CBS COPENHAGEN BUSINESS SCHOOL

ECONOMETRICS OF FIRM DATA

Introduction to exercises

Mette Suder Franck, msf.eco@cbs.dk
DVIP, Department of Economics

Agenda

- General introduction to exercise classes
- Econometric mindset and way of thinking
- Recap of key concepts from Ralfs' lectures that will be useful today
  - Interpretation of coefficient estimates
  - Inference and hypothesis testing
  - Goodness-of-fit
- Exercises!

General introduction

Teaching format & practicalities

- We have three exercise classes with more exercises to cover than time allows, therefore…
  - **Work with exercises ahead of class**, so you get an impression of which exercises you would like me to prioritize covering during class
  - **You are encouraged to present solutions** alone or in groups during class as a learning exercise
  - If no requests for specific exercises, I will choose what to cover
- Solutions for all exercises will be available after class including code for both STATA and R
  - In class, I will focus on intuition and switch between showing code in R and STATA
- For questions related to exercise classes reach out to me: msf.eco@cbs.dk
- For questions related to exam or lectures: Utilize Ralf’s office hours

Econometrics

- Used to **understand** relationships
- Emphasis on:
  - Hypothesis testing & model validation
  - Interpretation & significance of coefficients
- Use case: **Understand why customers default to inform policy-making**

Machine learning

- Used to **predict** behaviour/outcomes
- Emphasis on:
  - Predictive performance, accuracy & generalization
  - Hyperparameter tuning
- Use case: **Predict which customers will default to allow for proactive outreach**

Note: Illustrative and non-exhaustive

Interpretation of coefficient estimates

Units of measurement and transformation matters for interpretation

Model	*Dependent Variable*	*Independent Variable*	*Interpretation of* β_1
Level-level	*y*	*x*	$\Delta y = \beta_1 \Delta x$
Level-log	*y*	log(*x*)	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	log(*y*)	*x*	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	log(*y*)	log(*x*)	$\%\Delta y = \beta_1 \%\Delta x$

Source: Ralf's slide

Inference and hypothesis testing

What does our coefficient estimates indicate about statistical significance of the covariates?

- Hypothesis testing:
  - one sided alternatives $H_0 : \beta_j \leq 0 \text{ vs. } H_1 : \beta_j > 0$
  - two sided alternatives $H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$

    $H_0 : \beta_j = a_j \text{ vs. } H_1 : \beta_j \neq a_j$

  - testing for linear restrictions $H_0 : \beta_1 = \beta_2$

$$H_0 : \beta_i = \beta_j = \beta_k = 0 \text{ vs. } H_1 : H_0 \text{ is not true}$$

Here focus of **statistical significance**, but in most applications we are **also interested in practical or economic significance of estimates**: Is the estimated coefficient size small or large? What does it imply for real-life significance?

Source: Ralf's slide OLS_estimation_inference.

Note: There are many other types of hypotheses to test for a specific regression model. What tests are relevant depends on your research question at hand. There is a lot of useful information in Ralf's slides! Slide 60 in particular has some good practical advice on how to interpret coefficients.

Goodness-of-fit

R^2 and how to use it

$$R^2\text{=SSE/SST=1-SSR/SST}$$

- Which is between *0* and *1*.
- It increases if another independent variable is added to the model.
- It decreases if one regressor is removed from the model.
  - The intuition behind this is that is that the residuals will not decrease when we remove an independent variable.
  - This property makes the R^2 a poor tool to decide which set of regressors is "optimal".
  - Instead a variable should be included if there is a nonzero partial effect on *y* in the population.
- One exception: If we simply wish to forecast *y*, a high R^2 is desirable.
- A low R^2 does not mean that the regression is useless.

Total sum of squares (SST): $\text{SST}=\sum_i (y_i - \bar{y})^2$

Explained sum of squares (SSE): $\text{SSE}=\sum_i (\hat{y}_i - \bar{y})^2$

Residual sum of squares (SSR):

$$\text{SSR}=\sum_i \hat{u}_i^2$$

With SST=SSE+SSR

Source: Ralf's slide OLS_estimation_inference

Time for exercises!

Any requests for specific exercises?

PS 1 – Part 2, Exercise 3 a)

$$\log(salary) = 4.62 + .162\log(sales) + .107\log(mktval)$$

$$n = 177,\ \ R^2 = .299.$$

PS 1 – Part 2, Exercise 3 b)

$$\log(salary) = 4.69 + .161\log(sales) + .098\log(mktval) + .000036\,profits$$

$$n = 177, R^2 = .299.$$

PS 1 – Part 2, Exercise 3 c)

$$\log(salary) = 4.56 + .162\log(sales) + .102\log(mktval) + .000029\, profits + .012\, ceoten$$

$$n = 177, \quad R^2 = .318.$$

PS 1 – Part 2, Exercise 4 a)

$$\log(bwght) = 8.06 - .0032\ cigs + .0056\ npvis$$
$$n = 1{,}656, \quad R^2 = .0159$$

PS 1 – Part 2, Exercise 4 c)

A)

$$\log(bwght) = 8.06 - .0032\, cigs + .0056\, npvis$$
$$n = 1{,}656, \quad R^2 = .0159$$

C)

$$\log(bwght) = 8.12 - .0034\, cigs$$
$$n = 1{,}722, \quad R^2 = .0053$$

Problem set 2, 2b)

General idea behind White test for heteroskedasticity

White test

1. Estimate the model by OLS

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

2. Obtain the squared residuals, e_i^2
3. Estimate

$$e^2 = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \delta_3 X_1^2 + \delta_4 X_2^2 + \delta_5 X_1 X_2 + v$$

4. Do the whole model F-test, rejection indicates heteroskedasticity $H_0 : \delta_1 = \delta_2 = ... = \delta_5 = 0$

$$H_1 : not\ H_0$$

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48, pp. 817 - 838.

In other words the null hypothesis is that errors are homoskedastic

Problem set 2, 3b)

Testing Multiple Linear Restrictions

- For example we like to test whether a set of variables has no partial effect on a dependent variable:
  - Testing Exclusion Restrictions
  - In this case, we test for example if there are at least three independent variables:

$$H_0 : \beta_i = \beta_j = \beta_k = 0 \text{ vs. } H_1 : H_0 \text{ is not true}$$

Source: Ralf's slide OLS_estimation_inference

Problem set 2, 3b)

Testing Multiple Linear Restrictions

- Example: Baseball players' salaries (MLB1.dta)

$$\begin{aligned} log(salary) &= \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg \\ &+ \beta_4 hrunsyr + \beta_5 rbisyr + u \end{aligned}$$

with:

years: years in the league

gamesyr: average games played

bavg: career batting average

hrunsyr: home runs per year

rbisyr: runs batted in per year

  - We test: $H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \text{ vs. } H_1 : H_0 \text{ is not true}$

  - In this case we cannot test for "individual" significance of each variable. This could give us a misleading result.

Source: Ralf's slide OLS_estimation_inference

Problem set 2, 3b)

Testing Multiple Linear Restrictions

- <u>Example</u>: (cont.)
  - We estimate the baseball players' wage equation:

$$\begin{aligned} \widehat{log(salary)} &= 11.19 + 0.0689 years + 0.0126 gamesyr + 0.0098 bavg \\ &\quad (0.29) \quad (0.0121) \qquad (0.0026) \qquad\qquad (0.00110) \\ &\quad + 0.144 hrunsyr + 0.0108 rbsiyr \\ &\quad (0.0161) \qquad\quad (0.0072) \\ &\quad n = 353, \quad SSR = 183.186, \quad R^2 = 0.6278 \end{aligned}$$

  - None of the three variables in the *unrestricted* model has a statistically significant *t* statistic against a two sided alternative at the 5% level.
  - Now, we exclude the three variable and estimate the *restricted* model again:

$$\begin{aligned} \widehat{log(salary)} &= 11.22 + 0.0713 years + 0.0202 gamesyr \\ &\quad (0.11) \quad (0.0125) \qquad (0.0013) \\ &\quad n = 353, \quad SSR = 198.11, \quad R^2 = 0.5971 \end{aligned}$$

KAN-CEAPV2605U - 2025 70

CBS

Source: Ralf's slide OLS_estimation_inference

Problem set 2, 3b)

Testing Multiple Linear Restrictions

- We use a test statistic, which measures the relative increase in the SSR by imposing the exclusion restrictions:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

where *r*: restricted model and *ur*: unrestricted model.

- *F* is always nonnegative. Why?

- The numerator and denominator are divided by their degrees of freedom: $df_{ur} = n - k - 1, \; df_r = n - k - 1 + q$

  and therefore $df_r - df_{ur} = q$.
- The denominator is the unbiased estimator for σ^2 in the unrestricted model.
- The *F* statistic can be easily computed in an application.

Source: Ralf's slide OLS_estimation_inference

Problem set 2, 3c)

RESET test for model specification

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

- How do we know whether we have assumed the correct functional form?
  - For example: have we included all relevant quadratics and interaction terms?
- By noting that y^2 and y^3 are highly nonlinear functions of all regressors and their interactions, we could use the fitted values of the model above to compute $\hat{y}^2$ and $\hat{y}^3$.
- Then we estimate

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + u$$

and perform an F-test for joint significance of $\hat{y}^2$ and $\hat{y}^3$:

$$H_0 : \delta_1 = \delta_2 = 0$$

CBS

Source: Ralf's slide OLS_estimation_inference