Ekaterina Kazak
University of Manchester

The purpose of this session is to introduce you to the OLS inference and restricted least squares estimator. The code from the previous tutorial will be useful, keep it open.

Size and power properties of statistical tests

Simulate a bivariate linear regression model:

$$y_i = 1 - 3x_i + u_i, \tag{1}$$

where $u \sim \mathcal{N}(0, 0.25)$, $x \sim \mathcal{N}(5, 16)$ and $n = 500$.

a) Examine the size properties of the two-sided test for the slope coefficient:

   - Simulate the null rejection probability for a two-sided test $H_0 : \beta = -3$ over a 1000 simulation draws.
   - How does the sample size affect the Type I error? Plot the empirical null rejection probability over a grid of sample sizes $n = 100, ..., 5000$.
   - How does the number of simulation draws, or Monte Carlo iterations, influences the precision of the empirical size? Using the function `cumsum` plot the empirical size of the test against the number of simulations $mc = 10, ..., 10000$.
   - How does the error term variance influence the Type I error? Plot the empirical null rejection probability over a grid of $\sigma_u^2 = 0.1, ..., 10$ for $n = 1000$.

b) Plot power curves for $n = 500$ and $n = 1000$ (on one graph) of the two-sided test for the slope coefficient for a model defined in (1) with $\sigma_u^2 = 10$.

Restricted least squares

Consider a bivariate linear regression model:

$$y_i = 1 - 3x_i + u_i, \tag{1}$$

where $u \sim \mathcal{N}(0, 0.25)$, $x \sim \mathcal{N}(5, 16)$ and $n = 500$.

a) Use your code from the previous tutorial to create a function which computes the sum of squared residuals as a function of intercept and slope parameter α_0 and α_1. Create a 3D visualization with the help of `plotly` package, which plots the sum of squared residuals surface for a range of parameters $\alpha_0, \alpha_1 \in [-6, 6]$.

b) On the 3D surface mark the parameter combinations which satisfy the restriction $\beta_0 + \beta_1 = 0$. [1]

c) Create a function which estimates a bivariate restricted least squares problem with the restriction of the form $R\beta = r$. Use it to estimate the slope and the intercept which satisfy the restriction $\beta_0 + \beta_1 = 0$. Compare it with the estimates which satisfy the restriction $\beta_0 + \beta_1 = -2$ and $\beta_0 + \beta_1 = 9$.

[1] It might be useful to pre-install `install.packages("pracma", repos="http://R-Forge.R-project.org")`.

Ekaterina Kazak
University of Manchester

Exercises to solve at home

OLS Inference I

We look at the dataset in `attend.csv` again, containing the following variables:

Variable	Definition
`termgpa`	GPA performance measure for term, between 0 (worst) and 4 (best) Note: "Pass" corresponds to `gpa` = 1.0
`priGPA`	cumulative GPA prior to term
`ACT`	ACT score (university entrance test)
`attend`	percent classes attended
`soph`	= 1 if student is a sophomore (2nd-year), = 0 else

Consider two linear regression models

$$\textbf{termgpa}_i = \alpha + \beta_{\text{attend}}\,\textbf{attend}_i + \epsilon_i. \tag{1}$$

$$\textbf{termgpa}_i = \alpha + \beta_{\text{attend}}\,\textbf{attend}_i + \beta_{\text{ACT}}\,\textbf{ACT}_i + \beta_{\text{priGPA}}\,\textbf{priGPA}_i + \epsilon_i. \tag{2}$$

a) Based on model (2), test on a 95% confidence level the hypothesis that an increase in the attendance rate of 10% increases the performance by 0.2.

b) Test both models (1) and (2) for statistical significance.

c) You surmise that $2\beta_{\text{attend}} = \beta_{\text{ACT}}$ in model (2). Test this hypothesis.[2] Interpret.

d) What attendance rate would you predict for a first-year student with **termgpa** of 3.5, given her average performance in previous courses (**priGPA**) was one grade below average, but her **ACT** score was average?

e) Based on predictions from model (2), compare the performance of an individual with attendance rate of 90% and **priGPA** of 2.5 with the performance of an individual with 10% lower attendance rate, but a **priGPA** of 3.5, given **ACT** is at the mean for both individuals.

f) Compute the SST, SSR, SSE and R^2 for both models (1) and (2). Interpret.

g) Which of the models (1) and (2) would you choose based on the adjusted R^2? Why should you not use the R^2 to choose the model?

h) Draw a scatterplot of **termgpa** against **attend**. Add a regression line based on model (1). Also add a regression line based on model (2), given that all other regressors are at the mean.

Compare your results with the homework answers which will be available on Blackboard on Friday.

[2] Use the function `linearHypothesis` of the package `car`. First install the package `car` under 'Packages' in the lower right window, then load the package in the script via the line `library(car)`.

Ekaterina Kazak
University of Manchester

OLS Inference II

An economist is analysing house prices using the data in the Wooldridge dataset `hprice1`[3]. She is relating the logarithm of house prices (this variables is called *lprice* in the R data set) to some characteristics of the house using the following multiple regression model:

$$\log(price) = \beta_0 + \beta_1 bdrms + \beta_2 sqrft + u$$

where $sqrft$ is the size of the house in square feet and $bdrms$ is the number of bedrooms in the house while $price$ is the price of the house in thousands of dollars. Regress this model in R and use the output to test the following hypotheses at the 5% significance level:

a) $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

b) $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 > 0$

c) $H_0 : \beta_1 = 0.05$ vs $H_1 : \beta_1 \neq 0.05$

d) $H_0 : \beta_2 = 0.0005$ vs $H_1 : \beta_2 < 0.0005$

In each case be clear about the interpretation of the null and alternative hypotheses. Also make sure you write down an appropriate rejection rule. How would you use the p-values produced by R to do (i)? Can you use R's p-values to do any of the other tests?
Compare your results with the homework answers which will be available on Blackboard on Friday.

[3] in order to use the `library(wooldridge)`, make sure you have installed the package `install.packages("wooldridge")`. The data can be uploaded with `data("hprice1")`. Use `help(hprice1)` to get data description.