Logistic Regression

Theory and Application using Stata

Najib A. Mozahem

Contents

Logistic Regression – The Theory

Contingency Tables

Two-by-Two Tables

When the outcome that we are interested in can take on one of two values, the variable is referred to as a binary variable. As an example, consider the data shown in **Table 1**. The table shows the records for 31 students, where the first column indicates whether the student has withdraw from or completed a certain course, while the second column shows the major of the student. In this case, the outcome of interest is whether the student completed the course or whether he/she withdrew from the course. These are the only two possible outcomes. Hence, the variable is binary. The other variable is also binary since it also has two possible values: engineering and business. The question that we would like to ask is whether students from both colleges are equally likely to withdraw from the course. To find the answer, we create a two-by-two table. The table is shown in **Table 2**. This table sums up the results. We see that there is a total of 16 business students, four of whom withdrew from the course. We also see that there is a total of 15 engineering students, nine of whom withdrew from the course. This means that a proportion of $4/16 = 0.25$ of business students withdrew from the course, compared to a proportion of $9/15 = 0.6$ of engineering students. If an engineering student enrolls in the course, we calculate that the probability that he or she will withdraw from the course is 0.6. If a business student enrolled in the course, we calculate that the probability that he or she will withdraw from the course is 0.25. Therefore, we see that engineering students are more likely to withdraw from the course than business students.

The Odds Ratio

In order to better compare the two groups, we can use the concept of **odds ratios**. To do that, we need to calculate the odds that a student will withdraw from the course. This can be done using the equation:

$$odds = \frac{(Probability\ of\ withdrawal)}{1 - (probability\ of\ withdrawal)}$$

The odds of withdrawal for an engineering student is $0.6/(1\text{-}0.6) = 1.5$. The odds of withdrawal for a business student is $0.25/(1\text{-}0.25) = 0.33$.

The odds are never negative. They are zero or greater than zero. When the odds are equal to one, this means that the probability of both outcomes are equal ($0.5/(1\text{-}0.5) = 1$). In the case of engineering students, since the odds of withdrawal are 1.5, this means that the probability of withdrawal is 1.5 times the probability of finishing the course. For business students, the probability of withdrawal is 0.33 times the probability of finishing the course.

Table 1 Records of students

Outcome	**College**
Withdraw	Engineering
Withdraw	Engineering
Finish	Business
Finish	Business
Finish	Business
Finish	Engineering
Finish	Engineering
Finish	Engineering
Withdraw	Engineering
Finish	Business
Withdraw	Engineering
Finish	Business
Withdraw	Engineering
Withdraw	Engineering
Finish	Business
Finish	Business
Withdraw	Business
Withdraw	Business
Withdraw	Business
Finish	Business
Finish	Engineering
Withdraw	Engineering
Finish	Business
Finish	Business
Finish	Engineering
Withdraw	Business
Withdraw	Engineering
Withdraw	Engineering
Finish	Engineering
Finish	Business
Finish	Business

Now that we have the odds for each row in **Table 2**, we can calculate the odds ratio:

$$odds\ ratio = \frac{odds_{engineering}}{odds_{business}} = \frac{1.5}{0.33} = 4.5$$

Since the odds cannot be negative, the odds ratio cannot be negative as well. When the odds of an event are equal in both rows (**Table 2**), the odds ratio will be equal to one. When the odds of the numerator is greater than the odds of the denominator, the odds ratio will be greater than one. This means that the probability of the event is higher in the row that is associated with the numerator. In our case, the odds ratio is 4.5, which is larger than one. This means that the probability of the event,

Table 2 Cross classification of college and outcome

	Outcome		
College	Withdraw	Finish	Total
Business	4	12	16
Engineering	9	6	15

which is withdrawal in our case, is higher in the numerator, which is engineering students in our case. This means that engineering students are more likely to withdraw from the course than business students. If, on the other hand, the odds ratio is less than one, then this means that the probability of the event in the denominator are higher.

As you can see, the odds ratio allows us to compare the incidence of an event between groups. If the event is equally probable in both groups, then the odds ratio will be one. In such a case, we say that the event (withdrawal in our case) and the group (college in our case) are independent. This means that withdrawal does not depend on college.

Two-by-Three Tables

The above logic remains intact when instead of a binary variable such as college, we have a variable that divides students into the groups sophomore, junior, and senior. In this case, the outcome variable is binary, but the other variable is not, since it divides the students into more than two groups.

As an example, consider the data shown in **Table 3**. The odds of withdrawal for sophomores, junior, and senior students are:

$$odds_{sophomore} = \frac{12/32}{1 - (\frac{12}{32})} = 0.6$$

$$odds_{junior} = \frac{6/36}{1-(\frac{6}{36})} = 0.2$$

$$odds_{senior} = \frac{5/30}{1 - (\frac{5}{30})} = 0.2$$

Since the odds for all groups are less than one, then the probability of course withdrawal is less than the probability of finishing the course in each of them. We can also compute the odds ratios in order to compare the odds of each group:

$$\frac{odds_{sophomore}}{odds_{junior}} = \frac{0.6}{0.2} = 3$$

$$\frac{odds_{sophomore}}{odds_{senior}} = \frac{0.6}{0.2} = 3$$

$$\frac{odds_{junior}}{odds_{senior}} = \frac{0.2}{0.2} = 1$$

Table 3 Cross classification of standing and outcome

	Outcome		
Standing	Withdraw	Finish	Total
Sophomore	12	20	32
Junior	6	30	36
Senior	5	25	30

The results indicate that the probability of withdrawal are highest in sophomore students.

The above exercise is useful when we want to compare the probabilities of an event across certain groups. We saw that the probability of withdrawal is affected by the major of the student. In the second example, we saw that the probability of withdrawal was affected by whether the student was a sophomore, junior, or senior.

This type of analysis however will not take us very far. The reason is that usually, we are interested in studying the effect that several variables have on the outcome. What if we wanted to see whether the withdrawal rate was affected by the major, standing, and GPA, all at the same time? In this case, we need to use the statistical technique of logistic regression.

Logistic Regression

One Independent Variable

We will start by considering the simplest case in which there is a single independent variable. In linear regression, the model is represented by the linear equation:

$$y = ax + b$$

In the above equation, y is the dependent variable, x is the independent variable, a is the slope, and b is the y-intercept. One of the nice things about linear regression is how easy it is to interpret the relationship between the dependent variable and the independent variable. As an example, assume that we have the following linear equation:

$$y = 3x + 2$$

If x is equal to 2, y will be equal to 8, and if x is equal to 3, y will be equal to 11. Note that for every one unit increase in x, the value of y increases by 3, which is the value of the slope. This is the definition of the slope. It is the amount by which the dependent variable changes when the independent variable increases by one. The slope is important for two reasons. The first reason relates to the sign. If the slope is positive, then any increase in the independent variable will lead to an increase in the dependent variable. The more I eat, the heavier I get. If the slope is negative, then an increase in the independent variable will lead to a decrease in the dependent variable. The more I buy food, the less money I have.

The second reason relates to the magnitude of the slope. The larger the magnitude of the slope, the greater the effect that the independent variable has on the dependent variable. If the slope is 2, then a one unit increase in the independent variable will result in an increase of 2 in the dependent variable. If, however, the slope is 10, then a one unit increase in the independent variable will result in an

increase of 10 in the dependent variable. So the sign of the slope tells us about the direction of the relation and the magnitude tells us about the magnitude of the effect that one variable might have on the other.

Unfortunately, in logistic regression things are not that simple. The reason is that the logistic regression model has the following form:

$$\log(\frac{p}{1-p}) = ax + b$$

In the above equation, p is the probability that the event will happen. As we can see, instead of the left hand side of the equation being the dependent variable, what we have is a strange function that is the log of the odds. This function is called the logit function, hence the name logistic regression. This means that the interpretation of the slope a is that when x increases by one unit, the log of the odds increases by one. This doesn't make much sense. Fortunately, there is something that we can do to make the interpretation more intuitive. All we need to do is to take the exponential of both sides:

$$e^{\log(\frac{p}{1-p})} = e^{ax+b}$$

$$\frac{p}{1-p} = e^{ax+b}$$

There is nothing complicated in what we did. We know from algebra that an equality is maintained when we perform the same operation to both sides. In our case, we first took the exponent of both sides. We then took advantage of the rule $e^{\log(k)} = k$.

Why is the new form of the equation better? Because now instead of the log of the odds we have the odds on the left hand side of the equation. Therefore, if the slope a is positive, when x increases the term e^{ax+b} will increase. Since this term is equal to the odds, this means that the odds will increase. This means that when a is positive, the odds that the event will happen will increase with increasing values of x. On the other hand, when a is negative, when x increases the odds will decrease.

Let us take an example. Assume that we perform logistic regression where the dependent variable is whether a student withdraws from a course or not and the independent variable is the number of courses that the student is currently taking. Assume that once we fit this model we get the following equation:

$$\log(\frac{p}{1-p}) = 2.21(number\ of\ courses) - 11.25$$

What this means is that when the number of courses that a student is currently taking increases by one, the logit function increases by 2.21. Since, as we said, this is hard to understand, let's consider the more intuitive form:

$$\frac{p}{1-p} = e^{2.21(number\ of\ courses)-11.25}$$

Now consider two students, one currently taking four courses, and the other currently taking five courses. According to our model, the odds that each will withdraw from a course is:

Student taking four courses: $\frac{p}{1-p} = e^{2.21(4)-11.25} = 0.0898$

Student taking five courses: $\frac{p}{1-p} = e^{2.21(5)-11.25} = 0.8187$

This means that the odds that a student taking four courses withdraws from a course is 0.0898, while the odds that a student taking five courses withdraws from a course is 0.8187. This means that the student taking five courses is more likely to withdraw. How much more likely? In order to compare, we need to calculate the odds ratio:

$$odds\ ratio = \frac{odds_{five\ courses}}{odds_{four\ courses}} = \frac{0.8187}{0.0898} = 9.12$$

What this means is that a student who is taking one more course than another student has 9.12 times greater odds of withdrawing from a course. The great news is that 9.12 is actually $e^{2.21}$. We now have a very intuitive meaning for the slope a. When we fit a logistic regression model and obtain a value for the coefficient associated with an independent variable, we know that when the independent variable x increases by one unit, the odds of the event happening is multiplied by e^a. When a is positive, e^a>1, which means that the odds increase when x increases. When a is negative, e^a<1, which means that the odds decrease when x increases.

Although the above might seem complicated, it is actually very easy. As a recap, when we fit a logistic model, we are finding a line with the equation ax + b, just like in linear regression. The difference however is in the interpretation of the coefficient of x. In linear regression, when x increases by one unit, the dependent variable increases by the magnitude of a. In logistic regression, when x increases by one unit, the odds of an event happening are multiplied by e^a. If a is zero we have $e^0 = 1$, which means that the odds are multiplied by one, so they do not change. This means that x does not affect the odds. If a is greater than zero, then e^a>1, which means that the odds are multiplied by a number greater than one, so they increase. If a is less than zero, then e^a<1, which means that the odds are multiplied by a number that is less than one, so they decrease.

To see how simple the above is, assume that we fit a logistic model where the dependent variable is whether an individual has a heart problem or not, and where the independent variable is age. Once we fit the model, we get the following result:

$$\log(\frac{p}{1-p}) = 1.09(age) - 9.68$$

Here, *p* is the probability that a person has a heart problem. What does this output mean? Since the value of the coefficient associated with the independent variable, which is age, is 1.09, this means that when age increases by one year, the odds of having a heart condition is multiplied by $e^{1.09} = 2.97$. This means that a 40-year old individual has 2.97 times greater odds of having a heart condition than an individual who is 39-years old.

As another example, consider that we fit a logistic model where the dependent variable is whether a student goes out at night during the weekdays, and where the independent variable is the student's grades. The output of the model is the following:

$$\log(\frac{p}{1-p}) = -0.24(grades) + 17.84$$

Here, the coefficient is negative. Since $e^{-0.24} = 0.79$, the output indicates that the odds that a student goes out during the weekdays are multiplied by 0.79 (so they decrease) when grades increase by a single unit. This means that students with higher grades are less likely to go out during the weekdays.

As you can see, when the coefficient is positive, the odds increase, and when the coefficient is negative, the odds decrease. Since we are mostly interested in the exponential of the coefficient, and not the coefficient itself, statistical software packages usually display the value e^a instead of displaying the value of a. In that case, when e^a is greater than one, the odds increase, and when $e^a$is less than one, the odds decrease.

Binary Variables

So far, the independent variable has been numerical in nature. Sometimes however, including variables that are not numeric in nature is necessary. For example, what if we wanted to investigate whether the probability of withdrawing from a course could be explained by the gender of the students? Here, the variable gender is not numeric. It is categorical, in that it divides the observations into categories. Since biological gender is either male or female, there are two categories in which each student might fall.

In such a case, we can create a binary variable to represent the two categories. A binary number takes on the values of zero or one. We next assign each of these values to a category. Let us assign a zero to males and a one to females. The data is shown in **Table 4**. Note that the table is similar to **Table 1** except that we have added a new column which is gender.

Now that the variable gender has been quantified, it is possible to include it in a regression model. The result of running a logistic model would be again in the form:

$$\log(\frac{p}{1-p}) = ax + b$$

If we use a statistical software to run the model, we will get the following output:

$$\log(\frac{p}{1-p}) = -1.90(gender) + 0.51$$

We already know how to interpret the coefficients of continuous variables, such as age and grades. However, what does it mean that the coefficient of gender is -1.90? Remember that for males the value of gender is zero, while for females the value of gender is one. In order to calculate the odds for a male and a female student, we need to use the form:

$$\frac{p}{1-p} = e^{ax+b} = e^{-1.90(gender)+0.51}$$

Table 4 Records of students

Outcome	Gender	Binary
Withdraw	male	0
Withdraw	male	0
Finish	male	0
Finish	female	1
Finish	female	1
Finish	female	1
Finish	female	1
Finish	male	0
Withdraw	female	1
Finish	male	0
Withdraw	female	1
Finish	male	0
Withdraw	male	0
Withdraw	male	0
Finish	female	1
Finish	female	1
Withdraw	female	1
Withdraw	male	0
Withdraw	male	0
Finish	female	1
Finish	male	0
Withdraw	male	0
Finish	female	1
Finish	female	1
Finish	male	0
Withdraw	male	0
Withdraw	male	0
Withdraw	male	0
Finish	female	1
Finish	female	1
Finish	female	1

We can now calculate the odds for each student:

Male: $\frac{p}{1-p} = e^{-1.90(0)+0.51} = 1.67$

Female: $\frac{p}{1-p} = e^{-1.90(1)+0.51} = 0.25$

From these odds, we can calculate the odds ratio:

$$Odds\ ratio = \frac{odds_{female}}{odds_{male}} = \frac{0.25}{1.67} = 0.15$$

This means that males have higher odds to withdraw than females. The nice thing is that the number 0.15 happens to be $e^{-1.90}$. This means that when we are dealing with binary variables, the exponent of the coefficient is the odds ratio when we compare an individual who belongs to the group that is assigned a value of one and an individual who belongs to the group that is assigned the value zero. In our case, since males were assigned a value of zero, the exponent of the coefficient is the odds ratio that we obtain when we divide the odds of a female by the odds of males. In other words, since the coefficient is -1.90, the odds for females is 0.15 times the odds for males.

If you recall, we had actually calculated the odds ratio for the information shown in **Table 2**, which cross-classifies the variables outcome and engineering. When we did this manually, we found that the odds ratio was 4.5. If we run the logistic model, where the value zero is assigned to business and the value one is assigned to engineering, we will get the following output:

$$\log(\frac{p}{1-p}) = 1.50(college) - 1.10$$

Since $e^{1.5} = 4.5$, we conclude that the odds of withdrawal for engineering students are 4.5 times the odds of withdrawal for business students.

Let us take another example. Assume that we run a logistic regression model where the dependent variable is whether a visitor to our website subscribes to our services or not, and where the independent variable is whether the user accessed our website using a mobile device or using a desktop computer. The independent variable is binary, so we need to assign zero to a category and a one to the other category. In our case, let's say that we chose to assign a zero to users using a mobile device and a one to users using a desktop computer. We fit the model and get the following result:

$$\log(\frac{p}{1-p}) = 1.26(device\ type) - 1.01$$

This means that users who use a desktop computer have $e^{1.26} = 3.53$ times the odds of subscribing than users who use a mobile device. Since we are again mostly concerned with the exponent of the coefficient, statistical software packages tend to display the odds ratio directly, instead of displaying the value of the coefficient in the output.

Multiple Independent Variables

Now that we have seen how to interpret the output from logistic regression when there is a single independent variable, let us see what changes when there are two independent variables. **Table 5** shows the records for students. The table includes the dependent variable outcome and the independent variables college and courses. Therefore, we have one binary variable and one continuous variable. In this case, we want to see if the dependent variable, which is withdrawing from a course, depends on the college of the student and on the number of courses. The equation of this model is:

$$\log(\frac{p}{1-p}) = a_1x_1 + a_2x_2 + b$$

Each independent variable has its own coefficient now. If we run the model, the output will be:

Table 5 The case of two independent variables

Outcome	College	Courses
Withdraw	Engineering	6
Withdraw	Engineering	6
Finish	Business	4
Finish	Business	6
Finish	Business	4
Finish	Engineering	3
Finish	Engineering	5
Finish	Engineering	6
Withdraw	Engineering	6
Finish	Business	4
Withdraw	Engineering	5
Finish	Business	4
Withdraw	Engineering	6
Withdraw	Engineering	5
Finish	Business	4
Finish	Business	4
Withdraw	Business	5
Withdraw	Business	6
Withdraw	Business	5
Finish	Business	4
Finish	Engineering	5
Withdraw	Engineering	6
Finish	Business	4
Finish	Business	4
Finish	Engineering	5
Withdraw	Business	5
Withdraw	Engineering	6
Withdraw	Engineering	5
Finish	Engineering	4
Finish	Business	3
Finish	Business	4

$$\log(\frac{p}{1-p}) = -0.02(college) + 2.22(courses) - 11.27$$

Let us now calculate the odds for two students where both of them are currently taking five courses, but one is studying business and the other is studying engineering:

Business: $\frac{p}{1-p} = e^{-0.02(0)+2.22(5)-11.27} = 0.84$

Engineering: $\frac{p}{1-p} = e^{-0.02(1)+2.22(5)-11.27} = 0.83$

This means that the odds ratio is:

$$Odds\ ratio = \frac{0.83}{0.84} = 0.98$$

A simpler way to get this value is just to calculate the exponent of the coefficient, $e^{-0.02} = 0.98$. This shows that even when there are several independent variables, the coefficients retain their meanings. Therefore, to find the difference between two groups of students, just calculate e^{a_1}. The implication of this is that in a multiple regression model, where there are several independent variables, when we want to investigate the effect that an independent variable has on the dependent variable, we just need to take into consideration the coefficient of the independent variable, given that the rest of the variables do not change.

To further illustrate this, let us now calculate the odds for two engineering students, one of whom has three courses and the other has four courses:

Three courses: $\frac{p}{1-p} = e^{-0.02(1)+2.22(3)-11.27} = 0.00975476$

Four courses: $\frac{p}{1-p} = e^{-0.02(1)+2.22(4)-11.27} = 0.08981529$

This means that the odds ratio is:

$$Odds\ ratio = \frac{0.08981529}{0.00975476} = 9.21$$

This is also obtained by finding the exponent of the coefficient, $e^{2.22} = 9.21$. This same logic applies whether we have three independent variables, four independent variables, or even nine independent variables. It also doesn't matter whether the variables are binary or continuous. The coefficient of each independent variable gives us information about the relationship between the independent variable and the dependent variable. All we have to do is to take the exponent of the coefficient in order to calculate the effect that the independent variable has on the odds of the event happening.

Categorical Variables with more than Two Categories

When we included gender in the equation, we used a binary variable since gender can take on one of two values. What if we had a categorical variable that divided the observations into more than two groups? If you recall, **Table 3** presented a cross-classification of the variables outcome and standing, where students are classified as being in their sophomore year, junior year, or senior year (The table is reproduced in **Table 6**). In this case, we cannot use a binary variable because there are three groups instead of two. What we can do however, is to use more than one binary variable, as shown in **Table 7**.

If you look at the column for the variable x_1, you will notice that the variable takes a value of one for junior, and zero otherwise. The other binary variable, x_2, takes on a value of one for senior and zero otherwise. How did we know that we need three binary variables? The number of binary variables needed is the number of categories minus one. In our case, we have three categories, so it is 3 – 1 = 2. The logit equation now becomes:

Table 6 Cross classification of standing and outcome

	Outcome		
Standing	Withdraw	Finish	Total
Sophomore	12	20	32
Junior	6	30	36
Senior	5	25	30

$$\log(\frac{p}{1-p}) = a_1x_1 + a_2x_2 + b$$

For a sophomore student, x_1 and x_2 are zero. For a junior student, x_1 is one and x_2 is zero. For a senior student only x_2 is one and x_1 is zero. If we fit this model to the data, the output will be:

$$\log(\frac{p}{1-p}) = -1.10x_1 - 1.10x_2 - 0.51$$

Let us now calculate the odds for the three types of students:

Sophomore:

$$\frac{p}{1-p} = e^{-1.10(0)-1.10(0)-0.51} = 0.6$$

Junior:

$$\frac{p}{1-p} = e^{-1.10(1)-1.10(0)-0.51} = 0.2$$

Senior:

$$\frac{p}{1-p} = e^{-1.10(0)-1.10(1)-0.51} = 0.2$$

We can now calculate the odds ratios:

$$\frac{odds_{junior}}{odds_{sophomore}} = \frac{0.2}{0.6} = 0.33$$

$$\frac{odds_{senior}}{odds_{sophomore}} = \frac{0.2}{0.6} = 0.33$$

We can get the same values by calculating the exponents of the coefficients:

$$e^{-1.1} = 0.33$$

Table 7 Coding the categorical variable

	x_1	x_2
Sophomore	0	0
Junior	1	0
Senior	0	1

We see that the exponent of the coefficient for each variable produces the odds ratio when we compare the group associated with the variable to the base group, which is the group that is assigned the values of zero. In other words, in our example, sophomore students are the base, or referent, group, since they have a value of zero for both x_1 and x_2. Junior students have a value of one for x_1, which means that the exponent of the coefficient of x_1 is the odds ratio of junior students to sophomore students. Senior students have a value of one for x_2, which means that the exponent of the coefficient of x_2 is the odds ratio of senior students to sophomore students. Therefore, just like in the case of binary variables, the coefficient compares a group to another group. The only difference here is that there is more than one binary variable, where each is associated with a different group. In both cases, the referent group is the same.

Nonlinearity

In linear regression, the relationship between the independent variable and the dependent variable is expected to be linear. If it is not, then we need to account for the linearity by including a power term, such as the quadratic term. In the case of linear regression, detecting nonlinearity is easy, since all we have to do is to produce a scatter plot of the dependent variable against the independent variable. In the case of logistic regression, the equation is not y = ax + b. Instead, it is:

$$\log(\frac{p}{1-p}) = ax + b$$

This means that the logit function, which is the log of the odds, is linear with respect to the independent variable. When we have a continuous variable as an independent variable, we need to test this assumption of linearity. We can perform a graphical test and a non-graphical test.

Box-Tidwell Test

The non-graphical way is to use the Box-Tidwell test. As an example, assume that the dependent variable is whether a customer buys from our website or not (buy), and the independent variable is the previous number of visits of the customer to our website (visit). To test the assumption of linearity between the logit function and the independent buy using the Box-Tidwell test, we should create a new variable using the following formula:

$$new\ variable = visit\ X\ log(visit)$$

This means that the new variable is the product of the independent variable and the log of the independent variable. After we calculate this new variable, we should fit a new logistic regression model that includes both the variable (visit) and the new variable. If the new variable turned out to

have a p-value that is less than 0.05, i.e. if the variable was significant, then the assumption of linearity between the logit function and the independent variable is violated.

Graphical Tests

Lowess

Although the Box-Tidwell test is very useful, it does not inform us of the shape of the nonlinearity. It only tells us if the relationship is not linear. However, we would also like to know what sort of nonlinearity exists. In linear regression, the graphical method is basically a scatter plot. In linear regression, this graph does not inform us about the nonlinearity. To illustrate, let us take the example in **Figure 1**. The graph looks different from the scatter plots that we are used to. This is because the variable outcome, which is the dependent variable, can only take on two values, either a zero or a one. This is why the dots lie along one of two lines, the outcome equal zero line and the outcome equal one line. The graph, however, can be informative, as shown in **Figure 4**, which plots the loess curve of the data. A loess plot is simply a smoothed scatter plot. Therefore, it allows us to visualize the relationship when the scatter plot is not very clear. This is an extremely useful graph when the variable on the y-axis is binary. As you can see from **Figure 4**, the curve initially increases with increasing values of the dependent variable x, and then levels of when x reaches a value of four. Therefore, we conclude that the relationship between the logit function and the independent variable x is not linear.

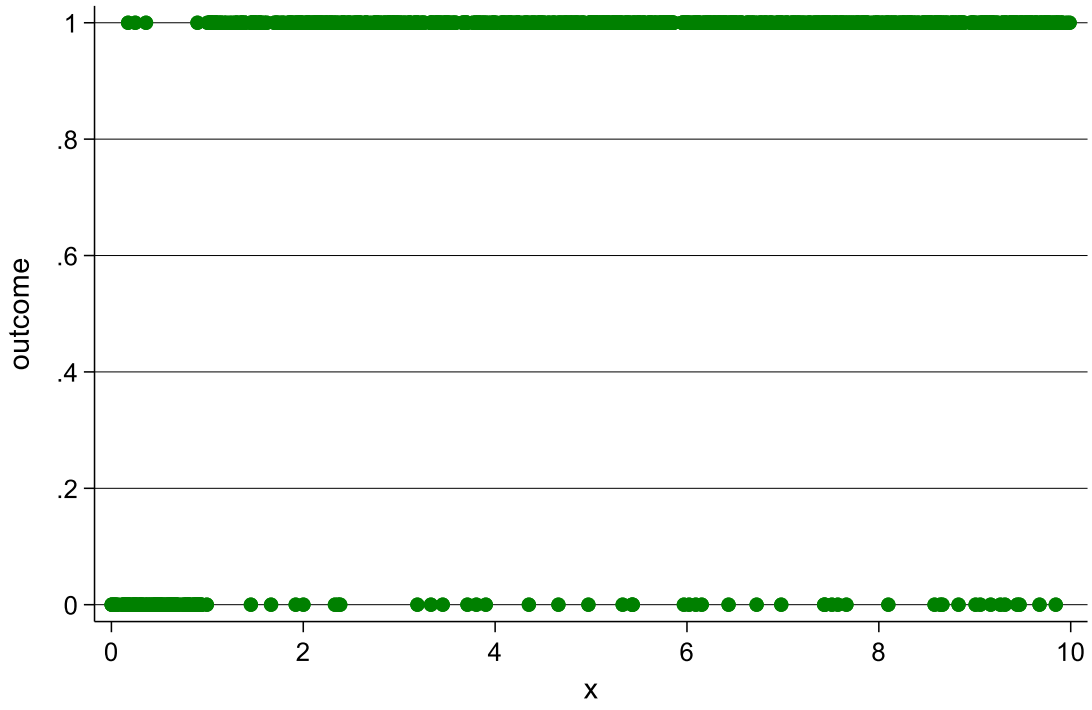

Figure 1 Scatter plot of a binary outcome and a continuous variable x

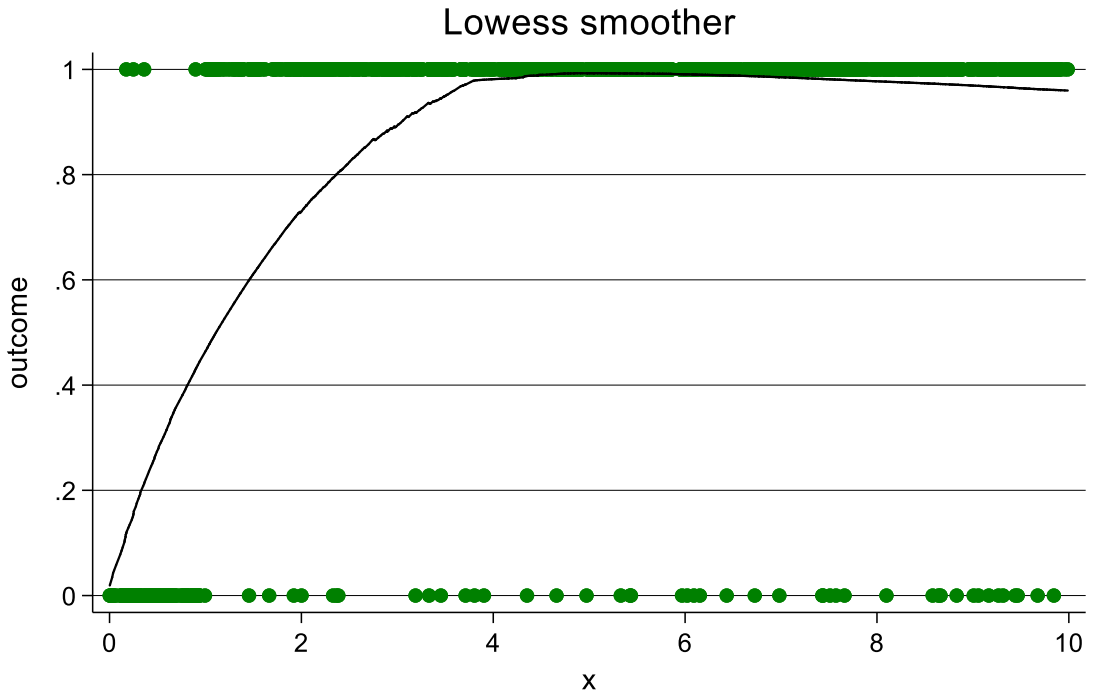

Figure 2 Loess graph of a binary outcome and a continuous variable x.

Linearity of Slopes

Another graphical test is the linearity of the slopes test. The idea behind this test is that the independent variable x is categorized. This means that instead of having a continuous independent variable, we end up with a categorical variable. Assume for example that we want to categorize the variable age, where age is between 18 years and 60 years. We create a new variable that takes on the value of zero when age is between 18 and 30, the value one when age is between 30 and 40, the value two when age is between 40 and 50, and the value three when age is greater than 50 (**Table 8**). Once we have our new categorized variable, we can fit a logistic regression with the categorized variable as the independent variable. Therefore, instead of having a continuous variable in the model we now have a categorical variable. We have already seen how to interpret the results obtained from including a categorical variable. Once the model is fit, we plot a graph of the predicted value of the logit function

Table 8 Categorizing a continuous variable

categorized age	age
0	18≤age<30
1	30≤age<40
2	40≤age<50
3	50≤age

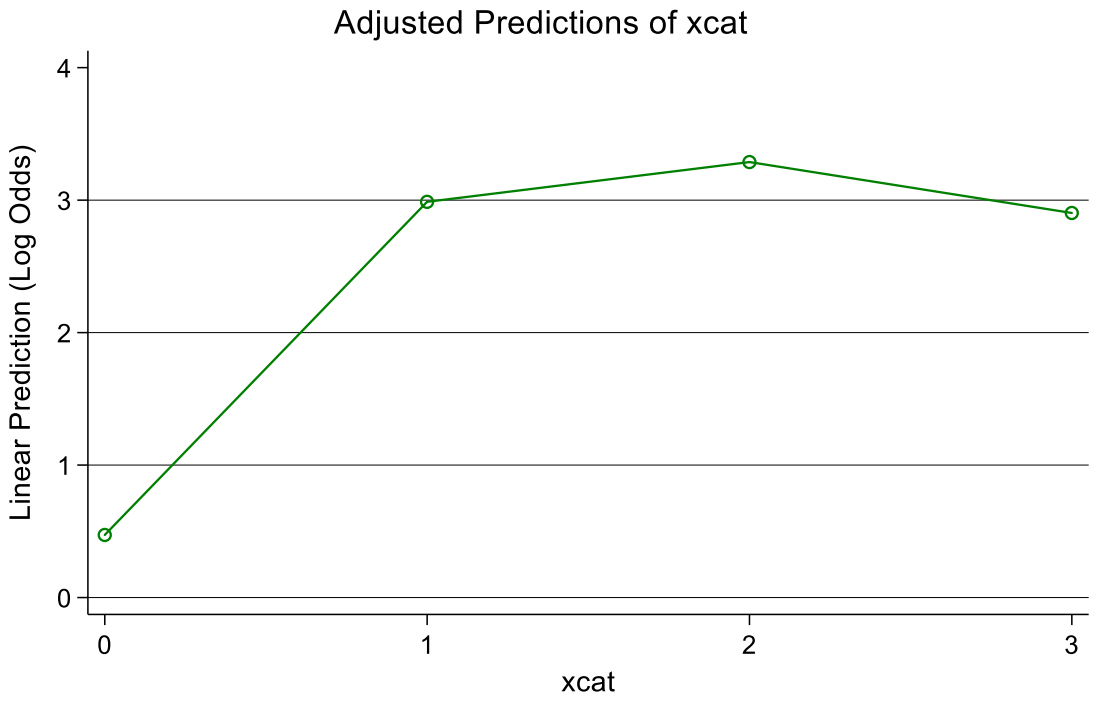

Figure 3 Linearity of slopes test.

and the different levels of the categorical variable. **Figure 5** shows the graph that is produced for the same data that was used to produce **Figure 2**, where the independent variable x has been categorized and named xcat. Since when we fit the logistic model we obtained the value of the coefficient for each level, we calculate the value of the logit function for each coefficient. If the relationship between the independent variable and the logit function is linear, then the graph should resemble a line. This is clearly not the case. In fact, there is a considerable amount of similarity between **Figure 4** and **Figure 5**.

Selection of Independent Variables

An important issue that we face when we have a number of independent variables is how to decide which variables to add to the model and in what order. There are generally four ways to do this. The first three all rely on an algorithm and you are advised not to trust them. This is a very important point. You should never let the computer pick the independent variables. However, the three methods will be described since many statistical packages allow the user to use them. In addition, I do not think that there is anything wrong with using them as an investigative tool, i.e. in order to get an idea of what independent variables are significant and which are not. The first selection method is referred to as forward selection. As the name suggests, this method adds independent variables one step at a time. Originally, we start with no independent variables. The algorithm then adds one of the variables. If the p-value of that variable turns out to be less than 0.05, the variable is kept in the model. The

algorithm then selects another variable and adds it to the model. These models are repeated until there are no further independent variables left.

The second selection method is referred to as backward elimination. As you can imagine, we start with a model that includes all possible independent variables. The algorithm them selects the least significant independent variable (the one with the highest p-value). If the p-value of the selected independent variable is greater than 0.05 (which means that it is not significant) the variable is removed. The algorithm then repeats and selects the least significant variables from the ones that are still in the model. These steps are repeated until all variables that are included in the model have p-values that are less than 0.05 (which means that they are all significant).

The third selection method is referred to as stepwise regression. This method is a combination of the previous two. The model starts in forward mode with no independent variables. The algorithm selects the most significant independent variable and adds in to the model. Next, the algorithm goes into backward mode by checking to see whether any variable can be eliminated. Next, the algorithm goes back into forward mode and selects a variable from the pool of remaining variables, and then it goes back into backward mode. This process continues until there are no more variables to be added or dropped.

As I said, the above three algorithms should not be used to find the final model. You can, however, initially use them in order to get an initial picture of which independent variables are selected and which are not. As an initial step, there is nothing wrong with doing this. Ultimately however, you need to rely on the fourth method to select when and how to add the variables, and that method is to use your knowledge. Any good research must be informed by theory. The better you understand the theory, the better you can determine which variables to include and which to ignore. In general, we prefer models in which the number of independent variables is as small as possible. In linear regression we can rely on the value of R-squared, or adjusted R-squared, when choosing between two models. Although some statistical packages display a statistic that is called pseudo R-squared when you run a logistic model, this statistic does not have the same meaning as R-squared does in linear regression, so you should not pay attention to it. We can, however, rely on the AIC and BIC statistics when comparing two models. These statistics can be easily calculated by statistical software. When comparing two models, we tend to favor the one with smaller values of both AIC and BIC statistics.

Prediction

In linear regression, because the left-hand side of the equation is the dependent variable y, we can easily calculate the predicted value of y and then plot it on the y-axis. In logistic regression however, the left hand side is not the dependent variable:

$$\log(\frac{p}{1-p}) = ax + b$$

Once we fit the logistic regression model, we are able to calculate the values of a and b. This means that the equation will have one unknown in it, and this unknown is p, which is the probability that the event will happen. Since we have one equation and one unknown, we can find the value of the unknown. This means that using logistic regression we can, for each observation, calculate the probability that the event will occur. Once we calculate the predicted probability, we can produce

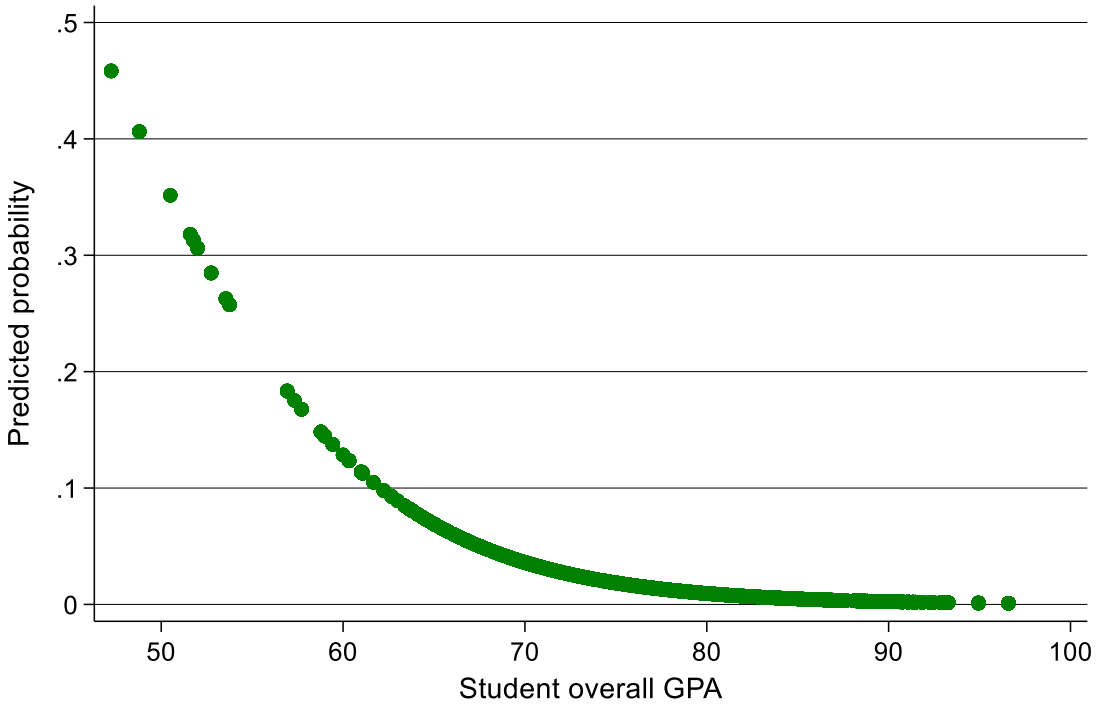

Figure 4 Visualizing the result of logistic regression: Relationship between the probability of withdrawing from a course and student GPAs.

graphs in which the predicted probability is plotted on the y-axis and any of the independent variables can be plotted on the x-axis. This will allow us to visualize how the probability of an outcome changes with changing values of the independent variable. An example is shown in **Figure 4**, where we can see that the probability of withdrawing from a course decreases as the GPA increases.

Goodness of Fit

Once we have chosen the variables that we wish to include in the model, we should test how effectively the model describes the outcome variable. There are several ways to do this. In order to illustrate each of these ways, consider the data shown in **Table 9**. The output of running a logistic model on this data will be:

$$\log(\frac{p}{1-p}) = 2.21(courses) - 11.25$$

By now we know that this means that when the number of courses increases by one, the odds of withdrawing is multiplied by $e^{2.21} = 9.12$. The odds, therefore, increase when the course loads increases.

Likelihood Ratio Test

This test compares our model with a constant-only model. In other words, this test checks whether the model with the chosen independent variables in significantly better than a model that contains no independent variables. If the result of the test is statistically significant ($p < 0.05$), then we reject the null hypothesis that both models are the same, and we conclude that the model with the independent variables is significantly better. Otherwise, if $p \geq 0.05$, we cannot reject the null hypothesis, thereby we conclude that the model with the added independent variables does not do significantly better than the model with no added variables. With regards to the dataset shown in **Table 9**, running a logistic

Table 9 The dependent variable is outcome and the independent variable is courses

Outcome	**Courses**
Withdraw	6
Withdraw	6
Finish	4
Finish	6
Finish	4
Finish	3
Finish	5
Finish	6
Withdraw	6
Finish	4
Withdraw	5
Finish	4
Withdraw	6
Withdraw	5
Finish	4
Finish	4
Withdraw	5
Withdraw	6
Withdraw	5
Finish	4
Finish	5
Withdraw	6
Finish	4
Finish	4
Finish	5
Withdraw	5
Withdraw	6
Withdraw	5
Finish	4
Finish	3
Finish	4

model will result in a p-value that is less than 0.05, thus indicating that the model does significantly better than a constant-only model.

Hosmer-Lemeshow GOF Test

The Hosmer-Lemeshow test is considered to be one of the best ways to assess the fit of a logistic model. What this test does is that it divides the dataset into groups (usually ten groups), and then compares the observed and fitted values within each group. If there is considerable discrepancy between the observed values and the fitted values, the Hosmer-Lemeshow statistic will be large, and this will result in a small p-value. What we ideally want to see is that the discrepancy between the observed and the fitted values is small, thereby resulting in a small Hosmer-Lemeshow statistic, which would result in a large p-value. This means that in this test, the null hypothesis is that the model fits. If the p-value is less than 0.05, then we reject the null hypothesis and we conclude that the model is not a good fit.

If we conduct this test on the data in **Table 9**, we will get a p-value of 0.0565 which is only slightly greater than the cut-off value of 0.05. Since the p-value is greater than 0.05, we cannot reject the null hypothesis that the model is a good fit.

Classification Tables

An intuitive way of determining whether the model is well-fit or not is to compare the predicted outcome with the actual observed outcome. However, before doing that, we need to determine the point at which the model predicts that the outcome will occur. We know that after fitting the logistic model we can calculate the probability that the outcome will occur for each observation. In order for us to be able to construct a classification table, we need to determine the probability above which we would consider that the outcome value has occurred. For example, if the predicted probability of the outcome for an observation is 0.88, do we consider that the model predicts that the outcome will occur? What about if the probability was 0.52? Usually, a cut-off value of 0.5 is used. If the predicted probability is greater than 0.5, then the model predicts that the dependent variable will have a value of one (outcome will occur).

A better way to determine the cutoff value is to actually let the data inform us of the best value to use. The idea here is to calculate the sensitivity and the specificity of the model. Sensitivity represents the probability that the model will correctly predict that the outcome has occurred. For example, if the outcome has occurred in 150 of the observations, and the model correctly predicts 140 of them, then the sensitivity is 140/150, which is 93.33%. Specificity on the other hand represents the probability that the model will correctly predict that the outcome has not occurred. For example, if the outcome has not occurred in 200 of the observations, and the model correctly predicts 170 of them, then the

Table 10 Classification table

	Observed		
Classified	Outcome = 1	Outcome = 0	Total
Outcome = 1	7	2	9
Outcome = 0	6	16	22
Total	13	18	31

sensitivity is 170/200, which is 85.00%. The ideal cutoff value is the one at which sensitivity and specificity are equal.

As an example, **Table 10** shows the classification table for the logistic regression model that is fit using the data in **Table 9**. We see that the model correctly predicts seven cases where the observed outcome variable is one, which means that the sensitivity is 7/13 which is 53.85%, and 16 of the cases where the outcome variable is zero, which means that the specificity is 16/18 which is 88.89%. In two cases, the model predicts a one where the observed value is a zero, and in six cases the model predicts a zero where the observed value is one. Therefore, the model correctly classifies (7+16)/31 = 74.19% of the observations. This is considered to be an acceptable value.

ROC Analysis

Another way to test the model fit is to use ROC curves, where we are interested in the area under the curve. This area, which ranges from zero to one, is a measure of the model's ability to discriminate between observations where the outcome of interest is experienced and observations where the outcome of interest is not experienced. The higher the value, the stronger the ability of the model to discriminate. As a general guideline:

- ROC = 0.5: No ability to discriminate
- ROC is between 0.7 and 0.8: Acceptable discrimination
- ROC between 0.8 and 0.9: Excellent discrimination
- ROC greater than 0.9: Outstanding discrimination

If we calculate the area under the ROC curve for the data in **Table 9**, we will find it to be 0.88, thus indicating that the model has an excellent ability to discriminate.

Residual Analysis

Residuals are the difference between the observed outcome and the model's predicted outcome. As you can imagine, a well-fit model should have small residual values. In linear regression, residual analysis is extremely important because linear regression makes strong assumptions about the residuals. In the case of logistic regression, we need to look at the size of the residuals in order to see whether there might be influential observations that are biasing our results.

There are many types of residuals that are used in logistic regression. However, the three most commonly used ones are the standardized residuals, deviance residuals, and the DeltaX residuals. Just like in linear regression, these statistics are plotted against the predicted variable in order to visualize the results.

Influential Observations

In linear regression, the three statistics that are used to measure influence are DFBETAS, DFFITS, and Cook's D statistics. High magnitudes of these statistics indicate that an observation is influential. In logistic regression, we use the hat diagonal statistic and the delta-beta influence statistic in order to measure influence. Just as in the case of the residual statistics described above, these statistics are plotted against the predicted probabilities in order to visualize the results.

Table 11 Guidelines for residual and influence statistics

Measure	Value above which there might be a problem
Deviance residual	Greater than two
DeltaX residual	Greater than four
Hat diagonal statistic	Greater than two times the average hat statistic
Delta-beta influence	Greater than one

Although there are no fixed-set of rules with regards to determining the values that determine whether an observation is an outlier or whether it is influential, **Table 11** offers some general guidelines that are useful in many situations.

Logistic Regression: Application

We now have the necessary tools that allow us to analyze a dataset where the dependent variable takes on two values. In this section, we will be looking at the dataset logistic_project.dta. This dataset contains the following variables:

- withdraw: this is the dependent variable which records whether the student withdrew from the course or finished the course (zero means continued, one means withdraw)
- college: whether the student is in the engineering school or the business school (zero means business, one means engineering)
- gender: whether the student is a male or a female (zero means female, one means male)
- gpa: overall GPA of the student
- semester: records whether the course was taken in the spring, fall, or summer semester (zero means fall, one means spring, two means summer)
- level: records whether the level of the course (zero means remedial, one means one-hundred level course, two means two-hundred level course, three means three-hundred level course, four means four-hundred level course, and five means five-hundred level course)

Univariable Tests

The first thing that we should do when conducting regression analysis is to perform univariate analysis, where we try and uncover whether there is a relationship between the dependent variable and each independent variable separately. Once we have a good idea about the nature of these individual relationships, we can start building the model.

Continuous Variables

In linear regression, when we have a continuous independent variable, we start our analysis by plotting a scatter plot. Graphs are also useful as a starting step in logistic regression, but their shape is different from what we are used to due to the nature of the dependent variable. For example, let us tell Stata to produce a scatter plot of the dependent variable withdraw and the continuous independent variable GPA:

```
scatter withdraw gpa
```

The graph produced by the command is shown in **Figure 5**. The reason that the graph looks different is that the variable withdraw can only take on two values, either a zero or a one. This is why the dots lie along one of two lines, the withdraw equal zero line and the withdraw equal one line. The graph, however, is informative. Notice, for example, that the students who have a value of one for the withdraw variable (these are the students who withdraw from the course) tend to have a GPA that is lower than 85. Students with higher GPAs do not seem to withdraw from courses. However, this does not mean that all students with a low GPA withdraw. If you look at the lower horizontal line, we see

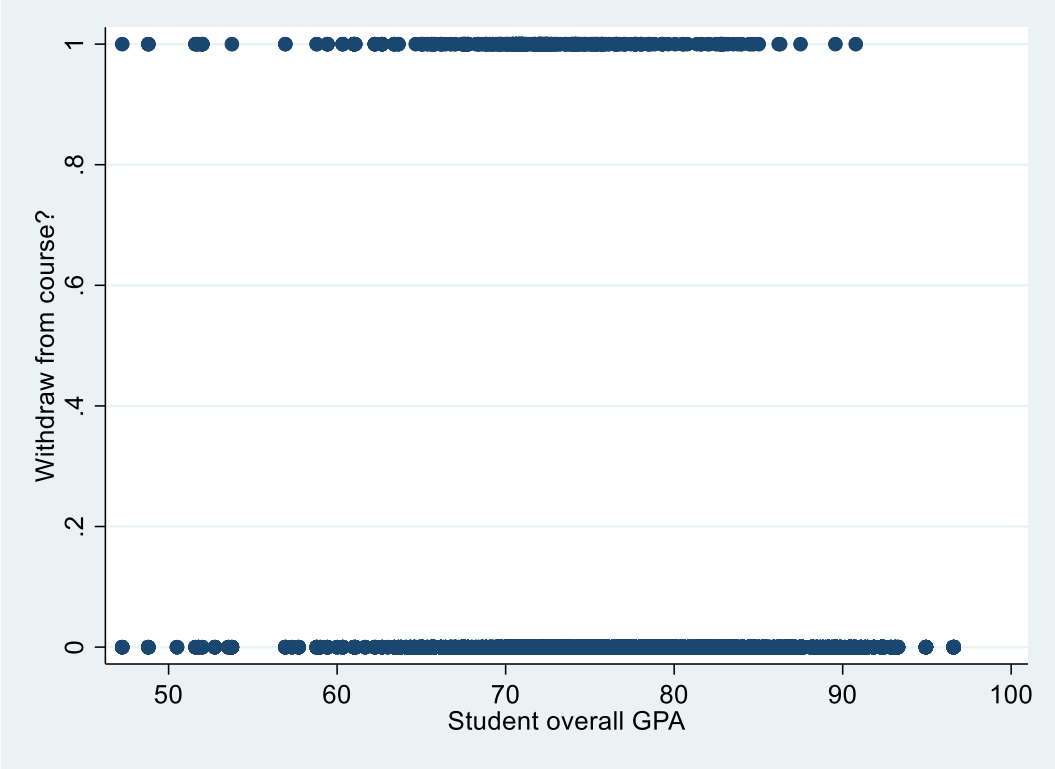

Figure 5 Scatter plot of withdraw and GPA.

that the range of GPAs is very wide. We have students who have very low GPAs but who didn't withdraw from the course, we have students with average GPAs who did not withdraw from the course, and we have students with very high GPAs who did not withdraw from the course. The difference between the two horizontal lines is that there is an absence of very high GPAs in the line at the top.

We can investigate this further by telling Stata to calculate the average GPA for students who withdraw from courses and then to compare it to the GPA of students who do not withdraw:

```
by withdraw, sort: summarize gpa
```

The **summarize** command tells Stata to display summary statistics about the variable GPA. We use the **by** command in order to tell Stata to produce the statistics separately for different groups. In our case, the groups are determined by the variable withdraw. So we are basically telling Stata to calculate the summary statistics of GPA twice, once for students who have a value of zero for withdraw, and another time for students who have a value of one for withdraw. The output of running the command is shown in **Figure 6**. We can see that the average GPA of students who did not withdraw is 77.38 while the average for students who did withdraw is 71.19.

We next fit a logistic model where the only independent variable is GPA:

```
-> withdraw = No withdraw
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gpa	24,656	77.38047	6.622922	47.25	96.59

```
-> withdraw = Withdraw
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gpa	504	71.18786	6.447055	47.25	90.78

Figure 6 Summary statistics of the variable GPA for students who withdraw and those who do not withdraw.

```
logistic withdraw gpa
```

The **logistic** command is used to tell Stata to fit a logistic regression model. The command is directly followed by the name of the dependent variable and then by the independent variable. The output of the command is shown in **Figure 7**. The output displays the odds ratio. The value of this odds ratio indicates that an increase of one-unit in the GPA causes the odds to be multiplied by 0.87. This means that the odds decrease by 13%.

We can tell Stata to display the coefficient by specifying the **coef** option:

```
logistic withdraw gpa, coef
```

```
Logistic regression                             Number of obs     =     25,160
                                                LR chi2(1)        =     425.42
                                                Prob > chi2       =     0.0000
Log likelihood = -2257.0651                     Pseudo R2         =     0.0861
```

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	.8718886	.0057075	-20.94	0.000	.8607737	.8831471
_cons	550.52	259.6654	13.38	0.000	218.4159	1387.592

```
Note: _cons estimates baseline odds.
```

Figure 7 Summary statistics of the variable GPA for students who withdraw and those who do not withdraw.

```
Logistic regression                                     Number of obs     =     25,160
                                                        LR chi2(1)        =     425.42
                                                        Prob > chi2       =     0.0000
Log likelihood = -2257.0651                             Pseudo R2         =     0.0861
```

withdraw	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	-.1370936	.0065461	-20.94	0.000	-.1499237	-.1242635
_cons	6.310863	.471673	13.38	0.000	5.386401	7.235325

Figure 8 Summary statistics of the variable GPA for students who withdraw and those who do not withdraw.

The output is shown in **Figure 8**. The value of the coefficient is -0.1371. We already know that the odds ratio is $e^{-0.1371} = 0.87$, which is the odds ratio displayed when we ran the model without the **coef** option.

The output also shows that the p-value of GPA is less than 0.05, indicating that the result is significant. Therefore, it seems that including the variable in the model is a good idea. However, as discussed in the theory section of this course, when we have continuous variables we need to test the assumption of linearity. We know that the form of the logistic model is:

$$\log(\frac{p}{1-p}) = ax + b$$

This means that the independent variable is linear with respect to the logit function. As also discussed in the theory section, there are three ways to test this assumption: the Box-Tidwell test, the loess curve, and the linearity of sloped test. We will perform each of these three tests.

Box-Tidwell Test

To perform this test we need to create a new variable and to include this variable in the logistic regression model:

```
gen boxtid = gpa*ln(gpa)

logistic withdraw gpa boxtid
```

The new variable is the product of GPA and the log function of GPA. When the variable is included in the **logistic** command, we get the output that is shown in **Figure 10**. Since the newly created variable is significant (the p-value is less than 0.05), the result indicates that the relationship between the logit function and the variable GPA is not linear.

Loess Curve

We next produce the loess curve of the outcome variable, which is withdraw, and the continuous variable, which is GPA:

```
Logistic regression                                   Number of obs     =     25,160
                                                      LR chi2(2)        =     442.63
                                                      Prob > chi2       =     0.0000
Log likelihood = -2248.4614                           Pseudo R2         =     0.0896
```

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	4.819065	2.114356	3.58	0.000	2.039387	11.38744
boxtid	.7209326	.0605389	-3.90	0.000	.6115284	.8499095
_cons	1.20e-07	6.88e-07	-2.78	0.005	1.59e-12	.0090799

Note: _cons estimates baseline odds.

Figure 10 Summary statistics of the variable GPA for students who withdraw and those who do not withdraw.

```
lowess withdraw gpa, logit
```

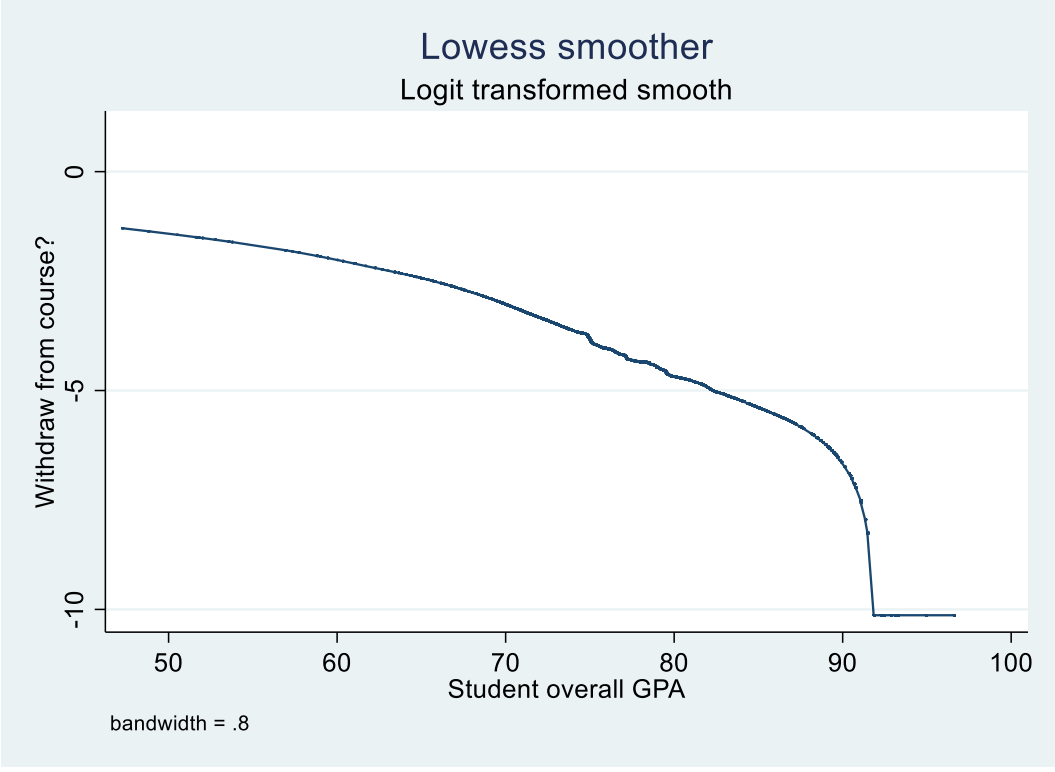

Figure 9 Loess curve of withdraw and GPA.

gpacat	Freq.	Percent	Cum.
40	19	0.08	0.08
50	154	0.61	0.69
60	2,009	7.98	8.67
70	14,912	59.27	67.94
80	7,212	28.66	96.61
90	854	3.39	100.00
Total	25,160	100.00	

Figure 11 Categorizing the continuous variable GPA.

Notice that we specified the **logit** option. This option tells Stata to produce a loess curve in terms of the log of the odds ratio, since we are testing the linearity of GPA with respect to the logit function. The result of running the command in shown in **Figure 9**. The curve clearly shows that the relationship is not linear, thus providing extra evidence.

Linearity of Slopes Test

In order to perform the linearity of slopes test, we need to categorize the continuous variable GPA. Stata provides a very useful command for this purpose:

```
egen gpacat = cut(gpa), at(40(10)100)
```

The **egen** command is an extension to the **generate** command. It allows the user to generate new variables using a set of functions. In this case, we are using the **cut()**function. The command is basically telling Stata to generate a new variable, which we named gpacat, by cutting the variable GPA into groups, which are specified in the **at()** option. This option is telling Stata to group observations with a GPA between 40 and 50 in one group, 50 and 60 in a second group, 60 and 70 in a third group, 70 and 80 in a fourth group, 80 and 90 in a fifth group, and finally 90 and 100 in a sixth group.

We can take a closer look at our newly created variable:

```
tabulate gpacat
```

The output of the command in shown in **Figure 11**. We see that there are four groups where each contains roughly the same number of observations. The groups are in order. This means that group zero contains the GPAs which are in the bottom quartile and group three contains the GPAs which are in the top quartile.

Now that we have our categorical variable, we include it by itself in a logistic model:

```
logistic withdraw i.gpacat
```

Notice that we use the **i** prefix. This is how we tell Stata that this is a categorical variable. Otherwise, Stata will treat the variable like a continuous variable that takes on values from zero to three. The output of this command is shown in **Figure 12**.

```
Logistic regression                                     Number of obs     =     25,160
                                                        LR chi2(5)        =     321.65
                                                        Prob > chi2       =     0.0000
Log likelihood = -2308.9538                             Pseudo R2         =     0.0651
```

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpacat						
50	.5597015	.3423477	-0.95	0.343	.1687756	1.856109
60	.2509294	.1430813	-2.42	0.015	.0820714	.7672047
70	.0814491	.0460666	-4.43	0.000	.0268817	.2467833
80	.0188127	.0110433	-6.77	0.000	.0059537	.0594453
90	.0043962	.0050468	-4.73	0.000	.0004634	.0417097
_cons	.2666667	.1500617	-2.35	0.019	.0885056	.8034643

Note: _cons estimates baseline odds.

Figure 12 Running a model with the categorized variable.

We would next want to produce a graph in order to see whether the relationship is linear or not. To do this we take advantage of Stata's excellent **margins** and **marginsplot** commands. These commands make producing graphs very easy. The **margins** command calculates the values while the **marginsplot** command plots them. Since we are running a logistic model, the **margins** command calculates the predicted probabilities. This, however, is not what we want, since we want to check for linearity by plotting the logit function against the categorical variable. Therefore, we need to tell Stata to calculate the value of the logit function at each value of the categorical variable by including the **predict(xb)** option:

```
margins gpacat, predict(xb)
```

What this command is doing is that it is predicting the value of the logit function at each level of the categorical variable gpacat. The output of the command is shown in **Figure 13**. We can see that the values for each level of the categorical variable is calculated and displayed in the column titled "Margin". We next use the **marginsplot** command to plot the values:

```
marginsplot, noci
```

This command is simply telling Stata to plot the values that it has already calculated when we ran the **margins** command. The graph is shown in **Figure 14**. Once again, the nonlinearity is evident in the graph as it curves downward.

```
Adjusted predictions                              Number of obs     =     25,160
Model VCE     : OIM

Expression    : Linear prediction (log odds), predict(xb)
```

	Delta-method					
	Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpacat						
40	-1.321756	.5627314	-2.35	0.019	-2.424689	-.2188225
50	-1.902108	.2397138	-7.93	0.000	-2.371938	-1.432277
60	-2.70434	.0920194	-29.39	0.000	-2.884694	-2.523985
70	-3.829533	.0567723	-67.45	0.000	-3.940804	-3.718261
80	-5.294978	.1670842	-31.69	0.000	-5.622457	-4.967499
90	-6.748759	1.000586	-6.74	0.000	-8.709871	-4.787647

Figure 13 The margins command calculates the linear prediction, which is the log of the odds, since the predict(xb) option was included.

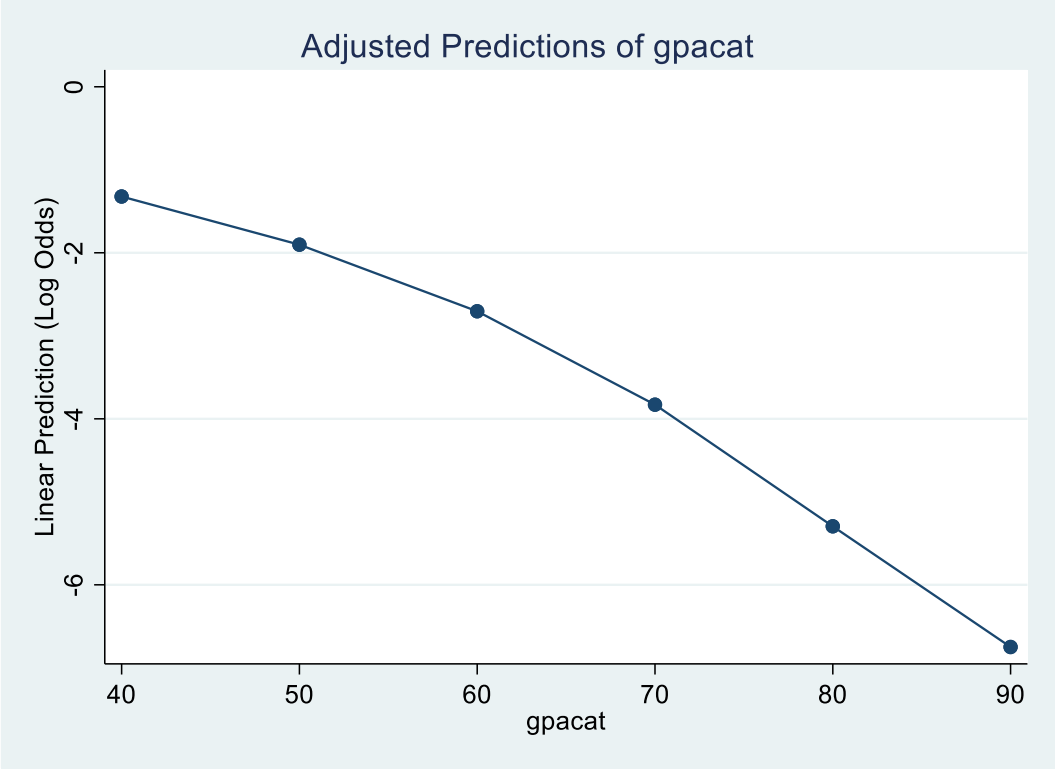

Figure 14 The output of the **marginsplot** command allows us to test the linearity of the slopes assumption.

Including a Quadratic Term

Now that we have seen that there is nonlinearity when it comes to the independent variable GPA, we need to do something to account for this nonlinearity. One way to include nonlinearity in a regression model is to add a quadratic term, where this term is the square of the variable. By including the original variable and the squared term, we will be modeling the following equation:

$$\log(\frac{p}{1-p}) = a_1x^2 + a_2x + b$$

There are two ways in which we can do this in Stata. Both produce the same equation, but one makes visualizing the result easy while the other complicates things. We first start with the method that you should not use.

Generating a New Variable

In the first method, we created a new variable that we named English2, and we included this variable in the model:

```
gen gpa2 = gpa*gpa

logistic withdraw gpa gpa2

margins, at(gpa=(40(1)100)) predict(xb)

marginsplot, noci
```

The first command generates a new variable which is the square of GPA. The second command includes both GPA and the newly created variable in the logistic model. The third command tells Stata to calculate the values of the logit function since we included the **predict(xb)** option. The values are calculated while varying the value of GPA from 40 to 100 in increments of one. The fourth command plots the calculated values. The result is shown in **Figure 15**.

The output is strange. Our linear model includes a squared term, yet the graph is a straight line. Why? The reason is simply because Stata does not know that the variable gpa2 is the square of the variable GPA. Stata includes the variable just like any other independent variable. When an independent variable that is included in the model is not included in the **at()** option of the **margins** command, Stata sets the value of the variable to the mean. Since the mean of the variable gpa2 is 6013.122, Stata is calculating the following:

GPA = 40: $\log(\frac{p}{1-p}) = 0.1878687(40) - 0.0023602(6013.122) - 4.754241$

GPA = 41: $\log(\frac{p}{1-p}) = 0.1878687(41) - 0.0023602(6013.122) - 4.754241$

GPA = 42: $\log(\frac{p}{1-p}) = 0.1878687(42) - 0.0023602(6013.122) - 4.754241$

.

.

.

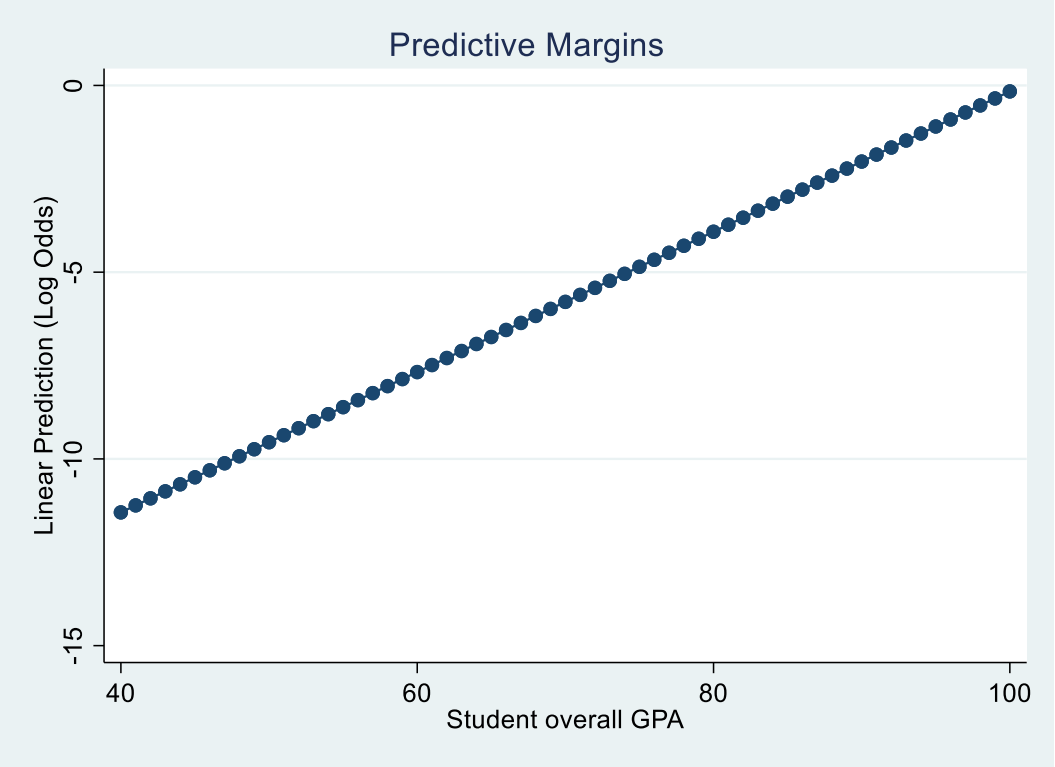

Figure 15 The graph produced when the quadratic term is included by generating a new variable.

GPA = 100: $\log(\frac{p}{1-p}) = 0.1878687(43) - 0.0023602(6013.122) - 4.754241$

Notice that the value of gpa2 is not changing. The variable is fixed at the mean. Stata does not know that when GPA is 40 gpa2 is 40*40, even though this is the formula that we used to generate the variable gpa2.

Using Interaction Terms

Let us now include the quadratic term using an interaction term:

```
logistic withdraw gpa c.gpa#c.gpa
```

Notice that this command does not use any new variable. Instead, it includes the term c.gpa#c.gpa. This term is how we tell Stata to include a variable that is GPA x GPA. The prefix c tells Stata that the variable GPA is continuous. The output from running this command is exactly the same as the output from running the command where the quadratic term was included by generating the variable gpa2. The only difference is in the name of the variable in the output. Notice that the interaction term is statistically significant (p-value < 0.05).

Now that we have fit the logistic model, let us generate the graph:

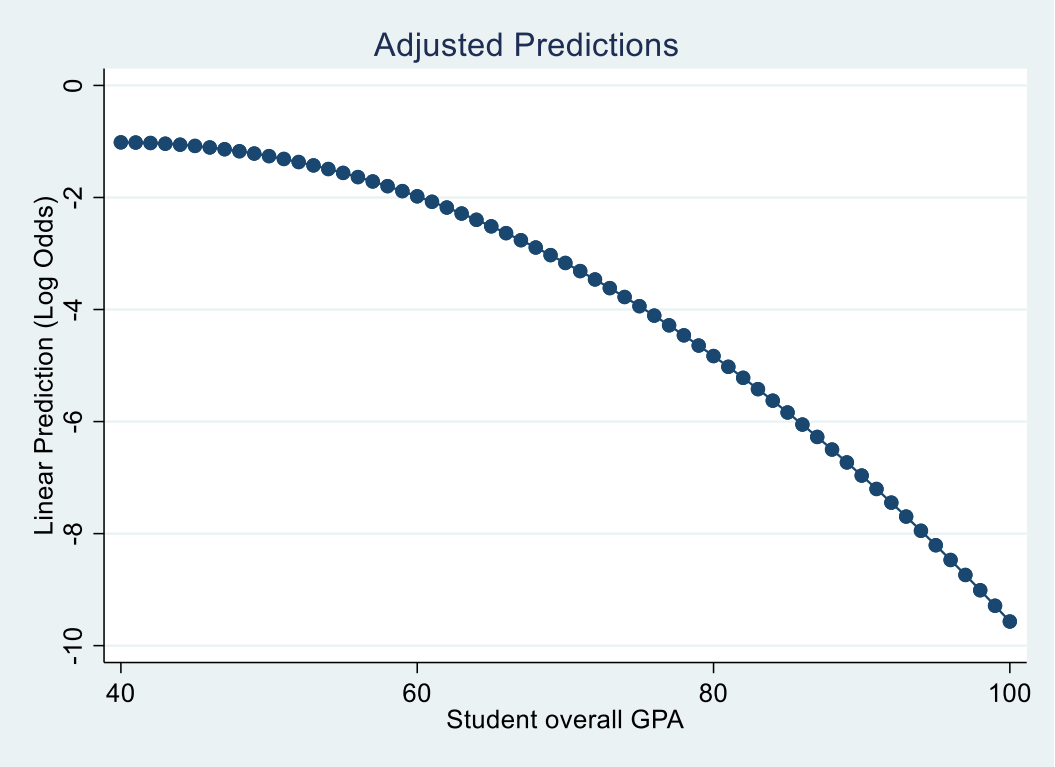

Figure 16 The graph produced when the quadratic term is included as an interaction term.

```
margins, at(gpa=(40(1)100)) predict(xb)

marginsplot, noci
```

The output is shown in **Figure 16**. The difference is clear. This correct graph shows how the logit function decreases with increasing values of GPA. By using the c.gpa#c.gpa notation, we have explicitly told Stata that the second term is the squared of GPA. Stata is now calculating the following:

GPA = 40: $\log(\frac{p}{1-p}) = 0.1878687(40) - 0.0023602(40^2) - 4.754241$

GPA = 41: $\log(\frac{p}{1-p}) = 0.1878687(41) - 0.0023602(41^2) - 4.754241$

GPA = 42: $\log(\frac{p}{1-p}) = 0.1878687(42) - 0.0023602(42^2) - 4.754241$

.

.

.

Gender	Withdraw from course? No withdr	Withdraw	Total
female	7,957	76	8,033
male	16,699	428	17,127
Total	24,656	504	25,160

Figure 17 Cross tabulation when the independent variable is also binary.

GPA = 42: $\log(\frac{p}{1-p}) = 0.1878687(100) - 0.0023602(100^2) - 4.754241$

This example illustrates why you should always use interaction terms.

Binary Variables

Now that we have seen how to analyze the relationship between the binary dependent variable and a continuous independent variable, we move onto other types of variables. Looking at our dataset, we notice that the variables gender and college are binary. Both take on two values. While graphs are used to investigate the relationship when a continuous variable is involved, contingency tables are used to investigate the relationship when the independent variable is binary.

We start by looking at the variable gender. The following command creates a cross tabulation of the variables withdraw and gender:

```
tabulate gender withdraw
```

The output is shown in **Figure 17**. We notice that the table includes frequencies. It would also be useful if we were able to see the percent of females who withdrew from a course and the percent of

Gender	Withdraw from course? No withdr	Withdraw	Total
female	7,957	76	8,033
	99.05	0.95	100.00
male	16,699	428	17,127
	97.50	2.50	100.00
Total	24,656	504	25,160
	98.00	2.00	100.00

Figure 18 Cross tabulation with percentages.

```
Logistic regression                             Number of obs     =     25,160
                                                LR chi2(1)        =      76.62
                                                Prob > chi2       =     0.0000
Log likelihood = -2431.4659                     Pseudo R2         =     0.0155
```

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gender						
male	2.683423	.3360166	7.88	0.000	2.099433	3.429857
_cons	.0095513	.0011008	-40.35	0.000	.0076201	.0119721

Note: _cons estimates baseline odds.

Figure 19 Logistic regression with a binary independent variable.

males who withdrew. This can be done by specifying the **row** option which tells Stata to calculate the percentages in each row:

```
tabulate gender withdraw, row
```

The output is shown in **Figure 18**. We see that 0.95% of females withdrew from courses as opposed to the 2.50% of males. This result indicates that there seems to be a difference between the two groups. To verify this, we fit a logistic model:

```
logistic withdraw i.gender
```

Notice that we include the **i** prefix in the binary variable. The output, which is shown in **Figure 19**, shows that the odds ratios is 2.68. This means that the odds of males are 2.68 times the odds of

College	Withdraw from course? No withdr	Withdraw	Total
Business	9,079	219	9,298
	97.64	2.36	100.00
Engineering	15,577	285	15,862
	98.20	1.80	100.00
Total	24,656	504	25,160
	98.00	2.00	100.00

Figure 20 Cross tabulation with percentages.

```
Logistic regression                             Number of obs     =     25,160
                                                LR chi2(1)        =       9.13
                                                Prob > chi2       =     0.0025
Log likelihood = -2465.2122                     Pseudo R2         =     0.0018

------------------------------------------------------------------------------
    withdraw | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |
Engineering |   .7584989   .0688913    -3.04   0.002     .6348102    .9062875
       _cons |   .0241216   .0016495   -54.47   0.000     .0210959    .0275813
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Figure 21 Logistic regression with a binary independent variable.

females. The result is significant since the p-value is less than 0.05. Therefore, when building our final model, it would make sense to include this variable.

We next perform the same analysis for the variable college:

```
tab college withdraw, row
```

```
logistic withdraw i.college
```

The output of both commands are shown in **Figure 20** and **Figure 21**. The odds ratio is 0.76 indicating that the odds for an engineering student to withdraw is less than the odds of a business student, since the odds ratio is less than one. The result is also statistically significant. It therefore seems that this variable merits inclusion in the model.

Categorical Variables with More than Two Groups

Our dataset also contains variables that are categorical in nature, but unlike binary variables, these variables contain more than one group. To see this, we can run the commands:

```
tabulate semester withdraw, row
```

```
tabulate level withdraw, row
```

The outputs of both commands are shown in **Figure 22** and **Figure 23**. From **Figure 22**, it seems that the largest percentages of withdrawals are in the spring semester. From **Figure 23**, it seems that the largest percent of withdrawals are in 200-level courses. It should be noted that 100-level courses are the courses that are taken during the freshman year before students have decided on their major. Once a student has enrolled in the major of his or her choice, they start taking the 200-level courses. During the third and fourth year of their studies, students take the 300-hundred and the 400-hundred level courses. For majors that extend beyond four years, students take 500-hundred level courses in their final year. Therefore, what this output shows is that the largest percentage of withdrawals takes place in the first year after students have enrolled in their major.

Semester the course was taken	Withdraw from course? No withdr	Withdraw	Total
Fall	10,447	191	10,638
	98.20	1.80	100.00
Spring	10,638	244	10,882
	97.76	2.24	100.00
Summer	3,571	69	3,640
	98.10	1.90	100.00
Total	24,656	504	25,160
	98.00	2.00	100.00

Figure 22 Cross tabulation of withdraw and semester.

We nest fit a logistic model with semester as the independent variable:

```
logistic withdraw i.semester
```

Notice that just like in the case of binary variables, we use the **i** prefix in order to tell Stata that the variable represents categories. Otherwise, Stata will treat the variable as continuous. The out of the command is shown in **Figure 24**. We see that the odds ratio for both categories is greater than one.

The level of the course	Withdraw from course? No withdr	Withdraw	Total
remedial	450	5	455
	98.90	1.10	100.00
100 level course	1,204	16	1,220
	98.69	1.31	100.00
200 level course	10,085	318	10,403
	96.94	3.06	100.00
300 level course	8,516	151	8,667
	98.26	1.74	100.00
400 level course	3,245	13	3,258
	99.60	0.40	100.00
500 level course	1,156	1	1,157
	99.91	0.09	100.00
Total	24,656	504	25,160
	98.00	2.00	100.00

Figure 23 Cross tabulation of withdraw and level.

```
Logistic regression                             Number of obs     =     25,160
                                                LR chi2(2)        =       5.69
                                                Prob > chi2       =     0.0581
Log likelihood =  -2466.931                     Pseudo R2         =     0.0012

------------------------------------------------------------------------------
    withdraw | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    semester |
      Spring |    1.25455   .1224308     2.32   0.020     1.036143    1.518995
      Summer |    1.05686   .1498511     0.39   0.697     .8004358    1.395431
             |
       _cons |   .0182828   .0013349   -54.81   0.000     .0158449    .0210957
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Figure 24 Logistic regression with semester as the independent variable.

Since the base category is fall, the odds ratio compare the odds of withdrawing in spring and summer to the odds of withdrawing in the fall semester. We can tell Stata to use another category as the base. This is done the following way:

```
logistic withdraw b1.semester
```

By using the **b1** prefix, we are telling Stata that the base category is the one which is coded with a value of 1. In our case this is the spring semester. The output of running the command in shown in **Figure 25**. We see that the spring semester is no longer in the output. This is because the category spring is coded using the value one, and we have told Stata to use this category as the base. This means that the

```
Logistic regression                             Number of obs     =     25,160
                                                LR chi2(2)        =       5.69
                                                Prob > chi2       =     0.0581
Log likelihood =  -2466.931                     Pseudo R2         =     0.0012

------------------------------------------------------------------------------
    withdraw | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    semester |
        Fall |   .7970984   .0777883    -2.32   0.020     .6583301    .9651174
      Summer |   .8424214   .1160132    -1.25   0.213     .6431422    1.103448
             |
       _cons |   .0229366   .0014851   -58.30   0.000      .020203    .0260402
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Figure 25 Changing the base category of the variable semester.

odds ratios in **Figure 25** compare the odds of withdrawal in the fall and summer semester with the odds of withdrawal in the spring semester. Both are less than one indicating that the odds of withdrawal in the spring semester are higher.

Going back to **Figure 24** in which the fall semester is the base, if we now look at the p-values we will notice that the p-value for the category spring is less than 0.05 while the p-value for the category summer is greater than 0.05. This means that the difference between the spring semester and the fall semester is significant, while the difference between the summer semester and the fall semester is not. Given this result, we might want to consider collapsing the variable semester. Since the odds ratio for summer when compared to fall is not significant, it might be better if we just treated these two as a single group. In other words, we can create a binary variable that takes a value of zero when the semester is fall or summer, and takes a value of one when the semester is spring. This can be accomplished using the following commands:

```
recode semester 0 2=0 1=1, gen(spring)

label define spring 0 "Fall or Summer" 1 "Spring"

label values spring spring

label variable spring "Spring semester"
```

The **recode** command lets us recode a variable. In our case, we are telling Stata to code the values 0 and 2 (which represent fall and semester) as 0, and to keep the value 1 as 1. This means that the new variable spring, which is generated because we used the **gen()** option, takes the value zero when the semester is fall or summer, and takes the value one when the semester is spring. We then label the values in the new variable and finally we label the variable itself. We can cross tabulate the new variable with the old one in order to see the result:

```
tabulate semester spring
```

The output is shown in **Figure 26**. We can see that all observations with a value of spring for the semester variable have a value of spring in the new variable. We also see that all observations with a

Semester the course was taken	Spring semester		
	Fall or S	Spring	Total
Fall	10,638	0	10,638
Spring	0	10,882	10,882
Summer	3,640	0	3,640
Total	14,278	10,882	25,160

Figure 26 Collapsing the three-group semester variable to a new two-group variable named spring.

```
Logistic regression                                   Number of obs     =     25,160
                                                      LR chi2(1)        =       5.54
                                                      Prob > chi2       =     0.0186
Log likelihood = -2467.0064                           Pseudo R2         =     0.0011
```

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
spring						
Spring	1.236638	.1113651	2.36	0.018	1.036544	1.475358
_cons	.0185476	.0011609	-63.71	0.000	.0164063	.0209683

Note: _cons estimates baseline odds.

Figure 27 Using the collapsed variable in the model.

fall or summer value have a value of "Fall or Summer" in the new variable. Therefore, we confirm that the new variable has been coded correctly.

We next include this new variable in the logistic model:

```
logistic withdraw i.spring
```

The output is shown in **Figure 27**. We see that the odds ratio is greater than one and is significant. We therefore conclude that for courses that are taken in the spring semester, the odds of withdrawal is 1.24 times the odds of withdrawal in the other two semesters. Which model should we use, the one

```
Logistic regression                                   Number of obs     =     25,160
                                                      LR chi2(5)        =     161.47
                                                      Prob > chi2       =     0.0000
Log likelihood = -2389.0396                           Pseudo R2         =     0.0327
```

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
level						
100 level course	1.196013	.6163273	0.35	0.728	.4356086	3.283792
200 level course	2.837878	1.286364	2.30	0.021	1.167234	6.899688
300 level course	1.59582	.7294871	1.02	0.307	.6514473	3.909203
400 level course	.3605547	.1906013	-1.93	0.054	.1279374	1.01612
500 level course	.0778559	.085396	-2.33	0.020	.009071	.6682342
_cons	.0111111	.0049966	-10.01	0.000	.0046023	.0268247

Note: _cons estimates baseline odds.

Figure 28 Including the variable level as an independent variable.

with the spring/fall/summer division or the one with the spring/not spring division? I usually prefer to use the model with the collapsed variable for the sake of simplicity.

We now perform the same analysis on the level variable:

```
logistic withdraw i.level
```

The output is shown in **Figure 28**. We see that the result for 200-level courses and 500-level courses is significant, with 200-level courses having odds that are 2.84 times higher than the odds of intensive courses (the base category), and 500-level courses having odds that are 0.08 times the odds of intensive courses. The other categories have p-values that are less than 0.05. Looking at this output, we might deduce that once students have reached the very end of their studies, the probability that they will withdraw from a course decreases significantly since such a decision will probably postpone their graduation. We can also deduce that students who have just enrolled in a major face the largest uncertainty in terms of not being sure whether this is the correct major for them, thus leading to a higher probability of withdrawal. Given that the other categories are not significant, we might choose to collapse this variable as well by creating a new three-group variable that contains the groups 200-level courses, 500-level courses, and the remaining courses. This can be done using the following commands:

```
recode level 0 1 3 4=0 2=1 5=2, gen(level3)

label define level3 0 "Other courses" 1 "200 level courses" 2 "500 level courses"

label values level3 level3
```

We use the **recode** command to tell Stata to create a new variable which we name level3. The command tells Stata that the values 0, 1, 3, and 4 are to be assigned the value 0, the value 2 is to be assigned the value 1, and the value 5 to be assigned the value 2. This means that the new variable, level3, will have values 0, 1, and 2. We then label the values using the **label define** and the **label**

The level of the course	RECODE of level (The level of the course)			Total
	Other cou	200 level	500 level	
remedial	455	0	0	455
100 level course	1,220	0	0	1,220
200 level course	0	10,403	0	10,403
300 level course	8,667	0	0	8,667
400 level course	3,258	0	0	3,258
500 level course	0	0	1,157	1,157
Total	13,600	10,403	1,157	25,160

Figure 29 Collapsing the six-group level variable to a new three-group variable named level3.

```
Logistic regression                             Number of obs     =     25,160
                                                LR chi2(2)        =     121.49
                                                Prob > chi2       =     0.0000
Log likelihood = -2409.0301                     Pseudo R2         =     0.0246
```

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
level3						
200 level courses	2.286495	.213561	8.85	0.000	1.904	2.745828
500 level courses	.062729	.0629272	-2.76	0.006	.0087817	.4480835
_cons	.0137905	.0010209	-57.87	0.000	.011928	.0159438

Note: _cons estimates baseline odds.

Figure 30 Using the collapsed variable in the model.

values commands. To make sure that everything went as expected, we can cross tabulate the old and the new variables:

```
tabulate level level3
```

The output is show in **Figure 29**. We see that the coding operation was successful. The remedial, 100-hundred level, 300-level, and 400-level courses all end up in the first group of the new variable. The 200-hundred level courses end up in the second group, and the 500-level courses end up in the third group.

We now include this new variable in the model:

```
logistic withdraw i.level3
```

The output is shown in **Figure 30**. Both groups of the variable are significant.

Multivariate Analysis

After looking at each independent variable by itself, we need to start building a more complex model. This means that we need a model that includes more than one independent variable. We start with a model that includes all the variables that were found to be significant when we conducted the univariate analysis:

```
logistic withdraw gpa c.gpa#c.gpa i.gender i.college i.spring i.level3
```

Notice that we include the quadratic term of the variable GPA, since we had uncovered that the logit function is not linear with respect to GPA. We also include the collapsed versions of the variables semester and level. The output of the model is shown in **Figure 31**. We see that all of the variables are significant except for the variable college. Therefore, it seems like a good idea to remove this variable from the model:

```
logistic withdraw gpa c.gpa#c.gpa i.gender i.spring i.level3
```

```
Logistic regression                             Number of obs     =     25,160
                                                LR chi2(7)        =     539.85
                                                Prob > chi2       =     0.0000
Log likelihood = -2199.8495                     Pseudo R2         =     0.1093
```

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	1.217514	.1043718	2.30	0.022	1.029211	1.440268
c.gpa#c.gpa	.9976925	.0006185	-3.73	0.000	.9964811	.9989054
gender						
male	1.528905	.2032302	3.19	0.001	1.178241	1.983931
college						
Engineering	1.011603	.0971431	0.12	0.904	.83805	1.221096
spring						
Spring	1.223419	.111875	2.21	0.027	1.022675	1.463569
level3						
200 level courses	1.979354	.1880548	7.19	0.000	1.643056	2.384485
500 level courses	.0726959	.0730008	-2.61	0.009	.0101564	.520333
_cons	.0016024	.0047537	-2.17	0.030	4.78e-06	.5369365

Note: _cons estimates baseline odds.

Figure 31 Including all variables in the model.

The output when we exclude college is shown in **Figure 32**. We now see that all the independent variables are significant. When we have several independent variables, it is quite difficult to make sense of the individual odds ratios, especially when it comes to quadratic terms. This is why Stata provides us with powerful graphical tools that allow us to visualize the effect that each independent variable has on the probability of withdrawal. Before interpreting the result of the model however, we need to check the goodness-of-fit of the model using the tests that were discussed in the theory section.

Analysis of Model Fit

Likelihood Ratio Test

This test is displayed whenever we run a logistic model. From the top right corner of **Figure 32**, we can see that the test yields a p-value that is less than 0.05, thereby indicating that our model does a significantly better job than a constant-only model.

Hosmer-Lemeshow Test

To run the Hosmer-Lemeshow test in Stata, we use the following command after we have fit the model:

```
estat gof, group(10)
```

Notice that we specify the **group(10)**. This is because the Hosmer-Lemeshow test divides the probability ranges into groups. The user can specify this number as they see fit, but the authors of the

Logistic regression — Number of obs = 25,160; LR chi2(6) = 539.84; Prob > chi2 = 0.0000; Log likelihood = -2199.8567; Pseudo R2 = 0.1093

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	1.217242	.1043142	2.29	0.022	1.029038	1.439868
c.gpa#c.gpa	.9976954	.0006179	-3.73	0.000	.9964851	.9989072
gender						
male	1.534498	.1985629	3.31	0.001	1.190752	1.977476
spring						
Spring	1.2234	.111873	2.20	0.027	1.022659	1.463545
level3						
200 level courses	1.9789	.1879736	7.19	0.000	1.642741	2.383848
500 level courses	.0729881	.073254	-2.61	0.009	.0102082	.5218603
_cons	.0016093	.0047734	-2.17	0.030	4.81e-06	.538842

Note: _cons estimates baseline odds.

Figure 32 Excluding the variable college since was not found to be significant.

test recommend that the data be divided into 10 groups, which is what we did in the command. The output of the command is shown in **Figure 33**. The Hosmer-Lemeshow statistics is 6.76, which is low, thus resulting in a p-value that is much larger than 0.05. This means that the model is a good, if not excellent, fit.

Classification Table

The classification table allows us to compare the observed outcome with the outcome as predicted by our model. Before producing the classification table, we first need to determine the cutoff probability. As discussed in the theory section, the cutoff value is the optimal probability value that separates the

Logistic model for withdraw, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

number of observations =	25160
number of groups =	10
Hosmer-Lemeshow chi2(8) =	6.76
Prob > chi2 =	0.5623

Figure 33 The Hosmer-Lemeshow test.

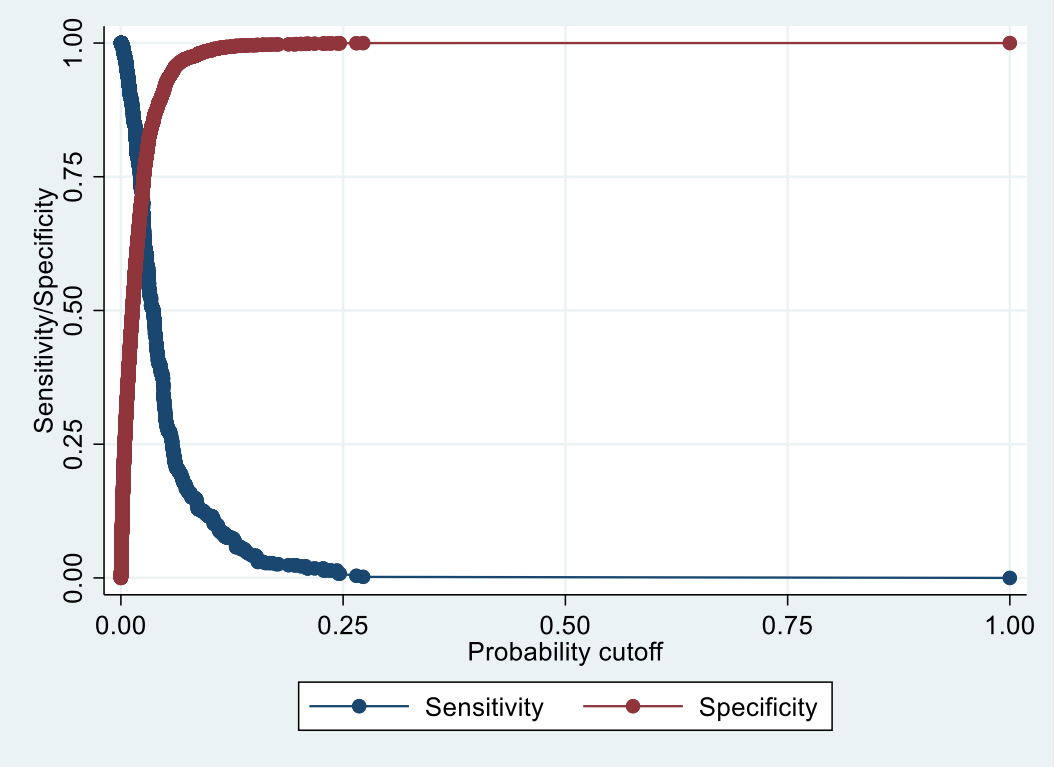

Figure 34 The sensitivity-specificity curves.

predicted and observed outcomes. This ideal cutoff value is the point at which the sensitivity and the specificity are equal. Stata has a command called **lsens** that allows us to produce a graph of the sensitivity and the specificity in order for us to see where the graphs intersect. The command also allows us to generate new variables that store the values of both sensitivity and specificity:

```
lsens, gense(se) gensp(sp) genp(cutp)
```

The **gense(se)** option tells Stata to generate a new variable called se, where the variable stores the values of sensitivity. The **gensp(sp)** option tells Stata to generate a new variable called sp, where the variable stores the values of specificity. Finally, the **genp(cutp)** option tells Stata to generate a new variable called cutp, where the variable stores the value of the cutoff point. The result of running the above command will be the generation of these new variables as well as the graph that is shown in **Figure 34**. We see that the sensitivity and the specificity curves intersect at a very small probability. To find the exact value of this probability we can generate a new variable that represents the difference between sensitivity and specificity, and then we can sort the dataset on this variable:

```
gen diff = abs(se - sp)

sort diff
```

	withdraw	id	gender	gpa	college	semester	level	spring	level3	sp	se	cutp
1.	No withdraw	345	female	73.08	Business	Fall	200 level course	Fall or Summer	200 level courses	0.716377	0.716270	0.024088
	diff											
	.0001075											

Figure 35 Sorting the data on the difference between sensitivity and specificity.

Notice that we use the **abs()** function, which returns the absolute value. This is because we are not interested in the sign of the difference. We are interested in the smallest magnitude. Once we sort the observations on the new variable, the observation with the smallest value will appear at the very top. We display this observation using the following command:

```
list in 1
```

The output of this command is shown in **Figure 35**. We see that the cutoff p-value is 0.024088. We can now produce the classification table:

```
estat class, cut(0.024088)
```

```
Logistic model for withdraw

                 True
Classified |         D          ~D |      Total
-----------+----------------------+-----------
         + |       361        6973 |       7334
         - |       143       17683 |      17826
-----------+----------------------+-----------
     Total |       504       24656 |      25160

Classified + if predicted Pr(D) >= .024088
True D defined as withdraw != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   71.63%
Specificity                     Pr( -|~D)   71.72%
Positive predictive value       Pr( D| +)    4.92%
Negative predictive value       Pr(~D| -)   99.20%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   28.28%
False - rate for true D         Pr( -| D)   28.37%
False + rate for classified +   Pr(~D| +)   95.08%
False - rate for classified -   Pr( D| -)    0.80%
-------------------------------------------------
Correctly classified                        71.72%
-------------------------------------------------
```

Figure 36 The classification table.

The table that is produced is shown in **Figure 36**. We see that the model correctly classifies 71.72% of the observations, which is an acceptable value.

ROC Curve

As discussed in the theory section, another way to test the model fit is to calculate the area under the ROC curve. To do that, we run the following command:

```
lroc
```

The resulting graph is shown in **Figure 37**. The output shows that the area under the curve is 0.7803. According to the set of rules that were mentioned in the theory section, this is considered acceptable discrimination.

Residual Analysis

When discussing the theory of logistic regression, it was mentioned that the three most commonly used ones are the standardized residuals, the deviance residuals, and the DeltaX residuals. We will now see how to calculate these in Stata.

Standardized Residuals

In order to calculate the standardized residuals in Stata, we use the following command right after we fit the model:

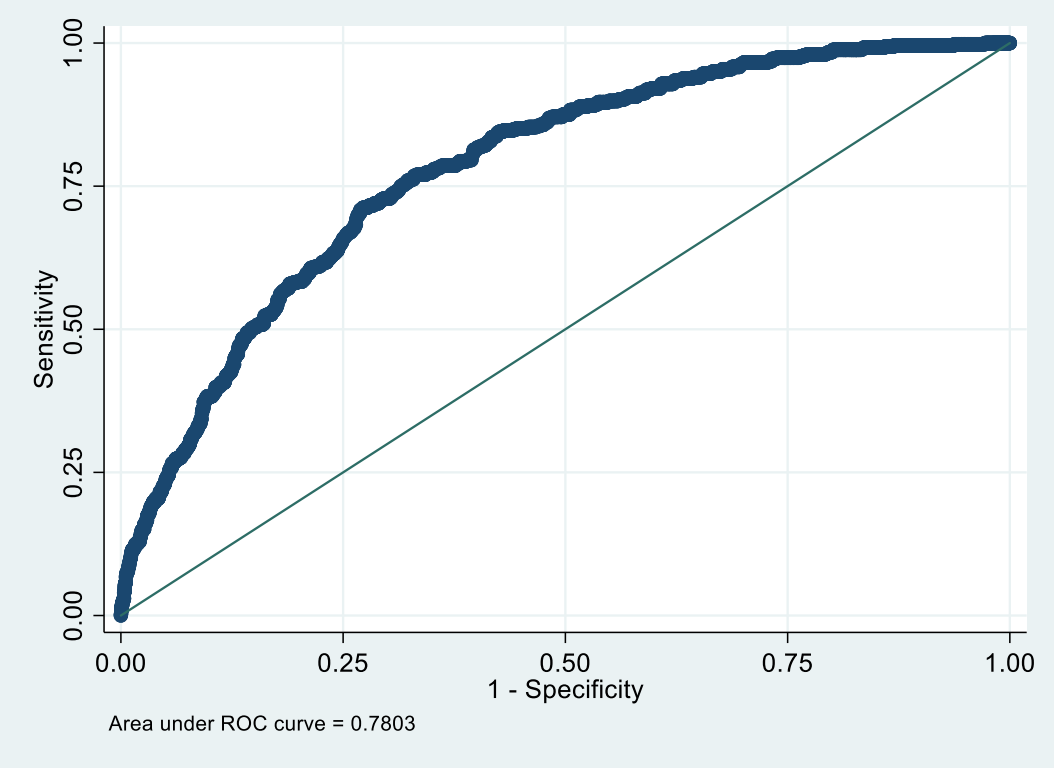

Figure 37 The ROC curve.

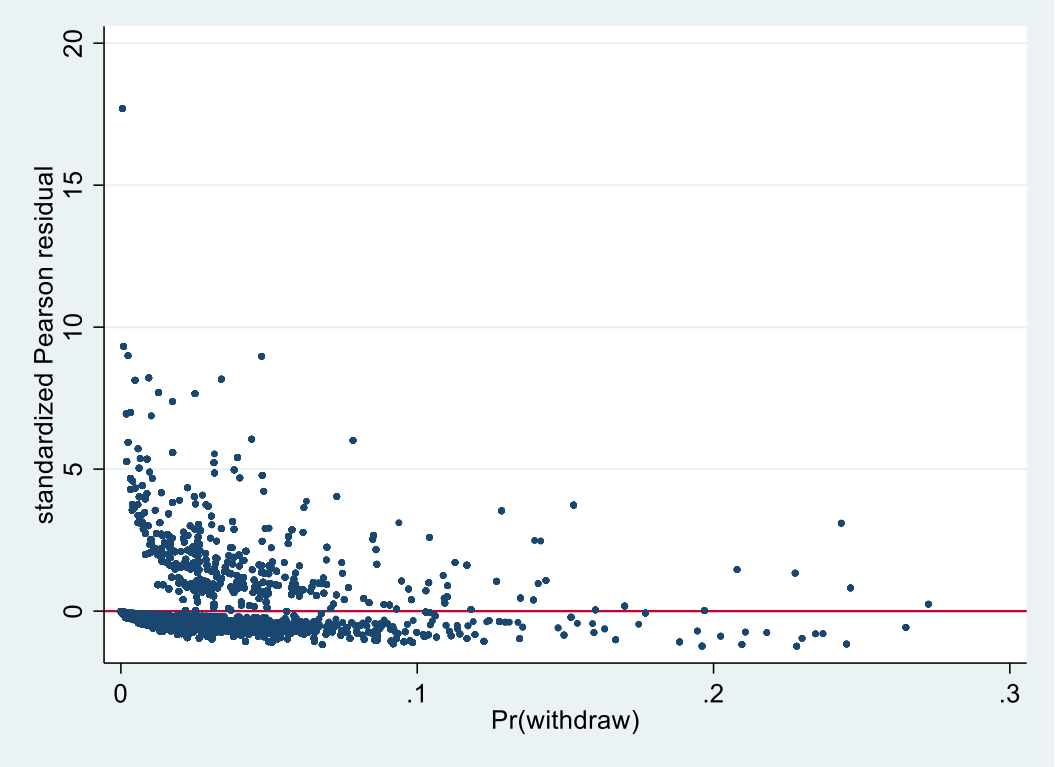

Figure 38 Plotting the standardized residuals against the predicted probabilities.

```
predict residuals, rstandard
```

Since we want to plot the residuals against the predicted values, we will also need to calculate the predicted probabilities:

```
predict prob
```

We can now plot the standardized residuals against the predicted probabilities:

```
scatter residuals prob, yline(0) msize(vsmall)
```

We use the **yline(0)** option in order to draw the line where y, which is the standardized residuals, are zero. This helps us visualize where the residuals are close to the value zero and where are they far from it. We also set the size of the dots to very small because there is a large number of observations. The graph is shown in **Figure 38**.

Deviance Residuals
To calculate the deviance residuals, we use the command:

```
predict dv, deviance
```

We then plot the statistic against the predicted probabilities, which we have already calculated:

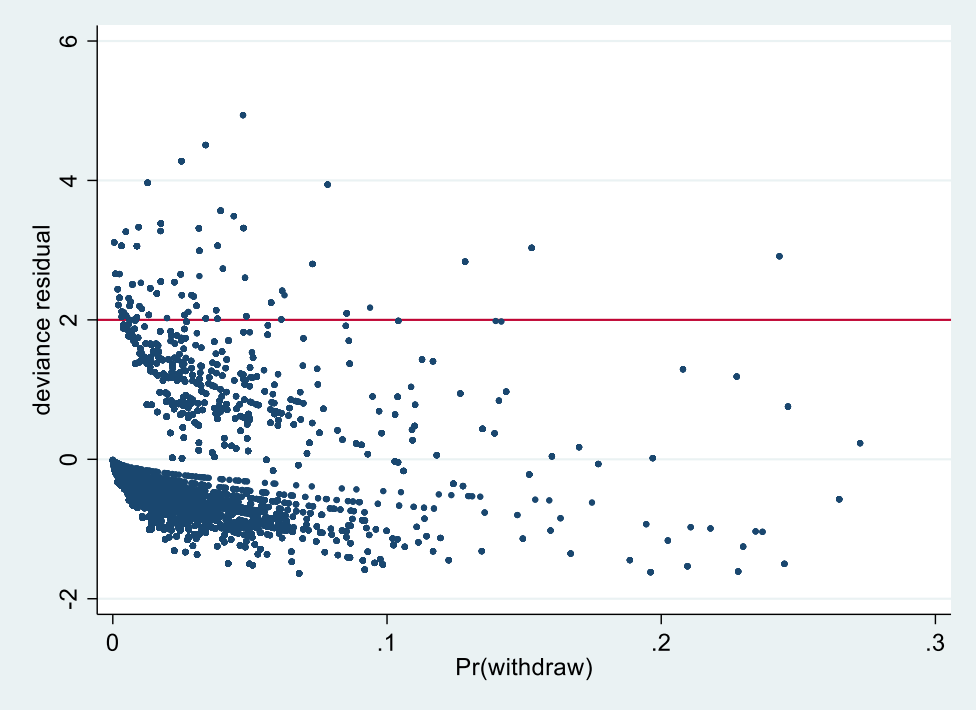

Figure 39 Plotting the deviance residuals against the predicted probabilities (values above four are considered to be outliers).

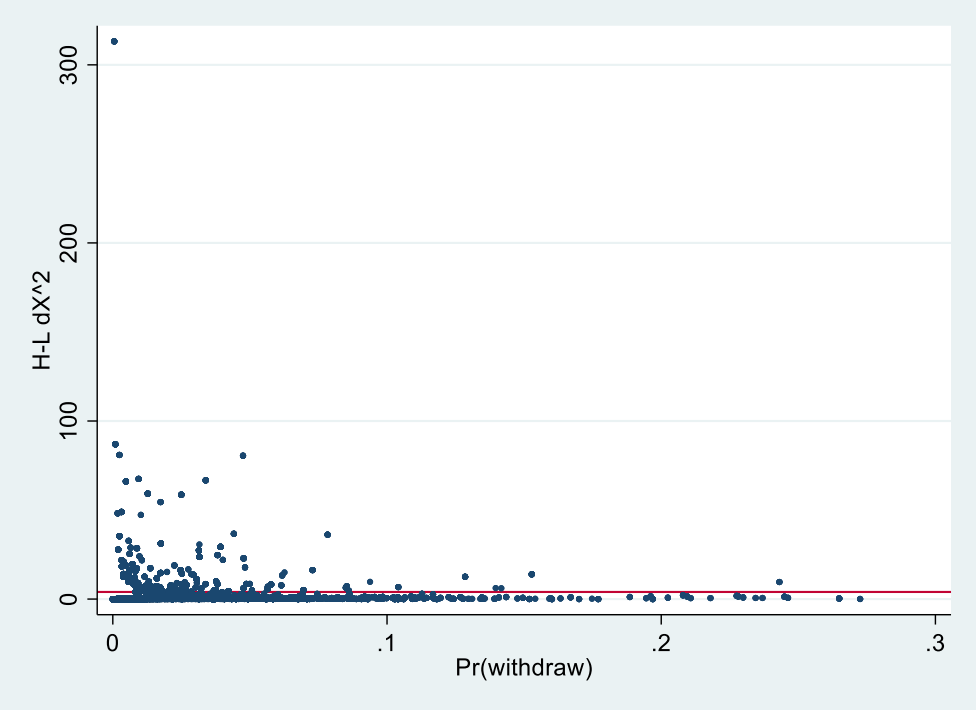

Figure 40 Plotting the DeltaX residuals against the deviance residuals (values above four are considered to be outliers).

```
scatter dv prob, yline(2) msize(vsmall)
```

The graph is shown in **Figure 39**. Both **Figure 38** and **Figure 39** show that there are some observations that have residual values that are large when compared to other observations. In general, when the sample size is large enough, as is the case in our dataset, an observation that has a deviance residual that is greater than two should raise a flag, which is why we drew a horizontal line at the point y equal to two when plotting the deviance residuals.

DeltaX
To calculate the DeltaX residuals, we use the following command:

```
predict dx2, dx2
```

We then produce the scatter plot against the predicted probabilities:

```
scatter dx2 prob, yline(4) msize(vsmall)
```

We draw a horizontal line at the y equal 4 point since values above four are considered to be outliers. The output is shown in **Figure 40**.

Influential Observations

The Hat Diagonal Statistic
To calculate the hat statistic, we use the following command:

```
predict hat, hat
```

We then plot the statistic against the predicted probability:

```
scatter hat prob, yline(0.00519402) msize(vsmall)
```

We draw a horizontal line at the y equal 0.00519402 point since values that are more than two times greater than the average are considered to be influential (the mean of the variable hat is 0.002597). The graph is shown in **Figure 41**.

Delta-Beta Statistic
To calculate the delta-beta influence statistic, we use the following command:

```
predict dbeta, dbeta
```

We then plot the values against the predicted probabilities:

```
scatter dbeta prob, yline(1) msize(vsmall)
```

We specify the **yline(1)** option because values greater than one indicate that the observation is influential. The graph is shown in **Figure 42**.

So far, the graphs that have been produced have indicated that we have both outliers and influential observations. At this point it would be instructive to see how we can combine these two findings together. What we want is to produce a single graph that will include information about the size of the residuals (being an outlier) and the size of the leverage (having influence). Such a plot can be produced by using the following command:

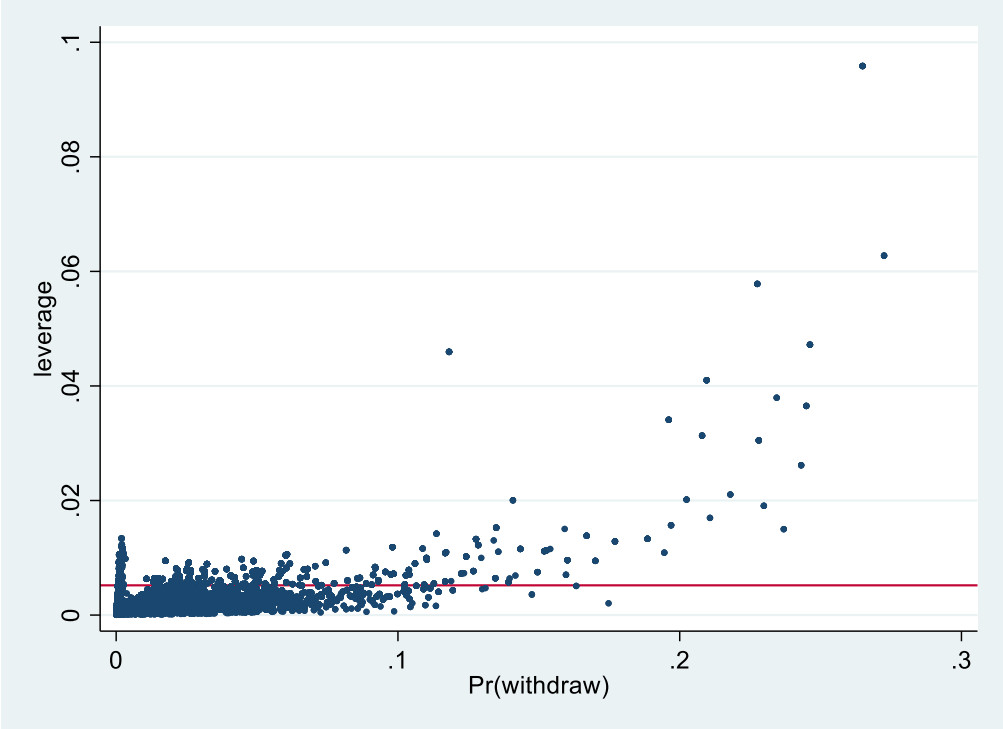

Figure 41 Plotting the DeltaX residuals against the predicted probabilities (values that are more than two times greater than the average are considered to be influential).

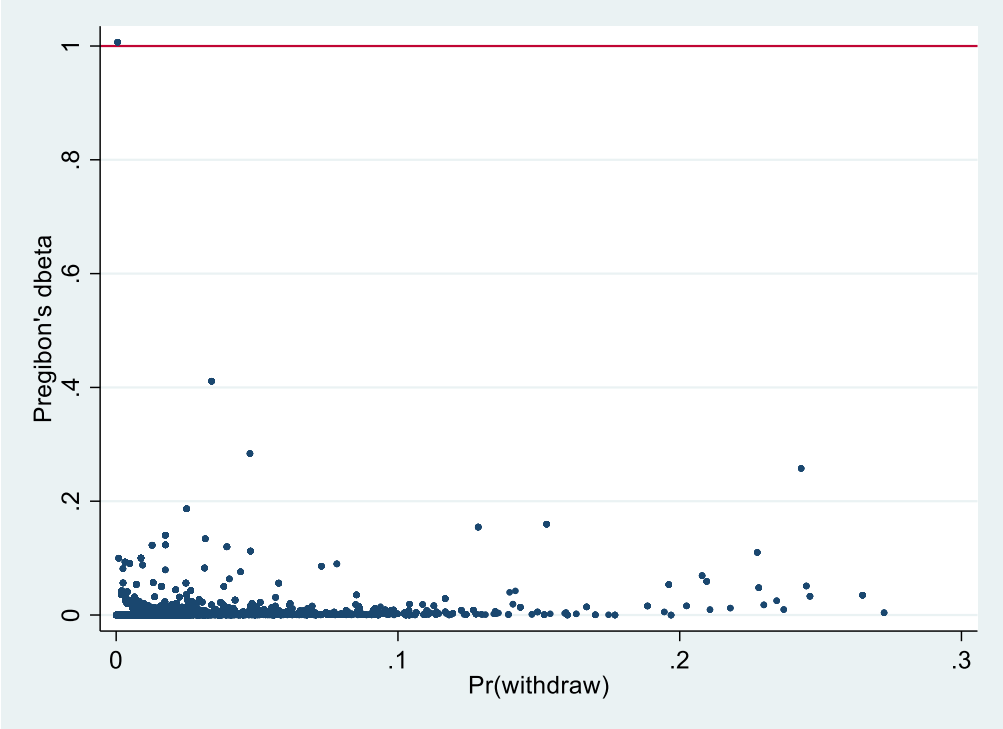

Figure 42 Plotting the Delta-Beta statistic against the predicted probabilities.

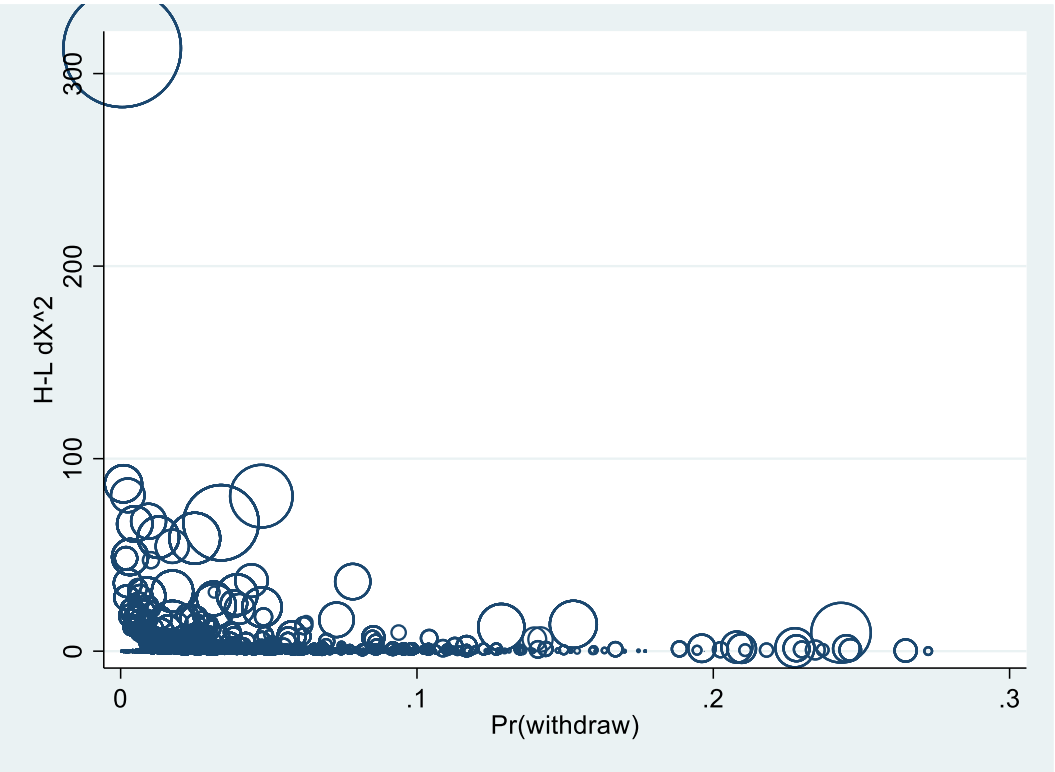

Figure 43 Plot of DeltaX versus estimated probability weighted by the variable delta-beta.

```
scatter dx2 prob [aweight=dbeta], msymbol(Oh)
```

What we are doing here is that we are producing a scatter plot of the DeltaX residuals against the predicted probabilities. We have actually already done this in **Figure 40**. This time however, the size of the dots are weighted by the value of the delta-beta statistic. Since DeltaX is a residual measure, the larger the value, the worse the fit of the observation, since residuals are a measure of the difference between the observed value and the predicted value. Since delta-measure is a measure of influence, and since we are weighing the dots by this variable, the larger the dot, the more influential it is. What this means is that when we produce this plot, the most problematic points are the large points in the upper left corner. This means that they are influential (hence their large size) and they are not a good fit with the model (high value of the residual which leads to them being near the top).

The graph produced by the above command in shown in **Figure 43**. The large circle in the top left-hand side of the graph raises concerns. We need to take a closer look at these values. To do that, we ask Stata to list all observations that have a DeltaX residual that is greater than 200 (since we want to look at the circle in the top left-corner):

```
list withdraw gender gpa college spring level3 if dx2 > 200
```

	withdraw	gender	gpa	college	spring	level3
16943.	No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
21037.	No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
23095.	Withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
23115.	No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
23127.	No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
23173.	No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses

Figure 44 Listing observations with a very large DeltaX value.

The output is shown in **Figure 44**. We see that there are six observations. We also notice that they all have a similar pattern: male engineering student with a GPA of 79.3, taking a 500-level course in a semester other than the spring semester. It seems that our model is not doing a good job of predicting the probability for observations with these covariate patterns. What would happen if we fit the model without including these observations? Let us see whether there will be a large change in the output:

```
logistic withdraw gpa c.gpa#c.gpa i.gender i.spring i.level3

estimates store model1

logistic withdraw gpa c.gpa#c.gpa i.gender i.spring i.level3 if dx2 <= 200

estimates store model2

esttab model1 model2
```

We first fit the model that included all observations and stored the results by using the **estimates store** command. These estimates were stored using the name model1. We then fit the model again, but this time we excluded the observations by including the **if dx2 < 200** condition. This means that the six observations that have a DeltaX residual that is greater than 200 are excluded. We then stored the results under the name model2. Finally, we use the **esttab** command to print the results from both model1 and model2 (if the command is not installed on your computer, you can install the package in which it comes by running the command **ssc install estout**). The output is shown in **Figure 45**. We see that neither values of the coefficients, nor the significance levels of any of the variables changes significantly. This means that our results are robust.

Interpreting the Results

Now that we have seen that the model fit is good, it is time to interpret the obtained model parameters. As usual, Stata allows us to do this using both graphical tools and non-graphical tools. It is important to note that logistic regression has the following linear form:

$$\log\left(\frac{p}{1-p}\right) = ax + b$$

Therefore, once we fit the model we can calculate the value of the logit function for each observation. From this logit function, we are able to calculate the individual probabilities. Ultimately, we are

	(1) withdraw	(2) withdraw
withdraw		
gpa	0.200*	0.197*
	(2.32)	(2.29)
c.gpa#c.gpa	-0.00233***	-0.00231***
	(-3.75)	(-3.73)
0.gender	0	0
	(.)	(.)
1.gender	0.425**	0.428***
	(3.29)	(3.31)
0.spring	0	0
	(.)	(.)
1.spring	0.206*	0.202*
	(2.25)	(2.20)
0.level3	0	0
	(.)	(.)
1.level3	0.682***	0.683***
	(7.18)	(7.19)
2.level3	0	-2.617**
	(.)	(-2.61)
_cons	-6.525*	-6.432*
	(-2.20)	(-2.17)
N	24003	25160

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure 45 Comparing the models before and after deleting certain observations.

interested in knowing the effect that each independent variable has on the probability of the event occurring. Does taking a course in the spring semester lead to an increase in the probability that a student might withdraw from the course? If so, what is the increase in the probability? Therefore, when we interpret the results, it is useful to know how the probability of the event occurring changes with changing values of the independent variables.

Non-Graphical Interpretation

There are several types of statistics that help us calculate the change in the probability relative to a change in the independent variable. Perhaps the most useful of these statistics is the discrete change, which is the change in probability given a change in the independent variable. To calculate discrete change in Stata, we will need to use the **mchange** command which is part of the **SPost package**. Therefore, you need to install the package in order to use the **mchange** command:

```
net from http://www.indiana.edu/~jslsoc/stata/

net install spost13_ado
```

Once you execute these commands, you can take advantage of the powerful tools that come with this package. To illustrate this, first fit the full logistic model:

```
logistic withdraw gpa c.gpa#c.gpa i.gender i.spring i.level3
```

Next, we will use the **mchange** command in order to calculate the discrete change in the probability when GPA changes by 10:

```
mchange gpa, delta(10)
```

We use the **delta(10)** option in order to tell Stata that we want to calculate the change in probability when GPA changes by 10. The output of this command is shown in **Figure 46**. This output shows

```
logit: Changes in Pr(y) | Number of obs = 25160

Expression: Pr(withdraw), predict(pr)

                |    Change    p-value
----------------+------------------------
gpa             |
           +1   |    -0.002      0.000
       +delta   |    -0.015      0.000
    Marginal    |    -0.003      0.000

Average predictions

                | No with~w   Withdraw
----------------+------------------------
  Pr(y|base)    |     0.980      0.020
```

Figure 46 The output from the **mchange** command.

```
logit: Changes in Pr(y) | Number of obs = 25160

Expression: Pr(withdraw), predict(pr)

                |    Change       From         To    p-value
----------------+------------------------------------------
gpa             |
           +1   |    -0.002      0.020      0.018      0.000
       +delta   |    -0.015      0.020      0.005      0.000
    Marginal    |    -0.003         .z         .z      0.000

Average predictions

                | No with~w   Withdraw
----------------+------------------------
  Pr(y|base)    |     0.980      0.020
```

Figure 47 The output from the **mchange** command with the stat() option.

that "on average, an increase in 10 points in the GPA decreases the probability of withdrawing from a course by 0.015." As you can see, this is a clear interpretation. We are no longer talking about logit functions, odds, or even odds ratio. We are talking about the change in the probability that a certain event will happen. We can also use the option **stat()** in order to see exactly how the probability changes:

```
mchange gpa, delta(10) stat(change from to pvalue)
```

We specified the **stat()** option in order to tell Stata that we want to see the change in probability, the value from which the probability changed, the value to which the probability changes when GPA increases by 10, and the p-value of the result. The output is shown in **Figure 47**. We see that, on average, the probability changes from 0.02 to 0.005 when GPA changes by 10. This is a decrease of 0.015.

We can also use the **mchange** command when the independent variable is binary:

```
mchange gender, stat(change from to pvalue)
```

Notice that we did not specify the **delta()** option. This is because the only meaningful change in the variable gender is from zero to one, given that the variable is binary. The output of this command is shown in **Figure 48**. The result shows that "on average, being a male increases the probability of course withdrawal from 0.014 to 0.022, which is a change of 0.007 ($p < 0.05$)."

The command can also be used when the independent variable has more than two categories:

```
mchange level3, stat(change from to pvalue)
```

```
logit: Changes in Pr(y) | Number of obs = 25160

Expression: Pr(withdraw), predict(pr)

                |    Change       From         To    p-value
----------------+----------------------------------------------
gender          |
 male vs female |     0.007      0.014      0.022      0.000

Average predictions

                | No with~w   Withdraw
----------------+----------------------
  Pr(y|base)    |     0.980      0.020
```

Figure 48 The output from the **mchange** command when the independent variable is binary.

```
logit: Changes in Pr(y) | Number of obs = 25160

Expression: Pr(withdraw), predict(pr)
```

	Change	From	To	p-value
level3				
200 level courses vs Other co~s	0.013	0.014	0.028	0.000
500 level courses vs Other co~s	-0.013	0.014	0.001	0.000
500 level courses vs 200 leve~c	-0.027	0.028	0.001	0.000

Average predictions

	No with~w	Withdraw
Pr(y|base)	0.980	0.020

Figure 49 The output from the **mchange** command when the independent variable is binary.

Here, we are looking at the change in probability depending on the level of the course. The output of the command is shown in **Figure 49**. The result indicates that, "on average, taking a 200-level course compared to taking courses classified as 'other' increases the probability of course withdrawal by 0.013 ($p < 0.05$). Taking a 500-level course compared to taking courses classified as 'other' decreases the probability of course withdrawal by 0.013 ($p < 0.05$). Taking a 500-level course compared to taking a 200-level course decreases the probability of course withdrawal by 0.027 ($p < 0.05$)."

Graphical Interpretation

As we saw in the previous section, interpreting the results in terms of the calculated probability is very intuitive and clear. In addition to calculating the discrete changes, it is also very useful if we can produce graphs that summarize the results. As an example, take the case of the independent GPA. We would like to know what is the change in the probability of withdrawing from a course when GPA changes. We use the **margins** command to tell Stata to calculate the probabilities of withdrawal at different values of GPA:

```
margins, at(gpa=(40(1)100))
```

We have already used this command when we were visualizing the effect of including a quadratic term. There are two differences. The first difference is that the logistic model includes other independent variables as well. The second difference is that we do not use the **predict(xb)** option. This is a very important point. We previously specified this option because we want to calculate the value of the logit function $\log(\frac{p}{1-p})$, since we were testing for linearity. What we now want to do is to calculate the value of p, the probability. By default, **margins** calculates probability. As we said, when visualizing the results, we are more interested in seeing the effect of the independent variables on the probability of the event occurring. **Figure 50** shows part of the output produced when we run the command.

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	.1696368	.074682	2.27	0.023	.0232628	.3160107
2	.1709642	.0703718	2.43	0.015	.033038	.3088904
3	.1716635	.0659951	2.60	0.009	.0423155	.3010115
4	.1717288	.0615925	2.79	0.005	.0510097	.292448
5	.1711597	.0572023	2.99	0.003	.0590452	.2832741
6	.1699608	.0528601	3.22	0.001	.0663569	.2735646
7	.1681423	.0485987	3.46	0.001	.0728906	.263394
8	.1657199	.0444483	3.73	0.000	.0786028	.2528371
9	.1627143	.0404363	4.02	0.000	.0834606	.241968
10	.1591515	.036587	4.35	0.000	.0874422	.2308607
11	.1550624	.0329219	4.71	0.000	.0905366	.2195882
12	.1504829	.0294595	5.11	0.000	.0927433	.2082224
13	.1454533	.0262149	5.55	0.000	.0940731	.1968335
14	.1400183	.0232	6.04	0.000	.0945472	.1854894
15	.1342264	.0204232	6.57	0.000	.0941977	.1742551
16	.1281292	.0178892	7.16	0.000	.093067	.1631914
17	.1217811	.0155989	7.81	0.000	.0912079	.1523544

Figure 50 Part of the output from the **margins** command.

What we see is that the command is calculating the probability at each value of the variable GPA. Since the other independent variables were not included in the **margins** command, they are set to their mean value.

We can tell Stata to plot these results using the **marginsplot** command:

```
marginsplot, noci
```

The graph is shown in **Figure 51**. Notice that the probability drops to almost zero for values of GPA that are higher than 80. We can also produce a more informative graph by including one of the categorical variables in the **margins** command:

```
margins, at(gpa=(40(1)100) level3=(0 1 2))
```

In this case, we are telling Stata to calculate the probabilities when GPA ranges from 40 to 100 for each value of the categorical variable level3. The output of this command is shown in **Figure 52**. This graph is interesting because it shows that for 500-level courses, the probability of withdrawal is close to zero no matter what the GPA is. The GPS has the largest effect on probability for 200-level courses. This makes sense since at this stage, students will still be uncertain about their choice of major. Getting a low GPA will raise a flag that perhaps they are enrolled in the wrong type of course.

A final example will illustrate how we can graph the results when all variables are categorical:

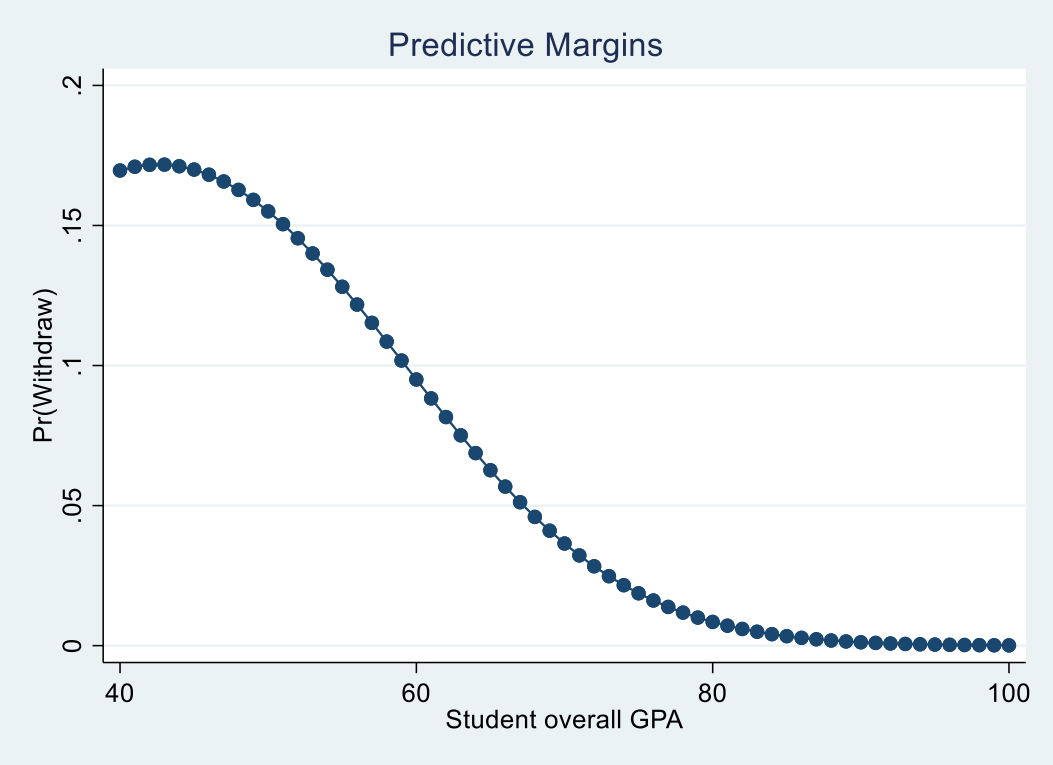

Figure 51 Graphing the probability of withdrawing from a course for each value of GPA.

```
margins, at(gender=(0 1) level3=(0 1 2))
```

Here, we are calculating the probabilities while varying the variables gender and level3. Note that since GPA was included in the logistic model but was omitted from the **margins** command, the calculations will be done by taking the mean value of GPA. We can visualize the result by using the **marginsplot** command. The result of running the command are shown in **Figure 53**. We see that the probabilities are calculated for females in the three different course levels, as well as for males. We also see that at the 500-level courses, the difference between females and males is minimal, while at the 200-level courses the difference is considerable.

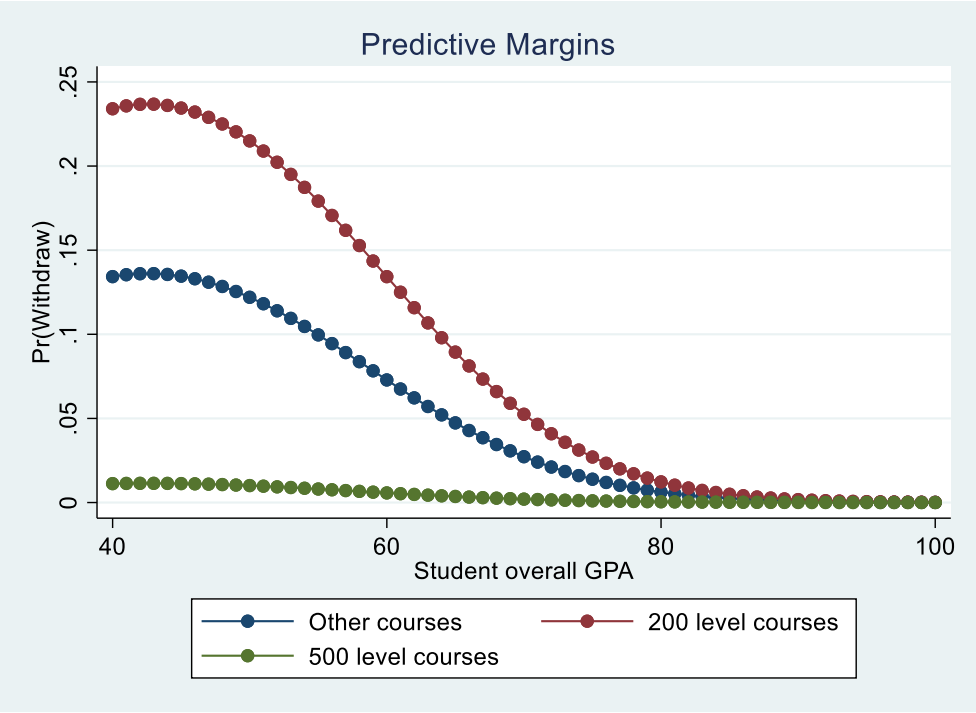

Figure 52 Graphing the probability of withdrawing from a course for different values of GPA for different level courses.

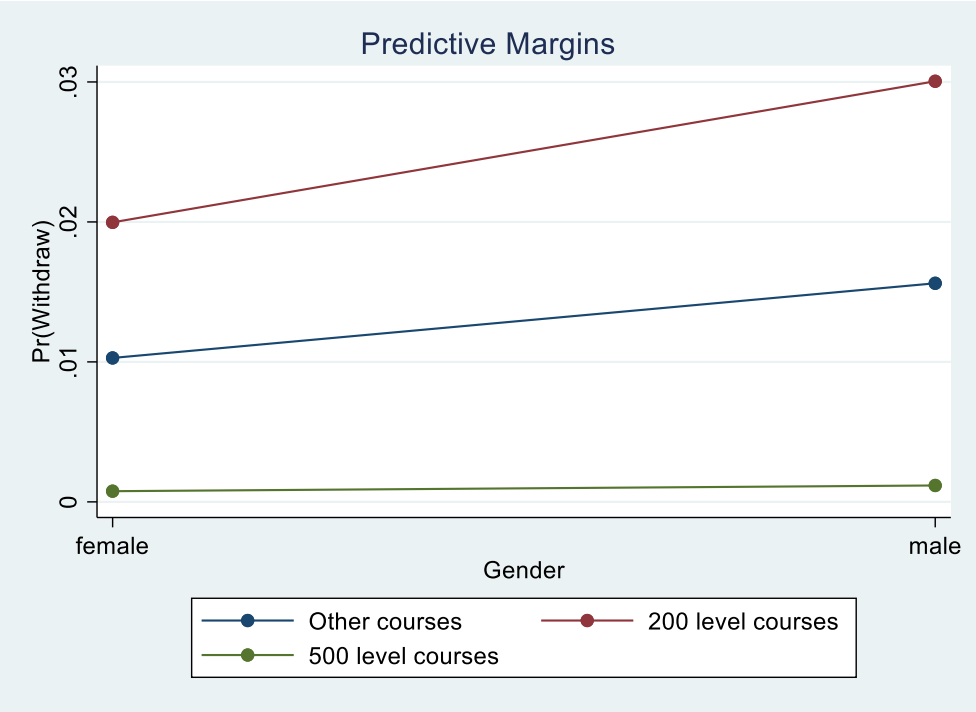

Figure 53 Graphing the probability of withdrawing from a course for different values of gender for different level courses.

References

Hilbe, J.M. (2009). Logistic Regression Models. CRC Press.

Hosmer, D.W. & Lemeshow, S. (2000). Applied Logistic Regression. 2nd edition. Wiley.

Long, J.S. & Freese, J. (2014). Regression Models for Categorical Dependent Variables using Stata. 3rd ediction. Stata Press.

Mitchell, M.N. (2012). Interpreting and Visualizing Regression Models using Stata. Stata Press.