Qualitative Variables in Regression
Unit 7

Prepared by

Aditya Korekallu Srinivasa

Scientist (Senior Scale)

ICAR-Indian Agricultural Research Institute
New Delhi

Contents

1 Types of Variables in Regression

1.1 Quantitative (Continuous) Variables

Variables measured on a ratio or interval scale, such as income, age, yield, and price, are quantitative. They can take on a wide range of numeric values and are suitable for standard regression analysis.

1.2 Qualitative (Categorical) Variables

- **Nominal variables**: Categories with no inherent order (e.g., gender, region, marital status).
- **Ordinal variables**: Categories with a logical order but not a measurable distance between categories (e.g., education level, satisfaction rating).
- **Binary (dichotomous) variables**: Special case of nominal variables with only two categories (e.g., adopted technology or not).

2 Problems with Qualitative Variables in Regression

2.1 Qualitative Variables as Independent Variables

- Cannot be directly entered into regression models due to their non-numeric nature.
- Need to be coded (e.g., using dummy variables) to be included as regressors.
- Interpretation and model specification require care to avoid pitfalls such as the dummy variable trap.

2.2 Qualitative Variables as Dependent Variables

- Standard linear regression is inappropriate for binary or categorical dependent variables.
- Issues include predicted probabilities outside [0,1], non-normal errors, and heteroscedasticity.
- Specialized models (e.g., LPM, Logit, Probit, Tobit, Fractional Probit) are needed.

3 Dummy Variables as Independent Variables in Linear Regression

3.1 How to Use Dummy Variables

- Represent qualitative characteristics by coding categories as 0 or 1 (or more generally, as a set of binary variables).
- For a qualitative variable with k categories, introduce $k - 1$ dummy variables.

- **Example:** In a study of adoption of agricultural technology, suppose we have a variable for gender (Male/Female). Define $D = 1$ if female, $D = 0$ if male.
- Regression model: $Y_i = \beta_0 + \beta_1 D_i + \beta_2 \text{Education}_i + \beta_3 \text{Age}_i + \beta_4 \text{Income}_i + u_i$

3.2 Interpretation of Dummy Variables

- The coefficient on a dummy variable measures the difference in the intercept (or slope, if interacted) between the reference category and the category represented by the dummy.
- **Example:** If $\beta_1 = 0.15$, it means that, ceteris paribus, the expected value of Y (e.g., probability of adoption) is 0.15 higher for females than for males.

3.3 Dummy Variable Trap and How to Avoid It

- Including k dummies for k categories along with an intercept causes perfect multicollinearity (the dummy variable trap).
- **Remedy:** Always include only $k-1$ dummies for k categories, or omit the intercept and include all dummies.

3.4 Extensions

- **Interaction terms:** Allow the effect of one variable to depend on the value of another (e.g., $D \times$ Education to see if the effect of education differs by gender).
- **Seasonal dummies:** For quarterly data, use three dummies to capture seasonal effects.
- **Piecewise linear regression:** Use dummies to allow different slopes in different ranges.

4 Qualitative Dependent Variables: Models and Detailed Examples

Consider the following practical example throughout this section:

Example: *Adoption of Agricultural Technology* (Y) as a function of Education (years), Age (years), Extension Contact (number of visits), and Income (in thousands).

Let $Y = 1$ if a farmer adopts the technology, $Y = 0$ otherwise.

4.1 Linear Probability Model (LPM)

Intuition and Model

- The LPM models the probability of adoption as a linear function of the predictors:

$$P(\text{Adopt} = 1|X) = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Age} + \beta_3 \text{Extension} + \beta_4 \text{Income}$$

- **Hypothetical Example:**

$$\hat{P}(\text{Adopt} = 1) = 0.20 + 0.04 \times \text{Education} - 0.01 \times \text{Age} + 0.10 \times \text{Extension} + 0.02 \times \text{Income}$$

- **Interpretation:** Each additional year of education increases the probability of adoption by 4 percentage points, holding other factors constant.

Estimation and Issues

- Estimated by OLS.
- **Problems:** Predicted probabilities can be less than 0 or greater than 1; error term is heteroscedastic and non-normal; R^2 is not meaningful.

When to Use

- For quick, rough analysis or as a benchmark; not recommended for final inference.

4.2 Logit Model

Intuition and Model

- Models the log-odds of adoption as a linear function of predictors:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Extension} + \beta_4\text{Income}$$

- **Probability function:**

$$P(\text{Adopt} = 1|X) = \frac{\exp(\beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Extension} + \beta_4\text{Income})}{1 + \exp(\beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Extension} + \beta_4\text{Income})}$$

Practical Example with Hypothetical Coefficients

Suppose the estimated logit model is:

$$\log\left(\frac{P}{1-P}\right) = -2.0 + 0.15 \times \text{Education} - 0.03 \times \text{Age} + 0.40 \times \text{Extension} + 0.08 \times \text{Income}$$

For a farmer with 10 years of education, age 40, 2 extension contacts, and income of 30:

$$\text{Linear index} = -2.0 + 0.15 \times 10 - 0.03 \times 40 + 0.40 \times 2 + 0.08 \times 30 = -2.0 + 1.5 - 1.2 + 0.8 + 2.4 = 1.5$$

$$P = \frac{\exp(1.5)}{1 + \exp(1.5)} \approx \frac{4.48}{5.48} \approx 0.82$$

So, the predicted probability of adoption is 82%.

Interpretation of Odds Ratios

- The coefficient for education is 0.15. The odds ratio is $\exp(0.15) \approx 1.16$.
- **Interpretation:** Each additional year of education increases the odds of adoption by 16%, holding other factors constant.
- For extension contact (0.40): $\exp(0.40) \approx 1.49$, so each additional extension contact increases the odds of adoption by 49%.

Marginal Effects

- Marginal effect at the mean (MEM): $\beta_j \times P(1 - P)$.
- For education: $0.15 \times 0.82 \times (1 - 0.82) \approx 0.15 \times 0.1476 \approx 0.022$.
- **Interpretation:** At the mean, each additional year of education increases the probability of adoption by about 2.2 percentage points.

4.3 Probit Model

Intuition and Model

- Models the probability of adoption using the cumulative standard normal distribution:

  $$P(\text{Adopt} = 1|X) = \Phi(\beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Extension} + \beta_4\text{Income})$$

  where $\Phi(\cdot)$ is the standard normal CDF.

Practical Example with Hypothetical Coefficients

Suppose the estimated probit model is:

$$P(\text{Adopt} = 1|X) = \Phi(-1.0+0.09\times\text{Education}-0.02\times\text{Age}+0.28\times\text{Extension}+0.05\times\text{Income})$$

For the same farmer (Education=10, Age=40, Extension=2, Income=30):

$$\text{Linear index} = -1.0 + 0.9 - 0.8 + 0.56 + 1.5 = 1.16$$

$$P = \Phi(1.16) \approx 0.877$$

So, the predicted probability of adoption is about 88%.

Marginal Effects

- Marginal effect for education: $\beta_1 \times \phi(\text{index})$ where $\phi(\cdot)$ is the standard normal PDF.
- At index = 1.16, $\phi(1.16) \approx 0.209$.
- So, marginal effect for education: $0.09 \times 0.209 \approx 0.019$.
- **Interpretation:** At the mean, each additional year of education increases the probability of adoption by about 1.9 percentage points.

4.4 Tobit Model

Intuition and Model

- Used when the dependent variable is censored (e.g., observed only above or below a certain threshold).
- **Example:** Suppose adoption intensity (proportion of land under new technology) is observed only for those who adopt (i.e., $Y^* > 0$).
- **Equation:**

$$Y^* = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Age} + \beta_3 \text{Extension} + \beta_4 \text{Income} + u$$

$$Y = Y^* \text{ if } Y^* > 0, \quad Y = 0 \text{ if } Y^* \leq 0$$

Estimation and Interpretation

- Estimated by maximum likelihood.
- Coefficients reflect the effect on the latent variable (Y^*), not directly on the observed Y.
- Marginal effects can be decomposed into effects on the probability of being uncensored and the expected value conditional on being uncensored.
- **When to use:** When the dependent variable is continuous but censored (e.g., expenditure with many zeros).

4.5 Fractional Probit Model

Intuition and Model

- Used when the dependent variable is a proportion or fraction bounded between 0 and 1 (e.g., proportion of land under technology).
- **Equation:** $E[Y|X] = \Phi(\beta_0 + \beta_1 X)$, $0 \leq Y \leq 1$
- **Estimation:** Quasi-maximum likelihood or generalized linear models.
- **When to use:** For modeling proportions or rates, especially when there are many observations at the boundaries.

5 Summary Table: Models for Qualitative Variables

Model	**When to Use**	**Equation/Link**	**Estimation**	**Interpretation/Example**
Linear Probability Model (LPM)	Binary dependent variable, quick analysis	$P(Y = 1\|X) = \beta_0 + \beta_1 X$	OLS	Each unit increase in education increases adoption probability by β_1 points; may predict probabilities outside [0,1]
Logit Model	Binary dependent variable, probabilities in [0,1]	$P(Y = 1\|X) = \frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}$	Maximum Likelihood	Odds ratio: $\exp(\beta_1)$; marginal effect: $\beta_1 \times P(1-P)$; e.g., odds of adoption increase by 16% per year of education
Probit Model	Binary dependent variable, latent variable interpretation	$P(Y = 1\|X) = \Phi(\beta_0 + \beta_1 X)$	Maximum Likelihood	Marginal effect: $\beta_1 \times \phi(\text{index})$; e.g., each year of education increases adoption probability by 1.9 percentage points
Tobit Model	Censored dependent variable (e.g., many zeros)	$Y^* = \beta_0 + \beta_1 X + u$, $Y = \max(0, Y^*)$	Maximum Likelihood	Used for adoption intensity; coefficients reflect effect on latent variable; marginal effects decomposed
Fractional Probit Model	Dependent variable is a proportion/rate, $0 \leq Y \leq 1$	$E[Y\|X] = \Phi(\beta_0 + \beta_1 X)$	Quasi-MLE/GLM	For proportions/rates; marginal effects on probability; e.g., proportion of land under technology
Dummy Variables in OLS	Qualitative independent variables	$Y = \beta_0 + \beta_1 D + u$	OLS	Coefficient measures mean difference; e.g., female farmers have 0.15 higher adoption probability than males

Glossary

- **Qualitative Variable**: A variable that categorizes or describes an element of a population (e.g., gender, region).
- **Dummy Variable**: A binary variable (0 or 1) used to represent categories in regression models.
- **Linear Probability Model (LPM)**: A linear regression model for binary dependent variables.
- **Logit Model**: A model for binary outcomes using the logistic function.
- **Probit Model**: A model for binary outcomes using the standard normal CDF.
- **Tobit Model**: A regression model for censored dependent variables.
- **Fractional Probit Model**: A model for dependent variables that are proportions between 0 and 1.
- **Dummy Variable Trap**: Perfect multicollinearity arising from including all categories of a qualitative variable as dummies.
- **Maximum Likelihood Estimation**: A method of estimating model parameters by maximizing the likelihood function.
- **Odds Ratio**: The ratio of the odds of an event occurring in one group to the odds in another group.
- **Marginal Effect**: The change in the predicted probability (for binary models) or expected value (for other models) for a unit change in a predictor.

Practice Questions

1. Distinguish between quantitative and qualitative variables. Give examples of each.
2. Explain the problems that arise when qualitative variables are used as independent variables in regression.
3. How are dummy variables constructed and interpreted in linear regression? Illustrate with an example.
4. What is the dummy variable trap? How can it be avoided?
5. Why is the linear probability model problematic for binary dependent variables?
6. Using the agricultural technology adoption example, interpret the coefficients, odds ratios, and marginal effects in logit and probit models.
7. What is the Tobit model? In what situations is it appropriate?
8. Discuss the intuition and application of the fractional probit model.

9. For each model (LPM, Logit, Probit, Tobit, Fractional Probit), state the basic equation, estimation method, and key issues.

10. How would you choose among different models for qualitative dependent variables in applied research?

References

- Gujarati, D. N., & Porter, D. C. (2010). *Basic Econometrics* (5th ed.). McGraw-Hill.

- Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). Cengage Learning.