

Introduction to Statistical Methods

SOC-GA 2332

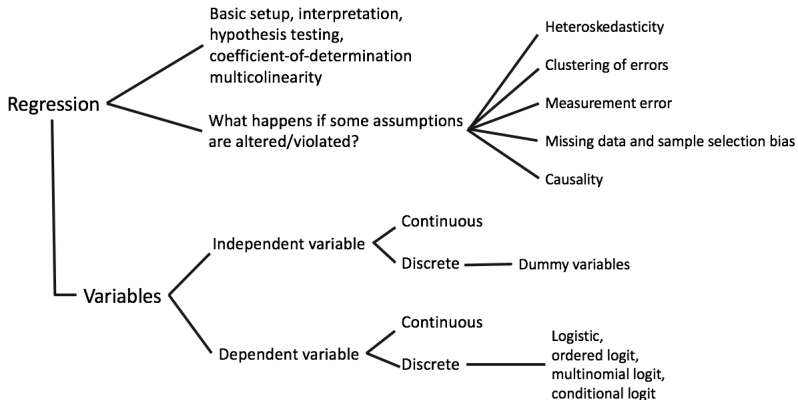
Lecture 9: Discrete Dependent Variable

Siwei Cheng



NEW YORK UNIVERSITY

Outline



Predicting Discrete Dependent Variables

- ▶ In previous lectures, we introduced regression models with continuous outcome variables. Today, we discuss the case in which you have a **discrete dependent variable**.
- ▶ Examples of **binary** dependent variables: voter's choice in presidential election, high school graduation, depression diagnosis, being married, being unemployed.
- ▶ Examples of **multicategory** dependent variables: marital status, education attainment, Likert scale variables (strongly agree, agree, neutral, disagree, strongly disagree).

Binary Response Variables

- ▶ We will focus on **binary** response variables in today's lecture. We denote a binary variable Y , with two categories 1 and 0.
- ▶ For example, if your outcome variable is high school graduation, you can code your data so that 1 denotes "graduated from high school" and 0 denotes "did not graduate from high school."
- ▶ When our outcome variable is a discrete dependent variable, we are interested in the probability of this variable taking a certain value, and what independent variables predict this probability.

Binary Response Variables

- ▶ For example, you may be interested in whether and the extent to which the likelihood of graduating from high school (1=graduated and 0=did not graduate) depends on the students' family background, neighborhood conditions, and race, etc.
- ▶ ~~When the effect of an explanatory variable is positive, you can interpret the finding as suggesting that having a higher value on the explanatory variable is associated with an increased likelihood of graduating from high school.~~

Bernoulli distribution

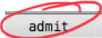
- ▶ The binary outcome variable y is assumed to follow the **Bernoulli distribution**.
- ▶ That is, Y takes value 1 with probability p . That is:

$$y \sim \text{Bernoulli}(p)$$

- .
- ▶ Note that Y can only take two values: 1 and 0, so that $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$.

Binary Response Variables

- Here is an example of how your data with a binary dependent variable look like:



admit	gre	gpa	rank
0	380	3.61	3
1	660	3.67	3
1	800	4	1
1	640	3.19	4
0	520	2.93	4
1	760	3	2
1	560	2.98	1
0	400	3.08	2
1	540	3.39	3
0	700	3.92	2
0	800	4	4
0	440	3.22	1
1	760	4	1
0	700	3.08	2
1	700	4	1

Two Types of Models

1. Linear Probability Model: directly modeling $P(Y=1)$.
2. Models with a *transformation* of $P(Y=1)$: Logit and Probit models.

The Linear Probability Model

- ▶ The **linear probability model** predicts $P(Y = 1)$ as a linear function of independent variables:

$$P(Y = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_k X_k + \epsilon \quad (1)$$

- ▶ The estimated regression coefficients tell us how much we expect the probability of $Y=1$ changes in response to one unit change in the corresponding independent variable:

$$\Delta \hat{P}(Y = 1) = \hat{\beta}_1 \cdot \Delta X_1 \quad (2)$$

- ▶ When we fit the model, we use the value of Y as the outcome variable.

The Linear Probability Model

- For example, we fit a linear probability model predicting the probability of admission using GPA, GRE, and institution rank:

```
> mylm <- lm(admit ~ gre + gpa + factor(rank), data = data)
> summary(mylm)
```

Call:

```
lm(formula = admit ~ gre + gpa + factor(rank), data = data)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.7022	-0.3288	-0.1922	0.4952	0.9093

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.2589103	0.2159904	-1.199	0.2314
gre	0.0004296	0.0002107	2.038	0.0422 *
gpa	0.1555350	0.0639618	2.432	0.0155 *
factor(rank)2	-0.1623653	0.0677145	-2.398	0.0170 *
factor(rank)3	-0.2905705	0.0702453	-4.137	4.31e-05 ***
factor(rank)4	-0.3230264	0.0793164	-4.073	5.62e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

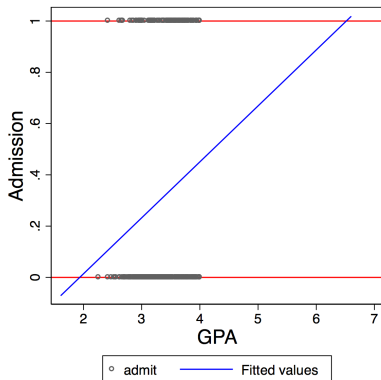
Residual standard error: 0.4449 on 394 degrees of freedom

Multiple R-squared: 0.1004, Adjusted R-squared: 0.08898

F-statistic: 8.795 on 5 and 394 DF, p-value: 6.333e-08

The Linear Probability Model

- ▶ The coefficient on GPA is pretty straightforward to interpret: a 1-unit increase in a student's GPA is associated with a 15.55 percentage point (or 0.1555) increase in the likelihood of being admitted.
- ▶ Below is the scatter plot of the data and the predicted probability.



The Linear Probability Model

- ▶ Linear probability models are easy to estimate and interpret.
- ▶ But the linear probability model suffers from several shortcomings:
 1. We can get predictions less than zero or greater than one. This doesn't make sense for a probability...
 2. The predicted effect of a set of explanatory variables can be greater than 1 in absolute value, as that wouldn't make sense.

The Linear Probability Model

- ▶ So, why do we still use LPM?
- ▶ The model will work OK when the value of independent variables is *near the average and not on the extreme.*
- ▶ For example, if we look at the GPA example, the values of GPA generally ranges between 2 and 4, and our predicted probability is between 0 and 1.
- ▶ Note that LPM predicts probability. If we want to predict the outcome, how shall we translate the predicted probability into our prediction of Y?
- ▶ There are many solutions. One common practice in literature is to predict $y_i = 1$ if $\hat{y}_i \geq 0.5$ and $y_i = 0$ if $\hat{y}_i < 0.5$.
- ▶ In the causal inference literature, the LPM makes it straightforward to interpret “causal effect.”



Methods that Rely on Transformation of p_i

Methods that Rely on Transformation of p_i

- ▶ Define a transformation function $g(\cdot)$.
- ▶ We can write the probability of $Y = 1$, denoted by p_i , by specifying:

$$p_i = g(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k) \quad (3)$$

- ▶ It thus follows that we can transform p_i by the inverse function of $g(\cdot)$ and get a linear function:

$$g^{-1}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (4)$$

- ▶ The transformation $g(\cdot)$ is often chosen as the CDF of a standard normal distribution (Probit model) or a logistic distribution (Logistic regression model/Logit model).
- ▶ The latter is more commonly used, so let's focus on the **Logistic Regression Model** for Binary Dependent Variables.

Logistic Regression Model for Binary Dependent Variables

$$g^{-1}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (5)$$

- ▶ As shown above, the intuition of the logistic regression model is that instead predict the probability as a linear function of X , we predict a **transformed probability** $g^{-1}(p_i)$ as a linear function of X .
- ▶ The transformation $g(z)$ is defined as the CDF of the logistic distribution:

$$g(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{e^z + 1} \quad \text{odds} \quad (6)$$

- ▶ The **inverse** of $g(z)$ ($g^{-1}(p_i)$) is a **logistic function**.

Logistic Regression Model for Binary Dependent Variables

- ▶ Now, we are going to introduce the logistic transformation (logit).
- ▶ First, we calculate the **odds** of $y_i = 1$ as:

$$odds_i = \frac{p_i}{1 - p_i} \quad (7)$$

- ▶ Next, we take the logarithm of the odds, to get the **log odds**, or **logit**:

$$logit(p_i) = \log \frac{p_i}{1 - p_i} \quad (8)$$

- ▶ Recall that $g^{-1}(p_i) = logit(p_i)$, so we have:

$$g^{-1}(p_i) = logit(p_i) = \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (9)$$

Logistic Regression Model for Binary Dependent Variables

- This is the logistic regression model:

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (10)$$

- The logit transformation of probability ($\text{logit}(p_i)$) corresponds to the probability p_i in the following way:

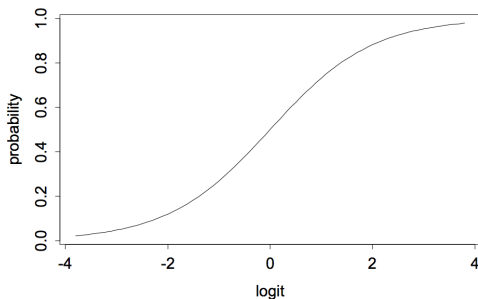


FIGURE 3.1: The Logit Transformation

Logistic Regression Model for Binary Dependent Variables

- From Equation (10) we can also derive the predicted probability as:

$$p_i = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)} \quad (11)$$

- As you can see, we have now modeled the probability of $Y=1$ as a *non-linear* function of the predictors.



Interpreting Coefficients in Logistic Regression

- ▶ How to interpret the coefficients from the logistic regression?
- ▶ The coefficient on X_1 is the effect of X_1 on *log odds*:

$$\text{logit}(p_i) = \log \frac{p_i}{1-p_i}$$

- ▶ But log odds is probably not very intuitive to interpret...
- ▶ Next, we further show that the exponential of β_1 , expressed as $\exp(\beta_1)$, can be seen as the **effect on odds**.

Interpreting Coefficients in Logistic Regression

- ▶ Note that the effect of an explanatory variable X_1 on **odds** is **NOT additive but multiplicative**.
- ▶ To see this, consider a bivariate regression case:

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 X_1 \quad (12)$$

- ▶ That is:

$$\text{Odds}_1 = \frac{p_i}{1 - p_i} = \exp(\beta_0 + \beta_1 X_1) \quad (13)$$

- ▶ If X_1 increases by one unit, then the odds becomes:

$$\text{Odds}_2 = \frac{p_i}{1 - p_i} = \exp(\beta_0 + \beta_1 (X_1 + 1)) \quad (14)$$

$$= \exp(\beta_0 + \beta_1 X_1) \cdot \exp(\beta_1) \quad (15)$$

Interpreting Coefficients in Logistic Regression

- ▶ That is, the coefficient on X tells us the amount by which we should multiply the odds for a given unit change in X . $\exp(\beta)$ is called the *odds ratio*.
 $e^{0.6} = 1.82$
- ▶ For example, if our coefficient $\beta_1 = 0.6$, then we calculate the exponential of β_1 as: $\exp(\beta_1) = \exp(0.6) = 1.82$. This means that if the value on X_1 increases by one unit, we will need to multiply the odds by 1.82.
- ▶ If $\beta > 0$, then $\exp(\beta) > 1$, and this means that the odds of success will be greater if X increases.
- ▶ If $\beta < 0$, then $\exp(\beta) < 1$, and this means that the odds of success will be lesser if X increases.
- ▶ If $\beta = 0$, then $\exp(\beta) = 1$, and this means that the odds of success does not change with X .

Interpreting Coefficients in Logistic Regression

- ▶ Do the effects of multiple explanatory variables depend on each other?
- ▶ Recall that when we introduced multivariate linear regression model, we knew that because we assume that Y is a linear function of $X_1, X_2 \dots X_k$ (with no interaction terms), the effect of each of the variables in X will not depend on the values of other variables.
- ▶ However, this is **not** the case with logistic regression, as the predicted probability is a non-linear function of the independent variables (see Equation (11)).
- ▶ The effect of each of the independent variables will depend on the value of other independent variables (if you are really curious, try taking the derivative of p with respect to X_1).

Extension: An Unobserved-Variable Formulation

- ▶ An alternative formulation posits an underlying regression for a continuous but unobserved variable u so that:

$$Y_i = \begin{cases} 0, & \text{if } u \leq 0 \\ 1, & \text{if } u > 0 \end{cases} \quad (16)$$

- ▶ We then assume that the latent u is a linear function of X :



$$u = \beta_0 + \beta_1 X_1 - \epsilon \quad (17)$$

- ▶ It thus follows that the probability of $Y = 1$ can be written as:

$$p_i = \Pr(u > 0) = \Pr(\epsilon < \beta_0 + \beta_1 X_1) \quad (18)$$

- ▶ It follows that, when ϵ follows a standard normal distribution, the model becomes a probit model; when it follows a logistic distribution, then the model becomes a logit model.

Estimation for the Logistic Regression Model

- ▶ The estimation of Logistic regression models are done by finding the maximum likelihood estimator. 
- ▶ As we mentioned in earlier lectures the maximum likelihood principle of estimation begins by defining a “likelihood function” that gives the probability of observing the data for any given set of parameters. Then, the maximum likelihood estimators are values of the parameters that maximize the likelihood function.
- ▶ In other words, the maximum likelihood method treats the observed data as fixed, and chooses parameter estimates that maximize the chance of observing those data.
- ▶ The maximum likelihood principle is used widely in obtaining estimates for regression models in which the dependent variable is not continuous. 



Example: Logistic Regression

- ▶ Let's look at an example. We use GRE, GPA, and institution rank to predict the likelihood of graduate school admission:
- ▶ In R, we can estimate the logistic regression model using glm:

```
> mylogit <- glm(admit ~ gre + gpa + factor(rank), data = data, family = "binomial")  
>  
> summary(mylogit)
```

Call:

```
glm(formula = admit ~ gre + gpa + factor(rank), family = "binomial",  
    data = data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6268	-0.8662	-0.6388	1.1490	2.0790

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.989979	1.139951	-3.500	0.000465 ***
gre	0.002264	0.001094	2.070	0.038465 *
gpa	0.804038	0.331819	2.423	0.015388 *
factor(rank)2	-0.675443	0.316490	-2.134	0.032829 *
factor(rank)3	-1.340204	0.345306	-3.881	0.000104 ***
factor(rank)4	-1.551464	0.417832	-3.713	0.000205 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Some Additional Notes

- ▶ Probit regression models, which uses another function to transform the probability. In most cases, results will be very similar using either model, and either is acceptable.
- ▶ **There is a pitfall in cross-group comparisons of logit or probit coefficients:** unlike linear regression coefficients, coefficients in these binary regression models are confounded with residual variation (unobserved heterogeneity).
- ▶ Differences in residual variance across groups/models can produce apparent differences in coefficients that are not indicative of true differences in effects.
- ▶ Next slide shows why.

Some Additional Notes

- ▶ To see this, suppose ϵ follows the standard logistic distribution, and suppose the latent variable u in the unobserved variable formulation is generated by:

$$u = \alpha_0 + \alpha_1 X_1 + \dots \alpha_k X_k + \sigma \epsilon. \quad (19)$$

where ϵ follows a logistic distribution.

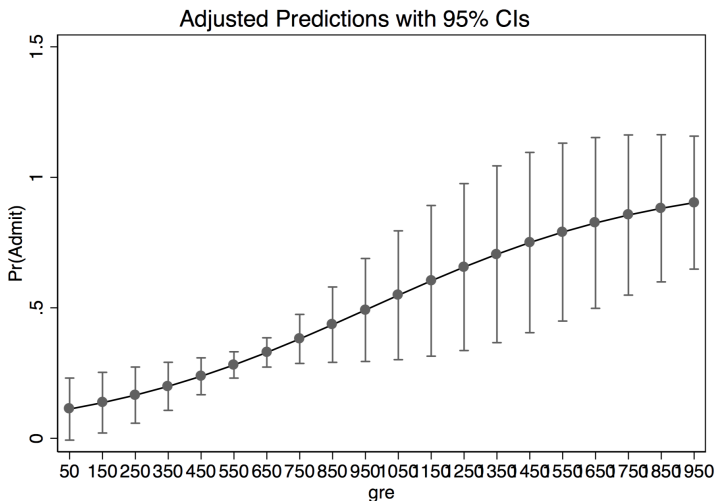
- ▶ It can be shown that, if we estimate a logistic regression model:

$$g(p_i) = \beta_0 + \beta_1 X_1 + \dots \beta_k X_k. \quad (20)$$

- ▶ Then, we have: $\beta_k = \alpha_k / \sigma$.
- ▶ Hence, comparison of the value of β_k across groups or models is problematic if σ differs by group ✱
- ▶ In practice, it's always more helpful to compare predicted probabilities across groups/models. ✱

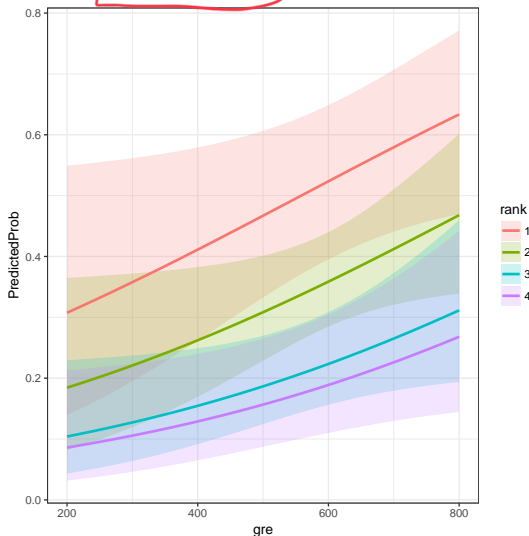
Example: Plotting Predicted Probability

- We can use the STATA **margins** command to get the predicted margins



Example: Plotting Predicted Probability

- R can also produce marginsplots like this:



Goodness-of-Fit Statistics

- ▶ In linear regression models, we calculated the R-squared as the proportion of total variance explained by the independent variables in the model.
- ▶ In Logistic regression models, a more common way of reporting the goodness-of-fit of a Logistic regression model is by looking at the **likelihood ratio test statistic**, which follows a chi-square distribution under the null hypothesis (that is, the model as a whole is not significant).
- ▶ The likelihood ratio test is performed by estimating two models and comparing the fit of one model to the fit of the other.
- ▶ The test compares the log likelihoods of the two models, if this difference is statistically significant, then the less restrictive model (the one with more variables) is said to fit the data significantly better than the more restrictive model.

An Example

- Finally, let's look at some examples from published studies. A study by Conley and Bennett (2000) examines the effect of low birth weight on the chances of high school graduation. Can you interpret the coefficients from the Logistic regression models?

Table 2. Coefficients from the Regression of Timely High School Graduation on Low Birth Weight, Maternal Socioeconomic Circumstances, and Other Independent Variables: Logistic and Fixed-Effects Conditional Logistic Models

Variables	Statistics for Models 1–3		Logistic Regression Models			Fixed-Effects Model	Statistics for Fixed-Effects Model	
	Mean	S.D.	Model 1	Model 2	Model 3		Mean	S.D.
<i>Perinatal Conditions</i>								
Low birth weight	.07	.26	-.626* (.247)	-.598* (.248)	-.577* (.250)	-2.024** (.764)	.07	.25
Income-to-needs ratio, ages 0–5	2.36	1.59	—	.114** (.039)	.088* (.044)	.050 (.181)	2.49	1.34
Maternal age < 18 years	.04	.19	—	-.384 (.333)	-.424 (.345)	-.904 (.696)	.03	.18
<i>Controls</i>								
Female child	.47	.50	—	—	.439*** (.119)	.363 (.204)	.48	.50
First-born child	.42	.49	—	—	.199 (.124)	.112 (.198)	.28	.45
African American	.316	.47	—	—	-.111 (.140)	—	—	—

