# Linear Regression 1

Instructor: Yuta Toyama

Last updated: 2020-03-30

# Section 1

## Framework

# Regression framework

- Let $Y_i$ be the dependent variable and $X_{ik}$ be k-th explanatory variable.
  - We have $K$ explantory variables (along with constant term)
  - $i$ is an index for observations. $i = 1, \cdots, N$.
  - Data (sample): $\{Y_i, X_{i1}, \ldots, X_{iK}\}_{i=1}^N$
- **Linear regression model** is defined as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

  - $\epsilon_i$: error term (unobserved)
  - $\beta$: coefficients

- **Assumptions for Ordinaly Least Squares (OLS) estimation**
  1. Random sample: $\{Y_i, X_{i1}, \ldots, X_{iK}\}$ is i.i.d. drawn sample
     - i.i.d.: identically and independently distributed
  2. $\epsilon_i$ has zero conditional mean

$$E[\epsilon_i | X_{i1}, \ldots, X_{iK}] = 0$$

  3. Large outliers are unlikely: The random variable $Y_i$ and $X_{ik}$ have finite fourth moments.
  4. No perfect multicollinearity: There is no linear relationship betwen explanatory variables.

- OLS estimators are the minimizers of the sum of squared residuals:

$$\min_{\beta_0,\cdots,\beta_K} \frac{1}{N}\sum_{i=1}^{N}(Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK}))^2$$

- Using matrix notation, we have the following analytical formula for the OLS estimator

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

where

$$\underbrace{X}_{N\times(K+1)} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1K} \\ \vdots & \vdots & & \vdots \\ 1 & X_{N1} & \cdots & X_{NK} \end{pmatrix}, \underbrace{Y}_{N\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \underbrace{\beta}_{(K+1)\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$$

# Theoretical Properties of OLS estimator

- We briefly review theoretical properties of OLS estimator.

1. **Unbiasedness**: Conditional on the explantory variables $X$, the expectation of the OLS estimator $\hat{\beta}$ is equal to the true value $\beta$.

$$E[\hat{\beta}|X] = \beta$$

2. **Consistency**: As the sample size $N$ goes to infinity, the OLS estimator $\hat{\beta}$ converges to $\beta$ in probability

$$\hat{\beta} \xrightarrow{p} \beta$$

3. **Asymptotic normality**: Will talk this later

# Section 2

## Specification

## Interpretation and Specifications of Linear Regression Model

- Remember that

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

- The coefficient $\beta_k$ captures the effect of $X_k$ on $Y$ **ceteris paribus (all things being equal)**
- Equivalently, if $X_k$ is continuous random variable,

$$\frac{\partial Y}{\partial X_k} = \beta_k$$

- If we can estimate $\beta_k$ without bias, can obtain **causal effect** of $X_k$ on $Y$.
  - This is of course very difficult task. We will see this more later.
- Several specifications frequently used in empirical analysis.
  1. Nonlinear term
  2. log specification
  3. dummy (categorical) variables
  4. interaction terms

# Nonlinear term

- We can capture non-linear relationship between $Y$ and $X$ in a linearly additive form
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i$$
- As long as the error term $\epsilon_i$ appreas in a additively linear way, we can estimate the coefficients by OLS.
  - Multicollinarity could be an issue if we have many polynomials (see later).
  - You can use other non-linear variables such as $log(x)$ and $\sqrt{x}$.

# log specification

- We often use `log` variables in both dependent and independent variables.
- Using `log` changes the interpretation of the coefficient $\beta$ in terms of scales.

| Dependent | Explanatory | interpretation |
|---|---|---|
| $Y$ | $X$ | 1 unit increase in $X$ causes $\beta$ units change in Y |
| $\log Y$ | $X$ | 1 unit increase in $X$ causes $100\beta\%$ incchangerease |
| $Y$ | $\log X$ | 1% increase in $X$ causes $\beta/100$ unit change in $Y$ |
| $\log Y$ | $\log X$ | 1% increase in $X$ causes $\beta\%$ change in $Y$ |

# Dummy variable

- A dummy variable takes only 1 or 0. This is used to express qualititative information
- Example: Dummy variable for race

$$white_i = \begin{cases} 1 & if\ \ white \\ 0 & otherwise \end{cases}$$

- The coefficient on a dummy variable captures the difference of the outcome $Y$ between categories
- Consider the linear regression

$$Y_i = \beta_0 + \beta_1 white_i + \epsilon_i$$

The coefficient $\beta_1$ captures the difference of $Y$ between white and non-white people.

# Interaction term

- You can add the interaction of two explanatory variables in the regression model.
- For example:

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 white_i + \beta_3 educ_i \times white_i + \epsilon_i$$

  where $wage_i$ is the earnings of person $i$ and $educ_i$ is the years of schooling for person $i$.
- The effect of $educ_i$ is

$$\frac{\partial wage_i}{\partial educ_i} = \beta_1 + \beta_3 white_i,$$

- This allows for heterogenous effects of education across races.

# Section 3

## Fit

# Measures of Fit

- We often use $R^2$ as a measure of the model fit.
- Denote **the fitted value** as $\hat{y}_i$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_K X_{iK}$$

  - Also called prediction from the OLS regression.
- $R^2$ is defined as

$$R^2 = \frac{SSE}{TSS},$$

  where

$$SSE = \sum_i (\hat{y}_i - \bar{y})^2, \ TSS = \sum_i (y_i - \bar{y})^2$$

- $R^2$ captures the fraction of the variation of $Y$ explained by the regression model.
- Adding variables always (weakly) increases $R^2$.

- In a regression model with multiple explanatory variables, we often use **adjusted** $R^2$ that adjusts the number of explanatory variables

$$\bar{R}^2 = 1 - \frac{N-1}{N-(K+1)}\frac{SSR}{TSS}$$

where

$$SSR = \sum_i (\hat{y}_i - y_i)^2 (= \sum_i \hat{u}_i^2),$$

# Section 4

## Inference

# Statistical Inference

- Notice that the OLS estimators are **random variables**. They depend on the data, which are random variables drawn from some population distribution.
- We can conduct statistical inferences regarding those OLS estimators: 1. Hypothesis testing 2. Constructing confidence interval
- I first explain the sampling distribution of the OLS estimators.

# Distribution of the OLS estimators based on asymptotic theory

- Deriving the exact (finite-sample) distribution of the OLS estimators is very hard.
  - The OLS estimators depend on the data $Y_i, X_i$ in a complex way.
  - We typically do not know the distribution of $Y$ and $X$.
- We rely on **asymptotic** argument. We approximate the sampling distribution of the OLS esimator based on the cental limit theorem.

- Under the OLS assumption, the OLS estimator has **asymptotic normality**

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V)$$

where

$$\underbrace{V}_{(K+1)\times(K+1)} = E[\mathbf{x}_i'\mathbf{x}_i]^{-1} E[\mathbf{x}_i'\mathbf{x}_i\epsilon_i^2] E[\mathbf{x}_i'\mathbf{x}_i]^{-1}$$

and

$$\underbrace{\mathbf{x}_i}_{(K+1)\times 1} = (1, X_{i1}, \cdots, X_{iK})'$$

- We can **approximate** the distribution of $\hat{\beta}$ by

$$\hat{\beta} \sim N(\beta, V/N)$$

- The above is joint distribution. Let $V_{ij}$ be the $(i, j)$ element of the matrix $V$.
- The individual coefficient $\beta_k$ follows

$$\hat{\beta}_k \sim N(\beta_k, V_{kk}/N)$$

# Estimation of Asymptotic Variance

- $V$ is an unknown object. Need to be estimated.
- Consider the estimator $\hat{V}$ for $V$ using sample analogues

$$\hat{V} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\hat{\epsilon}_i^2\right)\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}$$

  where $\hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \cdots + \hat{\beta}_K X_{iK})$ is the residual.
- Technically speaking, $\hat{V}$ converges to $V$ in probability. (Proof is out of the scope of this course)
- We often use the (asymptotic) **standard error** $SE(\hat{\beta}_k) = \sqrt{\hat{V}_{kk}/N}$.
- The standard error is an estimator for the standard deviation of the OLS estimator $\hat{\beta}_k$.

# Section 5

## Testing

# Hypothesis testing

- OLS estimator is the random variable.
- You might want to test a particular hypothesis regarding those coefficients.
  - Does x really affects y?
  - Is the production technology the constant returns to scale?
- Here I explain how to conduct hypothesis testing.

# 3 Steps in Hypothesis Testing

- Step 1: Consider the null hypothesis $H_0$ and the alternative hypothesis $H_1$

$$H_0 : \beta_1 = k, H_1 : \beta_1 \neq k$$

where $k$ is the known number you set by yourself.

- Step 2: Define **t-statistic** by

$$t_n = \frac{\hat{\beta}_1 - k}{SE(\hat{\beta}_1)}$$

- Step 3: We reject $H_0$ is at $\alpha$-percent significance level if

$$|t_n| > C_{\alpha/2}$$

where $C_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal distribution.
  - We say we **fail to reject** $H_0$ if the above does not hold.

## Caveats on Hypothesis Testing

- We often say $\hat{\beta}$ is **statistically significant** at 5% level if $|t_n| > 1.96$ when we set $k = 0$.
- Arguing the statistical significance alone is not enough for argument in empirical analysis.
- Magnitude of the coefficient is also important.
- Case 1: Small but statistically significant coefficient.
  - As the sample size $N$ gets large, the $SE$ decreases.
- Case 2: Large but statistically insignificant coefficient.
  - The variable might have an important (economically meaningful) effect.
  - But you may not be able to estimate the effect precisely with the sample at your hand.

# F test

- We often test a composite hypothesis that involves multiple parameters such as

$$H_0 : \beta_1 + \beta_2 = 0,\ \ H_1 : \beta_1 + \beta_2 \neq 0$$

- We use **F test** in such a case (to be added).

# Confidence interval

- 95% confidence interval

$$CI_n = \left\{ k : \left| \frac{\hat{\beta}_1 - k}{SE(\hat{\beta}_1)} \right| \leq 1.96 \right\} \quad (1)$$

$$= \left[ \hat{\beta}_1 - 1.96 \times SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96 \times SE(\hat{\beta}_1) \right] \quad (2)$$

- Interpretation: If you draw many samples (dataset) and construct the 95% CI for each sample, 95% of those CIs will include the true parameter.

# Homoskedasticity vs Heteroskedasticity

- So far, we did not put any assumption on the variance of the error term $\epsilon_i$.
- The error term $\epsilon_i$ has **heteroskedasticity** if $Var(u_i|X_i)$ depends on $X_i$.
- If not, we call $\epsilon_i$ has **homoskedasticity**.
- This has an important implication on the asymptotic variance.

- Remember the asymptotic variance

$$\underbrace{V}_{(K+1)\times(K+1)} = E[\mathbf{x}_i'\mathbf{x}_i]^{-1}E[\mathbf{x}_i'\mathbf{x}_i\epsilon_i^2]E[\mathbf{x}_i'\mathbf{x}_i]^{-1}$$

  Standard errors based on this is called **heteroskedasticity robust standard errors**/
- If homoskedasticity holds, then

$$V = E[\mathbf{x}_i'\mathbf{x}_i]^{-1}\sigma^2$$

  where $\sigma^2 = V(\epsilon_i)$.
- In many statistical packages (including R and Stata), the standard errors for the OLS estimators are calcualted under homoskedasticity assumption as a default.
- However, if the error has heteroskedasticity, the standard error under homoskedasticity assumption will be **underestimated**.
- In OLS, **we should always use heteroskedasticity robust standard error.**
  - We will see how to fix this in R.