Section 3: Binary dependent variables

1 Conditional Expectation of Binary Dependent Variables

When a variable takes one of only two values, 0 or 1, it is called a *binary* variable. And when a binary variable is placed on the left hand side of a regression specification, we call it a binary dependent variable. When we try to predict Y using some independent variable X, we will get predicted values of Y that are neither 0 nor 1. Since Y can only be 0 or 1, how do we interpret these predicted values?

We interpret them as probabilities, i.e. the predicted value of Y given X is an estimate of the probability that Y equals 1 for an observation with the given value of X. More precisely, the predicted value of Y given X is always the conditional expectation of Y given X, or $E[Y|X]$. In the case of binary dependent variables, since Y is either 0 or 1,

$$E[Y|X] = 1 \cdot Pr(Y = 1|X) + 0 \cdot Pr(Y = 0|X) = Pr(Y = 1|X). \tag{1}$$

Note that $0 \leq P(Y = 1|X) \leq 1$ because this is a probability.

Throughout, let's consider the example of trying to predict whether an individual has been hospitalized overnight in the past year, based on some information about them such as age, education and gender. We will focus on the population of Americans betweent the ages of 55-74 in the 1992-2003 National Health Insurance Survey.

Q1: What is the dependent variable, Y? What do its values indicate?

Q2: what are the independent variables, X?

In what follows, we will cover three common ways of estimating these conditional probabilities.

2 Linear probability model (LPM)

- The *LPM* is just the standard OLS framework regressing Y_i on X_i's:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki} + u_i.$$

Q1: How do we interpret the coefficients?

- Example: We want to predict hospitalizations using age (in years), education (in years) and gender.

```
. *Linear Probability Model
. reg inhosp age educ female, robust

Linear regression                               Number of obs     =    157,861
                                                F(3, 157857)      =     503.87
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0098
                                                Root MSE          =      .3301

------------------------------------------------------------------------------
             |               Robust
      inhosp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .004741   .0001496    31.68   0.000     .0044477    .0050343
        educ |  -.0040762   .0002443   -16.68   0.000    -.0045551   -.0035973
      female |  -.0189309   .0016794   -11.27   0.000    -.0222225   -.0156392
       _cons |  -.1172182   .0102799   -11.40   0.000    -.1373667   -.0970698
------------------------------------------------------------------------------
```

Q2: What is the predicted probability of having having been hospitalized in the past year for a female of age 58 with 12 years of education?

$\hat{P}(Y_i = 1) =$

Q3: What if we increase her age by 10 years?

$\hat{P}(Y_i = 1) =$

Q4: What's the change in predicted probabilities?

Q5: What are some possible problems with the linear probability model? How can we remedy these problems?

3 Probit model

- The *Probit* model fits a nonlinear model by assuming

$$P(Y_i = 1|X_i) = \Phi(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. $z = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$ is the z-score.

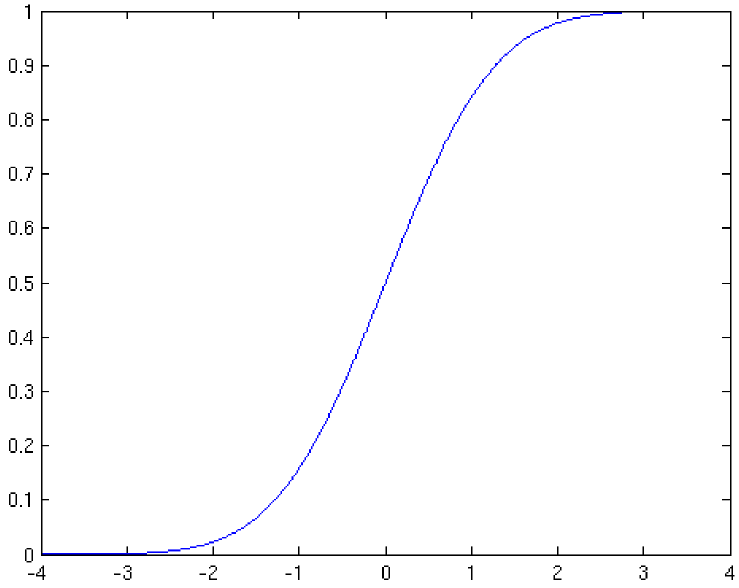


Figure 1: Standard Normal CDF

- This is not estimated using OLS. Stata fits the model using Maximum Likelihood Estimation (MLE)
- Interpretation of the coefficients:
  $\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$ is the **z-score** of the predicted probability.
  So β_1 is the effect on the **z-score** of a unit change in X_1 holding all the other regressors constant.

-
  ```
  . *Probit Model
  . probit inhosp age educ female, robust

  Iteration 0:   log pseudolikelihood = -59749.902
  Iteration 1:   log pseudolikelihood =   -58981.6
  Iteration 2:   log pseudolikelihood = -58979.705
  Iteration 3:   log pseudolikelihood = -58979.705

  Probit regression                               Number of obs     =    157,861
                                                  Wald chi2(3)      =    1549.18
                                                  Prob > chi2       =     0.0000
  Log pseudolikelihood = -58979.705               Pseudo R2         =     0.0129

  ------------------------------------------------------------------------------
               |               Robust
        inhosp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
           age |   .0230413   .0007199    32.01   0.000     .0216303    .0244523
          educ |  -.0195974   .0011345   -17.27   0.000     -.021821   -.0173737
        female |  -.0911007   .0081524   -11.17   0.000     -.107079   -.0751223
         _cons |  -2.345042   .0500071   -46.89   0.000    -2.443054    -2.24703
  ------------------------------------------------------------------------------
  ```

- Calculate changes in the predicted probabilities.

  Q6: What is the predicted probability of having having been hospitalized in the past year for a female of age 58 with 12 years of education?

  $\hat{P}(Y_i = 1) =$

  Q7: What if we increase her age by 10 years?

  $\hat{P}(Y_i = 1) =$

  Q8: What's the change in predicted probabilities?

4 Logit model

- The *Logit* model makes a different assumption about the CDF:

  $$P(Y_i = 1|X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki})}} \equiv F(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}), \tag{3}$$

  where $F(\cdot)$ is the CDF of the logistic distribution.

- Interpretation of the coefficients: β_1 is the effect on the Logit function input of a unit change in X_1 holding all the other regressors constant. That is, not very interpretable. We can use predicted effects to interpret.

-
  ```
  . *Logit Model
  . logit inhosp age educ female, robust

  Iteration 0:   log pseudolikelihood = -59749.902
  Iteration 1:   log pseudolikelihood = -58992.946
  Iteration 2:   log pseudolikelihood =  -58983.25
  Iteration 3:   log pseudolikelihood = -58983.249

  Logistic regression                             Number of obs     =    157,861
                                                  Wald chi2(3)      =    1561.14
                                                  Prob > chi2       =     0.0000
  ```

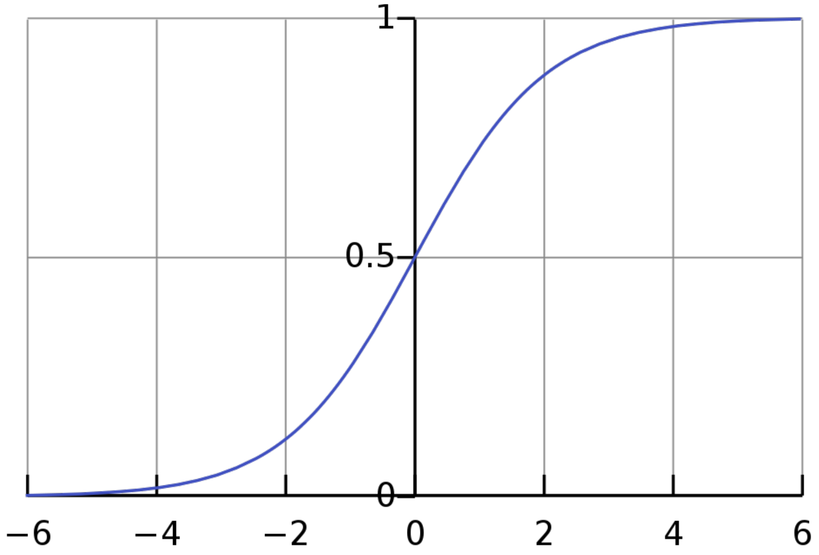


Figure 2: Logistic Function

```
Log pseudolikelihood = -58983.249               Pseudo R2         =     0.0128

------------------------------------------------------------------------------
             |               Robust
      inhosp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0432252   .0013511    31.99   0.000     .0405772    .0458733
        educ |  -.0358687   .0020826   -17.22   0.000    -.0399506   -.0317869
      female |  -.1710903   .0152838   -11.19   0.000     -.201046   -.1411346
       _cons |  -4.203517   .0941838   -44.63   0.000    -4.388113    -4.01892
------------------------------------------------------------------------------
```

- Calculate changes in the predicted probabilities.

  Q9: What is the predicted probability of having having been hospitalized in the past year for a female of age 58 with 12 years of education?

  $\hat{P}(Y_i = 1) =$

  Q10: What if we increase her age by 10 years?

  $\hat{P}(Y_i = 1) =$

  Q11: What's the change in predicted probabilities?

Q12: What are some advantages to the LPM? What are some advantages to Probit/Logit?

Q13: When are the predicted probabilities of an LPM, Probit, and Logit model the same?

5 Interactions

In some situations, we might think the effect of one regressor, call it x_1, could depend on the value of another regressor, x_2. For instance, we might be interested in whether small class size has different effects for girls and boys. In order to see these differential effects, we'll create an **interaction term** by multiplying an indicator for treatment with the other regressor.

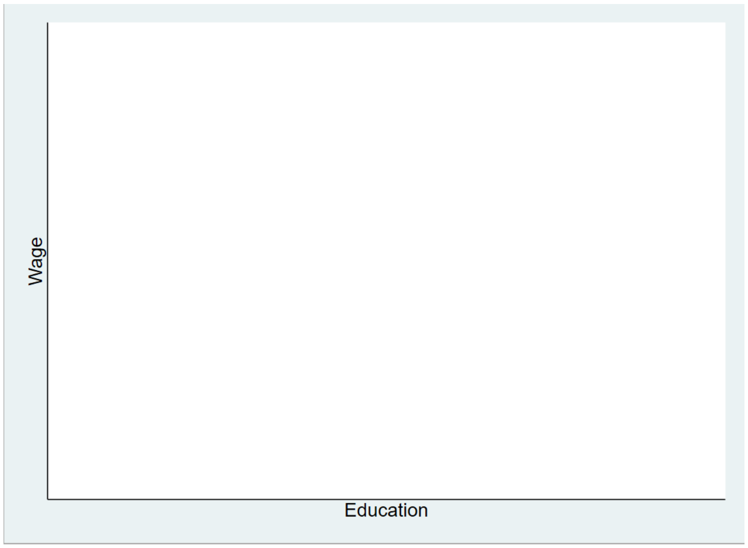
Consider our example above, where treatment $T_i = 1$ if student i was in a small class and 0 otherwise, and gender $F_i = 1$ if student i is female and 0 otherwise. We can estimate the following regression:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 F_i + \beta_3 T_i \cdot F_i + u_i$$

Q1: What is the treatment effect for females? For males?

However, sometimes our group isn't binary. For example, consider the effect of education on wages and whether or not years of education have a different effect on wages for women than for men. Let W_i be hourly wage and E_i be years of education. We might estimate the regression:

$$W_i = \gamma_0 + \gamma_1 E_i + \gamma_2 F_i + \gamma_3 E_i \cdot F_i + v_i$$



Q2: First, suppose that women earn more than men, and while everyone's wage increase as education increases, women's wages increase less with additional years of education than men's. What can we infer about the signs and relative magnitudes of γ_1, γ_2, and γ_3?

Q3: On the axes above, draw a line for $F_i = 0$ and for $F_i = 1$. Label where you used the coefficients from the regression on the graph.

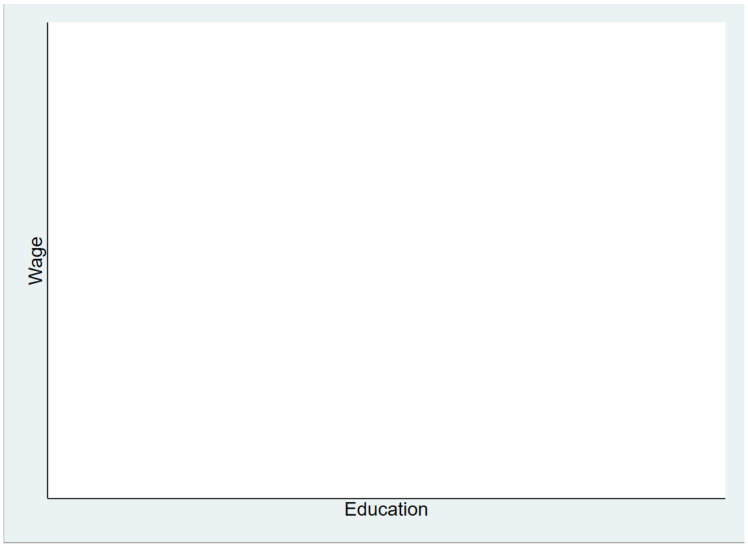
Q4: Now suppose that women earn more than men, and while everyone's wage increase as education increases, women's wages increase *more* with additional years of education than men's. What can we infer about the signs and relative magnitudes of γ_1, γ_2, and γ_3?

Q5: On the axes below, draw a line for $F_i = 0$ and for $F_i = 1$. Label where you used the coefficients from the regression on the graph.

6 An Introduction to (Sharp) Regression Discontinuity

Regression discontinuity (RD) designs are based on the assumption that individuals just below and just above a threshold value of some variable are similar, so that comparing outcomes for the individuals near such a cutoff value should provide valid estimates of a given variable's treatment effect.

This section note follows Lee (2008) to test for the existence of an incumbency advantage in elections for the US House of Representatives. That is, do the outcomes of majority rule elections (i.e. who wins) change discontinuously when the vote share in favor of a candidate reaches 50% in a prior election?



- There are three main types of right-hand side variables that are included in a regression utilizing RD design:

  1. The Treatment Variable (i.e. the Treatment Status)
  2. The Running Variable (often interacted with treatment variable)
  3. Other Control Variables

  For example, Lee (2008) utilizes the following variables, where **voteshare** represents the share of the votes given to the Democratic Candidate in the immediately previous election, and **demowin** represents whether or not the democratic candidate won the previous election. See the Stata code below:

```
// Create running variable.
gen difshare = voteshare - 50

// Create indicator for Democrat win
gen demowin = (difshare > 0)

// Interact difshare with Democrat win.
gen difshare_dwin = (difshare * demowin)

// Create quadratic difshare term
gen x2 = (difshare * difshare)

// Interact quadratic term with Democrat win
gen x2_demowin = (x2 * demowin)
```

Q1: What is the treatment variable in Lee (2008)? What values can the treatment value take on?

Q2: Which variable should Lee (2008) use as a running variable?

Q3: What should the value of the running variable be at the cutoff of the treatment variable? Why?

Q4: Write down the RD regression model with a linear control function.

Q5: Write down the RD regression model with a quadratic control function.

Q6: Estimate OLS, Probit, and Logit RD models with linear and quadratic control functions.

Q7: According to the OLS model with a quadratic control function, what is the RD estimate of the change in probability of a Democrat winning the next election at the vote share threshold?

Q8: According to the Probit model with a quadratic control function, what is the RD estimate of the change in probability of a Democrat winning the next election at the vote share threshold?

Q9: According to the Logit model with a quadratic control function, what is the RD estimate of the change in probability of a Democrat winning the next election at the vote share threshold?

7 Appendix: Odds-Ratio

In economics, reporting predicted effects, differences in predicted probabilities, and average derivatives are the most common way to report results from probit and logit regressions. However, in other fields, like medicine, *odds ratios* are more common. In statistics, odds are defined as the probability of an event occurring divided by the probability of it not occurring. For an event with an equal probability of occurring $P(Y = 1) = 0.50$ we would write the odds of that event as 1:1 or just 1. To get the odds of an event, they will be equal to $\frac{p}{1-p}$, to get back to the probability $\frac{odds}{1+odds}$. When Y is binary:

$$Odds = \frac{P(Y_i = 1|X_i)}{P(Y_i = 0|X_i)} = \frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)} \tag{4}$$

In some contexts, it's useful to know the odds of something happening given some event compared to the odds of it happening without that event. This is what motivates the usage of the odds ratio. Say, for example, we want to know the odds ratio of someone developing lung cancer if they smoke, compared to the odds of them developing lung cancer if they do not smoke.

Let's examine a simple example of how odds, odds ratios, and probabilties can all be useful when working with logit models.

- Remember, with the logit model the conditional expectation is modeled as:

$$P(Y_i = 1|X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i})}} \tag{5}$$

- However, it's possible to algebraically manipulate the conditional expectation equation to the following:

$$ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \beta_1 X_{1i} \tag{6}$$

Equation six will give us the log-odds of an event occurring. Notice that after this transformation, the coefficients have a much cleaner interpretation. A one unit change in X_{1i} is associated with a, on average, β_1 unit change in the log odds that $Y = 1$. This cleaner interpretation of the coefficients makes log-odds nice to work with.

Odds, probabilities, and log-odds are three different ways of describing the same events. But notice that there are nice relationship between the three different measures. Say for example we observe the following:

	Situation A	**Situation B**	**Situation C**
probability	0.20	0.50	0.80
odds	0.25	1.00	4.00
log-odds	-1.386	0	1.386

If you stare at the table, you might notice that moving from column a to column b, the odds increases by a factor of four each time, but the log-odds increase by a constant factor of 1.386. The odds ratio across all three columns is also a constant, 4.

If we had started out with the odds ratios, for example, we could have used those to find the difference in log-odds between columns. The difference in log-odds is equal to $log(oddsratio) = log(4) = 1.386$.

To go from the the difference in log odds to the odds ratio all we need to do is apply the exponential function. So, odds ratio $= e^{(\Delta log(odds))} = e^{1.386} = 4$. We have just uncovered something really interesting! notice that if we performed the exponential function on both sides of equation six, the left hand side would be equal to the odds that $Y = 1$ so we could say that a one unit increase in X_{1i} is associated with a, on average, e^{β_1} unit increase in the odds that $Y = 1$.

This simple exercise is helpful for a few reasons, it provides a new way to look at the logit model and also gives sheds insight on how econometric techniques are used in other fields.