Mohamed Khidher University

Faculty of economics, Commercial and Management Sciences
Department of commerce

Specialty: International Commerce & Finance
Module: Advanced Econometric
Academic year 2024/2025

Review

I. Basics of Regression Analysis

- **Objective of Regression**: help in estimating relationships between dependent and independent variables.
- **General formula of regression**: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_k X_k$
- **Key Concepts**:
  - Coefficients and their interpretation.
  - R^2 and F-statistics are used to check the Goodness of fit of the model.

1. Interpretation of Coefficients

1.1 Intercept (β_0):

- The intercept is the expected value of *Y* when all *X*-variables are equal to zero.
- **Example**: If β_0=5, it means that when all *X*-variables are zero, the predicted Y is 5.

1.2 Coefficients of Independent Variables ($\beta_1,\beta_2,\ldots,\beta_k$):

- The coefficient for an independent variable represents the **change in the dependent variable** for a **one-unit increase** in that independent variable, holding all other variables constant.

Example 1: Simple Linear Regression Y=2+3X

- Here, β1=3.
- Interpretation: For every **1-unit increase in X**, Y increases by **3 units**.

2. Understanding Statistical Significance

- Statistical significance is determined using **hypothesis testing** for each coefficient.
- **Null Hypothesis (H_0)**: The coefficient (β) is equal to 0 (X has no effect on Y).
- **Alternative Hypothesis (H_1)**: The coefficient (β) different from 0 (X has an effect on Y).

3. Key Metrics to Evaluate Significance

3.1 p-Value

- The p-value represents the probability of observing the data if H_0 is true.
- **Decision Rule**:
  - If p-value$<\alpha$ (e.g., 0.05), reject H_0; the variable is significant.
  - If p-value$\geq\alpha$, fail to reject H_0; the variable is not significant.
- **Example**:
  - A p-value of 0.03 indicates that the variable is significant at the 5% significance level.
  - A p-value of 0.15 indicates that the variable is not significant.

t-Statistic

- Measures how many standard errors the coefficient is away from 0.

- $t = \frac{\hat{\beta}}{SE\hat{\beta}}$
    - $\hat{\beta}$: Estimated coefficient.
    - $SE\hat{\beta}$: Standard error of the coefficient.
- **Decision Rule**:
    - Compare t-statistic to a t-statistic critical -critical value (based on the chosen confidence level and degrees of freedom).
- **Rule of Thumb**: If |t|>t-critical, the variable is typically significant at the 5% level otherwise the variable has no effect (insignificant).

II. Linear vs. Non-Linear Models

- **Linear Models**:
    - Assumption: The relationship between variables is linear.
    - Example: OLS regression.
- **Non-Linear Models**:
    - When linear models are inappropriate (e.g.,when the dependent variable is binary).
    - Transition into Probit and Logit models (non-linear model).

III. **Types of Variables by Nature**

1. **Dependent Variables (Response Variables):**

- **Continuous Variable:**
    - Measured on a continuous scale (e.g., income, temperature, weight).
- **Binary Variable:**
    - Takes two possible values (e.g., 0 or 1, Yes or No).
    - Example: Used in Probit or Logit models to represent decisions like purchase (Yes/No) or loan approval (approved/Denied).
- **Categorical Variable (Multinomial):**
    - Represents multiple categories without a natural order.
    - Example: Choice of a product (Product A, B, or C).
- **Ordinal Variable:**
    - Represents categories with a meaningful order.
    - Example: Customer satisfaction (Low, Medium, High).

2. **Independent Variables (Explanatory Variables):**

- **Continuous Variable:**
    - Example: Age, income, hours of study.
- **Categorical Variable:**
    - Example: Gender (Male/Female), region (North/South/East/West).
- **Dummy Variable (Binary):**
    - A special categorical variable coded as 0 or 1.
    - Example: 1 for "Urban" and 0 for "Rural."
- **Interaction Variable:**

- A product of two variables to study their combined effect.
- Example: Income × Education level.

IV. Why OLS Fails for Binary Data:

a. Predicted probabilities may lie outside [0,1].

b. Inefficiency and incorrect assumptions about errors, for that we use probability model (**Probit and Logit Models)**

3. Panel Data

- **What is Panel Data?**:
  - Combines cross-sectional and time-series data.
  - Example: Tracking income and purchasing behavior of individuals over several years.
- **Key Concepts**:
  - Fixed effects vs. Random effects.
  - Importance of controlling for unobserved heterogeneity.

6. Mathematical Foundation

- **Key Equations to Review**:
  - Linear regression: $Y=\beta_0+\beta_1X+\epsilon$
  - Probit: $\Pr(Y=1) = \Phi(\beta_0+\beta_1X)$, where Φ is the standard normal CDF.
  - Logit: $P(Y=1) = \frac{e^{\beta_0+\beta_1X_1+\varepsilon}}{1+e^{\beta_0+\beta_1X_1+\varepsilon}}$.
  - Panel data regression:
    - Fixed effects: $Y_{it}=\alpha_i+\beta X_{it}+\epsilon_{it}$.
    - Random effects: $Y_{it}= \alpha + \beta X_{it} + \nu_i + \epsilon_{it}$.

Exercise (1): choose the right answer

1. What is the primary goal of a regression analysis
   - To summarize data using descriptive statistics
   - To model the relationship between dependent and independent variables
   - To perform hypothesis testing
   - To measure the central tendency of variables
2. In a simple linear regression $Y=\beta_0+\beta_1X_1+\epsilon$, what does $\beta 1$ represent?
   - The predicted value of Y
   - The mean of the dependent variable
   - The slope of the regression line or the effect of X on Y
   - The random error term
3. Which of the following is NOT an issue in linear regression?
   - Multicollinearity
   - Heteroscedasticity
   - Model overfitting
   - Non-normal distribution of Y

5. Which of the following is an example of a binary dependent variable?
    - Annual income
    - Temperature (in Celsius)
    - Whether a customer purchased a product (Yes/No)
    - Number of ads clicked by a user
6. Why is Ordinary Least Squares (OLS) not ideal for binary dependent variables?
    - It requires non-linear relationships between variables
    - It assumes the dependent variable is continuous
    - It works only with panel data
    - It cannot handle large datasets
7. Which of the following models is most appropriate for analyzing binary dependent variables?
    - Linear Regression
    - ARIMA Model
    - Probit or Logit Model
    - Fixed Effects Model
8. A researcher is studying whether an individual will apply for a loan (Yes/No) based on income and education level. Which regression model should they use?
    - Probit or Logit
    - Simple Linear Regression
    - Time Series Analysis
    - Panel Data Model
9. What is panel data?
    - Data collected at one point in time for multiple entities
    - Data collected over time for a single entity
    - Data collected for multiple entities over multiple time periods
    - Data used only for time-series models
10. Which of the following is an example of panel data?
    - Monthly sales figures of a company over 5 years
    - Annual GDP of 10 countries in 2020
    - Grades of 100 students across 3 semesters
    - Daily stock prices of a company
11. What is one major advantage of using panel data models?
    - It requires fewer observations
    - It captures both cross-sectional and time-series variations
    - It does not require assumptions about unobserved effects
    - It is easier to estimate than simple regression