LECTURE #4

Econometrics I

MULTIPLE REGRESSION MODEL

Jiri Kukacka, Ph.D.
Institute of Economic Studies, Faculty of Social Sciences, Charles University

Summer semester 2024, March 12

In the previous lecture #3

- We derived basic OLS algebraic properties: $\sum_{i=1}^{n} \hat{u}_i = 0$, etc.
- We defined a **goodness-of-fit** measure: $R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$.
- We listed **four SLR assumptions** $\Rightarrow \mathbb{E}(\hat{\beta}_1) = \beta_1$.
- We added **SLR.5 Homoskedasticity:** $\text{Var}(u|x) = \sigma^2 \Rightarrow$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

- We finally estimated $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2$.
- **Standard error of** $\hat{\beta}_1$ is then $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$.
- Readings for lecture #4:
  - Chapter 3: 3.1–3.4, 3.6 (**3.1, esp. pg. 70**, and **3.4, sections 'Multicollinearity' and 'Misspecified models' mandatory**)

Home assignment #1

- Assigned on Thursday via SIS.
- Teams of two, one report.
- Matching spreadsheet.
- Delivered electronically in the .pdf format [5 MB max, .R or other formats can be attached in .zip] via the **Study group roster** app (Lecture JEB109) in SIS.
- Deadline: Thursday, March 28, 2024, 23:59:59.
- 'Academic integrity'; solo $\Rightarrow$ 0.
- IES guide: AI tools when studying at IES.

Outline

Outline

Outline

Model with k independent variables

- General **multivariate linear regression model** or **multiple regression model** can be written in population terms as

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

- General concepts of simple regression translate to multiple regression.
- In addition, the interpretation of $\beta_1, \ldots, \beta_k$ assumes the 'other factors fixed' concept to hold.
- Key assumption

$$\mathbb{E}(u|x_1, \ldots, x_k) = 0$$

  requires that
  - primarily, all other factors in u are are not related, on average, to (all combinations of) the explanatory variables,
  - it also means that we have selected a correct functional form.

Matrix setting

- We switch to matrix notation here.
- Population equation

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

can be rewritten as

$$Y = X\beta + u,$$

where:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{pmatrix}, u = \begin{pmatrix} u_1 \\ u_2 \\ . \\ . \\ u_n \end{pmatrix}.$$

- Check the dimensions!

Outline

Obtaining the OLS estimates

- Define the vector of residuals as $\hat{u} = Y - X\hat{\beta}$.
- Using the basic matrix algebra, and specifically, selected properties of matrix differentiation

M	$\frac{\partial M}{\partial \beta}$
$A\beta$	A^T
$\beta^T A$	A
$\beta^T \beta$	2β
$\beta^T A \beta$	$A\beta + A^T\beta$

where M, A are matrices and β is a vector, we finally get

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

- $\hat{\beta}$ is a vector of estimates $\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_k$.

Getting the OLS estimator

- Defining the residuals as $\hat{u} = Y - X\hat{\beta}$, we need to get an alternative to SSR.
- In matrix form, as $\hat{u}$ is a column vector, it is easy to see that we need $\hat{u}^T\hat{u}$:

$$\begin{aligned} \hat{u}^T\hat{u} &= (Y^T - \hat{\beta}^T X^T)(Y - X\hat{\beta}) = \\ &= Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}. \end{aligned}$$

- Then, the OLS estimator $\hat{\beta}$ is obtained via

$$\frac{\partial SSR}{\partial \hat{\beta}} = \frac{\partial \hat{u}^T\hat{u}}{\partial \hat{\beta}} = 0 \Longleftrightarrow 0 - X^T Y - X^T Y + X^T X\hat{\beta} + X^T X\hat{\beta} = 0 \Longleftrightarrow$$

$$2X^T X\hat{\beta} = 2X^T Y \Longleftrightarrow (X^T X)^{-1} X^T X\hat{\beta} = (X^T X)^{-1} X^T Y \Longleftrightarrow$$

$$\boxed{\hat{\beta}^{OLS} = (X^T X)^{-1} X^T Y}.$$

Interpreting the OLS regression equation

- Interpretation of the intercept $\hat{\beta}_0$ remains the same: the predicted value of y when $x_1 = \ldots = x_k = 0$.
- Estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$ have the **partial effect**, or the **ceteris paribus**, interpretation.
- From the OLS regression 'line', we have

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \ldots + \hat{\beta}_k \Delta x_k.$$

- This gives us an interpretation of

$$\hat{\beta}_j = \frac{\Delta\hat{y}}{\Delta x_j}$$

holding all other $x_{\neq j}$ **fixed**, i.e., after **controlling for** all variables $x_{\neq j}$ when estimating the effect of x_j on y.

- Ceteris paribus interpretation even for non-experimental data!

Example

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
log(*sales*)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*)	——	.112 (.050)	.100 (.049)
profmarg	——	−.0023 (.0022)	−.0022 (.0021)
ceoten	——	——	.0171 (.0055)
comten	——	——	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 (0.25)	4.57 (0.25)
Observations	177	177	177
R-squared	.281	.304	.353

Source: Wooldridge (2012)

Outline

Outline

Unbiasedness of OLS

Multiple linear regression (MLR) assumptions:

- **MLR.1 Linear in parameters:** We have the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

  where β_0 is the population intercept and $\beta_1, \ldots, \beta_k$ are the population slope parameters. The inclusion of β_0 implies $\mathbb{E}(u) = 0$.

- **MLR.2 Random sampling:** We have a random sample of size n following the population model.

- **MLR.3 No perfect collinearity:** In the sample and the population, none of the independent variables is constant, and there are no **exact linear** relationships among the independent variables. Mathematically, the matrix X must have full column rank.

- **MLR.4 Zero conditional mean:** The error u has an expected value of zero given any values of the independent variables, i.e., $\mathbb{E}(u|x_1, x_2, \ldots, x_k) = 0$.
  - it covers various misspecifications
  - endogenous vs. exogenous independent variables
  - it implies $Cov(x_j, u) = 0,\ j = 1, \ldots k$

Unbiasedness of the OLS estimators

Assuming MLR.1 through MLR.4, $\mathbb{E}(\hat{\beta}_j^{OLS}) = \beta_j, j = 0, 1, \ldots, k$. In other words, the **OLS estimators are unbiased** estimators of the population parameters.

Unbiasedness of the OLS estimator $\hat{\beta}$: Proof

- We use the same 'trick' as for the univariate case, i.e., we substitute for Y:

$$\boxed{\hat{\beta}} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + u) =$$

$$= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T u = \boxed{\beta + (X^T X)^{-1} X^T u}.$$

- We now need to show its expected value

$$\boxed{\mathbb{E}(\hat{\beta})} = \mathbb{E}(\beta + (X^T X)^{-1} X^T u) = \beta + (X^T X)^{-1} X^T \mathbb{E}(u) = \boxed{\beta}.$$

- OLS estimator $\hat{\beta}$ is thus unbiased.

Outline

Model misspecifications

- Including **irrelevant variables** (model **overspecification**) has a negligible effect on the estimates as these variables have no partial effect on y, and the OLS estimator remains **unbiased**.
- **Omitting** a relevant variable (model **underspecification**) causes troubles!
- Consider the following population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

  and assume MLR.1 through MLR.4 hold.
- We underspecify the model and estimate only parameters of

$$y = \beta_0 + \beta_1 x_1 + u.$$

- OLS estimator is then typically **biased**.

Bias in misspecified models

Again, we substitute for y_i. Note that the OLS estimator $\hat{\beta}_1$ is now based on a single independent variable, while the true population model has two:

$$\boxed{\hat{\beta}_1} = \frac{\sum y_i(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} = \frac{\sum (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i)(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} =$$

$$= \beta_0 \frac{\sum (x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} + \beta_1 \frac{\sum x_{1i}(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} + \beta_2 \frac{\sum x_{2i}(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} + \frac{\sum u_i(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} =$$

$$= 0 + \boxed{\beta_1 + \beta_2 \frac{\sum x_{2i}(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} + \frac{\sum u_i(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2}}.$$

$$\boxed{\mathbb{E}(\hat{\beta}_1)} = \beta_1 + \beta_2 \frac{\sum x_{2i}(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} + \frac{\sum \overbrace{\mathbb{E}(u_i)}^{=0 \ (MLR.1)} (x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} =$$

$$= \beta_1 + \beta_2 \frac{\sum x_{2i}(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} = \boxed{\beta_1 + \beta_2 \frac{\sum (x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2}}.$$

Omitted variable bias

In the simple case of estimating the simple regression when, in fact, we should have estimated the multiple regression with two independent variables, the following table summarizes the implied bias:

β_2	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias (up)	Negative bias (down)
$\beta_2 < 0$	Negative bias	Positive bias

Example again

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
log(*sales*)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*)	——	.112 (.050)	.100 (.049)
profmarg	——	−.0023 (.0022)	−.0022 (.0021)
ceoten	——	——	.0171 (.0055)
comten	——	——	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 (0.25)	4.57 (0.25)
Observations	177	177	177
R-squared	.281	.304	.353

Outline

Outline

Variance of the OLS estimators

Additional assumption:

- **MLR.5 Homoskedasticity:** The error u has the same variance given any values of the independent variables, i.e.,

$$\text{Var}(u|x_1, \ldots, x_k) = \sigma^2 \mathbb{I}.$$

- Under MLR.1 through MLR.5, conditional on the sample values of the independent variables,

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

for $j = 1, 2, \ldots, k$, where $SST_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j, and R_j^2 is the R^2 from regressing x_j on all other independent variables (and intercept).
In matrix form, it can be written as

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

Variance of the OLS estimator: Derivation

- We have $\hat{\beta} = \beta + (X^T X)^{-1} X^T u$.
- We also know that
  - variance of a constant is zero,
  - $\mathbb{E}(u) = 0$,
  - $\hat{\beta}$ is unbiased for β, i.e., $\hat{\beta} - \mathbb{E}(\hat{\beta}) = (X^T X)^{-1} X^T u$,
  - assumption MLR.3 implies a regular matrix $X^T X$.
- Now, we use the definition of variance and solve

$$\begin{aligned}
\boxed{\mathrm{Var}(\hat{\beta})} &= \mathbb{E}[(\hat{\beta} - \mathbb{E}(\hat{\beta}))(\hat{\beta} - \mathbb{E}(\hat{\beta}))^T] = \\
&= \mathbb{E}[(X^T X)^{-1} X^T u u^T X (X^T X)^{-1}] = \\
&= (X^T X)^{-1} X^T \mathbb{E}(u u^T) X (X^T X)^{-1} = \\
&= (X^T X)^{-1} X^T \sigma^2 \mathbb{I} X (X^T X)^{-1} = \boxed{\sigma^2 (X^T X)^{-1}}.
\end{aligned}$$

Estimating the error variance

- As for the simple regression model, σ^2 is not observed and needs to be estimated from data.
- Under the **Gauss-Markov assumptions** for cross-sectional regression, MLR.1 through MLR.5, $\hat{\sigma}^2$ **is an unbiased estimator of** σ^2 and it is similarly given as

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2.$$

- $n-k-1$ because we lose $k+1$ degrees of freedom due to $k+1$ restrictions on residuals:

$$\sum_{i=1}^{n}\hat{u}_i = 0,$$

$$\sum_{i=1}^{n}x_{ij}\hat{u}_i = 0, \quad j = 1,\ldots,k.$$

Estimating the error variance

- $\hat{\sigma}$ is still called the **standard error of the regression**.
- **Standard error of** $\hat{\beta}_j$ is then

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1 - R_j^2)}}.$$

In matrix form, it can be written as

$$se(\hat{\beta}_j) = \hat{\sigma}\sqrt{(X^T X)^{-1}_{j+1,j+1}}.$$

Outline

Variance of OLS in misspecified models

- Bias vs. variance trade-off
- Remember that
  - OLS is **unbiased** for an **overspecified** model,
  - OLS is usually **biased** for an **underspecified** model.
- What happens to variance if the model is over/underspecified?

  [Hint: Use the definition of R^2.]

Variance of OLS in misspecified models

- Recall that we have

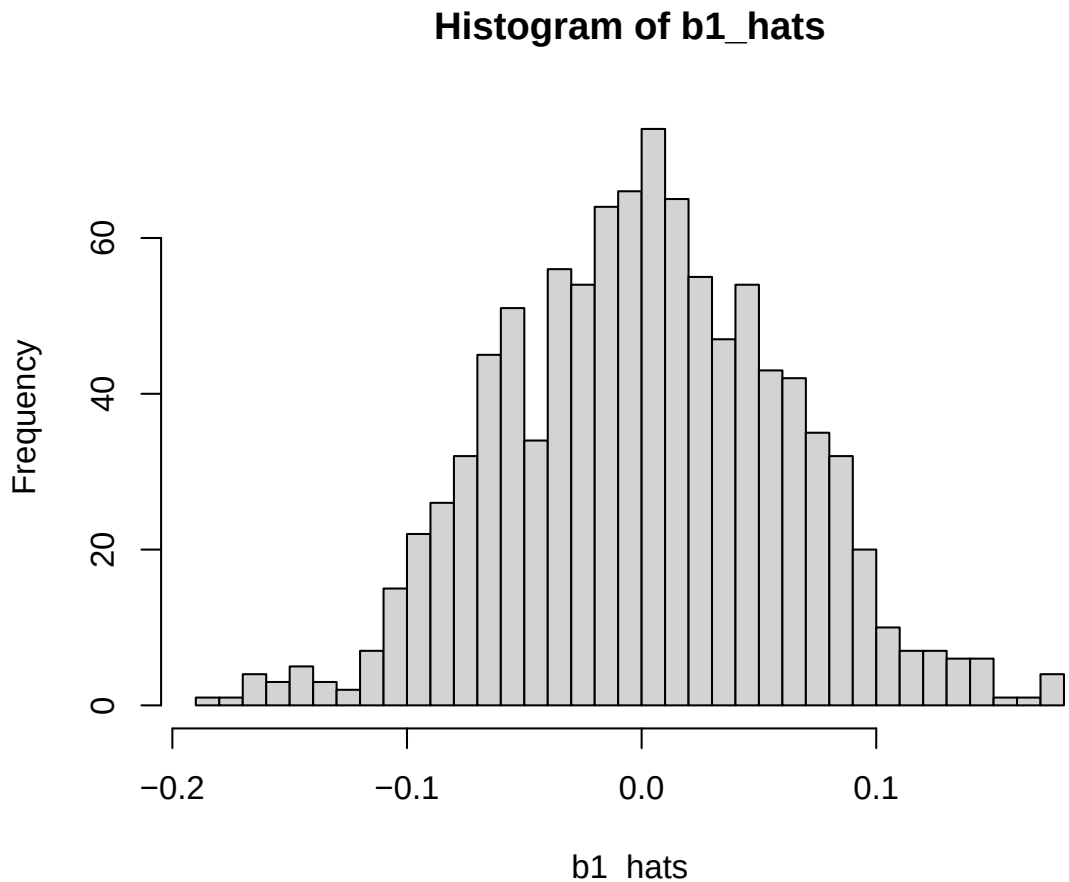
$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)} = \frac{\sigma^2}{SST_j} \cdot VIF_j$$

  and we assume σ^2 and SST_j do not change.
- Overspecified model, i.e., too many independent variables included in the model: $R_j^2 \nearrow \Rightarrow\ (1 - R_j^2) \searrow \Rightarrow\ \text{Var}(\hat{\beta}_j) \nearrow$.
- Underspecified model, i.e., a relevant independent variable is missing: $R_j^2 \searrow \Rightarrow\ (1 - R_j^2) \nearrow \Rightarrow\ \text{Var}(\hat{\beta}_j) \searrow$.

- Consult seminar #4 handout and **3.4, section 'Multicollinearity' (mandatory readings)** for details about the issue of multicollinearity and the Variance Inflation Factor (VIF_j) for β_j.

Simulation of an OLS bias

- Consider the population model $y = 1 + 1x_1 - 2x_2 + u$.
- $\text{Corr}(x_1, x_2) = 0.5$
- Estimate the underspecified model $y = \beta_0 + \beta_1 x_1 + u$.
- Repeat the simulation $1,000\times$ for $n = 1,000$.



Seminars and the next lecture

- Seminars:
  - multiple regression model: interpretation and prediction, R^2
  - bias and variance of a composite estimator
  - analysis of the expected omitted variable bias (TAKE HOME)
  - computer exercise: log-log, log-level, and quadratic functional forms (**3.1 mandatory readings**)
  - multicollinearity + VIF (**3.4 mandatory readings**)
- Next lecture #5:
  - efficiency and the sampling distribution of the OLS estimator: the Gauss-Markov theorem + normality
  - testing hypotheses
    - about a single population parameter: t test, confidence intervals, p-value
    - about a single linear combination of the parameters
- Readings for lecture #5:
  - Chapter 3: 3.5, Chapter 4: 4.1–4.4