# LECTURE #12

# Econometrics I

# REVISION OF KEY CONCEPTS

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

Summer semester 2024, May 21

# In the previous lecture #11

- We discussed the **functional form misspecification** $\Rightarrow$ MLR.4 assumption violated, OLS biased and inconsistent.
- We introduced the **RESET test**:

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

- **Proxy variables** were suggested as a remedy for OVB.
- Properties of OLS under **measurement error** were studied
  - in the dependent vs. in the independent variable.
  - **CEV** assumption: $\text{Cov}(x_1^*, e_1) = 0$ or $\text{Cov}(y_1^*, e_0) = 0$.
- Potential **violations of random sampling** (MLR.2) were briefly discussed: missing data, nonrandom samples, outliers.
- Readings for lecture #12:
  - your favorite book :-) or selected chapters/sections from Wooldridge (2012)

# Evaluation: A kind request

Please do not forget to fill in the electronic evaluation of our course Econometrics I (JEB109).

**No seminars** this week, the **first exam term** next week.

# Outline

Population models and OLS estimators

Unbiasedness, consistency, and variance of OLS

Hypothesis testing

Goodness-of-fit measures

Selection of explanatory variables

Heteroskedasticity

Functional form misspecification

Predictions

Qualitative variables

# Outline

Population models and OLS estimators

Unbiasedness, consistency, and variance of OLS

Hypothesis testing

Goodness-of-fit measures

Selection of explanatory variables

Heteroskedasticity

Functional form misspecification

Predictions

Qualitative variables

# Population model and related 'lines and functions'

- **Population model** of a **dependent variable** $y$ as a function of $k$ **independent variables** $x_j$ is given as

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

  with the intercept parameter $\beta_0$, slope parameters $\beta_j$, $j = 1, \ldots, k$, and the error term $u$ with $\mathbb{E}(u) = 0$.

- **Population regression function (PRF)** is given as

$$\mathbb{E}(y|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

- **OLS regression line** or the **sample regression function (SRF)** is given as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k.$$

- **Residual** is defined as

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_k x_{ki}.$$

- Note the essential differences:
  - parameter vs. estimator vs. estimate.
  - observations vs. expected values vs. fitted values.
  - PRF is fixed for the population but unknown.
  - in general, the PRF and SRF differ.
  - and for each sample of data, the SRF (OLS regression line) differs as well.

# OLS estimators

- For a **simple linear regression model**, the OLS estimator of $\beta_1$ is given as

$$\hat{\beta}_1^{OLS} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- For a **multiple linear regression model**, the OLS estimator of vector of $\beta_j$, $j = 0, 1, \ldots, k$, is given as

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

# Various 'sums of squares'

- Total sum of squares (SST)

$$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Explained sum of squares (SSE)

$$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Residual sum of squares (SSR)

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2$$

- It holds that

$$SST = SSE + SSR.$$

# Interpretation of the OLS regression equation

- Interpretation of the estimated intercept $\hat{\beta}_0$: 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 **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 $y$.
- From the perspective of economics, various logarithmic specifications are useful:

| Model | Dependent v. | Independent v. | Interpretation of $\beta_1$ |
|---|---|---|---|
| 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 | $\log(y)$ | $\log(x)$ | $\%\Delta y = \beta_1 \%\Delta x$ |

# Outline

# Multiple linear regression (MLR) assumptions (CLM)

- **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 $\beta_0$ is the population intercept and $\beta_1, \ldots, \beta_k$ are the population slope parameters. The inclusion of $\beta_0$ 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., $\mathbb{E}(u|x_1, x_2, \ldots, x_k) = 0$.
- **MLR.5 Homoskedasticity:** The error $u$ has the same variance given any values of the independent variables, i.e., $\text{Var}(u|x_1, \ldots, x_k) = \sigma^2 \mathbb{I}$.
- **MLR.6 Normality:** The population error $u$ is **independent** of the explanatory variables $x_1, \ldots, x_k$ and is **normally** distributed with zero mean and variance $\sigma^2$, i.e., $u \sim N(0, \sigma^2)$.

# Unbiasedness and consistency of OLS

- Assuming MLR.1 through MLR.4, $\mathbb{E}(\hat{\beta}_j) = \beta_j, j = 0, 1, \ldots, k$. In other words, the OLS estimators are **unbiased** estimators of the population parameters.
- Assuming MLR.1 through MLR.4, the OLS estimators are **consistent** estimators of the population parameters.
- In fact, only a weaker version of MLR.4 (**MLR.4′ Zero mean and zero correlation**, instead of mean independence) is sufficient for **consistency** of OLS.

# Variance of the OLS estimators

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

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

# Estimating the error variance

- Under MLR.1 through MLR.5, **the unbiased estimator of $\sigma^2$** is given as

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

- $\hat{\sigma}$ is 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

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

# Gauss-Markov theorem (BLUE) and BUE

- Under MLR.1 through MLR.5, the OLS estimator is the **best linear unbiased estimator (BLUE)**.
- Under MLR.1 through MLR.6, the OLS estimator is the **best unbiased estimator (BUE)**.

# Asymptotic normality of OLS

Under the Gauss-Markov assumptions MLR.1 through MLR.5:

- $\sqrt{n}(\hat{\beta}_j - \beta_j) \overset{a}{\sim} N(0, \textit{asymptotic}\ \text{Var}_j)$, i.e., $\hat{\beta}_j$ is **asymptotically normally distributed**.
- $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2 = \text{Var}(u)$.
- For each $j$,

$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \overset{a}{\sim} N(0,1)$$

  and

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \overset{a}{\sim} N(0,1).$$

# Outline

# $t$ distribution for the standardized estimators

- Under the CLM assumptions MLR.1 through MLR.6,
$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1} = t_{df}, \tag{1}$$
where $k+1$ is the number of unknown parameters in the population model (including the intercept), and $n-k-1$ is the *df*.
- Under $\quad H_0 : \beta_j = a_j, \quad$ equation (1) gives us the $t$ **statistic**
$$\boxed{t_{\hat{\beta}_j} \equiv \frac{\hat{\beta}_j - a_j}{se(\hat{\beta}_j)}}.$$
- $t$ **ratio** $\quad t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad$ identifies the statistically significant independent variables, i.e., the ones whose partial effect is statistically significantly different from zero. It is most commonly used for a **two-tailed** $t$ test under $H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$.
- For testing hypotheses about a single linear combination of parameters (e.g., $H_0 : \beta_1 = \beta_2$), we can still use the $t$ statistic, but we need to rewrite the model with the null hypothesis in mind.

# Confidence interval

- Under the CLM assumptions MLR.1 through MLR.6, we can easily construct a **confidence interval (CI)** for the population parameter $\beta_j$.
- Using the distribution of $\hat{\beta}_j$: $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$, we compute a $1-\alpha$ **confidence interval** as

$$\hat{\beta}_j \pm t_{n-k-1,1-\alpha/2} se(\hat{\beta}_j).$$

- For $n-k-1 > 100$, the 'rule of 2 (sigma)' for $\alpha = 5\%$ can be again used for a rough idea.

# Testing joint hypotheses

$F$ test:

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k-1)} = \frac{(R_U^2 - R_R^2)/q}{(1-R_U^2)/(n-k-1)} \sim F_{q,n-k-1},$$

v where $n-k-1$ is the degrees of freedom of the original unrestricted model, and $q$ is the number of restrictions.

$LM$ test:

1. $H_0$ and $H_1$ are the same as for the respective $F$ test, e.g.:

$$H_0 : \beta_1 = 0, \beta_2 = 0 \quad vs. \quad H_1 : H_0 \text{ does not hold.}$$

2. Estimate the **restricted model** and save the residuals $\tilde{u}$,
3. Run an **auxiliary regression:** regress $\tilde{u}$ on **all independent variables** and obtain $R^2$ of this regression, i.e., $R^2_{\tilde{u}}$ (intuition: if $H_0$ is true. $R^2_{\tilde{u}}$ is 'close' to zero).
4. Compute $\boxed{LM = nR^2_{\tilde{u}}}$.
5. Under the null hypothesis, $LM \overset{a}{\sim} \chi^2_q$.
6. If $LM > c$, we reject $H_0$ at the given significance level $\alpha$.

# Outline

# Goodness-of-fit measures

- **Coefficient of determination** $R^2$ and its adjusted version $\bar{R}^2$:

$$
\begin{aligned}
R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST} &= 1 - \frac{\frac{SSR}{n}}{\frac{SST}{n}}, \\
\bar{R}^2 &= 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}}.
\end{aligned}
$$

- Asymptotically, $\bar{R}^2 = R^2$.
- $R^2$ cannot decrease after adding an independent variable.
- $R^2$ can be used only for the same number of independent variables.
- $\bar{R}^2$ can also be used to compare various specifications (e.g., for the logarithmic vs. quadratic form of the explanatory variable) and controlling for too many explanatory variables.

# Outline

# Four important variable selection criteria

Does an explanatory variable belong to the model?

1. **Theory:** Is including a variable in the equation unambiguous and theoretically sound? Does intuition suggest that it should be included? Also, the modeling purpose is crucial:
   - prediction/explanation
   - vs. testing a specific theoretical/empirical relationship
2. **Omitted variable bias reduction:** Do estimated coefficients of other variables change considerably when the variable is added to the model? It is essential to avoid serious OVB.
3. **Adjusted $\bar{R}^2$:** Does the overall fit of the equation improve (enough) when the variable is added to the model?
4. ***t* test and *F* test:** Is its coefficient statistically significant in the expected direction? *F test* can help us when considering excluding multiple variables or for step-wise elimination.

# Outline

# Heteroskedasticity

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

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

- Violation of homoskedasticity is called **heteroskedasticity**.

**Consequences:**

1. OLS remains unbiased and consistent (under MLR.1–4).
   - estimated coefficients, $R^2$, and $\bar{R}^2$ remain unaffected.
2. True variance of the $\hat{\beta}^{OLS}$ distribution increases.
   - because the heteroskedastic error term explains a larger proportion of fluctuations of the dependent variable.

   $\Rightarrow$ OLS is no longer ~~B~~LUE, even not asymptotically efficient.

# Heteroskedasticity

3. But (!) estimators of $\mathrm{Var}(\hat{\beta}_j)$ are biased, usually down.
   - increase of the (true) variance is, however, 'masked' by OLS because it assumes a homoskedastic error.
   - OLS thus attributes the impact of the heteroskedastic error to the independent variables.

   $\Rightarrow$ **standard errors tend to be smaller** under heteroskedasticity, and statistical inference becomes unreliable and incorrect:

   $\Rightarrow$ $t$ statistics, CIs, $F$ statistics, and $LM$ statistics invalid even for large samples!

- Fortunately, the OLS standard errors can be modified to be asymptotically valid under MLR.1–4, i.e., without MLR.5.

# Heteroskedasticity

- We can use **White robust standard errors**, which are robust to heteroskedasticity of various forms.
- But the White robust standard errors work **only for large samples**, and even then, they are **only asymptotically valid**; no statements are made about bias, consistency, or efficiency.
- Testing for heteroskedasticity:
  - **Breusch-Pagan test:**

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_x + v.$$

  - **White test:**

$$\begin{aligned} \hat{u}^2 &= \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_1^2 + \delta_4 x_2^2 + \delta_5 x_1 x_2 + v, \\ \hat{u}^2 &= \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + v. \end{aligned}$$

# Weighted least squares

- **Weighted least squares (WLS)** estimation is a historically older method of treating heteroskedasticity compared to White standard errors.
- If we have a correctly specified form of heteroskedasticity, WLS is **unbiased** and **more efficient** than OLS, and it leads to $t$-distributed $t$ statistics and $F$-distributed $F$ statistics only under MLR.1 through MLR.4 (it is, in fact, **BLUE**).

# Outline

# Functional form misspecification

- **Functional form misspecification** occurs in a situation when we have selected a proper independent variable(s) but not a correct form of the relationship with the dependent one.
- This violates the MLR.4 assumption, i.e., the OLS procedure is biased and inconsistent:
- There are two popular tests:
  - **Ramsey RESET test** (for actual functional misspecification):

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

  - **Davidson-MacKinnon test** (for selecting between **nonnested models**):

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{\hat{y}} + \varepsilon.$$

# Outline

# Predictions in OLS cross-sectional data framework

- **Predicting** $\hat{y}_{n+1}$ simply means obtaining the fitted value for $x_{1,n+1}, \ldots, x_{k,n+1}$.
- Prediction uncertainty is represented by the **confidence intervals** for predictions and the **prediction intervals**.
- We distinguish between two types:
  - uncertainty about the **mean/average predicted value** of $y$ due to estimation variance (sampling variation).
  - **additional uncertainty** of the prediction for a **specific unit** such as an individual or a firm due to the error variance.

# Outline

# Dummy variables

- **Dummy** independent variables work the same way, from the technical point of view, as the 'standard' quantitative independent variables.
- Beware of the **dummy variable trap**.
- Specific qualitative characteristics can be combined to form new terms.
- **Base group** (the omitted one) is 'hidden' in the intercept.
- Dummy variables allow for **different slopes** (as part of an interaction term) and for **different intercepts** (intercept dummy).
- **Chow test** is frequently used to test the stability/equality of the parameters of the underlying population model for different groups.

# Linear probability model

- In the **linear probability model (LPM)**, the dependent variable $y$ is binary, i.e., only either 1 or 0.
- Under MLR.1–4, the OLS estimator is still unbiased and consistent.
- Importantly, as $y$ has the Bernoulli distribution,

$$\begin{aligned}\boxed{\mathbb{E}(y|X)} &= 1 \cdot \mathrm{P}(y=1|X) + 0 \cdot \big(1 - \mathrm{P}(y=1|X)\big) = \boxed{\mathrm{P}(y=1|X)}, \\ \mathrm{P}(y=1|X) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.\end{aligned}$$

- Previous derivations allow for the interpretation of $\beta$s as

$$\Delta p(X) = \Delta \mathrm{P}(y=1|X) = \beta_j \Delta x_j,$$

i.e., the change in the **probability of 'success'** (probability of $y$ being 1) when $x_j$ changes by one small unit. d

- Shortcomings:
  1. while the observed values are precisely 0 or 1, the estimated/predicted probability is **not bounded by 0 and 1**.
  2. usually **constant marginal effect** $\Delta x_j$ (often unrealistic).
  3. error term is inherently **heteroskedastic**.
  4. error term is **not normally distributed**.