Logistic Regression using Stata

NAJIB MOZAHEM

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
- 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.

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

Two-by-Two Tables

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

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.

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

The Odds

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

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

The Odds

- 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/(1-0.5) = 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.

The Odds Ratio

- Now that we have the odds for each row, we can calculate the odds ratio:

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

- This means that the probability of the event, which is withdrawal in our case, is higher in the numerator, which is engineering students.

Two-by-three Tables

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

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.

Two-by-three Tables (Odds)

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

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

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

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

Two-by-three Tables (Odds Ratio)

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

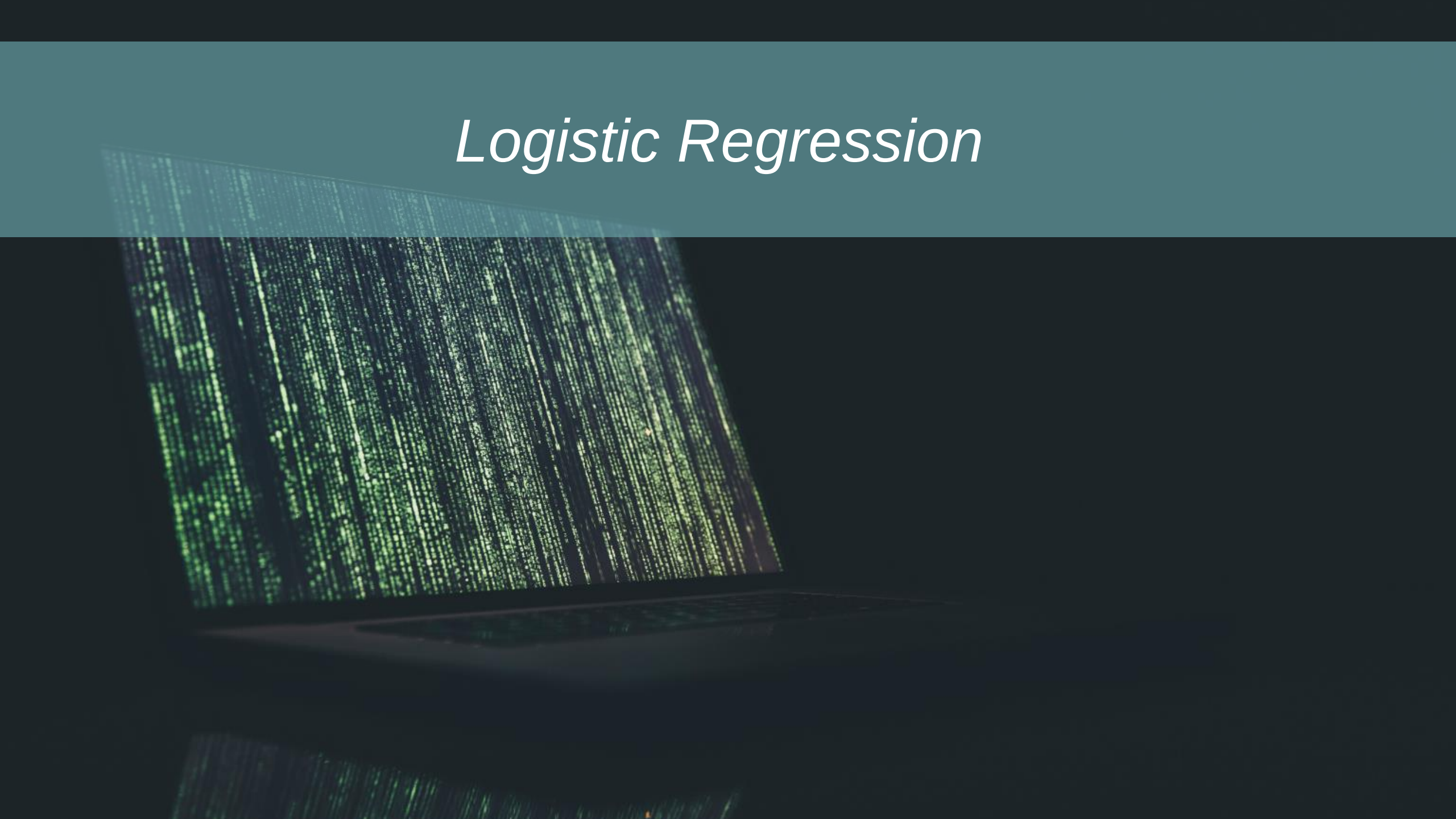
$$\frac{odds_{sophomore}}{odds_{junior}} = \frac{0.6}{0.2} = 3$$

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

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

- This exercise is useful when we want to compare the probabilities of an event across certain groups.
- This type of analysis however will not take us very far. 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?

Logistic Regression



Linear Regression

- 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

Linear Regression

- 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

$$y = 3x + 2$$

- 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

Logistic Regression

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

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

- In the above equation, *p* is the probability that the event will happen
- This function is called the logit function

Logistic Regression

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

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

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

Logistic Regression

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

- Now, instead of the log of the odds we have the odds on the left hand side of the equation
- 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

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:

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

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

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

The great news is that 9.12 is actually $e^{2.21}$

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

- 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

Example 1

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:

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

Here, *p* is the probability that a person has a heart problem.
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.

Example 2

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:

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

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.

Binary Variables

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

Gender

- 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
- 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

Logistic Regression with a Binary Variable

- Now that the variable gender has been quantified, it is possible to include it in a regression model:

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

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

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

Logistic Regression with a Binary Variable

- What does it mean that the coefficient of gender is -1.90?

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

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

- From these odds, we can calculate the odds ratio:

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

Logistic Regression with a Binary Variable

- 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.

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

Multiple Independent Variables

- The table includes the dependent variable outcome and the independent variables college and courses. Therefore, we have one binary variable and one continuous variable
- The equation of this model is:

$$\log\left(\frac{p}{1-p}\right) = a_1 x_1 + a_2 x_2 + b$$

Logistic Regression with Multiple Variables

- If we run the model, the output will be:

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

- 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:

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

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

Logistic Regression with Multiple Variables

- From these odds, we can calculate the odds ratio:

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

- 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}

Logistic Regression with Multiple Variables

- 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, $e^{2.22} = 9.21$

Categorical Variables

Standing

- What if we had a categorical variable that divided the observations into more than two groups?

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

- In this case, we cannot use a binary variable because there are three groups instead of two

Standing

	x_1	x_2
Sophomore	0	0
Junior	1	0
Senior	0	1

The number of binary variables needed is the number of categories minus one. In our case, we have four categories, so it is 3 – 1 = 2

$$\log(\frac{p}{1-p}) = a_1 x_1 + a_2 x_2 + b$$

Logistic Regression with a Categorical Variable

- If we fit this model to the data, the output will be:

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

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

Logistic Regression with a Categorical Variable

- 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:

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

Logistic Regression with a Categorical Variable

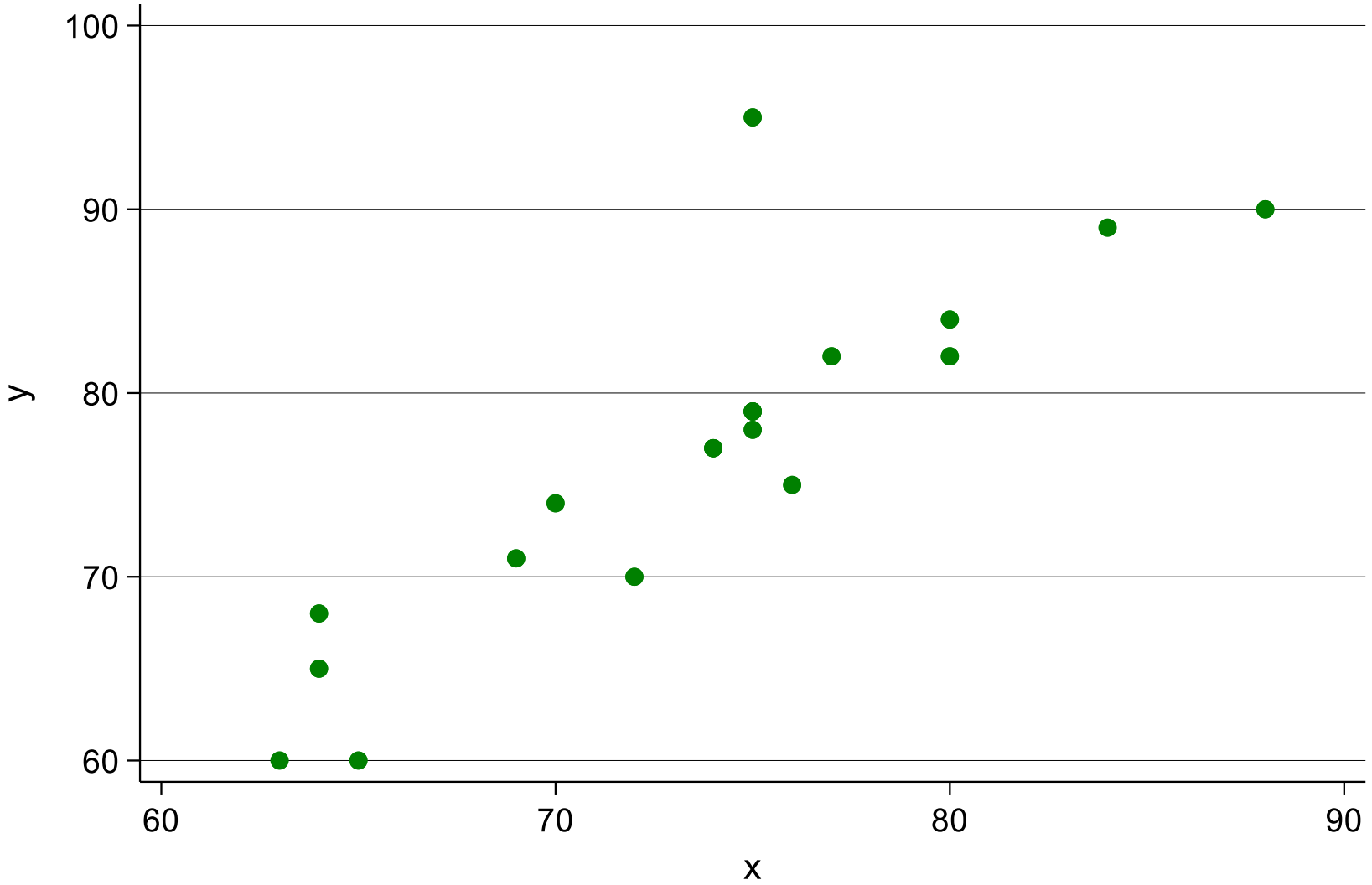
- We see that the exponent of the coefficient for each variable produces the odds ratio when we compare the group associated with the variable to the base group, which is the group that is assigned the values of zero.
- In other words, in our example, sophomore students are the base, or referent, group, 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.

Nonlinearity

Linear Regression

- 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

Scatter Plot



Logistic Regression

- 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

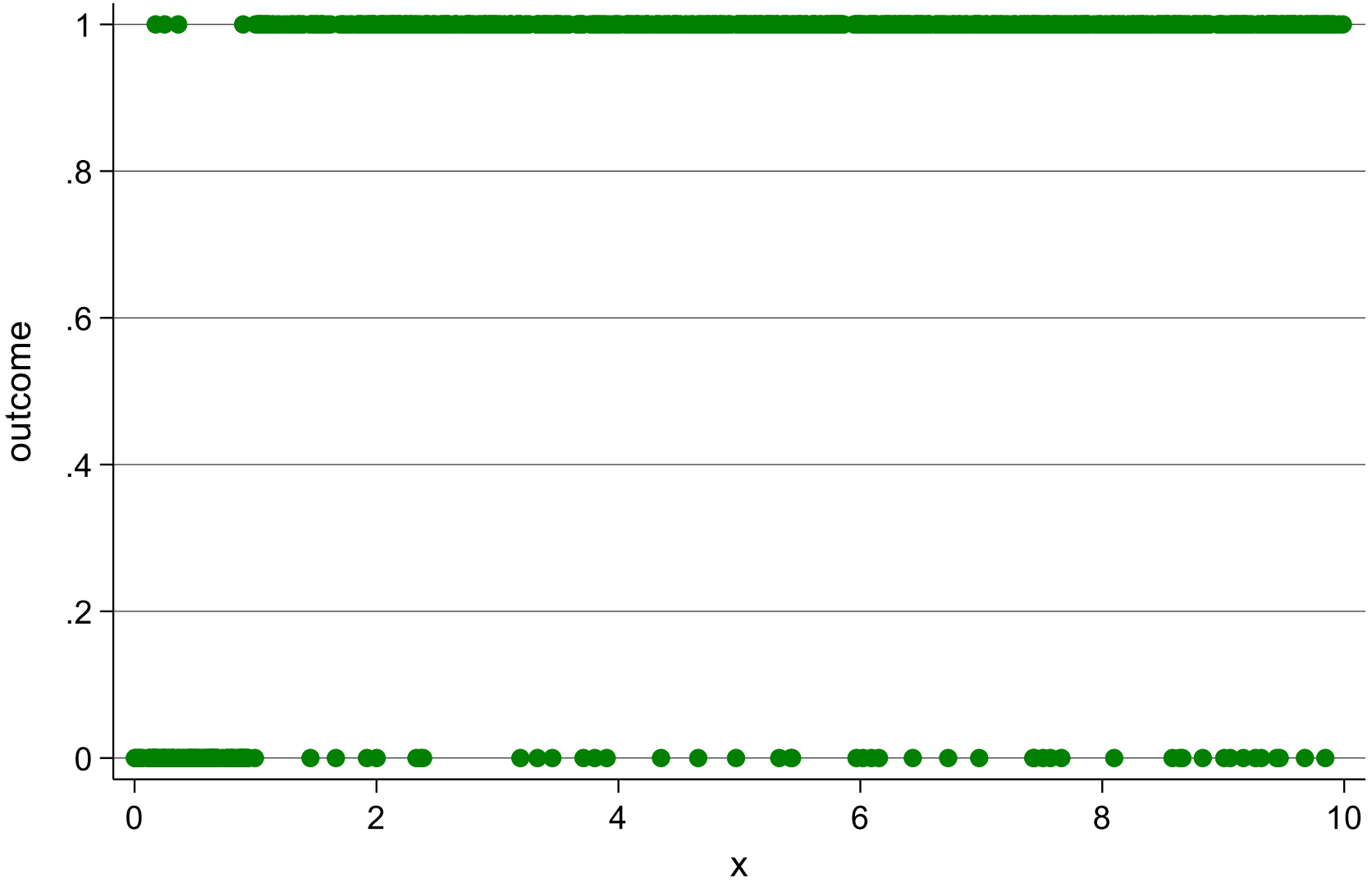
Box-Tidwell Test

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:

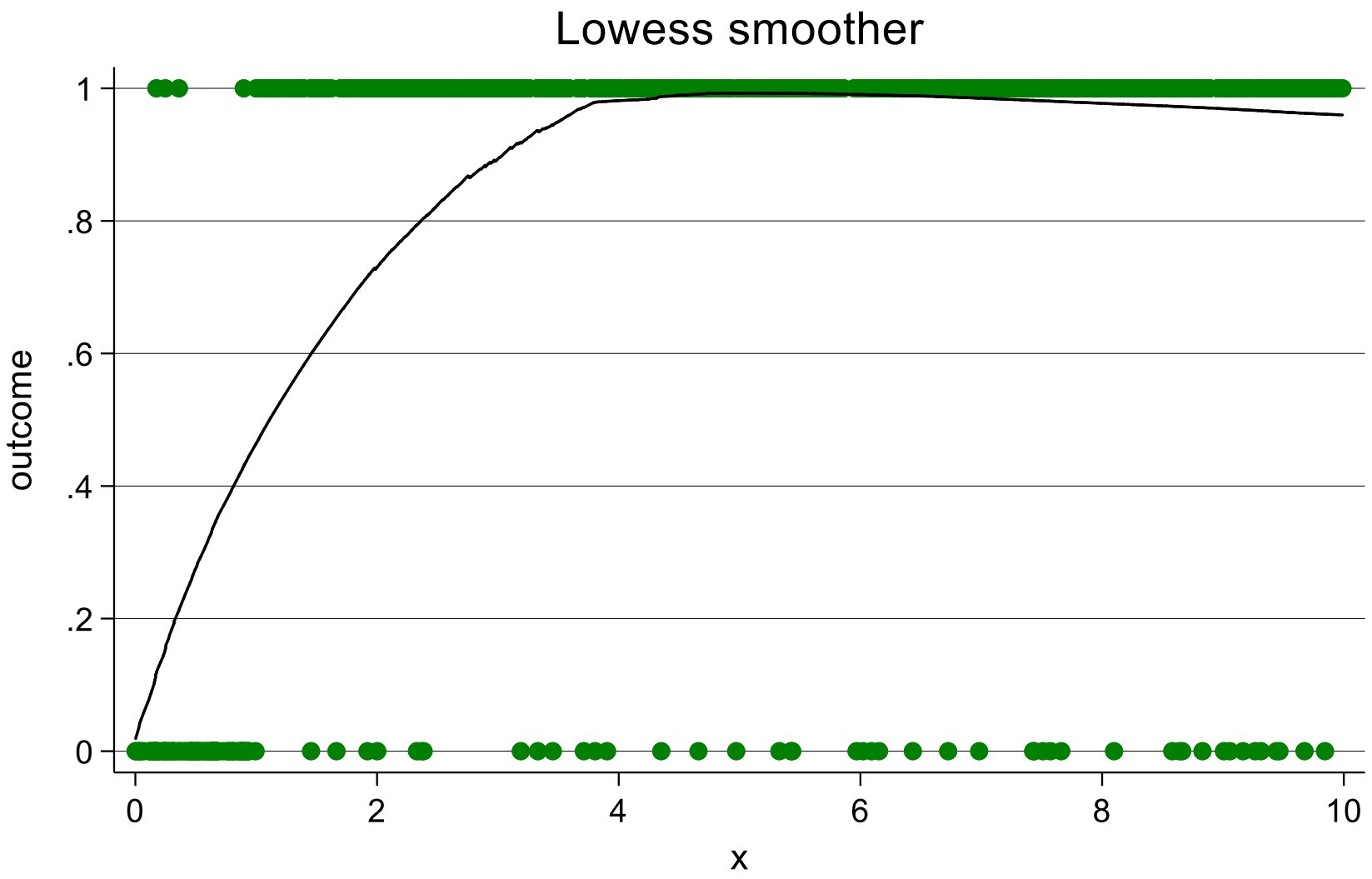
$$new\ variable = visit\ X\ log(visit)$$

After we calculate this new variable, we should fit a new logistic regression model that includes both the original variable and the new variable. If the new variable was significant, then the assumption of linearity between the logit function and the independent variable is violated

Graphical Test: Loess



Graphical Test: Loess



Graphical Test: Linearity of Slopes

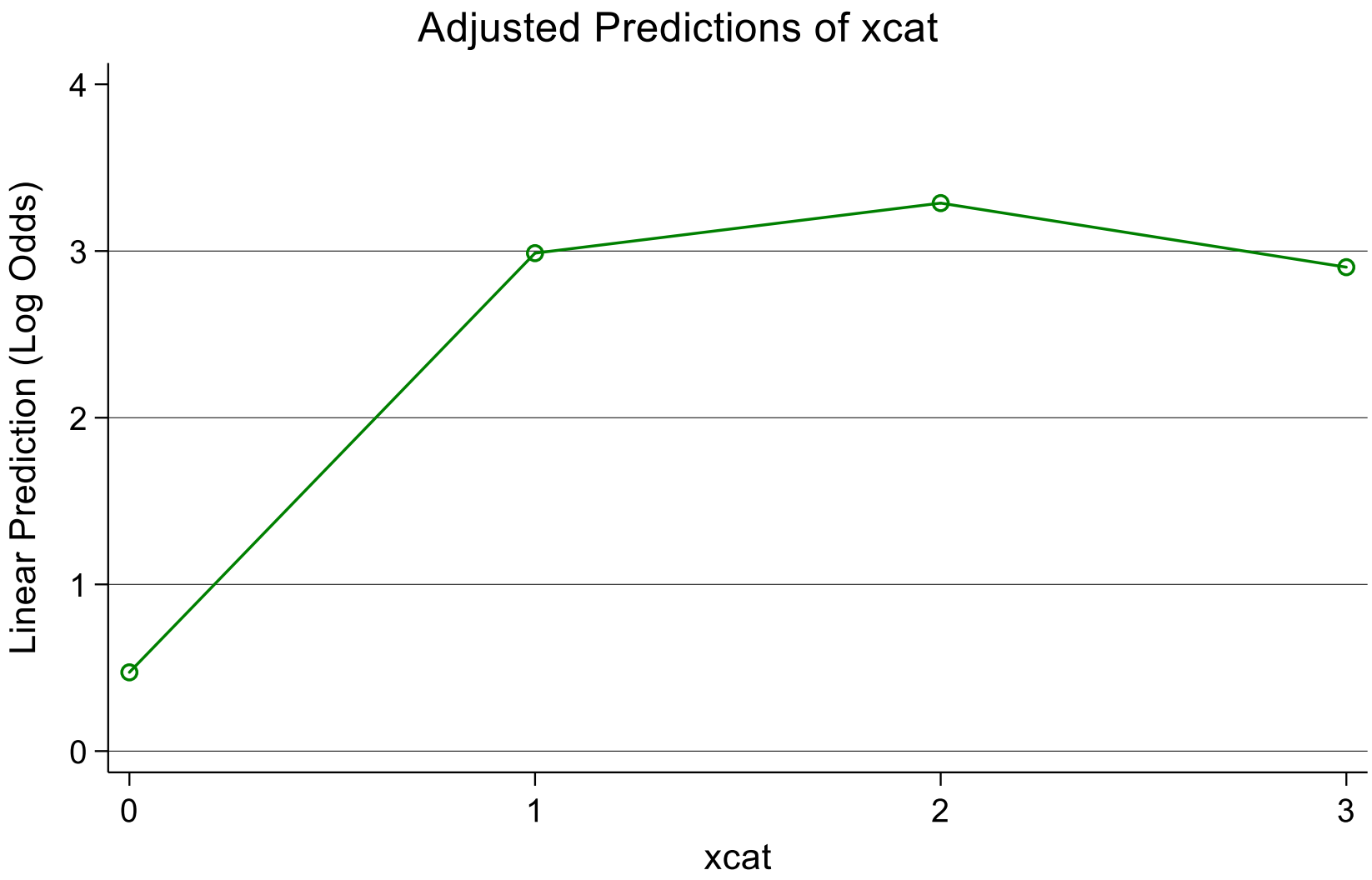
- The idea behind this test is that the independent variable x is categorized.
- Assume for example that we want to categorize the variable age, where age is between 18 years and 60 years. We create a new variable that takes on the value of zero when age is between 18 and 30, the value one when age is between 30 and 40, the value two when age is between 40 and 50, and the value three when age is greater than 50

Categorized age	Age
0	18≤age<30
1	30≤age<40
2	40≤age<50
3	50≤age

Graphical Test: Linearity of Slopes

- 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.

Graphical Test: Linearity of Slopes



Prediction and Model Fit

Prediction

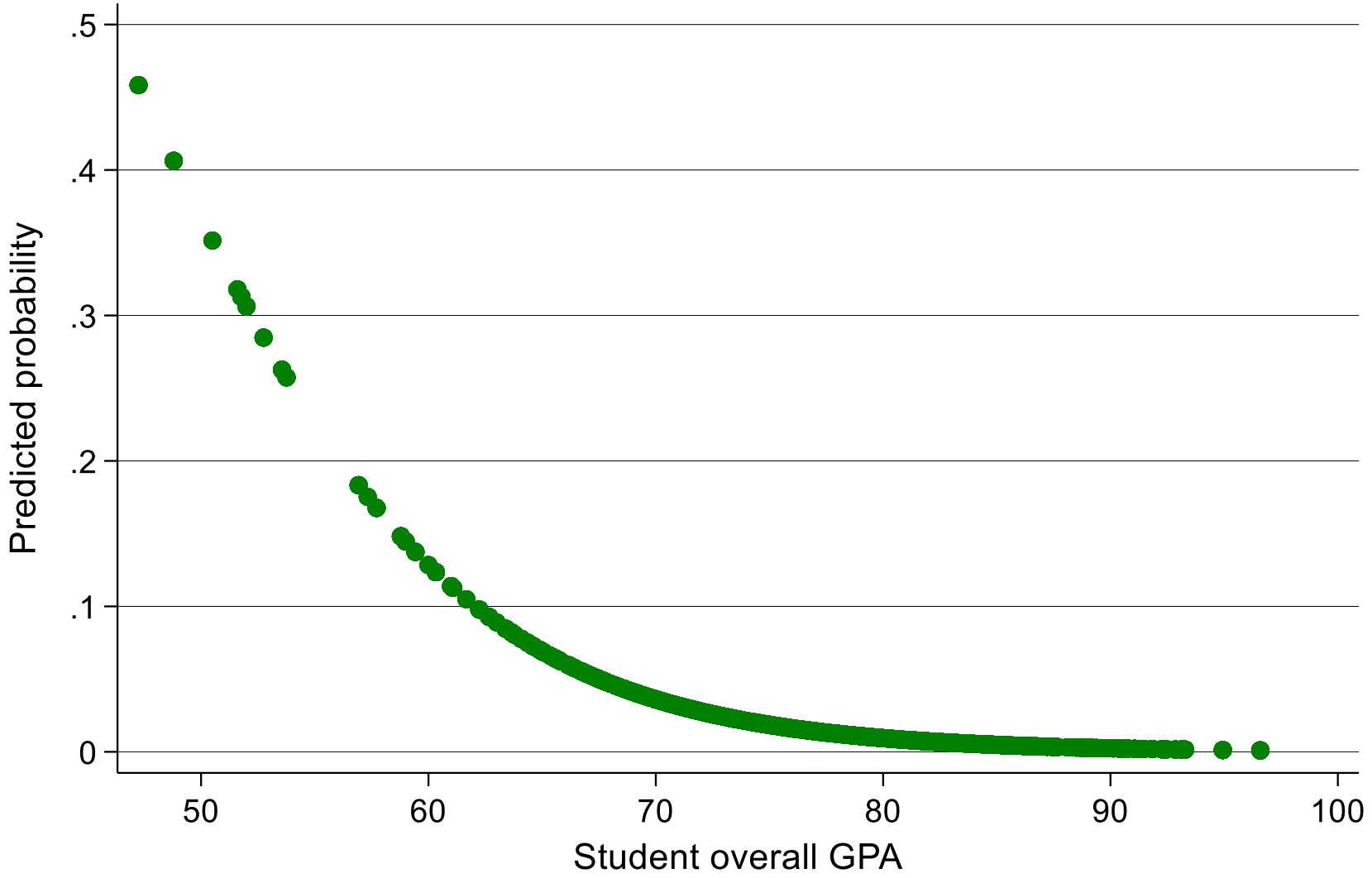
- In linear regression, because the left-hand side of the equation is the dependent variable y, we can easily calculate the predicted value of y and then plot it on the y-axis. In logistic regression however, the left hand side is not the dependent variable:

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

Prediction

- 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.
- This means that using logistic regression we can, for each observation, calculate the probability that the event will occur

Prediction – Calculating Probabilities



Model 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

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

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

Hosmer-Lemeshow GOF Test

- 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

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
- Usually, a cut-off value of 0.5 is used

Classification Tables

- 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
- Specificity on the other hand represents the probability that the model will correctly predict that the outcome has not occurred
- The ideal cutoff value is the one at which sensitivity and specificity are equal

Classification Tables (Example)

- 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%
- The model correctly classifies (7+16)/31 = 74.19% of the observations

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

ROC Analysis

- Another way to test the model fit is to use ROC curves, where we are interested in the area under the curve
- 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

Residuals

- 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 (which is probability in the case of logistic regression) in order to visualize the results

Influential Observations

- In logistic regression, we use the hat diagonal statistic and the delta-beta influence statistic in order to measure influence
- These statistics are plotted against the predicted probabilities in order to visualize the results

Guidelines for Detecting Residuals and Influential Observations

- 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, the below table offers some general guidelines that are useful in many situations

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

Plotting both Residuals and Influence