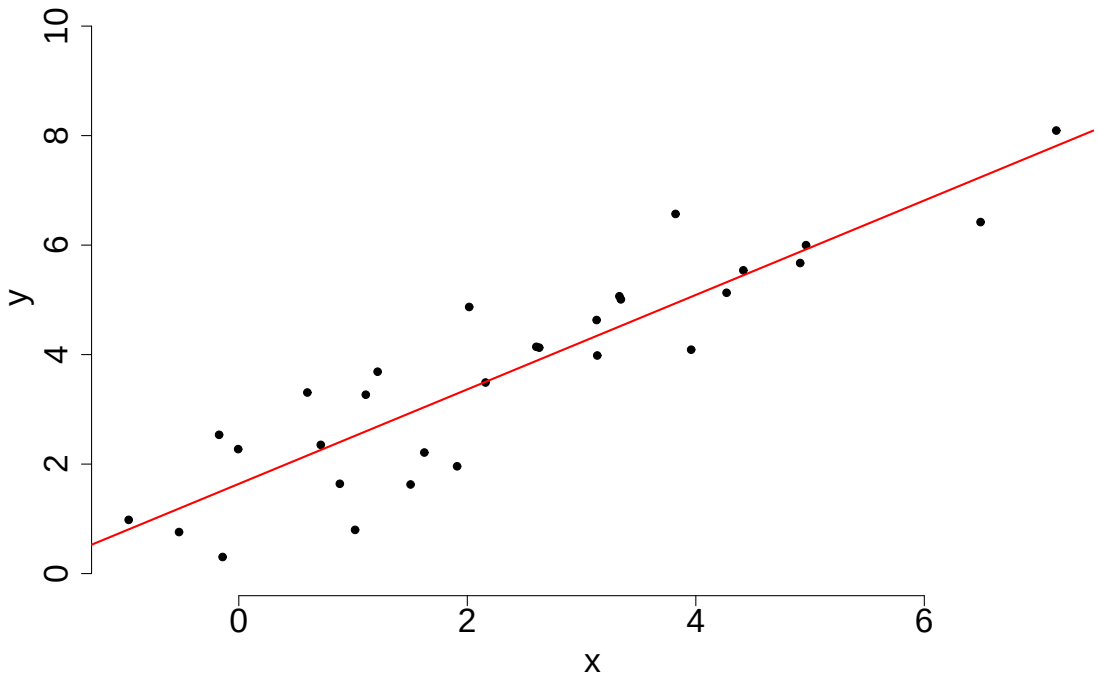


$$y = Intercept + Slope * x$$

Simple regression model

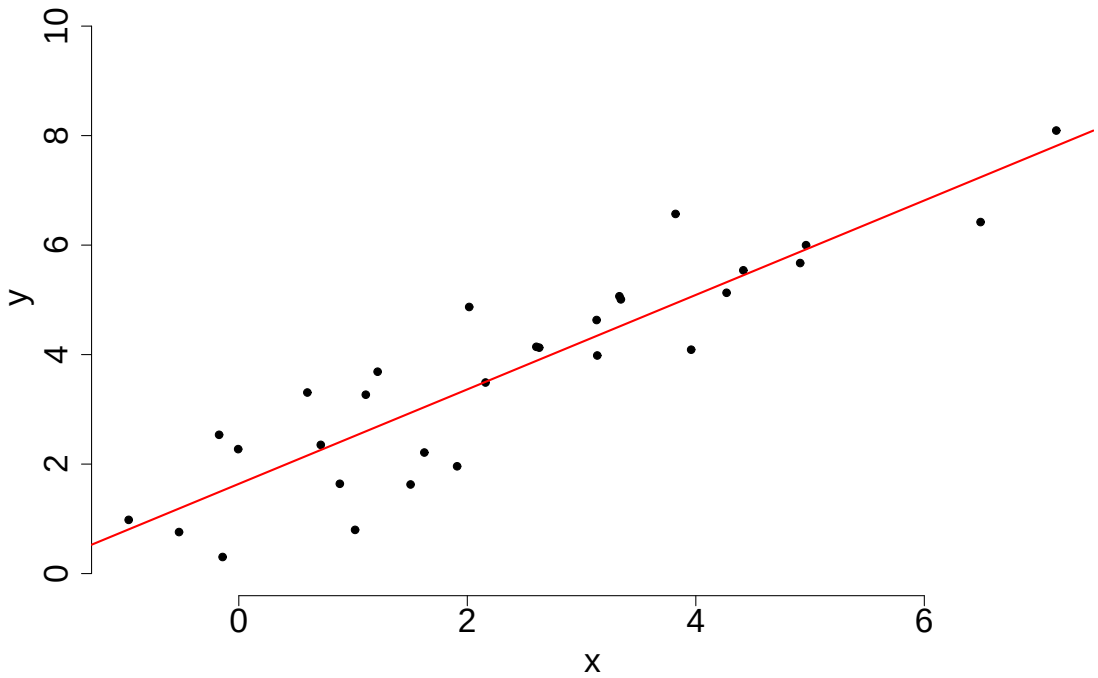


Simple regression model



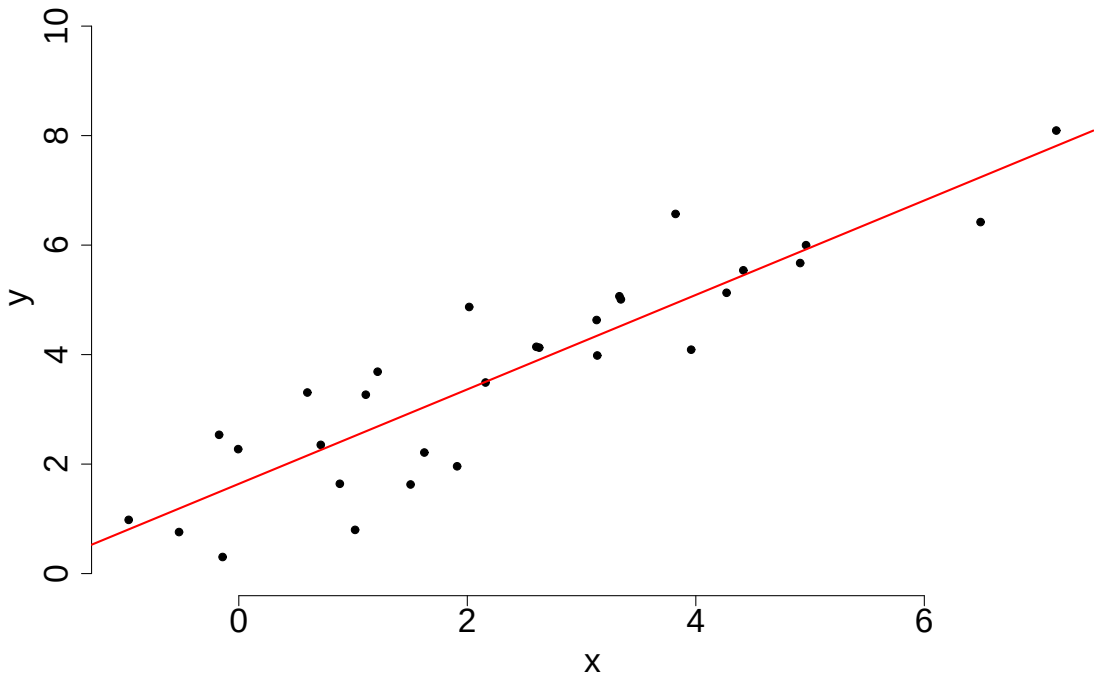
$$y = Intercept + Slope * x$$

Simple regression model



$$y = \beta_0 + \beta_1 x$$

Simple regression model



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

Notation

y_i	Value on the dependent variable for case i
x_i	Value on the independent variable for case i
$\bar{x}$	Mean value on the independent variable for case i
ε_i	The error for case i: $\varepsilon_i = y_i - \hat{y}_i$
β_k	True coefficient for variable k
$\hat{\beta}_k$	Estimated coefficient for variable k
$\hat{y}_i$	Predicted value on the dependent variable for case i

Ordinary Least Squares

"Quickly put, the regression line is chosen to minimize the RSS; it has slope $\hat{\beta}_1$, intercept $\hat{\beta}_0$, and goes through the point $(\bar{x}, \bar{y})$. Furthermore, the estimate for σ^2 is $\hat{\sigma}^2 = RSS/(n-2)$" (Verzani 2005: 280).

Outline

Breakdown of variance

Total Sum of Squares (TSS): $\sum_{i=1}^{N}(y_i - \bar{y})^2$

Explained Sum of Squares (ESS): $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$

Residual Sum of Squares (RSS): $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\varepsilon_i^2$

$$TSS = ESS + RSS$$

Breakdown of variance

Total Sum of Squares (TSS): $\sum_{i=1}^{N}(y_i - \bar{y})^2$

Explained Sum of Squares (ESS): $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$

Residual Sum of Squares (RSS): $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\varepsilon_i^2$

$$TSS = ESS + RSS$$

Sometimes the second is called "regression sum of squares" (RSS) and the third "errors sum of squares" (ESS), which might in fact be more accurate, since ε really represents errors, not residuals, in this specification. Beware the confusion!

R^2

How much of the variance did we explain?

R^2

How much of the variance did we explain?

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Can be interpreted as the *proportion of total variance explained by the model*.

R^2

How much of the variance did we explain?

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Can be interpreted as the *proportion of total variance explained by the model*.

For this interpretation, the model must include an intercept. Generally, one should not attach too much value to having a high R^2 - it is usually more important to understand whether x affects y and by how much, rather than to understand how much of y has not yet been explained.

R^2

How much of the variance did we explain?

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Can be interpreted as the *proportion of total variance explained by the model*.

For this interpretation, the model must include an intercept. Generally, one should not attach too much value to having a high R^2 - it is usually more important to understand whether x affects y and by how much, rather than to understand how much of y has not yet been explained.

For simple linear regression (i.e. one independent variable), R^2 is the same as the correlation coefficient, Pearson's r, squared.

Outline

Outline

Inference from regression

In linear regression, the sampling distribution of the coefficient estimates form a normal distribution, which is approximated by a t distribution due to approximating σ by s.

Inference from regression

In linear regression, the sampling distribution of the coefficient estimates form a normal distribution, which is approximated by a t distribution due to approximating σ by s.

Thus we can calculate a confidence interval for each estimated coefficient.

Inference from regression

In linear regression, the sampling distribution of the coefficient estimates form a normal distribution, which is approximated by a t distribution due to approximating σ by s.

Thus we can calculate a confidence interval for each estimated coefficient.

Or perform a hypothesis test along the lines of:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Inference from regression

To calculate the confidence interval, we need to calculate the standard error of the coefficient.

Inference from regression

To calculate the confidence interval, we need to calculate the standard error of the coefficient.

Rule of thumb to get the 95% confidence interval:

$$\beta - 2SE < \beta < \beta + 2SE$$

Inference from regression

To calculate the confidence interval, we need to calculate the standard error of the coefficient.

Rule of thumb to get the 95% confidence interval:

$$\beta - 2SE < \beta < \beta + 2SE$$

Thus if β is positive, we are 95% certain it is different from zero when $\beta - 2SE > 0$.

Inference from regression

To calculate the confidence interval, we need to calculate the standard error of the coefficient.

Rule of thumb to get the 95% confidence interval:

$$\beta - 2SE < \beta < \beta + 2SE$$

Thus if β is positive, we are 95% certain it is different from zero when $\beta - 2SE > 0$. (Or when the t value is greater than 2 or less than -2.)

Outline

Breakdown of variance

Total Sum of Squares (TSS): $\sum_{i=1}^{N}(y_i - \bar{y})^2$

Explained Sum of Squares (ESS): $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$

Residual Sum of Squares (RSS): $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\varepsilon_i^2$

$$TSS = ESS + RSS$$

F-test

In simple linear regression, we can do an F-test:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

F-test

In simple linear regression, we can do an F-test:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

$$F = \frac{ESS/1}{RSS/(n-2)} = \frac{ESS}{\hat{\sigma}^2} \sim F_{1,n-2}$$

with 1 and $n-2$ degrees of freedom.

F-test

In simple linear regression, we can do an F-test:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

$$F = \frac{ESS/1}{RSS/(n-2)} = \frac{ESS}{\hat{\sigma}^2} \sim F_{1,n-2}$$

with 1 and $n-2$ degrees of freedom.

For multiple regression, this would generalize to:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \sim F_{k-1,n-k}$$

Outline

Exercise

Open `oecd_1960.sav`.

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

1. Regress percentage in sector on income per capita.
2. Interpret the regression results.
3. Evaluate the model fit.
4. Interpret the t- and F-tests.

Exercise

Open `bes_class_data.sav` and investigate:

- whether left-right self-placement (`lr`) and explains trust in politics (`trustpol`)
- whether left-right self-placement influences attitude towards EU membership (`eumember`)