Understanding Regression

An Introduction to Predictive Analytics for Data Scientists

Najib A. Mozahem

Contents

Chapter 1

Linear Regression - The Theory

1.1 Simple Linear Regression

1.1.1 The Slope

In order to use linear regression, it is important for the student to understand the concept behind the technique. Fortunately, this can be accomplished without having to resort to complex mathematical equations. The important thing is to understand the idea.

I was once discussing with one of my colleagues whether universities should require students to attend classes. Some people argue that students who attend end up doing better, while others argue that this is not necessarily the case. In order to resolve this problem, we decided to look at the data. Table 1.1 displays the GPA and the attendance score of some students. The table isn't of much help, since it requires us to look at a large number

of columns and to compare these columns. This is why, whenever linear regression is involved, one of the first things that we should do is to produce a graph that will help us visualize the relationship. Figure 1.1 displays the scatter plot of the data points from Table 1.1.

Table 1.1: The data points.

GPA	**Attendance**
95	75
60	65
65	64
70	72
78	75
82	80
84	80
77	74
79	75
89	84
60	63
71	69
74	70
82	77
79	75
68	64
90	88
75	76
77	74

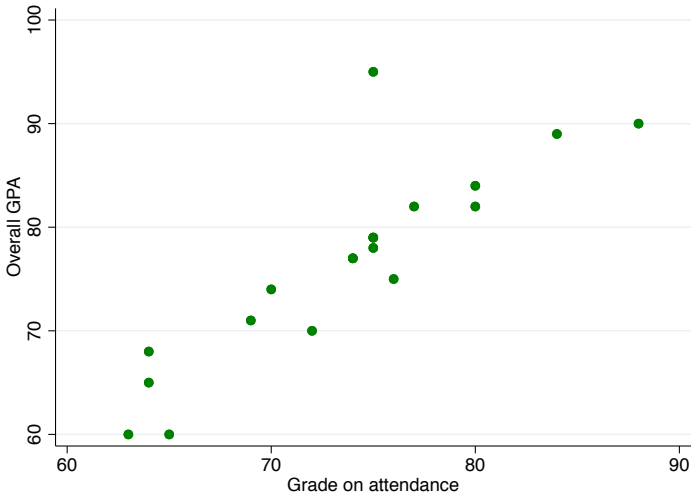

Figure 1.1: Scatter plot of the data points.

Looking at the graph, one might deduce that the higher the grade on attendance, the higher the overall GPA of the student. There seems to be some exceptions to this, most notably the student who has a 75 on attendance and a GPA of 95. However, most people would conclude that this seems to be the exception to the case. The scatter plot resembles a straight line, and the straight line has a positive slope. As you know, the equation of a straight line is:

$$y = ax + b$$

In our case, the y variable is GPA, and the x variable is attendance. The y variable is called the **dependent** variable because we believe that it's value depends on some other variables. The x variable is called the **independent** variable. Logically speaking, we would expect that the grade depends on the attendance level of the student. Therefore, our equation becomes:

$$GPA = a(attendance) + b$$

In this equation, the a represents the slope of the line, and the b represents the y-intercept. It is the value of the dependent variable when the independent variable is zero. The concept of the slope is very important, because it defines the relationship between the dependent and independent variables. 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 1. Now let us look at a case where the slope is negative:

$$y = -3x + 2$$

In this case, if x is equal to 2, y will be equal to -4, and if x is equal to 3, y will be equal to -7. Therefore, when x increases by 1, y will increase by -3, or in other words, y will decrease by 3.

Now you can see that 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 ate, 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.

In the case of our scatter plot, we saw that the graph has the shape of a line with a positive slope. However, what is the magnitude of the slope? In order to know, we use linear regression. Linear regression is the statistical tool that we use in order to find the equation of the best-fit line that represents the data. The word best-fit line is very important. There are an infinite number of lines that we can draw for any given scatter plot. What linear regression does is that it finds the line that fits the data the best. This is usually done by minimizing the square of the error terms. I do not want you to worry about this now. We will cover this in more detail later. For now, the most important thing to know is that we use linear regression in order to calculate the values of a and b in the equation:

$$GPA = a(attendance) + b$$

If we perform linear regression, the output will tell us that the following is

the equation of the best-fit line:

$$GPA = 1.22(attendance) - 13.20$$

You do not need to worry how we got these numbers. The statistical software will calculate them for us. Later on in this course, we will be seeing how to do this. For now, just look at the values. We see that the slope is 1.22. This means that if a student increases his or her attendance grade by one point, their GPA will increase by 1.22.

1.1.2 R-Squared

So far, we have seen how linear regression helps us calculate the slope, and how the slope helps us understand the nature of the relationship between the dependent variable and in the independent variable. It was also stated that the line which is calculated is the best-fit line. However, just because something is the best doesn't mean that it is good. If you got the best grade in your class on an exam, and that grade was a 40 out of 100, you still got a bad great, even though it was the best. The same logic applies to linear regression. The fact is that no matter what the relationship between the two variables is, if you ask any statistical software to calculate the best-fit line, the software will provide you with the equation of the line, even if the line was not a good fit. To illustrate this, look at the scatter plot shown in Figure 1.2.

Clearly the relationship does not resemble a line. Nonetheless, ask a statistical

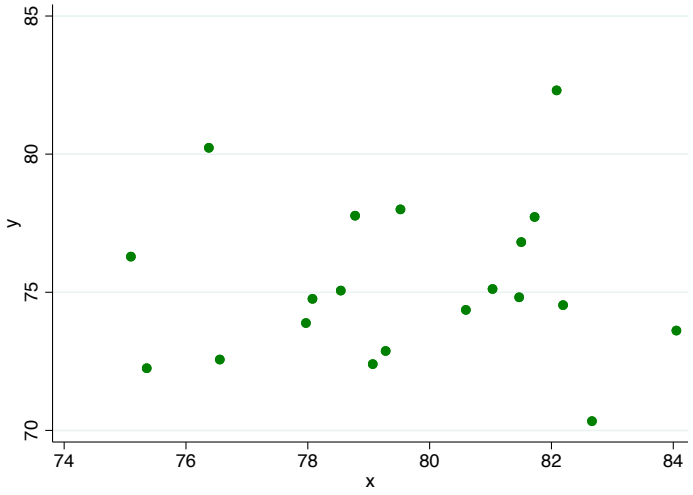

Figure 1.2: A scatter plot where there is no clear relationship between the variables y and x.

software to find the best-fit line, and the answer will be:

$$y = 0.019(x) + 75.05$$

As the above example illustrates, just because the statistical software gives us the equation of the best-fit line, we should not assume that the line actually fits the data well.

So what can we do in order to know if the best-fit line is actually a good line? We look at something that is called R-squared. This statistic calculates the proportion of the variation in the dependent variable that is explained by the line. This statement may seem complicated but it is actually easy to understand. In our original example, the dependent variable is GPA. Different students have different GPAs, and what we are trying to do is to explain how the value of GPA varies when we take into consideration

the grade on attendance. A line that fits the data well will do a good job in explaining the variation in the dependent variable with respect to the independent variable. A line that does not fit the data well will fail to explain this variation. R-squared is calculated by dividing the variation that is explained by the line by the actual variation that is observed in the independent variable:

$$R^2 = \frac{variation\,of\,the\,dependent\,variable\,explained\,by\,the\,line}{variation\,observed\,in\,the\,dependent\,variable}$$

If the line explains most of the observed variation, the value of R-squared will be close to 1 because the value of the numerator will be close to the value of the denominator. Otherwise, if the line fails to explain a large part of the variation, then the value of R-squared will be close to 0. In the figure above, the value of R-squared is 0.0005, which is very close to 0. This means that the best-fit line does not fit the data well. In the original dataset, which included the variables GPA and attendance, the value of R-squared is 0.75, which is considered to be good.

1.1.3 The P-value

We now come to one of the most important concepts in statistics, and it is the p-value. There is a saying that a broken clock is right twice a day. If my favorite TV show starts at 8:00, and the moment that it starts I look at my watch, and I see that it is 8:00, I assume that my watch is correct. However, this might not be the case. What if the watch was broken and it had stopped at 8:00. It just happened that I looked at it is 8:00. Although this might be the case, most people would not make that assumption. Instead, we assume

that the watch is working. Even though there is a probability that the watch can be broken, this probability is too small, so we go about our day as usual.

The p-value tells us about the probability that a certain observation was due to randomness and nothing else. An example will illustrate. Imagine a women who was sitting next to you and drinking tea. This women likes to drink her tea with milk. As she is drinking her tea, she suddenly turns to you and says that the tea tastes better when you pour the milk into the tea. She says that if you pour the tea into the milk the taste will not be the same.

This is a strange statement to make. Why should there be a difference? In order to test her, you conduct a small test. You blindfold her and tell her that you will give her a cup of tea that has milk in it. Her job is to identify whether you poured the milk into the tea or vice versa. You give her the first cup and she guesses correctly. Does this mean that she is right? Does it mean that she can tell the difference? No necessarily. Maybe it was a random lucky guess. After all, she has a probability of 0.5 to guess the correct answer. So you decide to try with another cup. Again she guesses correctly. Did she prove her point? The probability of her making two lucky guesses is 0.5 x 0.5 = 0.25. What if she guesses three cups in a row? The probability for her to do that purely out of luck is 0.125. The probability for her to guess four cups completely out of luck is 0.0625, and the probability for her to guess five cups is 0.03125.

How many guesses must she make in order for her to prove that what we are observing is not due to luck, or randomness? The common cut-off value for the probability is 0.05. If the probability of something happening out of randomness or luck is less than 0.05, then we reject the claim that what we are observing is purely due to luck or randomness. It would be safe for us

to conclude that our observation is in fact **significant**. In the case of the lady drinking tea, if she is able to answer correctly five times in a row, then we reject the claim that she was just lucky and we can conclude that she actually know what she is talking about.

The above explanation is not mathematically sound, but it does not matter. What matters is the idea. The example illustrates the idea. This brings us back to our line. As we saw, the statistical software has told us that the best-fit line is:

$$GPA = 1.22(attendance) - 13.20$$

The software has also told us that the value of R-squared is 0.75. Another piece of information that the statistical software gives us is the p-value of the slope. If the p-value of the slope is less than 0.05, then we reject the claim that the value that we have observed is due to randomness. Instead, we conclude that the value is significant. If, on the other hand, the p-value of the slope is greater than or equal to 0.05, then we cannot reject the claim that the value is due to randomness. In our case, the statistical software tells us that the p-value of the slope is less than 0.05, therefore we reject the claim that the value that we obtained for the slope might have been due to randomness and nothing more. We therefore conclude that the value of the slope is significant.

1.1.4 The Residuals

Now that we have the equation for the best-fit line, we can calculate how accurate the line is. This is done by predicting values. We predict values, when we enter the value of the independent variable into the equation in order to calculate the value of the dependent variable. What we want is for the predicted value to be as close to the observed value as possible. Table 1.2 shows the predicted values of GPA when we use the linear equation. It also shows the residuals, which are calculated using the equation $(actual\, GPA) - (predicted\, GPA)$. The residuals are very important for two reasons. The first reason is that they tell us how accurate our equation is. If the residuals are large, then this means that the predicted values are not close to the actual values. Therefore, what we want is for the residuals to be as small as possible. We also want almost half of the residuals to be negative and the other half positive. The reason for this is that if most residuals are negative, then this means that most predicted values are greater than the actual values. This implies that the line is always over predicting the values.

Graphically speaking, most of the points would lie below the line. If, on the other hand, most of the residuals are positive, then this means that the line is always under predicting the values. Graphically, this would mean that most of points would lie above the line. A well-fit line must pass between the points, which means that roughly half of the points are above the line and the other half below the line. If this is the case, the average of the residuals would be close to zero, since when we add the residuals the positive values will cancel out the negative values. If you calculate the average of the residuals in the above table, you will find that it is around 0.001, which is very close to zero. Figure 1.3 plots both the scatter plot and the best-fit line

Table 1.2: Calculating the predicted values and the residuals.

GPA	**Attendance**	**Predicted**	**Residuals**
95	75	78.18 16.82	
60	65	66	-6
65	64	64.78	0.22
70	72	74.53	-4.53
78	75	78.18	-0.18
82	80	84.27	-2.27
84	80	84.27	-0.27
77	74	76.96	0.04
79	75	78.18	0.82
89	84	89.15	-0.15
60	63	63.56	-3.56
71	69	70.87	0.13
74	70	72.09	1.91
82	77	80.62	1.38
79	75	78.18	0.82
68	64	64.78	3.22
90	88	94.02	-4.02
75	76	79.4	-4.4
77	74	76.96	0.04

on the same graph. We can see from the graph that the line passes through the points. Some of the points are above the line, and others are below the line. We also see that the points are in general close to the line. This means that the magnitude of the residuals is generally small.

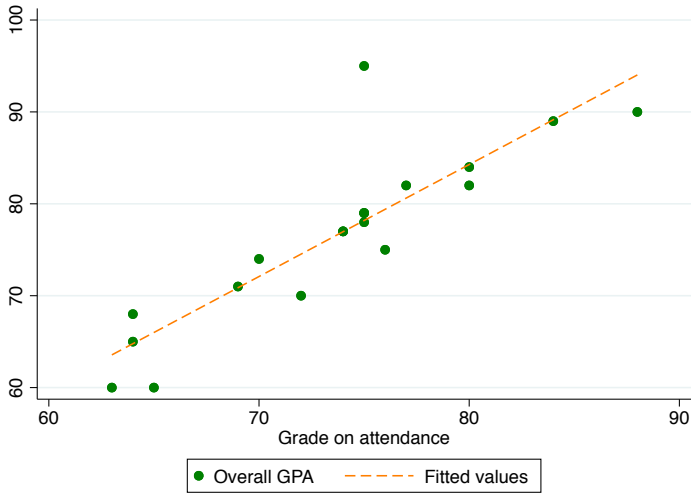

Figure 1.3: The best-fit line as computed by the simple regression model.

The second reason that residuals are extremely important is that after we fit a linear model, we need to test the validity of the assumptions that we have made. Residuals play a crucial role in this. This topic will be discussed in a later section. For now, what is important is that you understand how to calculate the predicted value and the values of the residuals.

1.2 Multiple Linear Regression

At this point, it would make perfectly good sense for someone to object to the fact that we have been using the grade on attendance to predict the overall GPA of the student. Surely there are other factors at play here. It cannot be that the only thing that affects students' GPA is how much they attend classes. This is a very valid concern. Actually, you would be hard pressed to see a publication where the author uses a model that includes only one independent variable. The reality is that the dependent variable

is influenced by several factors. We started with the simple case of a single independent variable in order to illustrate the concept of linear regression. Now that we understand the basic concept, expanding it to include more than one independent variable is quite easy.

1.2.1 The Slopes

When we have more than one independent variable, the equation becomes:

$$y = a_1x_1 + a_2x_2 + ... + b$$

The independent variables are represented by the variable x, and the slopes associated with each are represented by the variable a. There is nothing new in the equation except that we added extra terms. In order to understand what each slope represents, consider the following case:

$$y = 2x_1 + 3x_2 + 4$$

To calculate the value of y we will need to know the values of both x_1 and x_2. Assume that we start with x_1 equal to 2 and x_2 equal to 3. This means that y will 2(2) + 3(3) + 4 = 17. If the value of x_1 increased by one and became 3, the value of y will become 2(3) + 3(3) + 4 = 19. As you can see, the dependent variable increased by 2 points, which is the value of the coefficient that is attached to the independent variable that increased by one unit. This means that nothing has changed. The coefficient by which the independent variable is multiplied still tells us about the relationship between

the dependent and the independent variable. We know that if x_1 increased by one, y will increase by two.

What about the other independent variable, x_2? When x_1 equals 2 and x_2 equals 3, y is equal to 17. If the value of x_2 increases by one and becomes 4, y will become $2(2) + 3(4) + 4 = 20$. The value of the dependent variable increases by the value of the coefficient which is attached to the independent variable, which happens to be 3 in this case. There is nothing new here. This is just the same as when there was one independent variable. No matter how many independent variables there are, when we want to understand the relationship between any single independent variable and the dependent variable, we just look at the value of the coefficient that is associated with the independent variable.

An important point to note here is that since the slope of x_1 is 2 and the slope of x_2 is 3, we see that a change in x_2 results in a larger change in the dependent variable than a change in x_1. If x_1 increases by one the dependent variable will increase by 2, but if x_2 increases by one the dependent variable will increase by 3. We therefore conclude that the effect of x_2 on y is larger than the effect of x_1 on y.

Let us now see this concept in action. In our previous example, we gathered data about the students' GPAs and their grade on attendance. Since it is too simplistic to assume that GPA only depends on attendance, we go around and ask the students how many hours they studied over the past week. The results are shown in Table 1.3. The values of GPA and attendance are the same as before. The only new thing in the table is the column that records the number of hours studied by the student over the past week. So now what? Previously we created a scatter plot of the variables GPA and

Table 1.3: The data points.

GPA	**Attendance**	**Study**
95	75	45
60	65	19
65	64	15
70	72	22
78	75	28
82	80	33
84	80	40
77	74	30
79	75	28
89	84	37
60	63	10
71	69	29
74	70	31
82	77	36
79	75	38
68	64	25
90	88	30
75	76	30
77	74	30

attendance in order to see whether there was any type of pattern. Let us now do the same for the new variable study. Figure 1.4 shows the scatter plot for the variables GPA and study. Once again, we see evidence that students who have a higher GPA tend to study more than students who have a low GPA. Therefore, it would make sense to include this variable in our model.

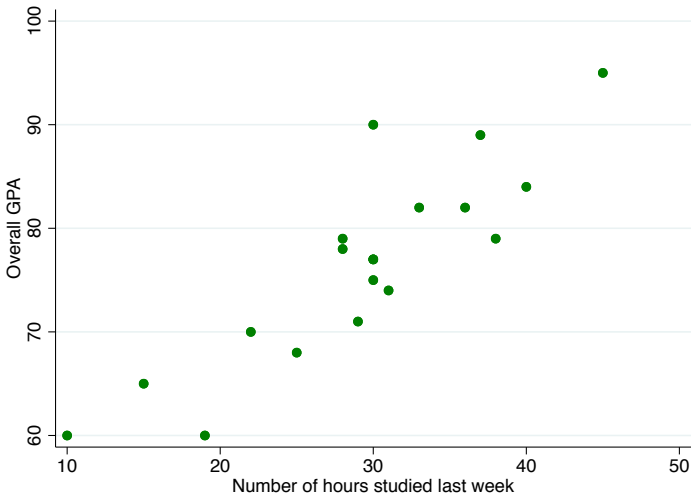

Figure 1.4: Scatter plot of GPA and study.

When we run a linear regression model that includes GPA as the dependent variable and the variables attendance and study as the independent variables, we find that the best-fit line has the following form:

$$GPA = 0.71(attendance) + 0.59(study) + 6.98$$

What does this mean? It means that if two students have the same level of attendance, but one student studies one hour more than the other, then that student will have a GPA that is 0.59 higher than the GPA of the other student. It also means that if two students study the same amount of time, but one student has one extra point on his or her attendance than the other, the that student will have a GPA that is 0.71 higher than the other student. We also see that the coefficient of the variable attendance is larger than the coefficient of the variable study. This means that attendance has a larger

effect on the GPA than studying.

1.2.2 R-squared

Which model is better? In the original model, we just had one independent variable. We now have two. How can we choose? There are several ways to test this. In this section, we will only look at one of these ways (there is a section later on dedicated to this topic). In the original mode, the value of R-squared was 0.75. In the new model, the value of R-squared is 0.90, which is much closer to one. As you recall, R-square is a measure of the proportion of the variation in the dependent variable that is explained by our model. The closer it is to one, the better our model at explaining the variation. Therefore, we see that by taking into consideration the variable study, the model now accounts for around 90% of the observed variation. This means that this model is better than the first model.

1.2.3 The P-values

The meaning of the p-values also remains the same, whether we have one independent variable or more than one. The only difference is that there is one p-value associated with each independent variable. In order to know whether an independent variable is significant, all we need to do is to look at its p-value. In the output of the model that includes both attendance and study, both p-values are found to be less than 0.05. As you recall, a p-value that is less than 0.05 means that we can reject the claim, or hypothesis, that what we are observing is just due to chance or randomness. This means that

the values for both coefficient are significant.

1.2.4 The Residuals

Now that we have our equation, we can use it to predict the values of GPA. Once we have the predicted values, we can calculate the residuals. Once again, there is no difference between having one independent variable or two. All we need to do is to plug in the values into our equation. The results are shown in Table 1.4. Take the first row for example. We know that the equation is:

$$GPA = 0.71(attendance) + 0.59(study) + 6.98$$

We replace the values of attendance and study to get:

$$GPA = 0.71(75) + 0.59(45) + 6.98$$

Therefore, the error is 95–86.78 = 8.22. This is a large error, but it seems to be the exception. If you calculate the average of the errors you will find it to be around -0.00053, which is very close to zero. As you recall, when the average is close to zero, what we have is that the positive errors and the negative errors are cancelling each other out, which is what we want since this shows that the line passes through the data as opposed to passing above or below the data.

Table 1.4: Calculating the predicted values and the residuals.

GPA	Attendance	Study	Predicted GPA	Residuals
95	75	45	86.78	8.22
60	65	19	64.37	-4.37
65	64	15	61.31	3.69
70	72	22	71.11	-1.11
78	75	28	76.77	1.23
82	80	33	83.27	-1.27
84	80	40	87.38	-3.38
77	74	30	77.24	-0.24
79	75	28	76.77	2.23
89	84	37	88.46	0.54
60	63	10	57.66	2.34
71	69	29	73.09	-2.09
74	70	31	74.98	-0.98
82	77	36	82.9	-0.9
79	75	38	82.65	-3.65
68	64	25	67.19	0.81
90	88	30	87.19	2.81
75	76	30	78.66	-3.66
77	74	30	77.24	-0.24

1.3 Binary Variables

So far, the independent variables have been numerical in nature. Both GPA and attendance levels are recorded as numbers. Sometimes however, including variables that are not numeric in nature is necessary. For example,

what if we wanted to investigate whether the variation in GPA could be explained by the gender of the students? We saw that students who attend more have higher GPAs, but what if we wanted to investigate whether males have higher GPAs than females? 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. Table 1.5 shows the result of this process. By assigning numbers to gender, we have quantified the variable. We can now include it in the regression model. The equation will be:

$$GPA = a_1(attendance) + a_2(study) + a_3(gender) + b$$

If we run the regression model, we will find that the value of the coefficients are as such:

$$GPA = 0.51(attendance) + 0.56(study) + 4.29(gender) + 21.18$$

We already know how to interpret the coefficients of the variables attendance and study. However, what does it mean that the coefficient of gender is 4.29? Remember that for males the value of gender is zero, while for females the value of gender is one. Take the following example. Calculate the predicted value of GPA for a student who has an attendance grade of 80, and who studied for 35 hours in the last week. Do this once for a male and once for

Table 1.5: Adding a binary variable to study the effect of gender.

GPA	**Attendance**	**Study**	**Gender**	**Binary**
95	75	45	female	1
60	65	19	male	0
65	64	15	male	0
70	72	22	male	0
78	75	28	female	1
82	80	33	female	1
84	80	40	female	1
77	74	30	male	0
79	75	28	female	1
89	84	37	female	1
60	63	10	male	0
71	69	29	male	0
74	70	31	male	0
82	77	36	female	1
79	75	38	male	0
68	64	25	male	0
90	88	30	female	1
75	76	30	male	0
77	74	30	male	0

a female:

Male: $0.51(80) + 0.56(35) + 4.29(0) + 21.18 = 81.58$

Female: $0.51(80) + 0.56(35) + 4.29(1) + 21.18 = 85.87$

What we see is that the female has a higher GPA, and that the GPA is higher

by 4.29. Therefore, the coefficient of the binary variable is the difference between an individual who belongs to the group that is assigned a zero value and an individual who belongs to the group that is assigned the value one.

At this point, you might ask what if we coded the variable differently? What if we assigned a value of zero to females and a value of one to males? If you do this, the output from linear regression will be:

$$GPA = 0.51(attendance) + 0.56(study) - 4.29(gender) + 25.47$$

The coefficients for attendance and study remain the same. Looking at the coefficient of gender, we notice that the magnitude is the same as before but that the sign has been reversed. We also notice that the value of the intercept term has changed. Let us now do the same calculations as before:

Male: $0.51(80) + 0.56(35) - 4.29(1) + 25.47 = 81.58$

Female: $0.51(80) + 0.56(35) - 4.29(0) + 25.47 = 85.87$

The predicted values remain the same. As you can see, it doesn't make a difference how we code the variable. What matters is that we be aware of the coding in order to properly interpret the value of the coefficient. In the first case, a positive coefficient indicated that the group which was assigned the value one (females) had a higher GPA. In the second case, the negative coefficient indicated that the group which was assigned the value one (males) had a lower GPA.

1.4 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? For example, students enroll in different majors. Assume that the students included in our dataset were majoring in business, engineering, biology, or philosophy. In this case, we cannot use a binary variable because there are four groups instead of two. What we can do however, is to use more than one binary variable, as shown in Table 1.6. If you look at the column for the variable

Table 1.6: Coding the categorical variable.

	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$
Business	0	0	0
Engineering	1	0	0
Biology	0	1	0
Philosophy	0	0	1

x1, you will notice that the variable takes a value of one for engineering, and zero otherwise. X2 takes on a value of one for biology and zero otherwise. X3 takes on a value of one for philosophy 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 four categories, so it is 4 – 1 = 3. The equation now becomes:

$$GPA = a_1(attendance) + a_2(study) + a_3(gender) + a_4x_1 + a_5x_2 + a_6x_3 + b$$

For a business student, x_1, x_2, and x_3 are zero. For an engineering student, only x_1 is one and the rest are zero. For a biology student only x_2 is one. For a philosophy student only x_3 is one. Assume that we ran the regression model, and that we got the following output:

$GPA = 0.47(attendance) + 0.43(study) + 4.13(gender) + 2.31x_1 + 2.17x_2 + -3.45x_3 + 23.02$

How do we interpret this result? It is actually simpler than it looks. The coefficient of x_1 is 2.31. This variable is one only when the student is an engineering student. Therefore, if a student is engineering we add 2.31 to the predicted GPA. The coefficients of x_2 and x_3 do not matter because the values of x_2 and x_3 for an engineering student are zero. So an engineering student has a GPA that is 2.31 points higher, but higher than who? Let us calculate the GPAs of female students, one from each major, who have a grade of 80 on attendance, and who have studied 35 hours the last week:

Business: $GPA = 0.47(80)+0.43(35)+4.13(1)+2.31(0)+2.17(0)+-3.45(0)+23.02 = 80$

Engineering: $GPA = 0.47(80) + 0.43(35) + 4.13(1) + 2.31(1) + 2.17(0) + -3.45(0) + 23.02 = 82.31$

Biology: $GPA = 0.47(80)+0.43(35)+4.13(1)+2.31(0)+2.17(1)+-3.45(0)+23.02 = 82.17$

Philosophy: $GPA = 0.47(80) + 0.43(35) + 4.13(1) + 2.31(0) + 2.17(0) + -3.45(1) + 23.02 = 76.55$

The difference between a business student and an engineering student is 2.31,

which is the coefficient of x_1. The difference between a business student and a biology student is 2.17, which is the coefficient of x_2. The difference between a business student and a philosophy student is negative 3.45, which is the coefficient of x_3. Therefore, as you can see, the coefficients of each binary variable represent the difference between individuals who are assigned a value of one for that variable and between the students who have been assigned zeros for all variables. This is why the group for which all the binary variables are zero is called the referent group. The coefficients compare each group with the referent group. In our case, the referent group is business.

1.5 Quadratic Terms

Assume that someone told you that a student's command of the English language also affects his or her GPA. The argument here is that students who are better at English, are in a better position to read more and to express themselves more. In addition, they might be more confident, and this would affect their performance. Table 1.7 displays the GPAs along with the grade obtained on the English course. The scatter plot of the above data is show in Figure 1.5.

We can see that students with a higher grade on English tend to have a higher GPA. If we fit a simple linear regression model, we get the following equation:

$$GPA = 0.80(english) + 21.95$$

Table 1.7: Data points for the variables GPA and English.

GPA	**English**
95	95
60	55
65	55
70	58
78	66
82	69
84	82
77	67
79	70
89	87
60	54
71	60
74	64
82	71
79	73
68	57
90	80
75	70
77	72

The output will also tells is that the p-value of the coefficient of the independent variable English is less than 0.05, so the result is significant. In addition, the R-squared value of the model is 0.89, which is very lose to one. Everything looks good. If we take a closer look at the scatter plot, we will notice that the dots don't seem to fall on a line. This is best illustrated by drawing the

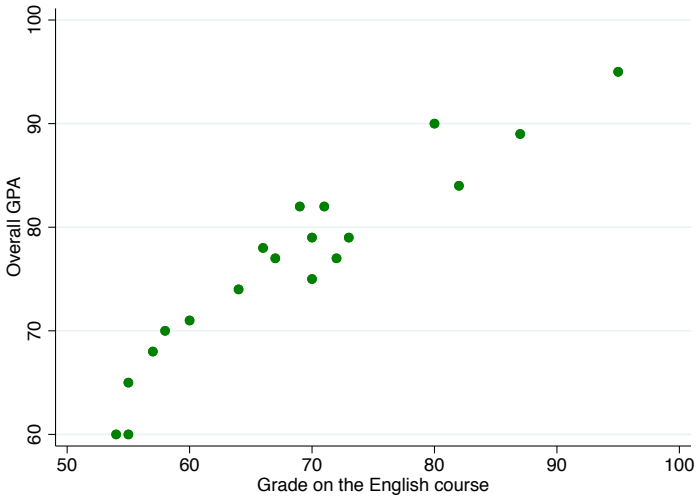

Figure 1.5: Scatter plot of the variables GPA and English.

best-fit line on the same curve. Figure 1.6 shows the result obtained.

We notice that there seems to be a steep rise in the dots initially, and that the rise tends to level off. When we suspect that the relationship between two variables might be non-linear, we can include a quadratic term in order to test our suspicion. If you recall from high school algebra, the equation of a quadratic formula is:

$$y = ax^2 + bx + c$$

This is the same as the linear equation only with an extra quadratic term, where the variable x is squared. We can do the same in our model. Instead of fitting this model:

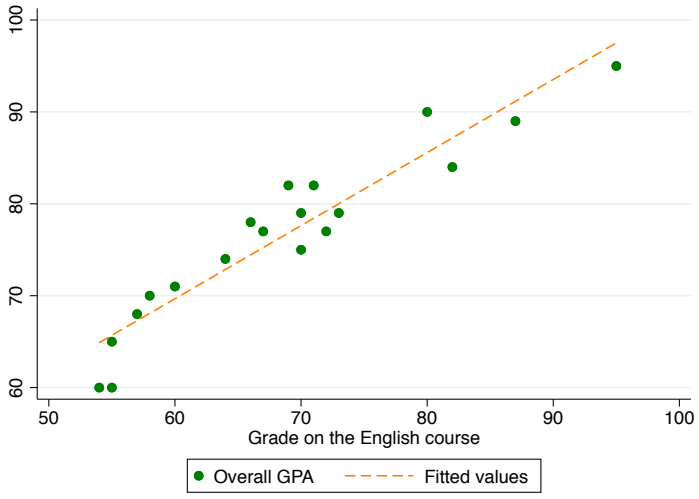

Figure 1.6: Drawing the best-fit line resulting from regressing GPA on English.

$$GPA = a(english) + b$$

We can fit this model:

$$GPA = a_1(english)^2 + a_2(english) + b$$

Since we already have the values of the variable English, we can just create a new column that contains the square of these values. Table 1.8 displays the result of squaring the variable English. We can next fit a linear regression model by including these two variables. The output of such a model will be:

$$GPA = -0.01(english)^2 + 2.35(english) - 32.82$$

Table 1.8: Data points after we add the square of the variable English.

GPA	**English**	**English^2**
95	95	9025
60	55	3025
65	55	3025
70	58	3364
78	66	4356
82	69	4761
84	82	6724
77	67	4489
79	70	4900
89	87	7569
60	54	2916
71	60	3600
74	64	4096
82	71	5041
79	73	5329
68	57	3249
90	80	6400
75	70	4900
77	72	5184

The output will also indicate that the p-value of the quadratic term is less than 0.05, which means that it is significant. The R-squared value of this model is 0.92, while the R-squared value of the model that did not contain the quadratic term was 0.89. Therefore, we can conclude that including the quadratic term is the right thing to do. This is why when you perform

linear regression, you should always start by producing plots. Plots are an excellent way for us to investigate what sort of relationship exists between the dependent variable and each independent variable. Any good regression analysis must start with graphs.

1.6 Checking Model Fit and Assumptions

Linear regression models make several assumptions about the data. The validity of the model depends on the validity of these assumptions. This is why, one of the most important topics in regression, is testing the assumptions. This is usually done after we do linear regression. We first find the best-fit model, and then we test the assumptions that linear regression makes using the best-fit model. In this section, we will go over some of the most important assumptions.

1.6.1 Prediction

The first thing that you should do after you fit a model is to see whether the values predicted by the model are close to the observed values. This can be easily accomplished by plotting the predicted values against the observed values. If the predicted values are similar to the observed values, then the scatter plot will lie along the diagonal line that represents the equation y = x. We had previously fit a model in which the dependent variable was GPA and the independent variables were attendance, study, and gender. The best-fit line turned out to be the following:

$$GPA = 0.51(attendance) + 0.56(study) + 4.29(gender) + 21.18$$

Figure 1.7 plots the predicted values against the observed values. The figure also shows the diagonal line along which the points should fall if the predicted values and the observed values are similar. In our case, the predicted values seem to be close to the actual values.

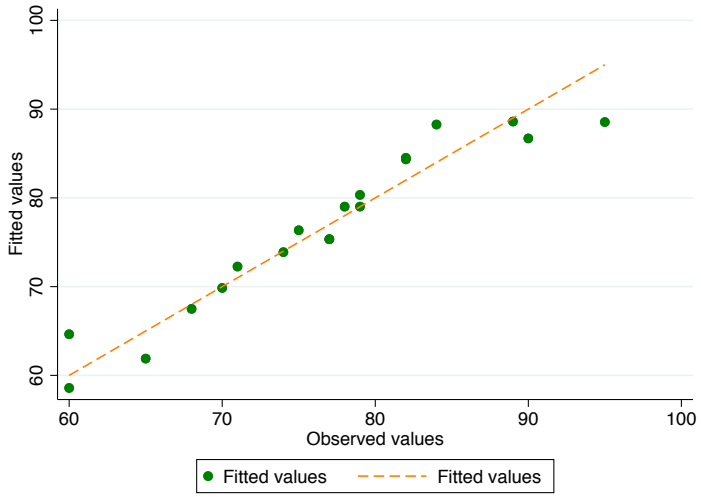

Figure 1.7: Checking model fit: Plotting the predicted values against the observed values (The dashed line represents the 45 degree diagonal along which the points should fall if the model was well fit).

1.6.2 Residuals

We have previously seen that the residuals are calculated using the following formula:

$$Observed\,value - Predicted\,value$$

Linear regression makes several assumptions about the distribution of the residuals. This is why after we fit a model, we need to calculate the residuals and test whether these assumptions are valid or not.

Normality

One assumption is that the residuals have a normal distribution. In order to test this, we can plot the histogram of the residuals. Let's look at an example. We had previously fit a model in which the dependent variable was GPA and the independent variables were attendance, study, and gender. The best-fit line turned out to be the following:

$$GPA = 0.51(attendance) + 0.56(study) + 4.29(gender) + 21.18$$

Figure 1.8 displays the histogram of the residuals with an overlaid normal distribution. Looking at the graph we can see that the residuals do not seem to follow a normal distribution, since the left tail of the histogram is cut abruptly. This result casts a shadow on our model and should make us question the results.

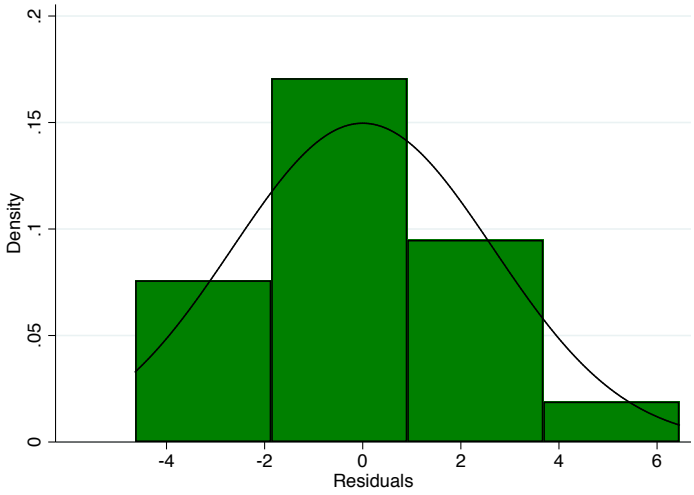

Figure 1.8: Checking the normality of the residuals: A histogram of the residuals overlaid with a normal curve.

Independence

Another assumption that is made by the linear regression model is that the residuals are independent. This means that if the residuals were plotted on the y-axis and any of the independent variables on the x-axis, we should see no pattern. Figure 1.9 for example is a plot of the residuals against the variable attendance. There doesn't seem to be a clear pattern, so it doesn't seem that the assumption of independence has been violated.

In order to see what sort of figure would indicate that the assumption of independence is violated, look at Figure **??** which plots the residuals against an independent variable named X. In the figure, the residuals display a pattern. We see that they tend to decrease and then start to increase again. This figure would raise a flag.

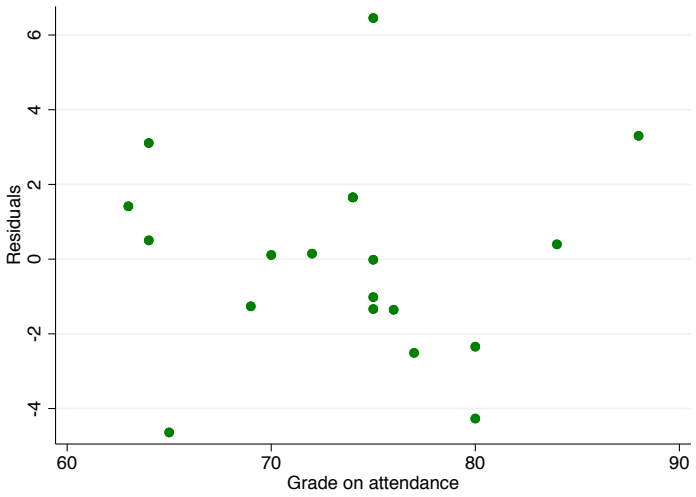

Figure 1.9: Testing for independence of the residuals: Plotting the residuals against the independent variable attendance.

If we have a multiple linear model that contains several independent variables, instead of plotting the residuals against each independent variable by itself we can just plot the residuals against the predicted values of the dependent variable. Figure 1.11 shows this graph for the model that contains the three independent variables attendance, study, and gender. Since there doesn't seem to be a patter, we can assume that the independence assumption is met. The graph however does show that there is another type of problem, which will be discussed next.

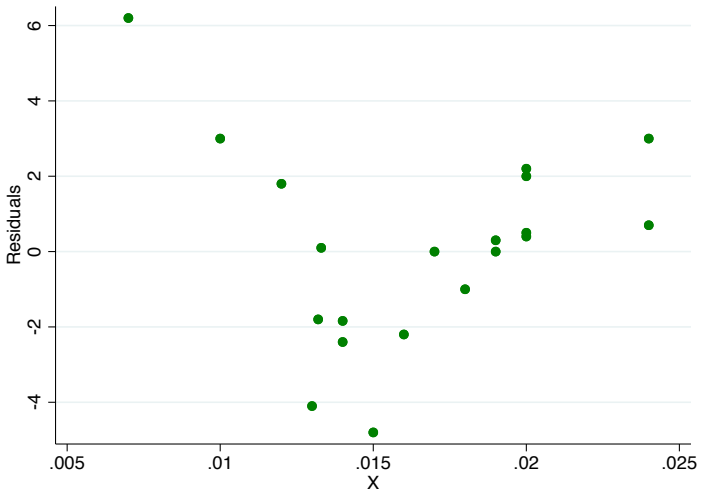

Figure 1.10: Testing for independence of the residuals: A case where the assumption is violated.

Constant Variance

In addition to being normal and independent, the residuals must also have a constant variance. This is called the homoskedasticity assumption. To investigate whether this assumption is valid or not, we again plot the residuals against an independent variable. This time however, instead of looking at whether there are patterns in the residuals, we would look at whether the variation of the residuals around the x-axis is constant. As you recall, some residuals are positive while others are negative. This is due to the fact that the line passes between the data points, so some of the data points are above the line while others are below the line. This means that for some points the observed value is greater than the predicted value, while in other cases the opposite is true. The residuals are therefore scattered on both sides of the y = 0 line, which is the x-axis. This can be seen from the two figures above. Homoskedasticity means that the variation of the errors around this

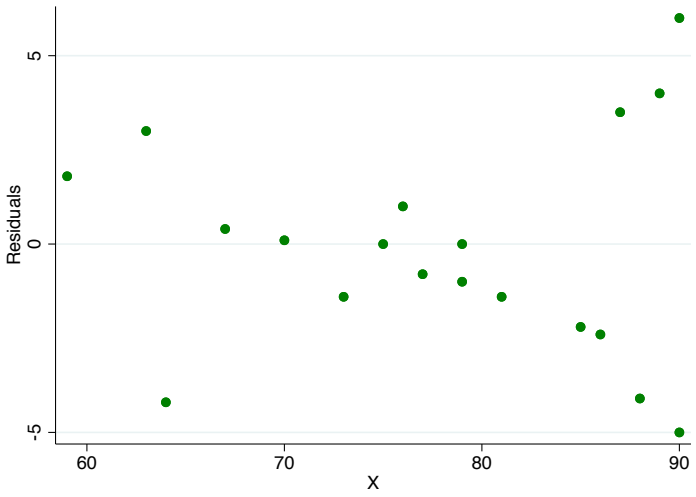

Figure 1.11: Testing for the independence of the residuals: Plotting the residuals against the predicted values of the dependent variable.

line should be constant. For example, Figure 1.12 shows a case where the assumption of homoskedasticity is clearly violated. We can see that the residuals tend to get further away from the x-axis as we move from left to right, thus indicating that their variance is increasing.

When we discussed the assumption of normality, it was stated that if we were running a multiple linear regression model where there were several independent variables, we can plot the residuals against the predicted values of the dependent variable. The same can be done here. In fact, we have already produced this graph (Figure 1.11). In this case the graph clearly shows that the assumption of homoskedasticity is clearly violated. The residuals near the two ends of the graph show a larger variability than the residuals near the middle.

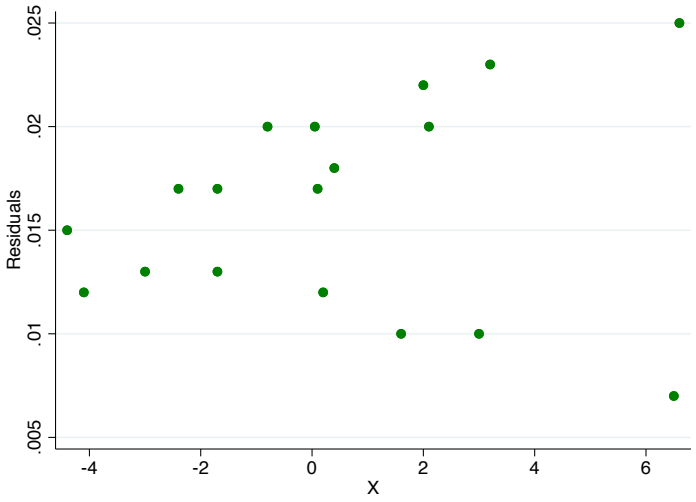

Figure 1.12: Testing for homoskedasticity: A case where the assumption is violated.

1.6.3 Multicollinearity

In the case of multiple linear regression we have more than one independent variable. An important assumption of multiple linear regression is that multicolinearity does not exist. This means that the independent variables should not be correlated with one another, and that no variable is a linear combination of other variables. This basically means that knowing the values of one or more of the independent variables should not allow us to predict the value of another independent variable.

If there are only two independent variables in the model, then testing for multicolinearity is easy. All we have to do is to calculate the correlation between the two independent variables. However, it is more often the case that there are more than two independent variables in any given model. Although one might think that we should calculate the correlation between

each pair of independent variables, this is not the best solution, since multicollinearity might be a result of an independent variable being a linear combination of two or more other independent variables. Therefore, to test for multicollinearity, we can calculate the variance inflation factor (VIF) for each independent variable. Multicollinearity exists if the value of the VIF for any variable is greater than 10. If this is the case, it might be necessary to eliminate the variable from the analysis.

1.7 Diagnostics

What if we checked the model fit and the assumptions of independence, normality, and homoscedasticity and found that some of these assumptions were violated? Does this mean that we should just through away our model? Fortunately, the answer is no. What we should do at this point is to take a closer look at the individual data points in order to see whether some points are responsible for the problems that we have uncovered. If this is the case, we will then have to decide what to do with these observations.

1.7.1 Outliers

An outlier is an observation that does not fit well with the rest of the point. It is quite easy to identify outliers because they are located very far from the rest of the points. Outliers by themselves are not a problem. There are some cases where an outlier can cause problems and other cases where it is not the case. Let us look at an example. Figure 1.13 shows a scatter plot of the exact same data that we have been using so far except that there is

an addition observation, which is colored in red in the figure. All the other points remain unchanged. This red dot represents a student who does not attend classes (he or she has an attendance grade of 30because the point that represents him or her is very far from the other points.

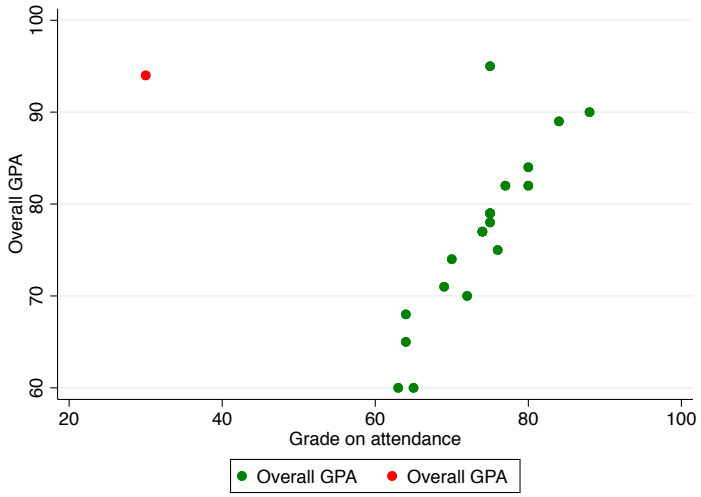

Figure 1.13: Outliers: The red dot is a new observation.

In order to see why this outlier may be a problem, look at Figure 1.14. The figure shows the scatter plot just like the one before it, but it also shows two lines. The line with short dashes is the best-fit line that is calculated when all the points (the original points and the new outlier point) are included. The line with long dashes is the best-fit line when we just include the original data points (the outlier is not included). What we see is that the outlier has a huge effect on the result.

When we include the outlier, the slope of the best-fit line decreases substantially,

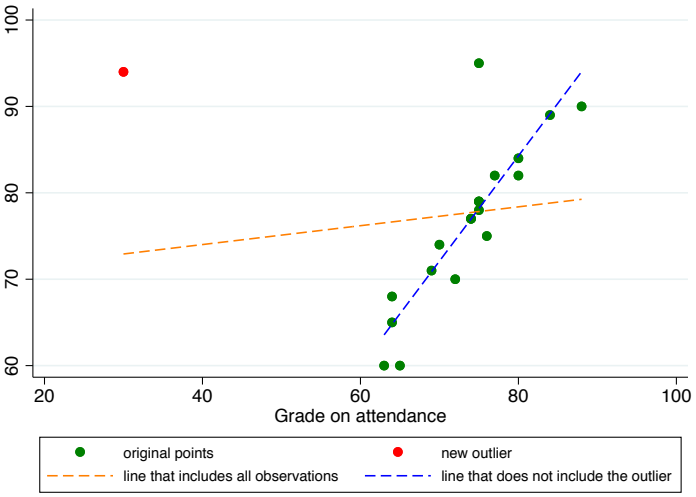

Figure 1.14: The effect that an outlier has: The short dashed line includes all observations, the long dashed line includes only the original observations (excludes the outlier).

thus weakening the effect that attendance has on the GPA. If you run both models, you will get the following:

Excluding outlier: $GPA = 1.22(attendance) - 13.20$

Including outlier: $GPA = 0.11(attendance) + 69.65$

Not only does the coefficient significantly drop, but the other statistics are also affected. In the first model the p-value for attendance is less than 0.05 while in the second model it is 0.59. Also, the R-squared value for the first model is 0.75 while in the second model it is 0.02. This is why, when we test our model for its goodness of fit and for the assumptions, looking at individual data points is very useful. If a single observation is causing too many problems, then a case can be made to exclude it. This is actually what I would do in this case. For the majority of students, there is a strong

relationship between attending and between GPA. However, there is a single student who is able to get very high grades without attending. This student however is the exception.

It is important to remember that not all outliers cause problems. Some outliers are more influential than others. This is why, in addition to checking whether there are outliers, we also need to calculate the influence of each observation.

1.7.2 Influential Observations

There are several statistics that are used in order to measure the influence that each observation has. The idea behind them is very similar. An observation is influential if the results obtained from running a regression model differ significantly when that observation is included and when it is excluded. Basically, a model that includes the observation is fit. The output from the model is recorded. Then another model is run this time excluding the observation. Again the output is recorded. We then check whether the output has differed significantly. If there is no significant change, the observation is not influential. If, on the other hand, there is a large change in the output, the observation is deemed significant and a flag is raised. The three most common statistics used to measure influence are DFBETAS, DFFITS, and Cook's D statistics. The difference between them is what change do they measure. DFBETAS measures the change in the regression coefficients when an observation is excluded. DFFITS and Cook's D statistic measure the change in the predicted values when an observation is excluded.

Table 1.9 shows the observations and the three statistics calculated for each

and every one of them. Notice that in all cases the magnitudes of the three statistics are small except in the case of the very last observation which is the outlier. As we can see, all three statistics agree that this is a very influential point.

1.8 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 (based on an F-test). 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 (again based on an F-test) 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

Table 1.9: Influence statistics: Calculating DFBETAS, DFFITS, and Cook's D for all observations.

GPA	Attendance	DFBETAS	DFFITS	Cook's D
95	75	.1279324	.4392341	.085373
60	65	.233627	-.4746602	.1004364
65	64	.1798848	-.3292105	.0529771
70	72	-.0073852	-.1699561	.0148148
78	75	.0011442	.0039286	8.17e-06
82	80	.0616364	.1036714	.0056474
84	80	.0961818	.1617761	.0136056
77	74	-.0035099	-.0165167	.0001444
79	75	.0079601	.0273297	.0003951
89	84	.27144	.368775	.0675833
60	63	.3048394	-.5127347	.1175016
71	69	.0303108	-.1426349	.0105424
74	70	.0095857	-.0740856	.0028884
82	77	.042779	.0991198	.0051558
79	75	.0079601	.0273297	.0003951
68	64	.1310648	-.2398641	.0291732
90	88	.3999571	.4874556	.116761
75	76	-.0259086	-.0710761	.0026616
77	74	-.0035099	-.0165167	.0001444
94	30	-11.61065	12.04787	16.59359

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.

As you recall, the R-squared calculates how well the model explains the variation that is observed in the dependent variable. Therefore, when we are comparing two different models, we should favor the one with a higher value of R-squared. An important point to note here is that there is another statistics that is a variation of the R-squared statistic and it is called the

adjusted R-squared. If we are comparing two models with the same number of independent variables, we can use R-squared to guide our decision. If, however, the models contain different numbers of variables, it would be better to look at the adjusted R-squared. Just like R-squared, the adjusted R-squared is between zero and one, and the closer it is to one, the better.

You can also rely on the AIC and BIC statistics when you are 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.

Chapter 2

Linear Regression - Case Study

We are now in a position to use what we have learned in a small case study. In this section, we will look at a dataset that contains information about students. What we are interested in is understanding differences in overall GPA. The dataset contains the following variables:

- gpa: overall GPA of the student (this is the dependent variable)
- english: the average grade on all English courses taken by the student (data is taken from a non-English speaking country where the language of instruction in university is English)
- college: whether the student is in the engineering school or the business school (zero means business, one means engineering)
- credits: the total number of credits completed so far by the student
- gender: whether the student is a male or a female (zero means female, one means male)

- attendance: attendance and participation grade last semester
- siblings: Number of brothers and sisters that the student has
- income: family income per year ($)
- work: records whether the student works full time, part time, or whether the student doesnt work at all.

2.1 Simple Regression

We start by building simple models in which we include the variables separately. This will give us a feel as to the nature of the relationship between each independent variable and the dependent variable GPA. There are two types of independent variables in our dataset, continuous and binary. We start by looking at the continuous variables.

2.1.1 Continuous Variables

The first continuous variable that we will look at is attendance, which is a measure of how much a student attends his or her assigned classes. We would like to see whether students who attend classes more often get higher GPAs.

When it comes to continuous variables, it is always best to start with a scatter plot of the dependent variables against the independent variable. The scatter plot os shown in Figure 2.1.

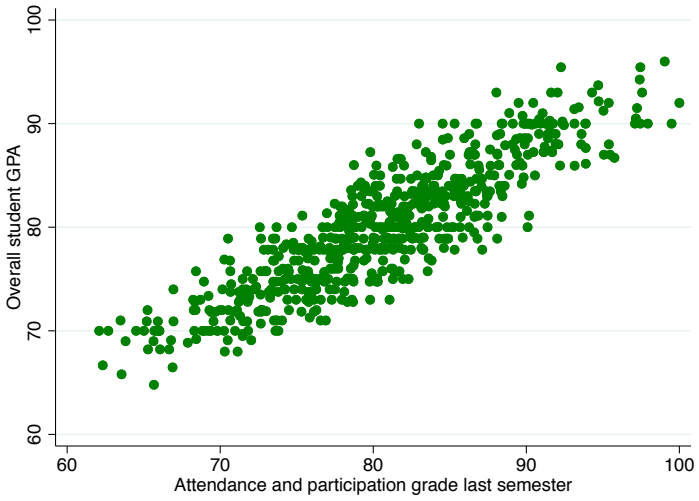

Figure 2.1: Scatter plot of GPA and attendance.

From the graph, we see strong evidence that the more a student attends, the higher the GPA. This result is further supported by fitting a simple linear regression model:

Source	SS	df	MS	Number of obs	=	666
				F(1, 664)	=	2305.64
Model	18376.0567	1	18376.0567	Prob > F	=	0.0000
Residual	5292.10793	664	7.97004207	R-squared	=	0.7764
				Adj R-squared	=	0.7761
Total	23668.1646	665	35.591225	Root MSE	=	2.8231

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.7406767	.0154253	48.02	0.000	.7103885	.7709649
_cons	20.27157	1.246345	16.26	0.000	17.82432	22.71882

We see that the coefficient of the independent variable attendance is both positive and signififcant (p-value is less than 0.05). The output tells us that when attendance increases by one-unit, GPA increases by 0.74 points.

Another continuous variable that we need to look at is the grade on English. We would expect that students with a better command of the English language would get higher GPAs betcause they will be able to both understand and communicate the material better. As usual, we start with a scatter plot. The scatter plot os shown in Figure 2.2.

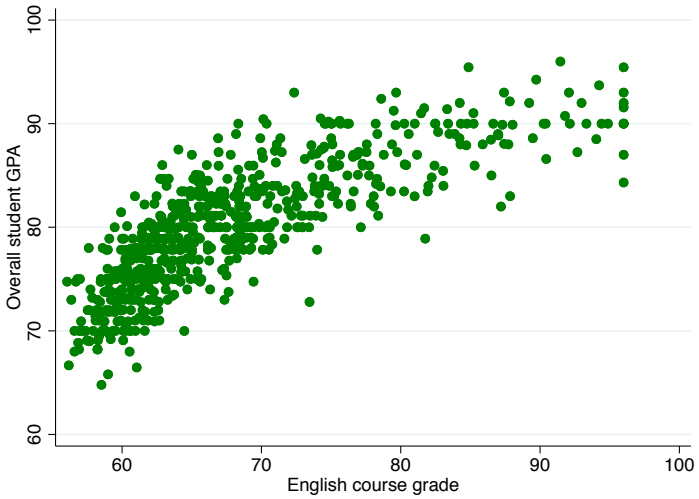

Figure 2.2: Scatter plot of GPA and english.

Looking at the scatter plot, we again see that the higher the grade on Ebglish, the higher the overall GPA. However, there is a difference between this scatter plot and the one shown in Figure 2.1. The scatter plot of GPA and attendance resembles a line to a large extent. In other words, the relationship seems to be extremely linear. This is not the same as the relationship between GPA and the grade on English. We see that the scatter plot starts to level off at a certain point. From Figure 2.2, we see that once the grade on English reaches around 80, further increases do not result in higher GPAs. What this means is that once you attain a certain higher level when it comes to

English, further gains mean little. A student who has an 85 on English is just as capable of understanding and communicating ideas as a student who has a 95.

What all of this means is that there is reason to suspect that the relationship between GPA and between English is non-linear. In order to see that, we can overlay the scatter plot with a line and a curve, as shown in Figure 2.3. I have made the scatter plot transparent in order to make the line and curve more visible.

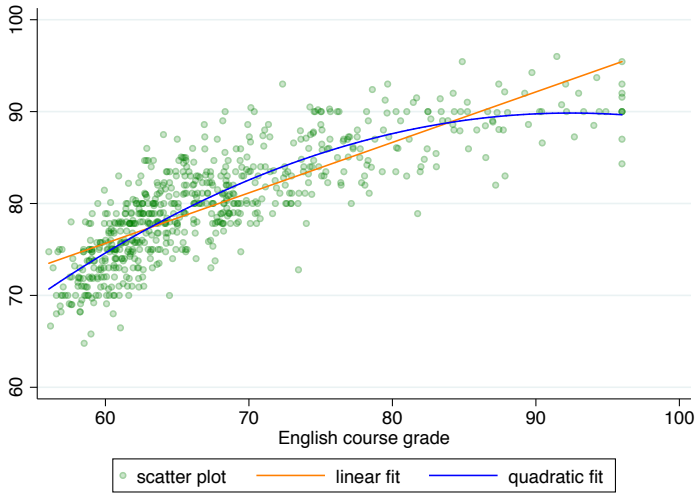

Figure 2.3: Scatter plot of GPA and english.

In this figure, we see that we have drawn a line and a curve in order to see which fits the data best. If you notice, both do a good job from the lower english grades up to the middle grades. However, when it comes to the high end of the English grades, we see that the line passes above the points while the curve passes through the points. This means that the curve is doing a

better job of fitting the data. This figure, leads us to suspect that we should include a quadratic term when we fit a regression model, as such:

Source	SS	df	MS	Number of obs	=	677
				F(2, 674)	=	809.55
Model	17101.4609	2	8550.73045	Prob > F	=	0.0000
Residual	7119.02181	674	10.5623469	R-squared	=	0.7061
				Adj R-squared	=	0.7052
Total	24220.4827	676	35.8291164	Root MSE	=	3.25

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	2.674998	.1910748	14.00	0.000	2.299824	3.050171
english2	-.0144694	.0012971	-11.16	0.000	-.0170162	-.0119227
_cons	-33.79013	6.925419	-4.88	0.000	-47.38812	-20.19214

From the output, we see that both english and the square of english are significant.

The third continuous independent variable is income, which records the family yearly income. We basically want to see if children from rich families do better academically, or if children from poorer fanilies do better. The scatter plot is shown in Figure 2.4. We very clearly see that the relationship is not linear. In fact, the graph is shaped like an inverted-U. It seems that the highest GPAs are obtained by students from middle income families. Children from very poor families and from rish families do not do as well as children from middle income families.

Given that the relationship between GPA and income is not linear, we fit a regression model that contains both the original income variable as well as

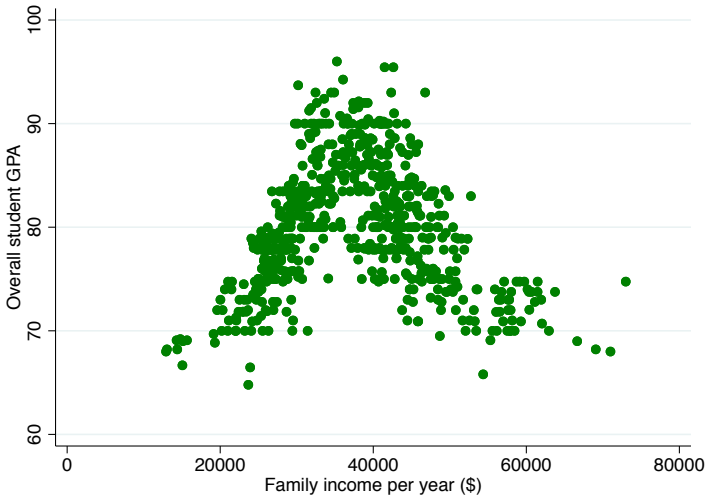

Figure 2.4: Scatter plot of GPA and income.

the squared value:

Source	SS	df	MS	Number of obs	=	677
				F(2, 674)	=	214.15
Model	9410.93931	2	4705.46966	Prob > F	=	0.0000
Residual	14809.5434	674	21.9726163	R-squared	=	0.3886
				Adj R-squared	=	0.3867
Total	24220.4827	676	35.8291164	Root MSE	=	4.6875

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0021508	.0001057	20.35	0.000	.0019433	.0023584
income2	-2.74e-08	1.32e-09	-20.69	0.000	-3.00e-08	-2.48e-08
_cons	40.61942	2.032082	19.99	0.000	36.62944	44.60939

Once again we see that both variables are significant.

The conclusion so far is that the variables attedance, english, and income are significant predictors of GPA. However, while attendance should be included as a linear term, our results indicate that we need to account for non-linearity

when including english and income.

There are still two more continuous variables and they are credits and siblings. We start with credits. We want to see whether students who have completed a higher number of courses have a higher GPA. Universities expect that students become better with time. The scatter plot is shown in Figure 2.5.

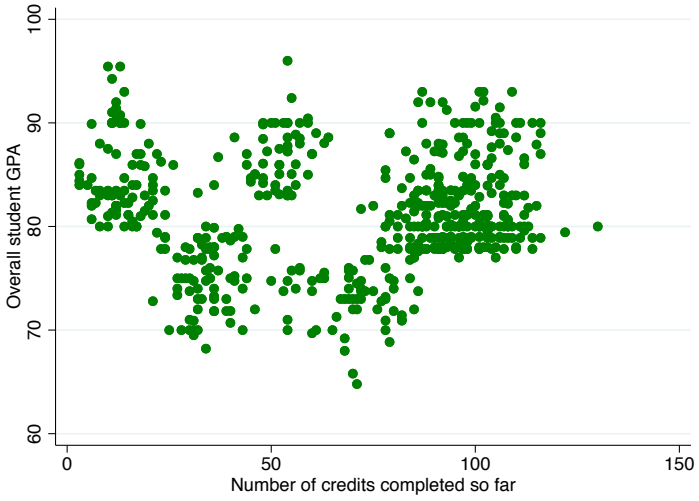

Figure 2.5: Scatter plot of GPA and credits.

The scatter plot does not really indicate that there is a relationship. Nonetheless, we fit a simple regression model:

Source	SS	df	MS	Number of obs	=	571
				F(1, 569)	=	0.00
Model	.02126165	1	.02126165	Prob > F	=	0.9793
Residual	17994.8615	569	31.6254157	R-squared	=	0.0000
				Adj R-squared	=	-0.0018
Total	17994.8828	570	31.5699698	Root MSE	=	5.6236

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]

credits	.0001823	.0070315	0.03	0.979	-.0136285	.0139931
_cons	81.13538	.5316944	152.60	0.000	80.09106	82.1797

The output of the model indicates that the variable credits is not significant. In fact, it is highly insignificant, given that the p-value is greater than 0.9. Therefore, we conclude that this variable does not contribute to our understanding of how GPA varies from student to student.

We next turn our attention to the last continuous variable, which is siblings. It would be interesting to see if the larger the number of children at home the lower the GPA, since this would mean that parents will have to divide their resources, in terms of money and time, amongst their various children. The scatter plot is shown in Figure 2.6.

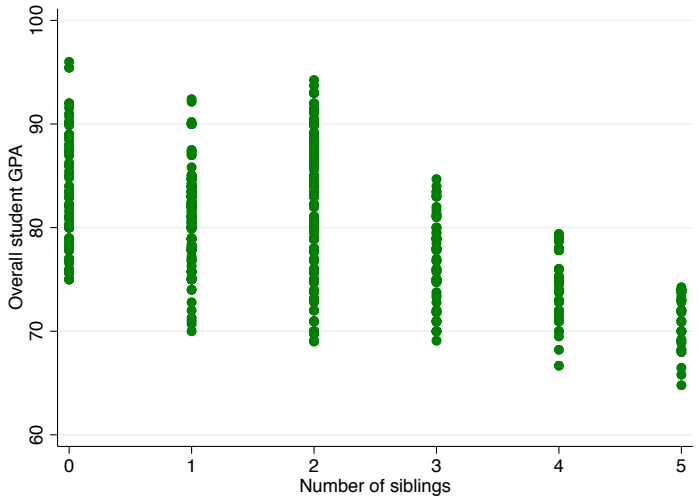

Figure 2.6: Scatter plot of GPA and siblings.

This scatter plot stands out from the previous ones. The reason is that the

independent variable siblings, in our dataset, takes on only 6 variables. The smallest value is takes is zero and the larges value it takes is five. This is why we see that the points tend to be divided along six vertical lines. The graph is not really a nice one, although we can see that, in general, there is a tendancy for GPA to decrease as we move from the left to the right. In such cases, it would be useful to produce a "smooth" graph. One of the most used techniques is to produce a loess curve. Such a curve is shown in Figure 2.7. We see that the smoothed plot is decreasing as the variable siblings increases.

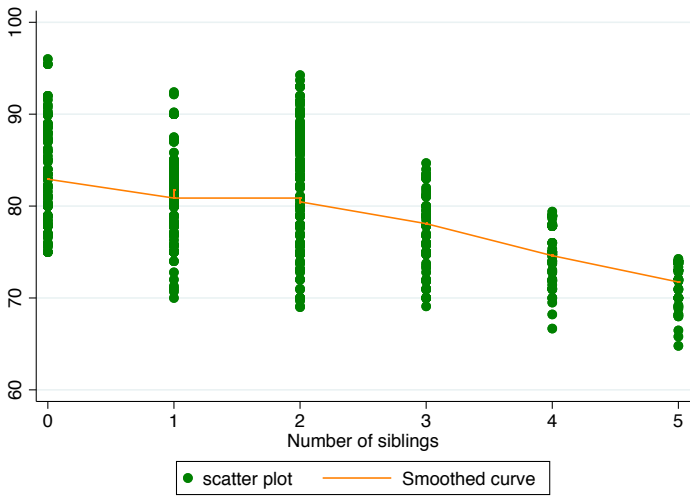

Figure 2.7: Scatter plot of GPA and siblings.

We now include the variable singlings in a simple regression model:

Source	SS	df	MS	Number of obs	=	677
				F(1, 675)	=	225.87
Model	6072.64126	1	6072.64126	Prob > F	=	0.0000
Residual	18147.8415	675	26.885691	R-squared	=	0.2507
				Adj R-squared	=	0.2496
Total	24220.4827	676	35.8291164	Root MSE	=	5.1851

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
siblings	-2.092668	.1392426	-15.03	0.000	-2.366069	-1.819267
_cons	83.96477	.3324845	252.54	0.000	83.31195	84.6176

Looking at the output, we see that the coefficient is both negative and significant. Therefore, it seems that the larger the number of siblings, the lower the GPA.

2.1.2 Binary Variables

We now turn our attention towards the binary variables. Our dataset contains two binary variables, and they are college and gender. We start by looking at college. Unlike continuous variables, we do not use graphs in order to investigate the relationship when the independent variable is binary (there are some curves that allow for a comparison but I personally do not find any utility in using them before fitting a regression model). We therefore fit a simple linear regression model where we include the variable college:

Source	SS	df	MS	Number of obs	=	677
				F(1, 675)	=	10.23
Model	361.527806	1	361.527806	Prob > F	=	0.0014
Residual	23858.9549	675	35.3465999	R-squared	=	0.0149
				Adj R-squared	=	0.0135
Total	24220.4827	676	35.8291164	Root MSE	=	5.9453

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
college						
Engineering	1.470097	.4596731	3.20	0.001	.5675363	2.372658
_cons	79.1506	.3421136	231.36	0.000	78.47886	79.82233

Looking at the output we see that the coefficient is positive and significant. We also notice in the output that the label "Engineering" is included and not business. This is because the base group is engineering (in the dataset the value zero is assigned to business). Therefore, the coefficient 1.47 is comparing engineering students to business students. What we see is that, on average, engineering students have a GPA that is 1.47 points higher than business students.

We next fit a regression model where we include gender:

Source	SS	df	MS	Number of obs	=	666
				F(1, 664)	=	37.39
Model	1261.55583	1	1261.55583	Prob > F	=	0.0000
Residual	22406.6088	664	33.7448927	R-squared	=	0.0533
				Adj R-squared	=	0.0519
Total	23668.1646	665	35.591225	Root MSE	=	5.809

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gender						
Female	2.810086	.4595897	6.11	0.000	1.907661	3.71251
_cons	78.76412	.2904518	271.18	0.000	78.19381	79.33444

From the output, we see that, ob average, female students have a GPA that is 2.81 points higher than their male colleagues. The result is significant.

2.1.3 Categorical Variables (more than two groups)

The dataset that we are using also contains the variable work. Unlike binary variables, this variable divides the observations into three groups: those that have a full time job, those that have a part time job, and those that have no job at all. Once again, we include the variable in a regression model:

Source	SS	df	MS	Number of obs	=	677
				F(2, 674)	=	94.05
Model	5284.63957	2	2642.31979	Prob > F	=	0.0000
Residual	18935.8431	674	28.0947228	R-squared	=	0.2182
				Adj R-squared	=	0.2159
Total	24220.4827	676	35.8291164	Root MSE	=	5.3004

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
work						
Part time	4.98294	.4317714	11.54	0.000	4.135161	5.830718
Full time	-2.616398	.6943671	-3.77	0.000	-3.979781	-1.253016
_cons	78.17105	.2940158	265.87	0.000	77.59375	78.74834

We see that the output shows the groups "Part time" and "Full time". This means that the referent group is the one that contains the students who do not work. Looking at the coefficients, we see that, on average, students who work on a part time basis have a GPA that is 4.98 points higher than students who do not work at all. We also see that, on average, students who have a full time job have a GPA that is -2.62 points lower than students who do not work at all. This means that student who do not work do better than students who have a full time job, and students who have a part time job do better than students who do not work. Both coefficients are significant.

2.2 Multiple Regression

From the previous section, it seems that we need a model that includes the variables attendance, english, the square of english, income, the square of income, siblings, college, gender, and work. We can now fit a multiple regression model that includes all of these variables:

Source	SS	df	MS	Number of obs	=	666
				F(10, 655)	=	499.43
Model	20923.9788	10	2092.39788	Prob > F	=	0.0000
Residual	2744.18585	655	4.18959671	R-squared	=	0.8841
				Adj R-squared	=	0.8823
Total	23668.1646	665	35.591225	Root MSE	=	2.0469

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.4184037	.018439	22.69	0.000	.382197	.4546103
english	.7920346	.1375871	5.76	0.000	.5218697	1.0622
english2	-.0037094	.0009115	-4.07	0.000	-.0054992	-.0019196
income	.0003589	.0000607	5.92	0.000	.0002398	.000478
income2	-4.63e-09	7.61e-10	-6.09	0.000	-6.13e-09	-3.14e-09
siblings	-.2505416	.0662263	-3.78	0.000	-.3805832	-.1205001
college						
Engineering	.5750044	.1653606	3.48	0.001	.2503035	.8997053
gender						
Female	-.2993557	.1764917	-1.70	0.090	-.6459134	.047202
work						
Part time	.9500899	.1860619	5.11	0.000	.58474	1.31544
Full time	-.5795809	.2849956	-2.03	0.042	-1.139196	-.0199657
_cons	3.349104	4.754391	0.70	0.481	-5.986581	12.68479

If you look at the output, you will notice something interesting, and that is

that the variable gender is no longer significant. This is a very important point. You will see that a variable that was signifiifant wne it was included by itself turns out to be not significant when included with other variables. Why is this the case? In our case, we saw that gender was significant by itself and now it is no longer significant. In out dataset, the average GPA for males is 78.76 and the average GPA for females is 81.57. So why did it turn out to be insignificant in the mutiple regression model? Digging deeper into the dataset would uncover that the average attendance grade for males is 78.62 and that the average attendance grade for females is 83.29. Therefore, it seems that the difference in GPAs between males and females is due to females attending more. When we included gender by itself in a simple model, the regression results told us that females do better. When we included the same variable in another model that also included attendance, the result turned out that the difference in GPAs between males and females can be explained by their different attendance records. Therefore, the variable gender is no longer contributing to the model.

Given the above, we can go ahead and fit a model that does not include gender:

Source	SS	df	MS	Number of obs	=	666
				F(9, 656)	=	553.02
Model	20911.9257	9	2323.5473	Prob > F	=	0.0000
Residual	2756.23895	656	4.20158377	R-squared	=	0.8835
				Adj R-squared	=	0.8819
Total	23668.1646	665	35.591225	Root MSE	=	2.0498

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.4100926	.0178014	23.04	0.000	.375138	.4450473
english	.8025509	.1376438	5.83	0.000	.5322754	1.072826
english2	-.0037703	.0009121	-4.13	0.000	-.0055613	-.0019794

income	.0003616	.0000607	5.95	0.000	.0002423	.0004808
income2	-4.66e-09	7.62e-10	-6.12	0.000	-6.16e-09	-3.17e-09
siblings	-.2452277	.0662468	-3.70	0.000	-.375309	-.1151464
college						
Engineering	.6364283	.1615772	3.94	0.000	.3191574	.9536991
work						
Part time	.9510869	.186327	5.10	0.000	.5852176	1.316956
Full time	-.5791741	.2854029	-2.03	0.043	-1.139587	-.0187607
_cons	3.370133	4.761171	0.71	0.479	-5.978839	12.71911

We now see that all variables are significant.

2.3 Model Fit

2.3.1 R-squared

It is now time to assess the goodness of fit of our model. The first piece of information is actually included in the regression output above. We see that the value of R-squared is 0.88, which is high. This means that the model is explaining around 88

2.3.2 Plotting predicted values against observed values

We can also visualise the goodness of fit by producing a figure that plots the observed values against the predicted values. If the model is a well fit model, we would expect to see that the predicted values are very close to

the observed values. Figure 2.8 shows this plot. If the model was a good fit, then we would expect that the dots would lie on the diagonal line, which represents y = x. We see that the tendancy for the points to be very close to the diagonal line, thereby illustrating that the model seems to be doing a good job.

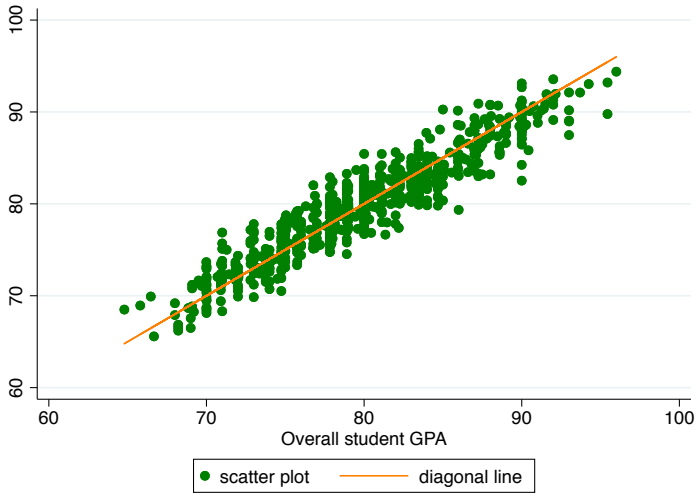

Figure 2.8: Comparing predicted values to observed values.

2.4 Assumptions

As was discussed in the theory section, regression models make some assumptions, and it is important that we test whether these assumptions are met or not.

2.4.1 Normality of the Residuals

The first assumption is that the residuals are normal. Figure 2.9 shows a histogram of the residuals. The histogram is overlayed with a normal curve in order to help us compare the two distributions. We can see that the residuals appear to be normal.

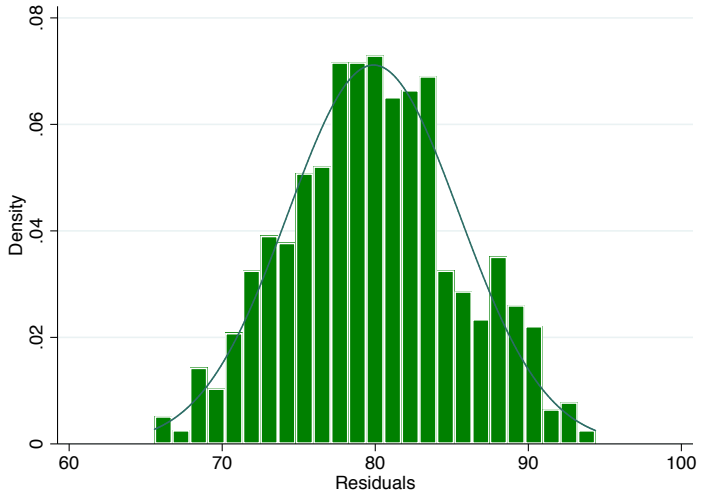

Figure 2.9: Comparing predicted values to observed values.

Another graphical way to test the assumption of normality is to produce a quantil-normal plot of the resilduals. This plot compares the distribution of the residuals with a normal distribtion. Figure 2.10 shows the plot. Since most of the dots are on the line, the plot supports the finding that the residuals are normal.

We can also use non-graphical tests to test the asusmption of normality. One

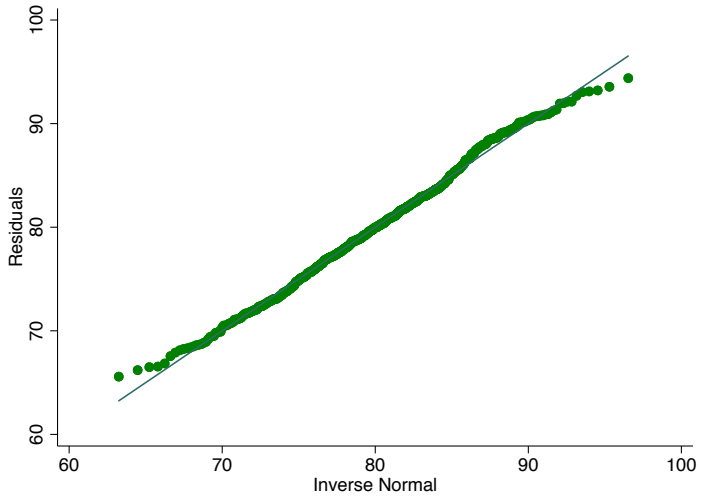

Figure 2.10: Comparing predicted values to observed values.

way to do this is to use the skewness-kurtosis test:

Skewness/Kurtosis tests for Normality

				joint	
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
residuals	666	0.4072	0.0162	6.46	0.0396

We see that the p-value is 0.0396, which is less than 0.05. This means that the null hypothesis, which is that the resilduals are normal, is rejected. What this means is that we cannot safetly conclude that the residuals are normal.

Another test is the Shapiro-Wilk test:

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
residuals	666	0.99441	2.438	2.170	0.01501

Once again, we see that the p-value produced by this test is less than 0.05, thereby leading us to reject the null hypothesis that the resilduals are normal.

These results indicate that we have a problem, since the assumption of linearity is not met, despite the fact that the model seems to be a well fit model. We will get back to this once we test the other assumptions.

2.4.2 Homoscedasticity

Another assumption is that the residuals have a constant variance. In order to test this assumption, we can use the Breusch-Pagan test:

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of gpa

         chi2(1)      =     5.75
         Prob > chi2  =   0.0165
```

Since the p-value is less than 0.05, we reject the null hypothesis that the residuals are homoscedastic. Again, this result is problematic to us.

2.4.3 Addressing Assumption Violations

Given that we have rejected the assumptions of normality and homoscedasticity, does this mean that we disregard our regression resukts? Fortunately no. As mentioned in the theory part, what we can do in this case is to fit the model while telling the statistical software to use robust standard errors. This way, the assumptions are relaxed and we can have more faith in the resulting

model:

```
Linear regression                               Number of obs     =        666
                                                F(8, 656)         =          .
                                                Prob > F          =          .
                                                R-squared         =     0.8835
                                                Root MSE          =     2.0498
```

gpa	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.4100926	.0183261	22.38	0.000	.3741078	.4460775
english	.8025509	.1302303	6.16	0.000	.5468324	1.058269
english2	-.0037703	.0008659	-4.35	0.000	-.0054707	-.00207
income	.0003616	.0000638	5.67	0.000	.0002363	.0004868
income2	-4.66e-09	8.15e-10	-5.72	0.000	-6.27e-09	-3.06e-09
siblings	-.2452277	.0652367	-3.76	0.000	-.3733256	-.1171299
college						
Engineering	.6364283	.1615658	3.94	0.000	.3191797	.9536768
work						
Part time	.9510869	.1933017	4.92	0.000	.5715221	1.330652
Full time	-.5791741	.2641098	-2.19	0.029	-1.097777	-.0605715
_cons	3.370133	4.477199	0.75	0.452	-5.421235	12.1615

The output of the model wshows that all varibales retain their significance. Therefore, it is safe to continue using our model while specifying the robust option.

2.4.4 Multicollinearity

The final assumption that we will test is that of multicollinearity. This is a very important assumoption which states that no independent variable is

a function of one or more other variables. This basically means that the independent variables should not be correlated with one another or with a combination of variables. To test this assumption, we can produce the VIF statistics:

Variable	VIF	1/VIF
attendance	2.53	0.395829
english	228.58	0.004375
english2	217.20	0.004604
income	59.98	0.016674
income2	60.35	0.016570
siblings	1.43	0.699116
1.college	1.02	0.978012
work		
1	1.34	0.748793
2	1.21	0.823429
Mean VIF	63.74	

As mentioned in the theory part, values that are greater than 10 are problematic. In our case, we see that english, english2, income, and income2 have values that are greater than 10. However, this is not a problem because we know that english2 is the square of english and income2 is the square of income. It is therefore expected that these variables be correlated with one another. What we care about is that the vif for other variables, the ones that do not include a quadratic term, are less than 10, and this is indeed the case. Therefore, this result indicates that multicollinearity is not an issue in our model.

2.5 Diagnostics

The next step is to investigate whether there are outliers and influential observations in the dataset.

2.5.1 Outliers

In order to identify whether there are outliers, we can plot a scatter plot of two variables. Any observation that is located away from most of the points is considered to be an outlier. The problem is that this method works when we just have one independent variable. We can easily plot a scatter plot of the dependent variable vs the independent variable. However, in our model, there are several independent variables. Fortunately, there is a tool that allows us to work around this problem, and this tool is the added-variable plot. What these plots do is that they produce a scatter plot of the dependent variable against each independent variable while accounting for the presence of the other independent variables. The main thing to understand is that if we have six independent variables, we produce six added variable plots, one for each independent variable.

In our case, there are eight independent variables. However, two of them (the quadratic terms) are just the square of two other terms. If an observation is an outlier with regards to the variable income for example, it will also be an outlier with regards to the square of income. Therefore, there is no need to look at the added value plots of the quadratic terms. This means that we need to look at six added variable plots. The plots are displayed in Figure 2.11. What we want to look for are points that stray away from the rest.

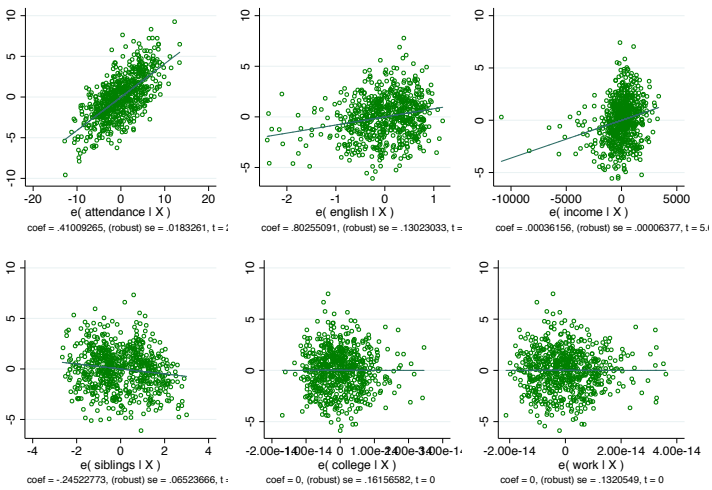

Figure 2.11: Added variable plots for each independent variable.

Looking at Figure 2.11, it doesnt seem that there are significant outliers in the case of the variables attendance, english, siblings, college, and work. With regards to income, we see that seems to be some outliers. Specifically, there is a point on the left side of the plot that seems to be floating alone. In addition, on the right hand side there are two dots on the very top and soem few other dots at the bottom. However, I do not see any reason to worry because the general result is that there are no significant outliers despite the fact that there are is a relatively large number of observations.

2.5.2 Influential Observations

We next investigate whether there are any particularly influential observations in our dataset. As discussed in the theory part, we can do this by calculating the DFBETAS, DFFITS, and Cook's D statistic. A useful exercise would

be to plot the DFFITS and Cook's D on the same plot. This would help us identify whether there are points with high values of both of these statistics. This graph is shown in Figure 2.12.

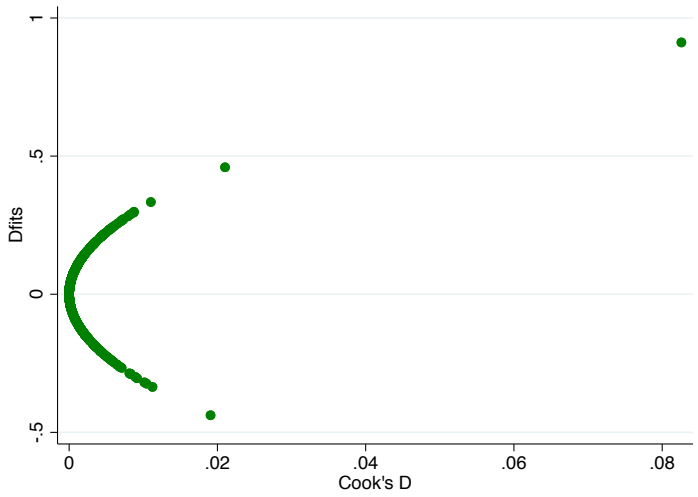

Figure 2.12: Plotting DFFITS and Cook's D.

Looking at the figure, we see that there is a single point that seems to be problematic since it has a higher than average values of both statistics. We note that this data point is the only one that has a Cook's D that is greater than 0.08.

When discussing the outliers, we noted that there seems to be some outliers with respect to the independent variable income (Figure 2.11). It would be interesting to look at this graph again, but this time while we are paying attention to the value of Cook's D.

Take a look at Figure 2.13. This figure is the added variable plot of the

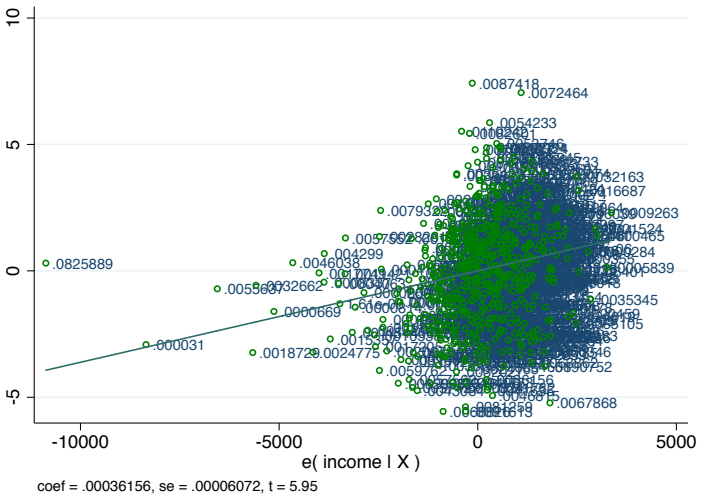

Figure 2.13: Added variable plot with Cook's D used as labels.

variable income, except that we have labelled each dot with the value of Cook's D. We now see that the outlier on the left hand side is actually the point that also has a Cook's D that is greater than 0.08. This means that this point is not only an outlier, but it is also influential. What do we do with it? The best thing to do about these points is to fit two models, one that includes all observations, and one that excludes these problematic observations. We can then compare the results. If there is no difference, then we conclude that the points are not causing any problems. If, however, there are significant differences between the two models, then we might consider eliminating these points all together.

Table 2.1 shows the results of this comparison. We see that the results do not change significantly. The level of significance of the variables is still the same, and the signs of the coefficients is also still the same. Therefore, we conclude that we do not have a problem with this particular point.

Table 2.1: Comparing estimates of both models

	(1)	(2)
Attendance and participation grade last semester	0.410***	0.406***
English course grade	0.803***	0.784***
english2	-0.00377***	-0.00364***
Family income per year - U.S. dollars	0.000362***	0.000409***
income2	-4.66e-09***	-5.29e-09***
Number of siblings	-0.245***	-0.239***
Business	0	0
Engineering	0.636***	0.617***
No	0	0
Part time	0.951***	0.938***
Full time	-0.579*	-0.571*
Constant	3.370	3.607
Observations	666	665

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2.6 Visualizing the Result

The regression output is useful because it allows us to quantify the relationship between the dependent variable and the independent variables. However, graphs are a more intuitive tool to use when trying to understand things. This is why an extremely useful thing to do after you have fit a regression model is produce graphs that visualise the results. The logic is simple. We need to plot the value of the dependent variable for various values of the independent variables.

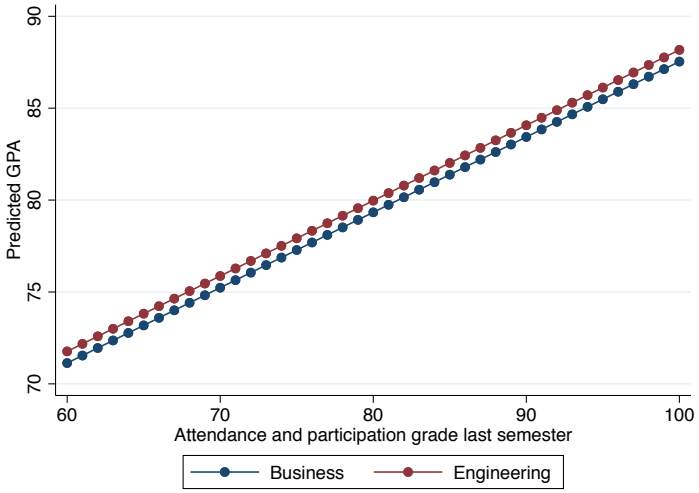

Figure 2.14: Visualizing how GPA varies with varying levels of the variables attendance and college.

As an example, consider Figure 2.14. This figure shows, according to our model, how the predicted GPA of both business and engineering students changes with attendance. The graph gives us two pieces of information. First, engineering students in general have a slightly higher GPA. Second,

the higher the level of ettendance, the higher the GPA.

Let us now look at Figure 2.15. This figure shows how the predicted GPA changes as the grade on English changes for students who do nto work, who have a full time job, and those that have a part time job. We see that unlike in the case of attendance, the relationship is not linear. We also see that students who work part time, on average, have a higher GPA, and that those who work full time have the lowest GPA.

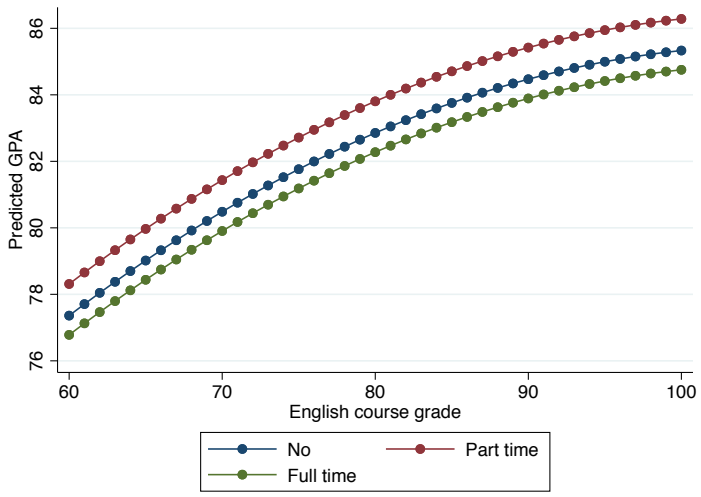

Figure 2.15: Visualizing how GPA varies with varying levels of the variables english and work.

Finally, consider Figure 2.16. This figure shows us how there is a constant decrease in GPA as the number of siblings increases.

As you can see from this section, looking at graphs makes understand models much easier, especially if you want to present your results to audience members who do not have a background in regression.

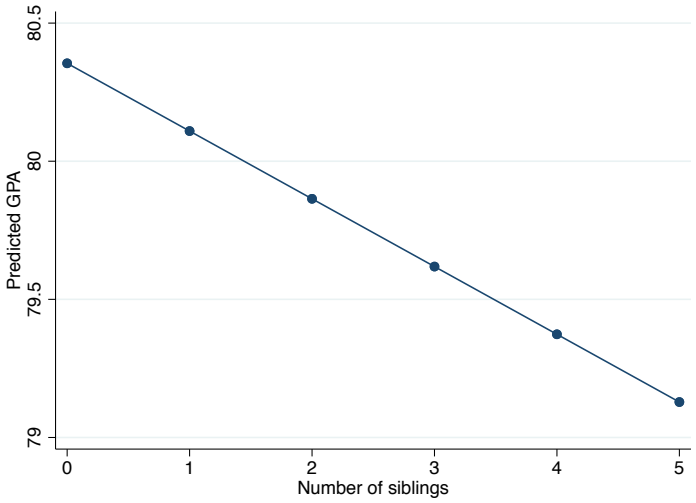

Figure 2.16: Visualizing how GPA varies with varying numbers of siblings.

Chapter 3

Logistic Regression - The Theory

3.1 Contingency Tables

3.1.1 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 3.1. 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.

Table 3.1: 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

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 3.2. 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.

Table 3.2: Cross classification of college and outcome.

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

3.1.2 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/(1-0.6) = 1.5. The odds of withdrawal for a business student is 0.25/(1-0.25) = 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/(1-0.5) = 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.

Now that we have the odds for each row in Table 3.2, 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 3.2), 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, 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.

3.1.3 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 3.3. 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:

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

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

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

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

Table 3.3: Cross classification of standing and outcome.

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

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.

3.2 Logistic Regression

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.

3.2.1 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 3.4. Note that the table is similar to Table 3.1 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:

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

Table 3.4: Records of students.

Outcome	**Gender**	**Binary**
Withdra	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 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:

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

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 3.2, 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.

3.2.2 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 3.5 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:

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

Table 3.5: 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

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.

3.2.3 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 3.3 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 3.6). 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 3.7. If you look at the column for the variable x_1, you will notice that the

Table 3.6: Cross classification of standing and outcome.

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

Table 3.7: Coding the categorical variable.

	$\mathbf{x_1}$	$\mathbf{x_2}$
Sophomore	0	0
Junior	1	0
Senior	0	1

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:

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

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:

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

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.

3.2.4 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:

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

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 3.1.

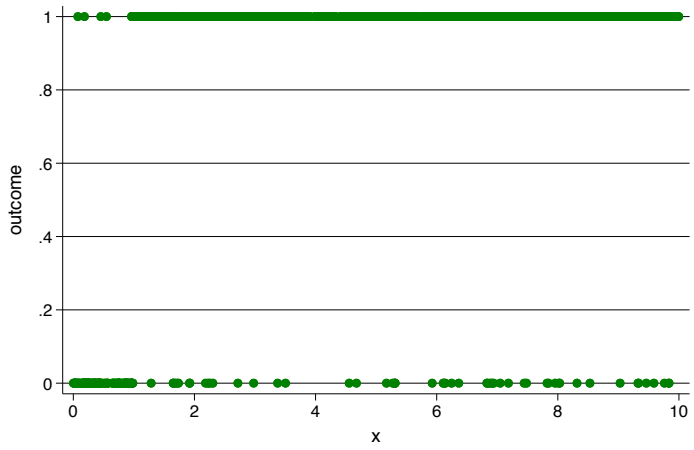

Figure 3.1: Scatter plot of a binary outcome and a continuous variable x.

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 3.2, 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.

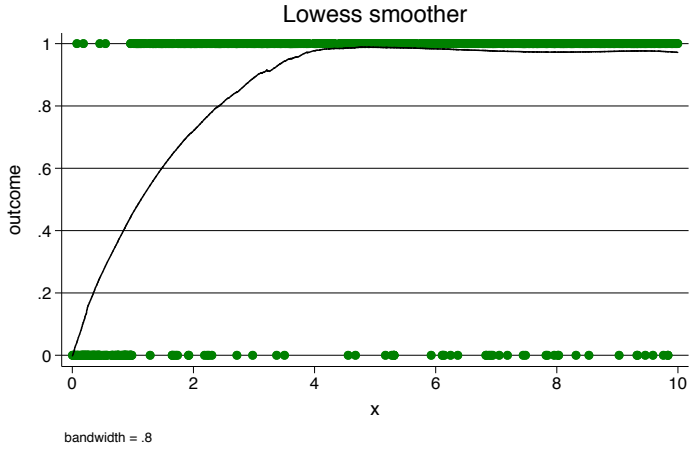

Figure 3.2: Loess graph of a binary outcome and a continuous variable x.

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 3.8). Once we have our new categorized variable, we can fit a

Table 3.8: Categorizing a continuous variable.

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

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 and the different levels of the categorical variable. Figure 3.3 shows the graph that is produced for the same data that was used to produce Figure 3.2, 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 3.2 and Figure 3.3.

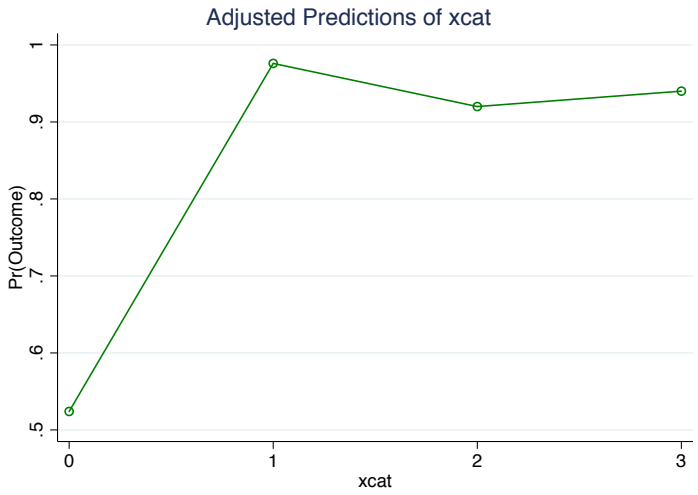

Figure 3.3: Linearity of slopes test.

3.3 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.

3.4 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 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 3.4, where we can see that the probability of withdrawing from a course decreases as the GPA increases.

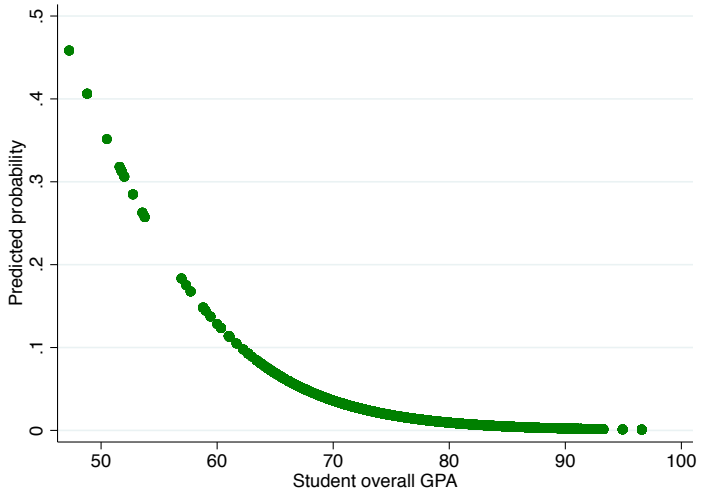

Figure 3.4: Visualizing the result of logistic regression: Relationship between the probability of withdrawing from a course and student GPAs.

3.5 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 3.9. 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.

Table 3.9: 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

3.5.1 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 3.9, running a logistic 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.

3.5.2 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 3.9, 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.

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

Table 3.10: Classification table.

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

As an example, Table 3.10 shows the classification table for the logistic regression model that is fit using the data in Table 3.9. 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.

3.5.4 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 3.9, we will find it to be 0.88, thus indicating that the model has an excellent ability

to discriminate.

3.5.5 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.

3.5.6 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.

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 3.11 offers some general guidelines that are useful in many situations.

Table 3.11: 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

Chapter 4

Logistic Regression - Case Study

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 a dataset that 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)

4.1 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.

4.1.1 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 produce a scatter plot of the dependent variable withdraw and the continuous independent variable GPA.

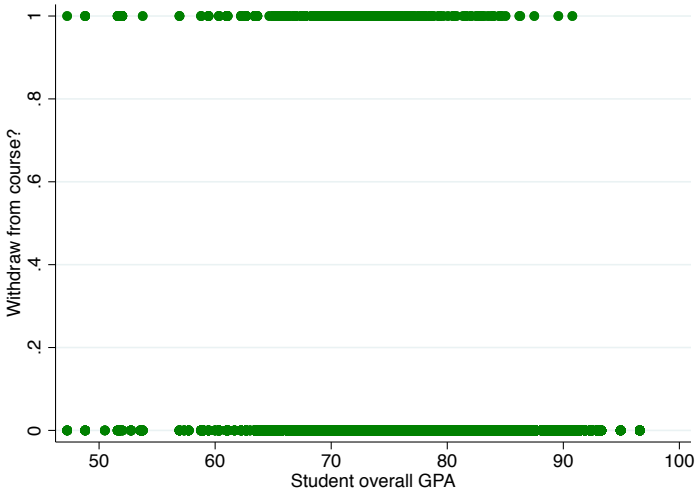

Figure 4.1: Scatter plot of withdraw and GPA.

The graph produced is shown in Figure 4.1. 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 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 calculating the average GPA for students who withdraw from courses and then to compare it to the GPA of students who do not withdraw:

```
-> 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
```

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:

```
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.
```

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 the statistical program to display the coefficient by specifying the coef option:

```
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
------------------------------------------------------------------------------
```

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:

```
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.
```

The new variable is the product of GPA and the log function of GPA. 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. The result is shown in Figure 4.2. The curve clearly shows that the relationship is not linear, thus providing extra evidence.

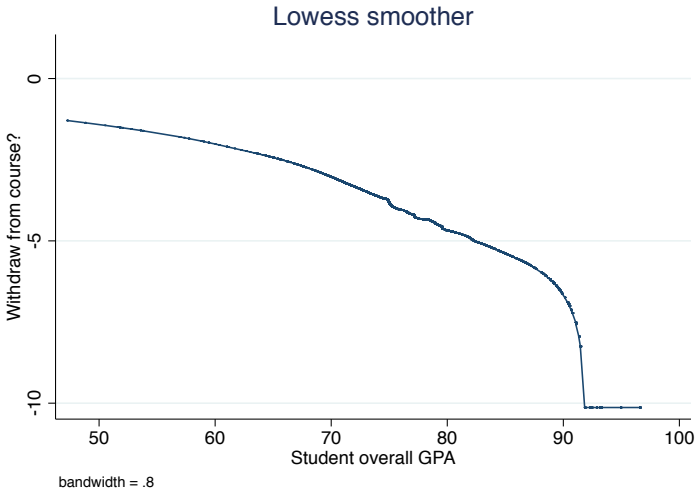

Figure 4.2: Loess curve of withdraw and GPA.

Linearity of Slopes Test

In order to perform the linearity of slopes test, we need to categorize the continuous variable GPA. In this case, we will generate a new variable, which we named gpacat, by cutting the variable GPA into groups: 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:

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	

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 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.

We would next want to produce a graph in order to see whether the relationship is linear or not.

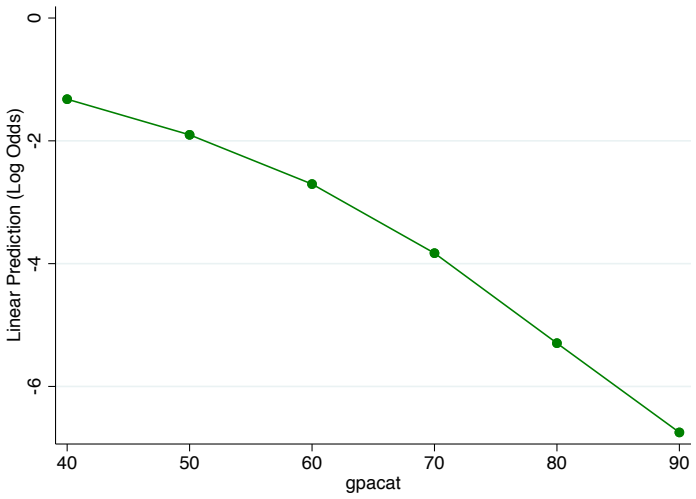

Figure 4.3: Testing the linearity of the slopes assumption.

The graph is shown in Figure 4.3. Once again, the nonlinearity is evident in the graph as it curves downward.

4.1.2 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:

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

```
Logistic regression                             Number of obs     =     25,160
```

```
                                                LR chi2(2)      =     441.91
                                                Prob > chi2     =     0.0000
Log likelihood = -2248.8207                     Pseudo R2       =     0.0895
```

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	1.206675	.1030634	2.20	0.028	1.020677	1.426568
gpa2	.9976426	.0006178	-3.81	0.000	.9964324	.9988542
_cons	.0086151	.0253454	-1.62	0.106	.000027	2.750891

Note: _cons estimates baseline odds.

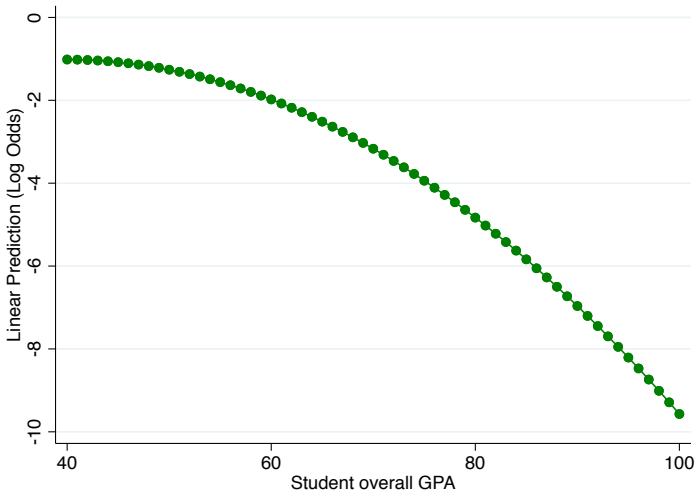

Figure 4.4: Including a quadratic term.

The result is shown in Figure 4.4.

4.1.3 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.

Key
frequency *row percentage*

	Withdraw from course?		
Gender	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

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 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.

The output shows that the odds ratios is 2.68. This means that the odds of males are 2.68 times the odds of 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:

Key
frequency *row percentage*

	Withdraw from course?		
College	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

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.

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.

4.1.4 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.

Key
frequency *row percentage*

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

Key
frequency
row percentage

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

It seems that the largest percentages of withdrawals are in the spring semester. It addition, it also 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.

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

```
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.

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.

If you look at the output in which the fall semester is the base, you 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:

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

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 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 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.

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 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 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.

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.

The level of the course	RECODE of level (The level of the course)			
	Other cou	200 level	500 level	Total
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

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 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.

Both groups of the variable are significant.

4.2 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 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	1.440269
gpa2	.9976925	.0006185	-3.73	0.000	.9964811	.9989054
gender						
male	1.528905	.2032302	3.19	0.001	1.178241	1.983931
college						
Engineering	1.011602	.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	.5369355

Note: _cons estimates baseline odds.

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. 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 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
gpa2	.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	.538841

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

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.

4.3 Analysis of Model Fit

4.3.1 Likelihood Ratio Test

This test is usually displayed whenever we run a logistic model:

```
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
        gpa2 |   .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	.538841

Note: _cons estimates baseline odds.

From the top right corner of the output, 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.

4.3.2 Hosmer-Lemeshow Test

We next perform the Hosmer-Lemeshow test:

```
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
```

Hosmer-Lemeshow recommend that the data be divided into 10 groups, which is what we did in the command. 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.

4.3.3 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 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.

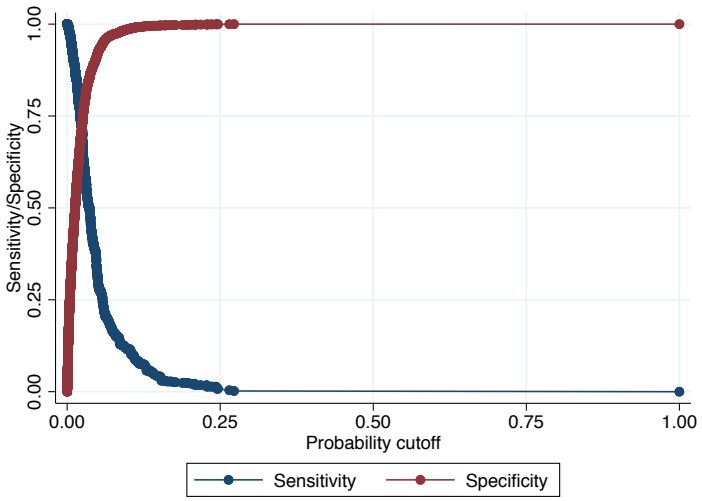

Figure 4.5: The sensitivity-specificity curves.

Figure 4.5 shows the point of intersection of the specificity and the sensitivity. We see that the sensitivity and the specificity curves intersect at a very small probability. The exact value of the point of intersection is 0.024088. We can now produce the classification table:

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%

We see that the model correctly classifies 71.72% of the observations, which is an acceptable value.

4.3.4 ROC Curve

As discussed in the theory section, another way to test the model fit is to calculate the area under the ROC curve, which is shown in Figure 4.6. 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.

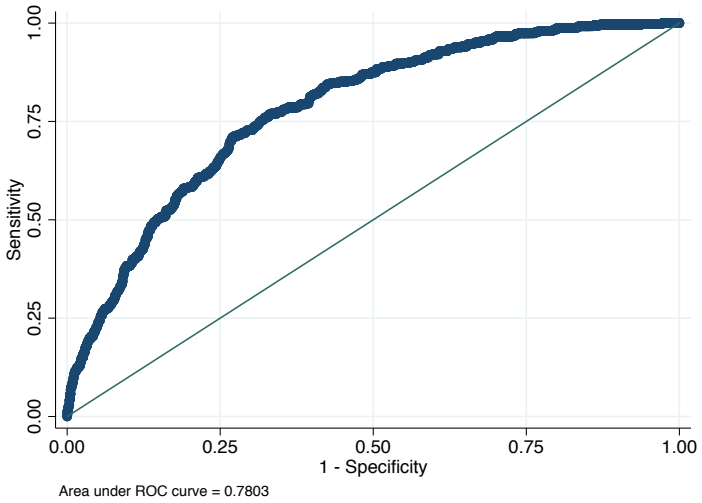

Figure 4.6: The ROC curve.

4.3.5 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. At this point in the analysis, we need to calculate these residuals and produce the appropriate plots.

Standardized Residuals

We start with the plot of the standardized residuals against the predicted probabilities. The graph is shown in Figure 4.7.

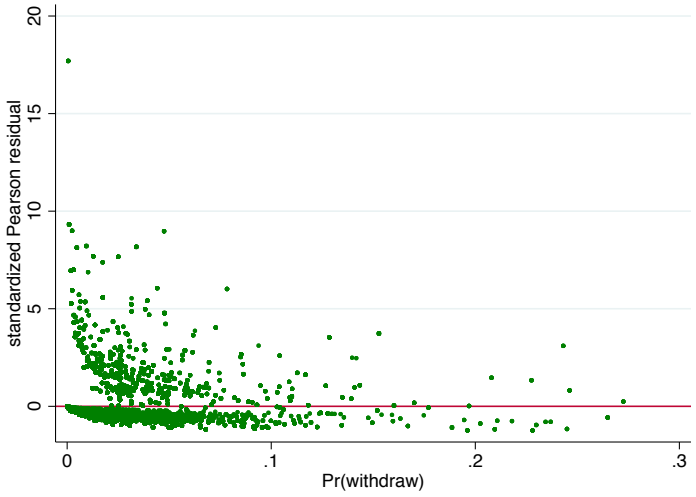

Figure 4.7: Plotting the standardized residuals against the predicted probabilities.

Deviance Residuals

We now do the same for the deviance residuals. The graph is shown in Figure 4.8. Both Figure 4.7 and Figure 4.8 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.

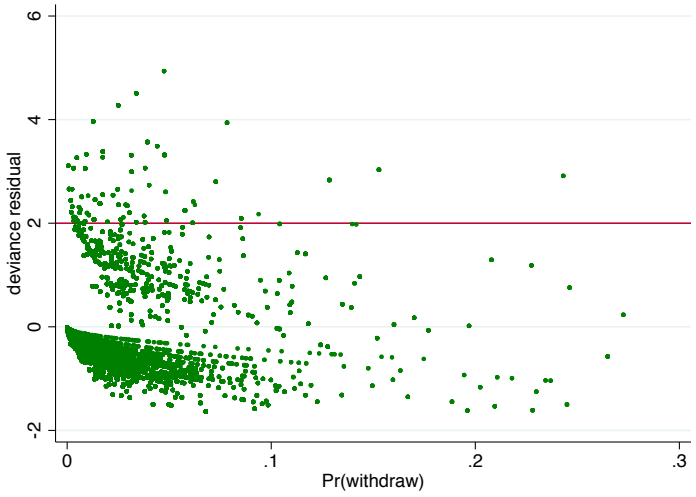

Figure 4.8: Plotting the deviance residuals against the predicted probabilities (values above two are considered to be outliers).

DeltaX

We now produce the scatter plot of the DeltaX residuals against the predicted probabilities. The output is shown in Figure 4.9. We draw a horizontal line at the y equal 4 point since values above four are considered to be outliers.

4.3.6 Influential Observations

The Hat Diagonal Statistic

Figure 4.10 shows the hat statistic plotted against the predicted probability. 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

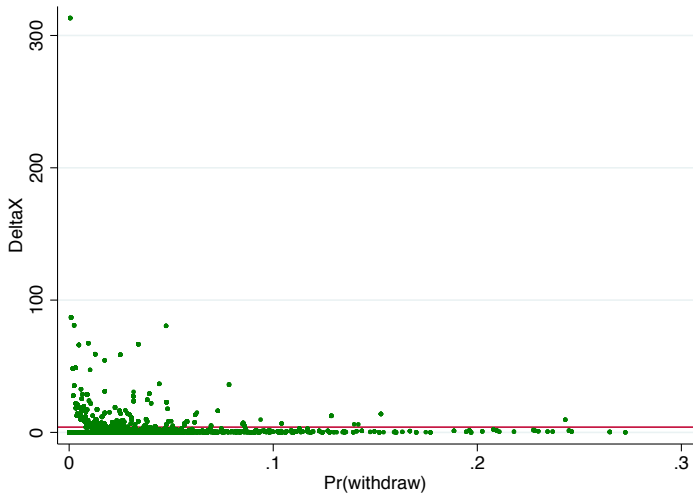

Figure 4.9: Plotting the DeltaX residuals against the deviance residuals (values above four are considered to be outliers).

(the mean of the variable hat is 0.002597).

Delta-Beta Statistic

Figure 4.11 shows the delta-beta influence statistic plotted against the predicted probability. We draw a horizontal line at the y = 1 point because values greater than one indicate that the observation is influential.

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 4.9. 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).

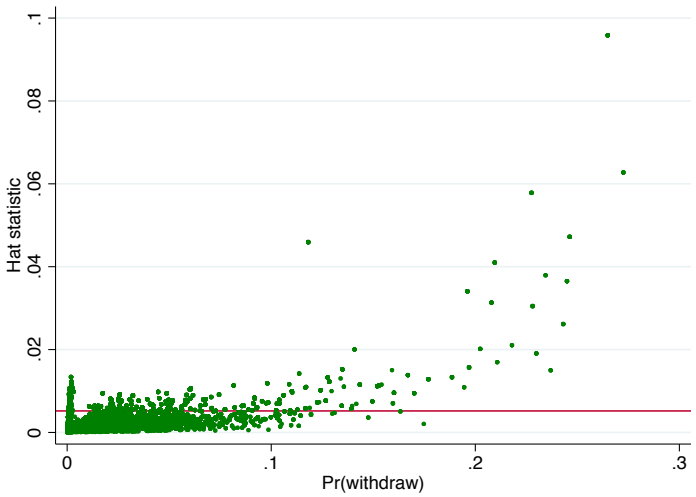

Figure 4.10: Plotting the hat statistic residuals against the predicted probabilities (values that are more than two times greater than the average are considered to be influential).

The large circle in the top left-hand side of the graph raises concerns. We need to take a closer look at these values, so we display the values of the variables for the observations that have DeltaX statistics that are greater than 200.

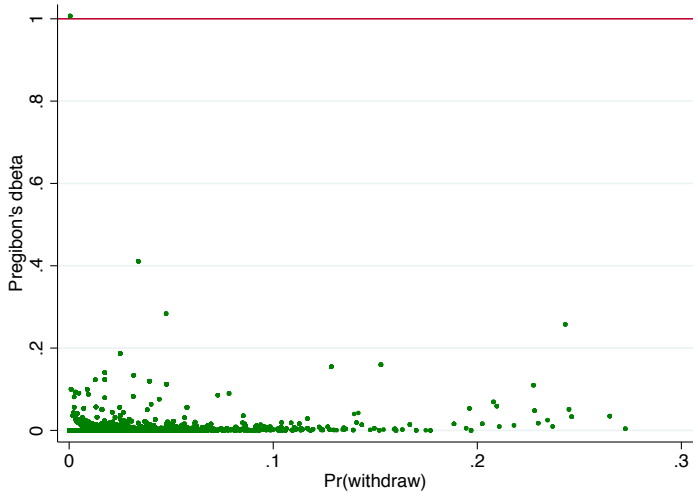

Figure 4.11: Plotting the Delta-Beta statistic against the predicted probabilities.

withdraw	gender	gpa	college	spring	level3
No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
Withdraw	male	79.3	Engineering	Fall or Summer	500 level courses

We see that there are six observations (the noobs option tells Stata not to list the observation numbers). 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

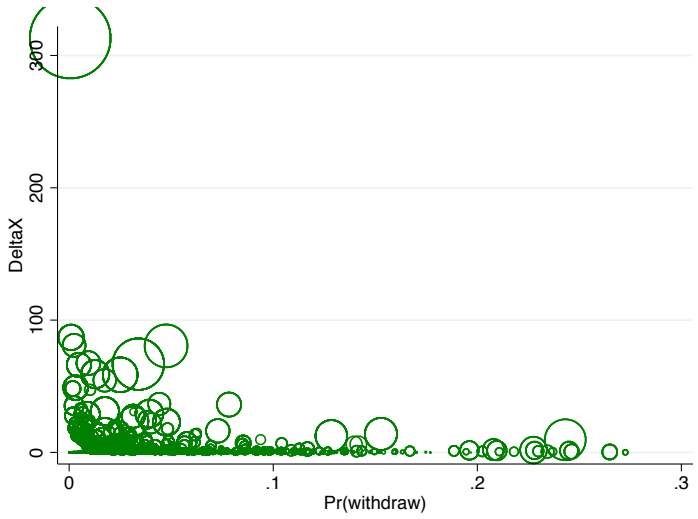

Figure 4.12: Plot of DeltaX versus estimated probability weighted by the variable delta-beta.

including these observations?

Table 4.1 shows a side by side comparison of both models. The table displays the output from two models, one that included all observations, and a second that excluded the observations that have a dx2 value greater than 200. This means that the six observations that have a DeltaX residual that is greater than 200 are excluded. 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.

Table 4.1: Comparing estimates of both models

	(1)	(2)
Withdraw from course?		
Student overall GPA	0.197*	0.200*
Student overall GPA × Student overall GPA	-0.00231***	-0.00233***
Gender	0.428***	0.425**
Spring semester	0.202*	0.206*
Other courses	0	0
200 level courses	0.683***	0.682***
500 level courses	-2.617**	0
Constant	-6.432*	-6.525*
Observations	25160	24003

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.4 Interpreting the Results

Now that we have seen that the model fit is good, it is time to interpret the obtained model parameters. It is important to note that logistic regression has the following linear form:

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

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 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.

4.4.1 Graphical Interpretation

One of the most useful tools to use to understand the results of regression models are graphs. As an example, take the case of the independent variable GPA. We would like to know what is the change in the probability of withdrawing from a course when GPA changes. The graph is shown in Figure 4.13. Notice that the probability drops to almost zero for values of GPA that are higher than 80.

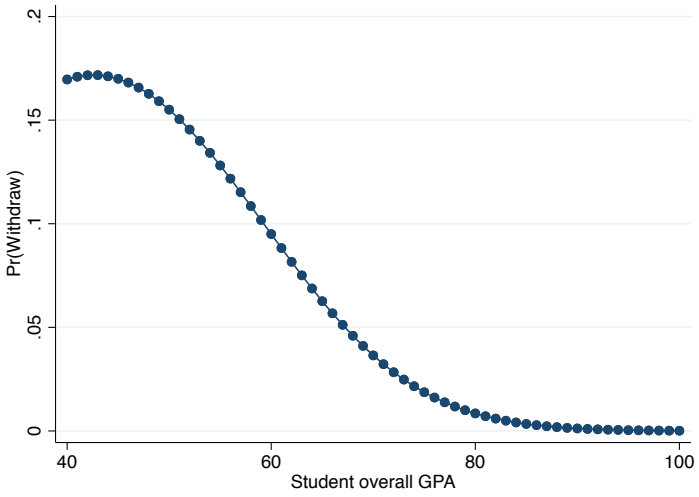

Figure 4.13: Graphing the probability of withdrawing from a course for each value of GPA.

We can also produce a more informative graph by including one of the categorical variables, as illustrated in Figure 4.14. 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 GPA 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.

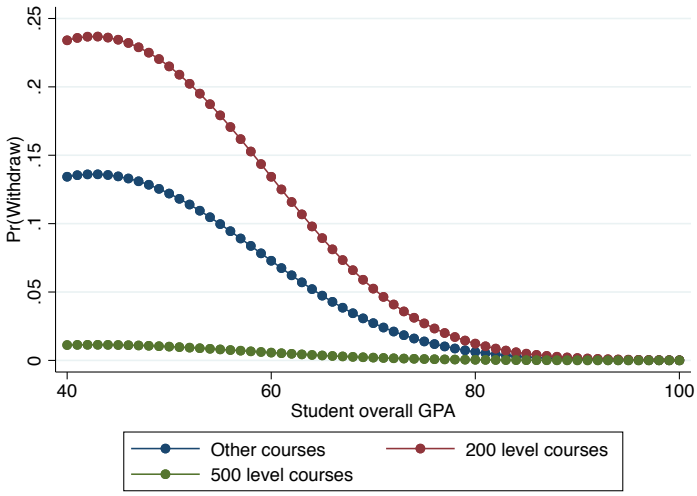

Figure 4.14: Graphing the probability of withdrawing from a course for different values of GPA for different level courses.

A final example will illustrate how we can graph the results when all variables are categorical as illustrated in Figure 4.15. Here, we are calculating the probabilities while varying the variables gender and level3. 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.

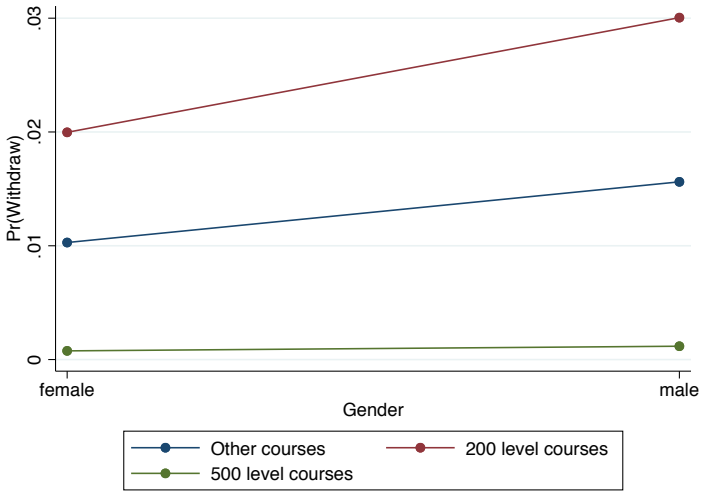

Figure 4.15: Graphing the probability of withdrawing from a course for different values of gender for different level courses.

Chapter 5

Count Models - The Theory

5.1 Introduction

Perhaps the most known and used regression technique is linear regression, where the dependent variable is continuous in nature. For example, if the dependent variable was salaries, then we can use linear regression, because a salary can take on any value in a certain interval. Linear regression can also be used when the dependent variable is student grades where the grade can be any value between zero and 100, including decimal values. As you know, any regression model makes certain mathematical assumptions. If these assumptions are violated, then the use of the regression model is questioned. This is why we cannot use linear regression when the dependent variable is not continuous in nature. When the dependent variable can take on only two values for example (i.e. a student either passes or fails a course) then linear regression cannot be used. In such a case we can use logistic regression. What about when the dependent variable represents a count? For example,

what if the dependent variable is the number of courses that a student has failed in? In this case the dependent variable cannot be negative since the minimum value is zero. In addition, the dependent variable cannot take on decimal values since students cannot fail in 2.5 courses for example. Another issue is that the maximum number of failed courses has a ceiling. A student cannot fail in 100 courses since the entire program does not have 100 courses. When the dependent variable is created by counting the times that a certain event has happened, we use regression techniques that are referred to as count models. In order to understand how these techniques work, let us first consider a simple count table. As an example, consider the data displayed in Table 5.1. The table contains the records of twenty students. The data includes the total number of courses in which each student has failed, the total number of courses in which the student has passed, and the college in which the student is enrolled. What we want to see is whether students in engineering fail in more courses than business students. This means that the variable that we intend to study is the number of courses in which the student has failed. This is an example of a count variable. To help us answer our question we can create a count table that summarizes the data.

5.2 Count Tables

Consider Table 5.2, which summarizes the data shown in Table 5.1. We see that out of a total of 243 courses taken by business students, there are 35 failing grades and 208 passing grade. For engineers, out of a total of 189 courses, there are 31 failing grades and 158 passing grades. Therefore, the count of failed courses for engineers is actually smaller than the count for

Table 5.1: Records of students.

Number of failed courses	Number of passed courses	College
3	20	Business
0	9	Business
5	25	Business
7	21	Business
1	26	Engineering
2	15	Business
2	13	Business
0	18	Business
4	23	Engineering
2	8	Engineering
5	23	Engineering
3	14	Engineering
8	19	Business
0	24	Business
1	5	Engineering
2	13	Engineering
3	17	Business
4	21	Engineering
5	27	Business
9	25	Engineering

business (31 for engineers and 35 for business). Does this mean that the failure rate in business is higher? The answer is no because we did not take into account the total number of courses. In other words, we need to calculate the proportion of courses which resulted in a failed grade. This means that

Table 5.2: Count table showing number of failed and not failed courses for students in the business and engineering schools.

	College		
Failed	**Business**	**Engineering**	**Total**
No	208	158	366
Yes	35	31	66
Total	243	189	432

we need to look at the risk.

5.2.1 Risk

Now that we have seen Table 5.2, we would like to calculate the risk that a student in business would fail a course and to compare it to the risk that a student in engineering would fail a course. Risk here indicates the probability of the event happening. For business students for example, out of a total of 243 courses, 35 resulted in a failed grade. This means that the probability of failure, or the risk of failure, is $35/243 = 0.144$. For an engineering student, the risk is $31/189 = 0.164$. Therefore, we see that this risk of failure for engineers is greater than the risk of failure of business students.

5.2.2 Incidence-rate Ratio

We now know that the risk of failure for business students is larger than the risk of failure of engineering students. To directly compare the two risks we can calculate the incidence-rate ratio, which is simply the ratio of the two

numbers:

$$Riskratio = 0.144/0.164 = 0.878$$

What does this value mean? Simply that the likelihood of a business student failing a course is 0.878 times the likelihood of an engineering student failing the course. We could have calculated the incidence-rate ratio by dividing the risk for engineers by the risk for business students:

$$Riskratio = 0.164/0.144 = 1.139$$

This means that the likelihood of an engineering student failing a course is 1.139 times the likelihood of a business student failing a course.

5.2.3 2x3 Tables

The above logic is maintained even when we have more than two groups. Consider Table 5.3 for example. This table contains the same information as Table 5.2 with the addition of a new college which is the life sciences college. We already know that the risk of failure for business and engineering students are 0.144 and 0.164 respectively. The risk for students in the life sciences school is $20/202 = 0.099$, which is smaller than the other two risks. Using these risks, we can also calculate the risk ratios. What is different is this case is that we can calculate two different risk ratios:

$$Riskratio_1 = 0.144/0.164 = 0.878$$
$$Riskratio_2 = 0.099/0.164 = 0.604$$

The first ratio is comparing the risk of business students to the risk of engineering students, while the second ratio is comparing the risk of life sciences students to the risk of engineering students. In both cases, the

Table 5.3: Count table showing number of failed and not failed courses for students in the business, engineering, and life sciences schools.

	College			
Failed	**Business**	**Engineering**	**Life Scences**	**Total**
No	208	158	182	548
Yes	35	31	20	86
Total	243	189	202	634

referent group is engineering students. It is up to you to pick and choose the referent category that suits your goals. In our case, the two risk ratios above are less than one, which indicates that the risk for both business and life sciences students is smaller than that of engineering students.

The above exercise is useful when we want to compare the risk across certain groups. 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 probability of the outcome. What if we wanted to see whether the risk of failure was affected by the college, gender, and the GPA, all at the same time? In this case, we need to use regression models.

5.3 Poisson Regression

First, you need to understand that there are several count models to choose from. The choice of the model depends on the data that we are analyzing. Usually, we start the analysis by assuming a Poisson model since this is considered to be the basic count model. In this type of regression, we are interested in the number of occurrences of a certain event, i.e. how many

times a student will fail for example. As such, the dependent variable μ refers to the rate of occurrence or the expected number of times an event will occur. In order to visualize the distribution of a variable that follows the Poisson distribution, consider Figure 5.1 which shows the probability distribution functions for different average rates. The y-axis represents the probability that a certain event will happen a certain number of times. For example, looking at the graph for mean = 1, we see that the probability of the event not happening at all (zero) or happening once is high, while the probability of the event happening four times is very low.

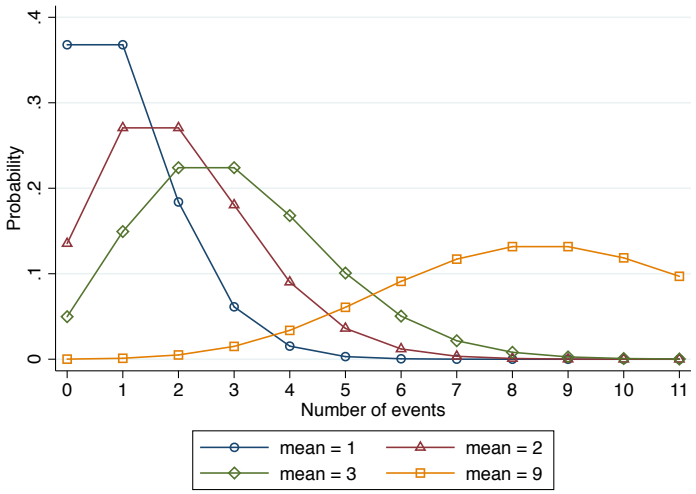

Figure 5.1: The Poisson distribution for different means.

If the expected average increases, i.e. the mean increases, then the probability of events happening more frequently also increases. This is why as the mean increases the graph starts to rise on the right side while dropping at the left side. The purpose of Poisson regression is to look at the factors that would increase the probability of an event happening more frequently.

In linear regression, the relationship between the dependent variable and the independent variable is formulated as:

$$y = ax + b$$

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 Poisson regression things are not that simple. The reason is that the Poisson regression model has the following form:

$$\ln(\mu) = ax + b$$

As already mentioned, μ is the rate of occurrences, which is the dependent variable. The above equation is linear, but instead of having the dependent variable on the left hand side we have the natural logarithm of the dependent variable. This means that the slope a represents the amount by which $\ln \mu$ increases when x increases by one unit. As you can see, this is not a natural way of interpreting things. 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:

$$e^{\ln(\mu)} = e^{ax+b}$$

$$\mu = e^{ax+b}$$

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 $e^{\ln(k)} = k$.

This new form is better because now the dependent variable, which is the rate μ, is on the left side. For example, if a is positive, when x increases the term e^{ax+b} will increase. Since this term is equal to the rate of occurrence

of the event, this means that the number of times that the event is expected to occur will also increase. On the other hand, when a is negative, when x increases the expected number of occurrences will decrease.

5.3.1 Continuous Variables

Let us take an example. Assume that we perform Poisson regression where the dependent variable is the number of courses in which a student has failed and the independent variable is the GPA of the student. Basically, we want to see if having a higher GPA predicts fewer course failures. Assume that once the model was fit that we get the following equation:

$$\ln(\mu) = -0.099(GPA) + 7.091$$

What this means is that when the GPA of the student increases by one, the function $\ln(\mu)$ decreases by -0.099. Since, as we said, this is hard to understand, let's consider the other more intuitive form:

$$\mu = e^{-0.099(GPA)+7.091}$$

Now consider two students, one with a GPA of 77 and the other with a GPA of 78. According to our model, the expected number of withdrawals for each is:

Student with a GPA of 77: $\mu = e^{-0.099(77)+7.091} = 0.587$

Student with a GPA of 78: $\mu = e^{-0.099(78)+7.091} = 0.532$

This means that the expected number of failed courses for a student with a GPA of 77 is 0.587, and the expected number of failure courses for a student with a GPA of 77 is 0.532. To compare these two numbers, we can divide them in order to find the incidence-rate ratio:

Incidence-rate ratio: $0.532/0.587 = 0.906$

What this means is that the expected count for a student with a GPA of 78 is 0.906 times the expected count of a student with a GPA of 77. The great news is that 0.906 is actually $e^{-0.099}$. We now have a very intuitive interpretation of the slope a. When we fit a Poisson 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 expected number of occurrences is multiplied by e^a. When a is positive, $e^a > 1$, which means that the expected number of occurrences increases when x increases. When a is negative, $e^a < 1$, which means that the expected number of occurrences decreases when x increases.

As a recap, when we fit a Poisson 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 Poisson regression, when x increases by one unit, the expected number of occurrences are multiplied by e^a. If a is zero we have $e^0 = 1$, which means that the expected number of occurrences are multiplied by one, so they do not change. This means that x does not affect the expected number of occurrences. If a is greater than zero, then $e^a > 1$, which means that the expected number of occurrences 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 expected number of occurrences are multiplied by a number that is less than one, so it decreases.

As another illustration, assume that we fit a Poisson model where the dependent variable is the number of customers that entered the store today, and where the independent variable is the number of advertisements that were ran on radio the preceding day. Once we fit the model we get the following results:

$$\ln(\mu) = 0.447(ads) + 0.241$$

Here, μ is the expected number of customers that will enter the shop. What does this output mean? Since the value of the coefficient associated with the independent variable, which is ads, is 0.447, this means that when ads increases by one, the expected number of customers is multiplied by $e^{0.447} = 1.564$. This means if the shop runs five radio ads the expected number of customers that will come is 1.564 times the expected number of customers if it runs four ads.

As another example, consider that we fit a Poisson regression model where the dependent variable is the number of times that a student goes out with his or her friends during the week and the independent variable is the student's grades. The output of the model is the following:

$$\ln(\mu) = -0.073(grades) + 6.629$$

Here, the coefficient is negative. Since $e^{-0.073} = 0.930$, the output indicates that the expected number of times that a student goes out during the week are multiplied by 0.93 (so they decrease) when grades increase by a single

unit. This means that students with higher grades go out fewer times during the week.

As you can see, when the coefficient is positive, the expected count increases, and when the coefficient is negative, the expected count decrease. Since we are mostly interested in the exponential of the coefficient, and not the coefficient itself, statistical software packages allows us to directly display the value e^a instead of displaying the value of a. In that case, when e^a is greater than one, the expected count increases, and when e^a is less than one, the expected count decreases.

5.3.2 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 count of failed courses 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 5.4. Now that the variable gender has been quantified, it is possible to include it in a regression model. The result of running a Poisson model would be again in the form:

Table 5.4: Records of students.

Number of failed courses	**Gender**	**Binary**
2	male	0
0	male	0
3	male	0
0	female	1
2	female	1
3	female	1
8	female	1
0	male	0
5	female	1
7	male	0
5	female	1
3	male	0
1	male	0
1	male	0
2	female	1
9	female	1
5	female	1
4	male	0
4	male	0
2	female	1

$$\ln(\mu) = ax + b$$

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

$$\ln(\mu) = 0.495(gender) + 0.916$$

We already know how to interpret the coefficients of continuous variables, such as grades and number of advertisements. However, what does it mean that the coefficient of gender is 0.495? Remember that for males the value of gender is zero, while for females the value of gender is one. In order to calculate the expected count for a male and a female student, we need to use the form:

$$\mu = e^{0.495(gender)+0.916}$$

We can now calculate the expected count for each student:

Male: $\mu = e^{0.495(0)+0.916} = 2.499$

Female: $\mu = e^{0.495(1)+0.916} = 4.1$

From these expected counts, we can calculate the incidence-rate ratio:

Incidence-rate ratio: $4.1/2.499 = 1.64$

This means that females have higher expected count than males. The nice thing is that the number 1.64 happens to be $e^{0.495}$. This means that when we are dealing with binary variables, the exponent of the coefficient is the incidence-rate 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 incidence-rate ratio that we obtain when we divide the expected count of females by the expected

count of males. In other words, since the coefficient is 0.495, the expected count for females is 1.64 times the expected count for males.

Let us take another example. Assume that we run a Poisson regression model where the dependent variable is the number of goals a player scores, and where the independent variable is whether the player got a good night sleep the night before or not. The independent variable is binary (either you get a good night sleep or not), so we need to assign zero to a category and a one to the other category. In our case, let's assign a zero to not getting a good night sleep and a one to getting a good night sleep. We fit the model and get the following result:

$$\ln(\mu) = 1.104(sleep) + 0.357$$

This means that the expected number of goals scored by those who get a good night sleep is $e^{1.104} = 3.016$ times the expected number of goals scored by those who did not get a good night sleep.

5.3.3 Multiple Independent Variables

Now that we have seen how to interpret the output from Poisson regression when there is a single independent variable, let us see what changes when there are two independent variables. Table 5.5 shows the records for students. The table includes the dependent variable which is the number of courses in which the student has failed and the independent variables gender and GPA. Therefore, we have one binary variable and one continuous variable. In this case, we want to see if the dependent variable, which is the number of failed

Table 5.5: The case of two independent variables.

Number of failed courses	**GPA**	**Gender**	**Binary**
2	80	male	0
0	95	male	0
3	77	male	0
0	90	female	1
2	75	female	1
3	72	female	1
8	60	female	1
0	82	male	0
5	74	female	1
7	69	male	0
5	69	female	1
3	79	male	0
1	81	male	0
1	78	male	0
2	83	female	1
9	62	female	1
5	72	female	1
4	70	male	0
4	71	male	0
2	87	female	1

courses, depends on the gender of the student and on the GPA of the student. The equation of this model is:

$$\ln(\mu) = a_1 x_1 + a_2 x_2 + b$$

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

$$\ln(\mu) = -0.086(GPA) + 0.033(gender) + 7.464$$

Let us now calculate the expected count for two students where both of them have a GPA of 74, but one is male and the other is female. First, we use the more intuitive form of the equation:

$$\mu = e^{-0.086(GPA)+0.033(gender)+7.464}$$

Male: $\mu = e^{-0.086(74)+0.033(0)+7.464} = 3.004$

Female: $\mu = e^{-0.086(74)+0.033(1)+7.464} = 3.105$

This means that the incidence-rate ratio is:

Incidence-rate ratio: $3.105/3.004 = 1.034$

A simpler way to get this value is just to calculate the exponent of the coefficient, $e^{0.033} = 1.034$. 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 expected count for two female students, one of whom has a GPA of 79 and another who has a GPA of 80:

GPA of 79: $\mu = e^{-0.086(79)+0.033(1)+7.464} = 2.02$

GPA of 80: $\mu = e^{-0.086(80)+0.033(1)+7.464} = 1.853$

This means that the incidence-rate ratio is:

Incidence-rate ratio: $1.853/2.02 = 0.917$

This is also obtained by finding the exponent of the coefficient, $e^{-0.086} = 0.917$. Therefore, we see that when GPA increases by one, the expected count is multiplied by 0.917, which means that the expected count decreases.

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.

5.3.4 Categorical Variables with more than Two Categories

If you recall, Table 5.1 presented data that included the variable college, which took on two values, business and engineering. In such a case, when we perform Poisson regression, we use a binary variable since the variable college can take on one of two values. What if we had a categorical variable

that divided the observations into more than two groups? In this case, we cannot use a single binary variable because there are three groups instead of two. As an example, consider the data displayed in Table 5.6. This is the same data that we used in Table 5.1 except that nine new records have been added for students in the college of life sciences (the count table for Table 5.6 was actually used in Table 5.3). This means that the variable college is no longer binary, since it can take on more than two values.

What we can do in this case, is to use more than one binary variable, as illustrated in Table 5.7. If you look at the column for the variable x_1, you will notice that the variable takes a value of one for business, and zero otherwise. The other binary variable, x_2, takes on a value of one for life sciences 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. Table 5.8 displays the result of this coding exercise. The regression equation now becomes:

$$\ln(\mu) = a_1 x_1 + a_2 x_2 + b$$

For an engineering student, x_1 and x_2 are zero. For a business student, x_1 is one and x_2 is zero. For a life sciences student only x_1 is zero and x_2 is one. If we fit this model, the output will be:

$$\ln(\mu) = -0.079 x_1 - 0.438 x_2 + 1.237$$

Let us now calculate the expected number of occurrences for each student.

Table 5.6: Records of students.

Number of failed courses	**College**
3	Business
0	Business
5	Business
7	Business
1	Engineering
2	Business
2	Business
0	Business
4	Engineering
2	Engineering
5	Engineering
3	Engineering
8	Business
0	Business
1	Engineering
2	Engineering
3	Business
4	Engineering
5	Business
9	Engineering
0	Life sciences
1	Life sciences
1	Life sciences
3	Life sciences
5	Life sciences
4	Life sciences
3	Life sciences
1	Life sciences
2	Life sciences

Table 5.7: Coding the categorical variable.

	$\mathbf{x_1}$	$\mathbf{x_2}$
Engineering	0	0
Business	1	0
Life sciences	0	1

As usual, we use the more intuitive form of the equation:

$$\mu = e^{-0.079(x_1)-0.438(x_2)+1.237}$$

Engineering: $\mu = e^{-0.079(0)-0.438(0)+1.237} = 3.445$

Business: $\mu = e^{-0.079(1)-0.438(0)+1.237} = 3.184$

Life Sciences: $\mu = e^{-0.079(0)-0.438(1)+1.237} = 2.223$

We can now calculate the incidence-rate ratios in order to be able to compare different groups:

Business/Engineering=3.184/3.445 = 0.924

Life sciences/Engineering=2.223/3.445 = 0.645

The above means that the expected number of failed courses for business students is 0.924 times the expected number of failed courses for engineering students, and that the expected number of failed courses for life sciences students is 0.645 times the expected number of failed courses for engineering students. We can get the same values by calculating the exponents of the coefficients:

Table 5.8: Records of students.

Number of failed courses	**College**	$\mathbf{x_1}$	$\mathbf{x_2}$
3	Business	1	0
0	Business	1	0
5	Business	1	0
7	Business	1	0
1	Engineering	0	0
2	Business	1	0
2	Business	1	0
0	Business	1	0
4	Engineering	0	0
2	Engineering	0	0
5	Engineering	0	0
3	Engineering	0	0
8	Business	1	0
0	Business	1	0
1	Engineering	0	0
2	Engineering	0	0
3	Business	1	0
4	Engineering	0	0
5	Business	1	0
9	Engineering	0	0
0	Life sciences	0	1
1	Life sciences	0	1
1	Life sciences	0	1
3	Life sciences	0	1
5	Life sciences	0	1
4	Life sciences	0	1
3	Life sciences	0	1
1	Life sciences	0	1
2	Life sciences	0	1

$$e^{-0.079} = 0.924 \, and \, e^{-0.438} = 0.645$$

We see that the exponent of the coefficient for each variable produces the incidence-rate 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, engineering students are the base, or referent group, since they have a value of zero for both x_1 and x_2. Business students have a value of one for x_1, which means that the exponent of the coefficient of x_1 is the incidence-rate ratio of business students to engineering students. Life sciences students have a value of one for x_2, which means that the exponent of the coefficient of x_2 is the incidence-rate ratio of life sciences students to engineering 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.

5.3.5 Exposure

There is an issue here which you have probably not noticed. In the previous section, we calculated the incidence-rate ratios when the categorical variable took on three values, business, engineering, and life sciences. This was done using Poisson regression. However, earlier in this book, we had calculated the incidence rate ratios for the exact same data using the count table (Table 5.3). In both cases, although the same data was used, the results are different. To make comparison easier, Table 5.9 shows the different

Table 5.9: Comparison of incidence-rate ratios when using count tables and when using Poisson regression.

	Count tables	**Poisson regression**
Business / Engineering	0.878	0.924
Life sciences / Engineering	0.604	0.645

values for the incidence-rate ratios obtained using each method. Why are the results different? The answer is actually simple. When we calculated the incidence-rate ratios using the count tables, we took into consideration the total number of courses taken by each group of students. In the count tables section, we did not compare the total number of failed courses for business students with the total number of failed courses for engineering students. We compared the proportion of failed courses for business students with the proportion of failed courses for engineering students (this is why we were dividing the number of failed courses by the total number of courses). This is an important point because the larger the number of courses taken by a group, the larger the expected number of failures. For example, a student who has taken twenty courses in university has a higher probability of having failed one of these courses than a student who has only taken two courses so far. We say that the first student was exposed to the risk of failure for a longer period time. Using the same logic, a player who spends more time on the field is expected to score more goals than a player who does not spend less time on the field. Someone who has smoked for thirty years is expected to have been hospitalized more than someone who has been smoking for one year.

Because the concept of exposure is important, we need to tell the statistical package to take it into account when calculating the regression equation. So

far, we have not been doing that. All what we have been doping is telling the statistical package to compare the number of occurrences while taking into consideration one or more independent variables.

Going back to the last regression model that we fit, we found that the equation was:

$$\ln(\mu) = -0.079x_1 - 0.438x_2 + 1.237$$

The above equation was obtained without taking into consideration the concept of exposure. Let us now tell the statistical software to take into account the exposure in each observation. In this case, the exposure is the total number of courses taken by each student. The data that includes the exposure variable is shown in Table 5.10. Now we need to tell the statistical software to run a Poisson regression model using the number of failed courses as a dependent variable and college as an independent variable while taking into account that different students have gone through a different number of courses. The following is the output of this model:

$$\ln(\mu) = -0.130x_1 - 0.505x_2 - 1.808$$

We see that the equation has changed. Let us now calculate the expected number of occurrences for each student using the new output. As usual, we use the more intuitive form of the equation:

$$\mu = e^{-0.130(x_1)-0.505(x_2)-1.808}$$

Table 5.10: Records that contain the exposure variable.

Number of failed courses	College	Total number of courses
3	Business	23
0	Business	24
5	Business	30
7	Business	28
1	Engineering	6
2	Business	17
2	Business	15
0	Business	18
4	Engineering	27
2	Engineering	15
5	Engineering	28
3	Engineering	17
8	Business	27
0	Business	9
1	Engineering	27
2	Engineering	10
3	Business	20
4	Engineering	25
5	Business	32
9	Engineering	34
0	Life sciences	17
1	Life sciences	16
1	Life sciences	20
3	Life sciences	28
5	Life sciences	32
4	Life sciences	26
3	Life sciences	17
1	Life sciences	18
2	Life sciences	28

Engineering: $\mu = e^{-0.130(0)-0.505(0)-1.808} = 0.164$

Business: $\mu = e^{-0.130(1)-0.505(0)-1.808} = 0.144$

Life Sciences: $\mu = e^{-0.130(0)-0.505(1)-1.808} = 0.099$

We can now calculate the incidence-rate ratios in order to be able to compare different groups:

Business/Engineering: $0.144/0.164 = 0.878$

Life sciences/Engineering: $0.099/0.164 = 0.604$

These are the exact same values we got when we calculated the incidence-rate ratios using the count tables (refer to Table 5.10). This exercise should illustrate why when running a count model, you need to account for the different exposure times in which each subject was exposed to the risk of the event happening.

5.4 Negative Binomial Regression

Although the Poisson regression model is the basic count model, it actually rarely fits the data because of what is referred to as overdispersion. An important characteristic of the Poisson probability density function is that the mean and the variance are equal. This means that as the mean increases, so does the variability in the data, a characteristic that is called equidispersion. When this assumption is violated, we say that the data displays overdispersion.

In order to address this issue, a negative binomial regression model is used.

This model accounts for overdispersion by adding the parameter alpha to the equation. To illustrate the difference, look at Figure 5.2. The figure shows the probability density function of both the negative binomial and the Poisson distributions (alpha is 1.5 in all plots). We see that in all plots, the negative binomial model allocates a higher probability for smaller counts, specifically the probability of a count of zero. Therefore, you can think of the negative binomial model as a correction to the under prediction of zero, or low, counts.

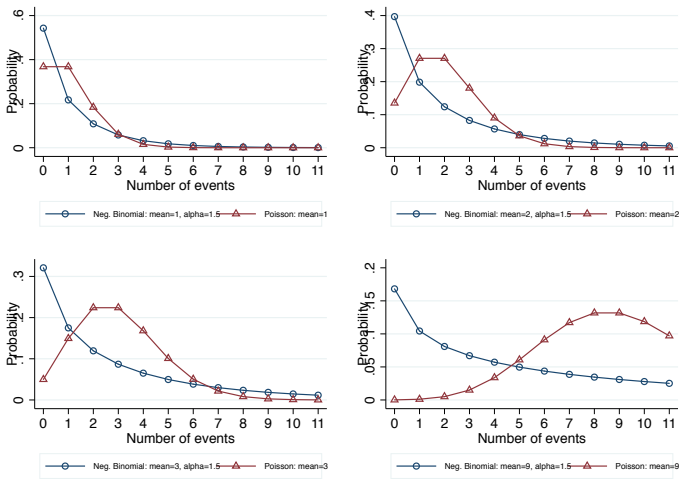

Figure 5.2: Poisson vs negative binomial at alpha = 1.5.

What implication does this have for us? Fortunately, very little. Since we are not interested in the math that goes on behind the scenes, all you need to know is that when we fit a Poisson regression model, we should always follow it up with a negative binomial regression model in order to test whether overdispersion exists. The beauty of it all is that everything we have covered with regards to the meaning of the coefficients when adding the independent

variables still applies the exact way. The output of a negative binomial model is very similar to the output of a Poisson model. We get the coefficients of the variables, and when we calculate e^a we get the incidence-rate ratio. Nothing has changed.

So how do we test whether overdispersion exists? This is done by a likelihood-ratio test that tests the null hypothesis that alpha (which is the extra parameter that is included in the negative binomial model) is equal to zero. If we fail to reject the null hypothesis, we conclude that alpha is zero, and when alpha is zero we end up with the Poisson model. This means that there is no reason to believe that there is overdispersion. If, on the other hand, we reject the null hypothesis that alpha is zero, we conclude that the overdispersion exists and that the negative binomial model should be used instead of the Poisson model.

As an example, consider the data shown in Table Table 5.10. We have already fit a Poisson regression model on this data while taking into consideration the different exposure of each subject. The resulting model was:

$$\ln(\mu) = -0.130x_1 - 0.505x_2 - 1.808$$

If we fit a negative binomial model to the same dataset (while accounting for exposure), we get the following equation:

$$\ln(\mu) = -0.130x_1 - 0.505x_2 - 1.808$$

This is the exact same equation as before. The likelihood-ratio test that is produced automatically by the statistical software tells us that the p-value

of the null hypothesis is 0.5, which is much larger than the cut-off value of 0.05. This means that we cannot reject the null hypothesis (that alpha is zero). The conclusion is that we should use the Poisson model.

Usually, the difference between the parameters of the Poisson model and the negative binomial model is in the p-vales of the coefficients and not the coefficients themselves. The p-values produced by a negative binomial model are larger than those produced by a Poisson model. The result is that a variable that is found to be statistically significant using a Poisson model will turn out not to be significant when a negative binomial model is used even though the value of the coefficient will be almost the same.

5.5 Truncated Models

Sometimes the data that we collect does not contain records with zero counts. An example is a dataset where the dependent variable is the number of semesters that a student spends in university. All students included in the dataset would have at least spent one semester in the university. As another example, I once received a dataset that contained the number of goals scored by each striker is various football (or soccer) leagues around the world. The data only contained records for players who had appeared on the scoresheet at least once, i.e. if a player never scored a goal, he was excluded from the list. This means that the minimum possible number of events was one and not zero. Figure 5.3 shows the histogram of the dependent variable, which is the total number of goals scored. In this case, the minimum is not one, but five. This means that the variable is truncated at the goals = 4 point.

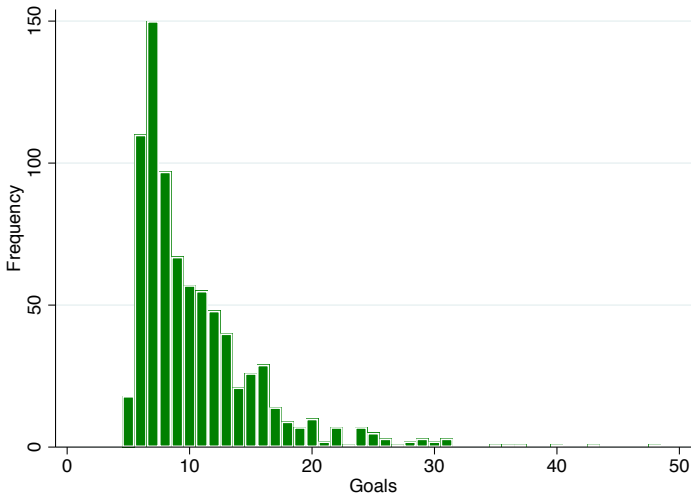

Figure 5.3: Histogram of a truncated variable.

In such cases, whether the data is zero-truncated or truncated at any other point, we use what is referred to as truncated models. These models take into account that a count that is less than a certain value is not possible. Once again, there is a truncated Poisson model and a truncated negative binomial model. Everything that we have said previously about the Poisson and the negative binomial models applies to the truncated models: the meaning of the coefficients, the incidence-rate ratios, and testing for overdispersion.

5.6 Zero-Inflated Models

When we discussed the negative binomial model, it was noted that the model corrects for the underprediction of zero counts. This is why in Figure 5.2 we saw that the probability of low counts, specifically zero, in the negative

binomial model is greater than the probability of these counts in the Poisson model.

Sometimes the number of zeros in the dataset is much larger than what both the Poisson and the negative binomial model assume. In such a case, we say that the number of zeros is inflated, i.e. it is greater than usual. Why would the number of zeros be inflated? This might be due to an underlying mechanism that is acting like a hurdle. As an example, assume that we want to model the number of heart attacks that men under 45 have suffered. In such a case, we would expect that most men at such a young age would not have suffered from a heart attack. A normal male under 45 years of age should not have suffered from a heart attack. This means that the dataset would contain a disproportionately large number of zeros. In this case, we can think of the dataset as containing two different types of men: healthy men who have a zero count, and men with health issues who have a count that is greater than zero.

When we have this type of situation, we use a zero-inflated model to account for the large number of zero counts. The math behind these models is not simple, so we will not get into it. However, it is very important to understand the idea behind these models, and this is what we will do now using the example of heart attacks for males under the age of 45.

- Step 1: If a male is healthy and living under normal conditions, we would expect that he has had no heart attacks, i.e. the count is zero. For males that are unhealthy for their age, or who lead a very stressful life, we would expect that the count would be greater than zero. This means that we can divide the observations into two groups.

The members of the first group have a count of zero and the members of the second group have a count that is greater than zero. To model this situation, we can think of a dependent variable that is binary: an individual is either in the first group or in the second group. This is done by using a binary model that predicts the probability that the event has never occurred as opposed to it having occurred at least once.

- Step 2: After predicting whether an individual is in the first group (where the count is zero) or the second group (where the count is greater than zero), the analysis moves on to predicting the counts for those in the second group. This is done by using a count model, such as a Poisson model or a negative binomial model.

As can be seen above, zero-inflated models are thus made up of two parts. The first part is a logistic model that predicts whether someone has never experienced the event or has experienced it at least once. The second part models the number of times that the event has been experienced by those who have experienced the event at least once. This is why the output of zero-inflated models is divided into two parts, one part for each model. The output helps us understand what are the independent variables that lead to someone being in either of the two groups, and what are the independent variables that increase the frequency of the count for those who have experienced the event. The two sets of the variables need not be the same. This means that the variables that determine to which group an individual belong can be different than the variables that lead to a higher count.

Going back to our heart attack example, assume that we have a variable that indicates how much someone smokes, since exposure to tobacco is one of the causes of heart attacks. If someone has never smoked, this would increase

the probability that they will belong to the group that contains the zero counts. It might also be the case that the variable smoke also leads to a higher count of heart attacks, but this is not necessary the case. We might find that smoking increases the probability of someone suffering at least once from a heart attack but that smoking does not increase the frequency of heart attacks for those who have suffered from it at least once.

As another example, consider a dataset that contains the count variable visits which records the number of times that a patient visits the doctor in the past year. Assume that this variable has an unusually large number of zeros. This means, that many of the patients have had no need to visit the doctor during the past year. Also, assume that recently the hospital has modified its internal policies in order to increase the efficiency of their patient care. The purpose of these reforms is to make sure that the hospital staff are able to respond quickly and efficiently to the needs of the patients thereby reducing the number of subsequent visits from the same patient. In this case, we would expect that the new reforms would reduce the frequency of the counts, but they would not have an effect on whether a patient initially visits or not. In other words, the reforms do not make people healthier, so people will continue to visit their doctor. However, the reforms make sure that when a patient visits, there will be less need for subsequent visits in the short term.

To make this clearer, consider the output that is shown in Table 5.11. This is a sample output from running a zero-inflated model where the dependent variable is the number of doctor visits during the past year. The variable male is a binary variable that indicates whether the individual is a male or not. The variable age records the age of the individual. Finally, the variable

Table 5.11: Sample output of a zero-inflated model.

	Coefficient	**$e^{\text{Coefficient}}$**	**P-value**
Count			
male	0.20	1.22	0.07
age	0.30	1.35	0.00
reform	-0.19	0.83	0.00
Inflate			
male	0.64	1.90	0.01
age	-0.55	0.58	0.00
reform	0.20	1.22	0.12

reform is a binary variable that indicates whether the observation is from the period before the reforms or after the reforms. We see that the output is divided into two parts, with the first part title "count" and the second part titled "inflate". The "count" part is the regression output from modeling the count variable. The "inflate" part is the regression output from modeling the binary variable. The same three variables have been included in both parts in order to see which is significant and which is not. Starting with the "count" part, we see that all three variables are significant (p-value is less than 0.05). Looking at t heir coefficients, we see that the coefficients of male and age are positive. This output is just regular count model output. As we were doing previously, all we need to do is to calculate $e^{Coefficient}$ in order to find the incidence-rate ratio. We see that males have an expected count that is 1.22 times that of females. This means that among those who visited the hospital, males visit the doctors more frequently. We also see that when age increases by one, the expected count is multiple by 1.35, so it increases. The coefficient of the variable reform is negative, with the incidence-rate ratio

being 0.83. This means that among those who visited the hospital, patients who have visited the doctor after the reforms have an expected count that is 0.83 times the expected count of those who have visited the doctor before the reform. This means that the reforms have decreased the frequency of visits.

We now move onto the second part of the output which is titled "inflate". The reason why it is called inflate is that we are investigating which variables are responsible for inflating the number of zeros. Looking at the p-values of the three variables, we see that only male and gender are significant. Since the "inflate" section presents the results of a binary regression, when we calculate $e^{Coefficient}$ we are finding the odds ratio (not the incidence-rate ratio). Looking at the table, we see that males have an odds that is 1.90 times the odds of females of not visiting the doctor. I have written "not visiting" in bold because what the second part is doing is testing which variables increase the odds of being in the zero group (this inflating the number of zeros) compared to the odds of being in the non-zero group. Since the coefficient of male is positive, it means that this variable increases the odds of being in the zero-group. Looking now at the coefficient of age, we see that it is negative. This means that when age increases, the odds of being in the zero-group decrease. Specifically, the odds of being in the zero-group are multiplied by 0.58 when age increases by one unit. This means that the older the patient is, the smaller the probability of him or her being in the zero group. Finally, the variable reform has a positive coefficient with an odds ratio of 1.22. This would mean that after the reforms, the odds of being in the zero-group are multiplied by 1.22 (so they increase). The result however is not significant.

As a recap, the results in Table 5.11 indicate that for the count model:

- Among those who have visited the hospital during the past year, the expected number of visits for males is 1.22 times the expected number of visits of females
- Among those who have visited the hospital during the past year, the expected number of visits are multiplied by 1.35 when age increases by one year
- Among those who visited the hospital during the past year, the expected number of visits for those who came after the reform is 0.83 times the expected number of visits of those who came after the reforms.

With regards to the logistic model, the results indicate that:

- The odds of males not visiting the doctor is 1.90 times the odds of females
- The odds of not visiting the doctor in the past year are multiplied by 0.58 when age increases by one year
- The reforms have no effect on the probability of a patient having visited the doctor during the past year

This example illustrates how the variables that are used in both parts of the model can be different. The variable that will increase the probability that we end up in the group with counts greater than zero might or might not be also responsible for increasing the frequency of the counts.

5.7 Model Comparisons

As you see, count data present an interesting dilemma, in that there are several models to choose from. So now the question becomes, how do we determine whether to use a Poisson model, a negative binomial model, a zero-inflated Poisson model, or a zero-inflated negative binomial model? Fortunately, there are several tests that we can use in order to help us make these decisions.

5.7.1 Comparing Predicted Values with Observed Values

In linear regression, one of the ways to see whether the model has a good fit or not is to plot the actual observed values of the dependent variable and the predicted values on the same graph. Ideally what we want is to see that the predicted values are very close to the observed values. We can use the same tool in count models. However, instead of plotting the observed probabilities and the predicted probabilities, we can plot the difference between them. This means that when we run the four models (Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial), we calculate the probabilities as predicted by each model and then calculate the difference between these probabilities and between the observed probabilities. We then plot these differences on the same graph to see which model results in the smallest differences. Figure 5.4 presents such a graph.

We see that the Poisson regression model (PRM) produces large differences with the observed probabilities for the counts zero, one and two. Clearly, this is not the right model to use. Looking at the other three models, we

see that the zero-inflated negative binomial model (ZINB) and the negative binomial regression model (NBRM) produce the smallest differences, since their graphs are the closest to the zero axis.

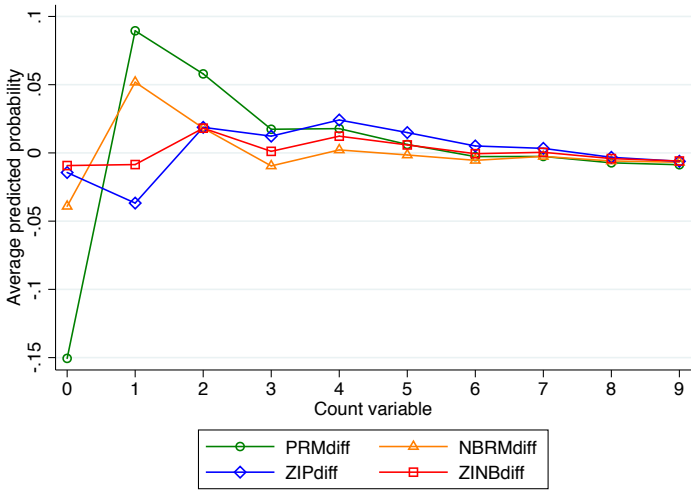

Figure 5.4: Plotting differences between average observed and predicted counts after fitting all models.

5.7.2 Likelihood-Ratio Test of Alpha

It was previously stated that a likelihood-ratio test helps us decide whether we should use a Poisson model or a negative binomial model. This test can be used to compare the Poisson model to the negative binomial model, and it can also be used to compare the zero-inflated Poisson model to the zero-inflated negative binomial model. If the test results in a p-value that is less than 0.05, then we reject the null hypothesis that alpha = 0 and we conclude that overdispersion exists, thus justifying the use of the negative binomial

model.

5.7.3 Vuong Test

The likelihood-ratio test for alpha allows us to compare the Poisson model to the negative binomial model. What if we wanted to compare the Poisson model to the zero-inflated model? In this case we use the Vuong test. Like the likelihood-ratio test, if the Vuong test produces a p-value that is less than 0.05, we conclude that the zero-inflated model should be used.

5.7.4 AIC and BIC Statistics

Another group of tests that can be used to compare two models are the information criteria fit tests. These statistics are only used to compare models. This means that calculating these statistics for a single model does not inform us about the goodness of fit of the model. Instead, we calculate the statistics for the two or more models that we wish to compare. The most commonly used form of statistics in this group are the AIC and BIC. 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.

5.8 Prediction

Once we have determined the best-fit model, it is time to calculated predicted values. In linear regression we use the model in order to predict the dependent variable. In logistic regression, we use the model in order to predict the probability of an event happening. In count models, we can do both. First, we can predict the value of the dependent variable by predicting the number of events for certain values of the independent variables. For example, if we are modeling the number of courses in which a student has failed, we can use the best-fit model in order to predict the number of courses in which a student fails for several values of the independent variables. Take Figure 5.5 as an example. This figure shows the predicted number of events, which is course failures in this case, for students with different grades. As you can see, students with lower grades tend to fail more. As grades increase, the number of failed courses decrease. We also see that the graph starts to level off when the grades are above 80.

Second, we can use count models in order to predict the probabilities for several values of the count variable. For example, instead of predicting the number of events for certain values of a variable, we can predict the probability that the number of events will be a specific value. Figure 5.6 shows an example of this. Unlike Figure 5.5, Figure 5.6 shows how the probability of failing in exactly four courses changes as the students' grades change.

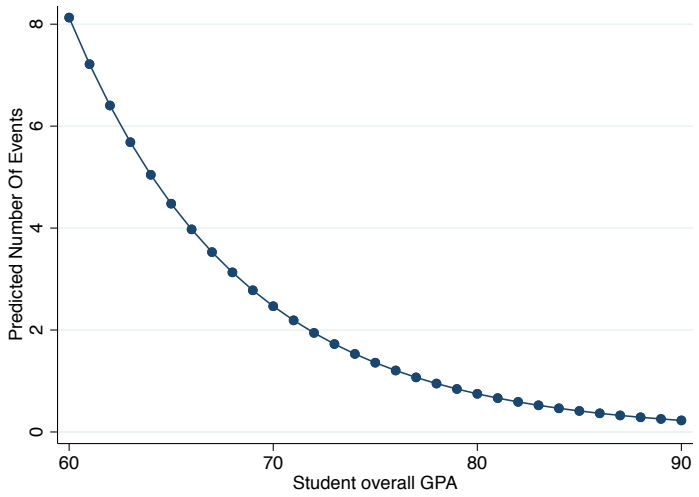

Figure 5.5: Predicting the number of events.

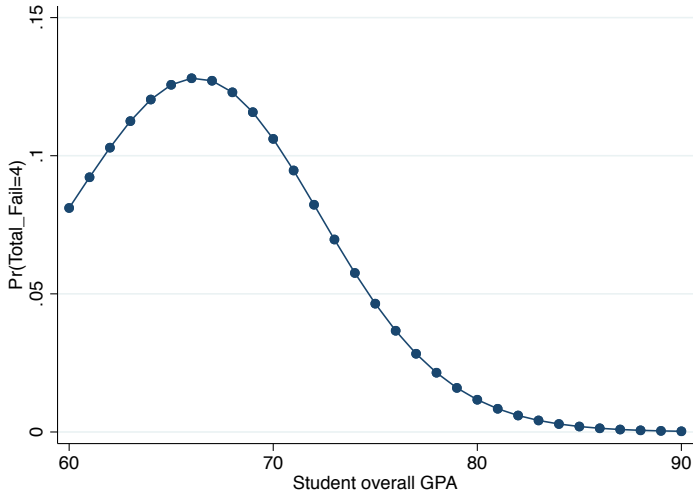

Figure 5.6: Predicting the probability that an event will occur four times.

Chapter 6

Count Models - Case Study

We now have the necessary tools that allow us to analyze a dataset where the dependent variable is a count. In this section, we will be looking at the dataset that contains the following variables:

- id: unique student identifier
- gpa: overall GPA of the student
- total_fail: the total number of courses in which the student has failed (this is the dependent variable)
- college: whether the student is in the engineering school or the business school (one means business, two means engineering)
- gender: whether the student is a male or a female (one means female, two means male)
- english: the average grade on all English courses taken by the student

(data is taken from a non-English speaking country where the language of instruction in university is English)

- total_courses: the total number of courses taken by the student so far in the university

6.1 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. In the case of count data, it is always a good idea to look at the histogram of the dependent variable in order to get an idea of the variable that we are dealing with. The histogram is displayed in Figure 6.1.

As we can see, the variable is not normally distributed, an outcome that is expected when looking at count data. We also see that there is a large number of zeros, which leads us to suspect that perhaps a zero-inflated model should be eventually used.

6.1.1 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

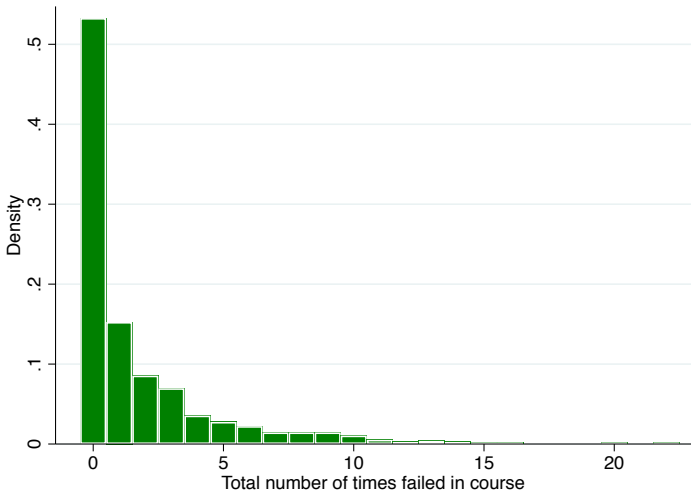

Figure 6.1: Histogram of the dependent variable.

starting step in count models, but their shape is different from what we are used to due to the nature of the dependent variable. For example, let us produce a scatter plot of the dependent variable total_fail and the continuous independent variable GPA.

The output is shown in Figure 6.2. The graph isn't visually appealing. The reason for this is that the outcome can only take on specific values. This is why we see that the points tend to cluster along the horizontal lines. However, if we look closely we see that when the GPA is around 80 and above, almost all of the points lie on the x-axis (the horizontal line that represents an outcome of zero). Therefore, it seems that students with high GPAs do not fail in courses.

In order to make things clearer, we can produce a smoothed scatter plot on top of the scatter plot. The resulting figure is shown in Figure 6.3.

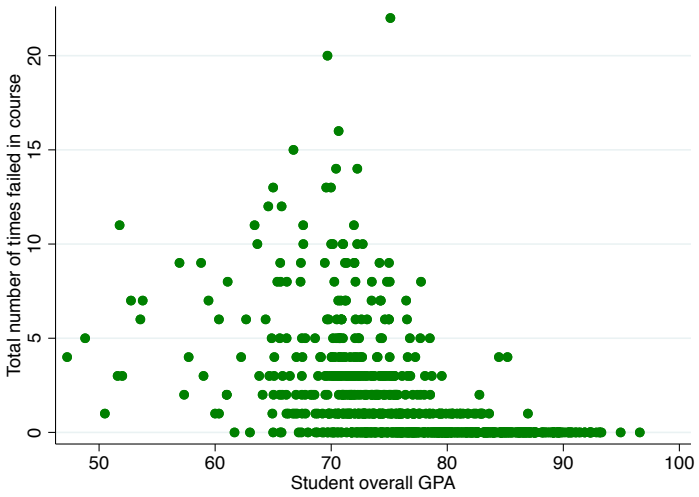

Figure 6.2: Scatterplot of total_fail and GPA.

As can be seen in the figure, as the GPA increases, there is a visible drop in the count variable.

We next fit a Poisson model where we specify total_fail as the dependent variable and gpa as the independent variable.

```
Iteration 0:   log likelihood = -1495.7527
Iteration 1:   log likelihood = -1495.6504
Iteration 2:   log likelihood = -1495.6504

Poisson regression                              Number of obs     =        760
                                                LR chi2(1)        =     817.86
                                                Prob > chi2       =     0.0000
Log likelihood = -1495.6504                     Pseudo R2         =     0.2147

------------------------------------------------------------------------------
  total_fail |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |  -.0952663   .0030879   -30.85   0.000    -.1013185   -.0892141
       _cons |   7.528266   .2187085    34.42   0.000     7.099605    7.956927
------------------------------------------------------------------------------
```

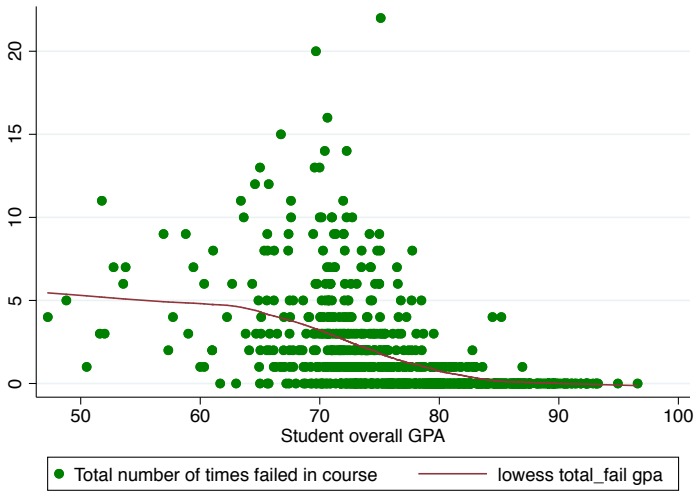

Figure 6.3: Smoothed scatterplot.

Looking at the output, we see that the coefficient of the gpa is -0.095. Recalling what we covered in the theory section, this means that when GPA increases by one, the expected number of occurrences is multiplied by $e^{-0.095} = 0.909$.

Most statistical packages allow us to display the incidence-rate ratio instead of displaying the value of the coefficient.

```
Iteration 0:   log likelihood = -1495.7527
Iteration 1:   log likelihood = -1495.6504
Iteration 2:   log likelihood = -1495.6504

Poisson regression                              Number of obs     =        760
                                                LR chi2(1)        =     817.86
                                                Prob > chi2       =     0.0000
Log likelihood = -1495.6504                     Pseudo R2         =     0.2147

------------------------------------------------------------------------------
  total_fail |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

gpa	.9091308	.0028073	-30.85	0.000	.9036452	.9146498
_cons	1859.877	406.7711	34.42	0.000	1211.488	2855.284

Note: _cons estimates baseline incidence rate.

We see that now the value 0.909 is displayed in the column titled "IRR". We also see that the p-value is less than 0.05, hence the result is significant.

We have however, disregarded one important factor so far, and it is the exposure time. As you recall from the theory part of the course, we need to account for the fact that different subjects were exposed to the probability of the event occurring for different period of time. In our dataset, the variable total_courses contains the total number of courses taken by the student. Statistical packages allow us to iclude an exposure variable. The output after dincluding total_courses as ab exposure variable is shown below.

```
Iteration 0:   log likelihood =  -1146.808
Iteration 1:   log likelihood = -1146.7986
Iteration 2:   log likelihood = -1146.7986

Poisson regression                              Number of obs     =        760
                                                LR chi2(1)        =    1345.83
                                                Prob > chi2       =     0.0000
Log likelihood = -1146.7986                     Pseudo R2         =     0.3698
```

total_fail	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	.868657	.0029586	-41.34	0.000	.8628775	.8744752
_cons	1694.273	408.1211	30.87	0.000	1056.681	2716.584
ln(total_~s)	1	(exposure)				

Note: _cons estimates baseline incidence rate.

As we can see, the value of IRR for the variable gpa is now slightly different.

We also see that there is a new row in the regression table that contains the elements ln(total_courses). From this point forward, we will be including the variable total_courses as an exposure in all the output.

We next take a look at the other continuous variable in our dataset, which is the variable english:

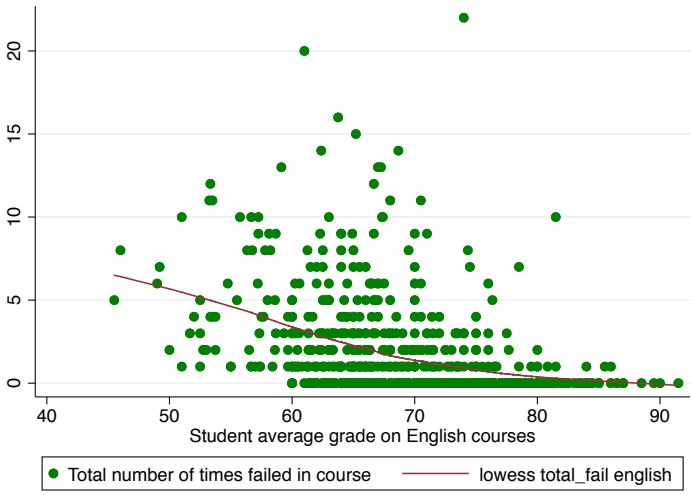

Figure 6.4: Smoothed scatterplot of total_fail and english.

The scatter plot of the variable english and the dependent variable is shown in Figure 6.4. Once again we see evidence that as the value of english increases that the number of events decreases. This makes sense because if students don't have a good grasp of the English language then they will have difficulties in passing subjects that are taught in English.

We next include the variable in a Poisson model:

```
Iteration 0:   log likelihood = -1406.8576
Iteration 1:   log likelihood = -1406.8565
```

```
Iteration 2:   log likelihood = -1406.8565

Poisson regression                              Number of obs     =        755
                                                LR chi2(1)        =     801.08
                                                Prob > chi2       =     0.0000
Log likelihood = -1406.8565                     Pseudo R2         =     0.2216

------------------------------------------------------------------------------
  total_fail |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     english |   .8893731   .0036963   -28.21   0.000      .882158    .8966472
       _cons |   131.3423   35.34142    18.13   0.000     77.51119    222.5589
ln(total_~s) |          1  (exposure)
------------------------------------------------------------------------------

Note: _cons estimates baseline incidence rate.
```

We see that when english increases by one, that the expected number of failed courses is multiplied by 0.889.

6.1.2 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. Let us include each of these two variable separately:

```
Iteration 0:   log likelihood = -1813.5521
Iteration 1:   log likelihood = -1813.5521

Poisson regression                              Number of obs     =        760
                                                LR chi2(1)        =      12.33
                                                Prob > chi2       =     0.0004
Log likelihood = -1813.5521                     Pseudo R2         =     0.0034

------------------------------------------------------------------------------
  total_fail |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

college						
Engineering	.8180037	.0465096	-3.53	0.000	.7317423	.9144341
_cons	.0571091	.0024783	-65.97	0.000	.0524525	.062179
ln(total_~s)	1	(exposure)				

Note: _cons estimates baseline incidence rate.

Looking at the output, we see that the IRR is 0.818. What this means is that the expected number of course withdrawals for engineers is 0.818 times the expected number of course withdrawals for the reference group, which is business students in this case. We can also see that the result is significant at the p ¡ 0.05 level.

We next include the binary variable gender:

```
Iteration 0:   log likelihood = -1712.8214
Iteration 1:   log likelihood = -1712.8145
Iteration 2:   log likelihood = -1712.8145

Poisson regression                              Number of obs     =        760
                                                LR chi2(1)        =     213.80
                                                Prob > chi2       =     0.0000
Log likelihood = -1712.8145                     Pseudo R2         =     0.0587
```

total_fail	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gender						
male	2.845421	.2288983	13.00	0.000	2.430369	3.331356
_cons	.0224076	.0016702	-50.96	0.000	.019362	.0259322
ln(total_~s)	1	(exposure)				

Note: _cons estimates baseline incidence rate.

We see that the expected number of failed courses for males is 2.845 times the expected number of failed courses for females.

6.2 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:

```
Iteration 0:   log likelihood = -1099.3037
Iteration 1:   log likelihood = -1099.2339
Iteration 2:   log likelihood = -1099.2339

Poisson regression                              Number of obs     =        755
                                                LR chi2(4)        =    1416.32
                                                Prob > chi2       =     0.0000
Log likelihood = -1099.2339                     Pseudo R2         =     0.3918

------------------------------------------------------------------------------
  total_fail |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   .8876014   .0041204   -25.68   0.000     .8795622     .895714
     english |   .9663382      .0051    -6.49   0.000      .956394    .9763859
             |
     college |
Engineering  |   1.113613   .0669225     1.79   0.073     .9898769    1.252815
             |
      gender |
        male |   1.413999   .1209597     4.05   0.000     1.195731    1.672109
       _cons |   2418.674   780.7668    24.14   0.000     1284.703    4553.568
ln(total_~s) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Looking at the model, we see that all variable retain their significance levels except for the variable college.

6.3 Negative Binomial Regression

Now it is time to see whether the data displays overdispersion. As you recall, when there is evidence that overdispersion exists, we will need to fit a negative binomial model that will estimate the new parameter alpha.

```
Negative binomial regression                    Number of obs     =        755
                                                LR chi2(4)        =     540.91
Dispersion     = mean                           Prob > chi2       =     0.0000
Log likelihood = -987.46614                     Pseudo R2         =     0.2150
```

total_fail	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	.8562923	.0082241	-16.15	0.000	.8403241	.8725639
english	.9626258	.0078959	-4.64	0.000	.9472738	.9782266
college						
Engineering	1.221273	.1176047	2.08	0.038	1.011218	1.474961
gender						
male	1.327607	.1531766	2.46	0.014	1.058912	1.664484
_cons	39032.16	26223.31	15.74	0.000	10460.5	145644
ln(total_~s)	1	(exposure)				
/lnalpha	-.6676535	.1361952			-.9345912	-.4007157
alpha	.5129107	.069856			.3927464	.6698404

```
Note: Estimates are transformed only in the first equation.
Note: _cons estimates baseline incidence rate.
LR test of alpha=0: chibar2(01) = 223.54              Prob >= chibar2 = 0.000
```

Of primary importance for us is the last line in the output, the one which reports the result of the likelihood-ratio test with regards to the value of alpha. As you recall, a likelihood-ratio test is used to test the null hypothesis that alpha is equal to zero, which means that there is no overdispersion.

Looking at our output, we see that the p-value is less than 0.05, thus leading to the rejection of the null hypothesis. We therefore conclude that there is overdispersion and that the negative binomial regression model should be used instead of the Poisson model.

6.4 Zero-Inflated Models

When we plotted the histogram of the dependent variables, we noted that the number of zeros seems to be too high. This means that perhaps a zero-inflated model would be of better use. The zero-inflated Poisson model is fit using the zip command and the zero-inflated negative binomial model is fit using the zinb command. Since we have found that there is overdispersion in the data, it would make sense for us to fit the zero-inflated negative binomial regression model:

```
Zero-inflated negative binomial regression      Number of obs     =        755
                                                Nonzero obs       =        351
                                                Zero obs          =        404

Inflation model = logit                         LR chi2(4)        =     253.69
Log likelihood  = -957.9385                     Prob > chi2       =     0.0000
```

total_fail	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
total_fail						
gpa	-.1057381	.0090785	-11.65	0.000	-.1235317	-.0879445
english	-.0304713	.0076049	-4.01	0.000	-.0453766	-.015566
college						
Engineering	.1139091	.0943508	1.21	0.227	-.071015	.2988332
gender						
male	.2515483	.1220756	2.06	0.039	.0122845	.4908122

_cons	6.801989	.66255	10.27	0.000	5.503415	8.100564
ln(total_~s)	1	(exposure)				
inflate						
gpa	.3518453	.0541493	6.50	0.000	.2457145	.4579761
english	.0159122	.0350812	0.45	0.650	-.0528457	.0846701
college						
Engineering	-.3981901	.4403973	-0.90	0.366	-1.261353	.4649727
gender						
male	-.0967932	.4380667	-0.22	0.825	-.9553881	.7618018
_cons	-28.72857	3.672004	-7.82	0.000	-35.92556	-21.53157
/lnalpha	-1.320756	.1988137	-6.64	0.000	-1.710424	-.9310886
alpha	.2669333	.05307			.1807891	.3941244

As you recall from the theory part, the zero-inflated model is made up to two parts. One part is binary, in that it predicts whether an individual will be in the zero group or in the non-zero group. The other part is the count part, where the expected number of occurrences is predicted. What I have done here is that I have included all independent variables in both parts of the models in order to see which part do these variables affect. In other words, we want to see which variables might be causing the inflated number of zeros. We have included all variables because at the point we do not know which variables will be significant and which will not be significant.

Notice that the output is divided into two parts. The top part represents the count model regression, while the bottom part, titled "inflate", represents the binary regression where we are trying to predict group membership (the zero-group vs the non-zero group). The following is very important. The bottom part is trying to see which variables are causing the inflated number

of zeros. What this means is that if a variable has a positive value for the coefficient, then increasing values of this variable will increase the probability that the individual will end up in the zero-group. If you look at the variable gpa in the "inflate" section, you can see that the coefficient is 0.352, which is positive. This basically means that higher values of GPA will lead to an increase in the probability that the student will remain in the zero-group. This makes sense. Students with higher GPAs are more likely to fail in zero courses. We also see that the coefficient of english is positive. The same logic applies here. Higher grades on the English courses will increase the probability that the student will remain in the zero group.

Looking at the variable college, we see that the coefficient is negative. As you know, this is a binary variable, and the reference group here is business. What this coefficient means is that engineering students are less likely than business students to remain in the zero group. How much less likely? In count models, when we find $e^{coefficient}$ we are finding the incidence-rate ratio. In logistic regression, when the dependent variable is binary, calculating $e^{coefficient}$ gives us the odds ratio, which is the ratio of the odds that the event happens in one group compared to the reference group. In our case $e^{-0.398} = 0.672$. What this means is that the odds of an engineering student being in the zero group are 0.672 times the odds that a business student being in the zero group.

In some statistical packages, it is possible to request that the exponentiated coefficients be diplayed. The following table is an example:

```
zinb (N=755): Factor change in expected count

  Observed SD:  2.8663

Count equation: Factor change in expected count for those not always 0
```

	b	z	P>|z|	e^b	e^bStdX	SDofX
gpa	-0.1057	-11.647	0.000	0.900	0.450	7.555
english	-0.0305	-4.007	0.000	0.970	0.800	7.338
college						
Engineering	0.1139	1.207	0.227	1.121	1.058	0.494
gender						
male	0.2515	2.061	0.039	1.286	1.125	0.467
constant	6.8020	10.266	0.000	.	.	.
alpha						
lnalpha	-1.3208	.	.	.	.	.
alpha	0.2669	.	.	.	.	.

Binary equation: factor change in odds of always 0

	b	z	P>|z|	e^b	e^bStdX	SDofX
gpa	0.3518	6.498	0.000	1.422	14.269	7.555
english	0.0159	0.454	0.650	1.016	1.124	7.338
college						
Engineering	-0.3982	-0.904	0.366	0.672	0.821	0.494
gender						
male	-0.0968	-0.221	0.825	0.908	0.956	0.467
constant	-28.7286	-7.824	0.000	.	.	.

We see that the output continues to show the coefficients in the first column, in addition to the values $e^{coefficient}$ in the column title e^b. If you look at the variable college in the bottom part, you will see that the coefficient is -0.3982 and that the odds ratio is 0.672, just like we calculated.

We are interested in the p-values of the inflate part since we originally included all the independent variables in order to see which ones might

be inflating the number of zeros. We had previously seen that that in the "inflate" part, only the variable gpa is significant. This would indicate that we might be better off fitting a model that only included gpa in the "inflate" part:

```
Zero-inflated negative binomial regression      Number of obs     =        755
                                                Nonzero obs       =        351
                                                Zero obs          =        404

Inflation model = logit                         LR chi2(4)        =     259.07
Log likelihood  = -958.4367                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
  total_fail |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
total_fail   |
         gpa |  -.1059945   .0090624   -11.70   0.000    -.1237565   -.0882324
     english |  -.0316688   .0073566    -4.30   0.000    -.0460876   -.0172501
             |
     college |
Engineering  |   .1447259   .0880362     1.64   0.100    -.0278219    .3172737
             |
      gender |
        male |   .2656364   .1119332     2.37   0.018     .0462513    .4850214
       _cons |    6.86581   .6550236    10.48   0.000     5.581988    8.149633
ln(total_~s) |          1  (exposure)
-------------+----------------------------------------------------------------
inflate      |
         gpa |    .360155   .0459039     7.85   0.000      .270185    .4501251
       _cons |  -28.61654   3.626404    -7.89   0.000    -35.72416   -21.50892
-------------+----------------------------------------------------------------
    /lnalpha |  -1.312343   .1968799    -6.67   0.000     -1.69822   -.9264651
-------------+----------------------------------------------------------------
       alpha |   .2691887   .0529978                      .183009    .3959509
------------------------------------------------------------------------------
```

The output shows that gpa is significant both as an inflation variable and as a count variable. The output also shows that the variable college is no longer significant.

6.5 Comparing Count Models

So which is it? Is it the negative binomial model or the zero-inflated binomial model? This is where we need to compare the four models: Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial.

As mentioned in the theory section, one way to compare the models is to produce a graph that displays the error produced by each model. What we do is that we calculate the predicted number of counts for each model and we compare these predicted number of counts to the observed counts in the dataset. Such a graph is displayed in Figure 6.5.

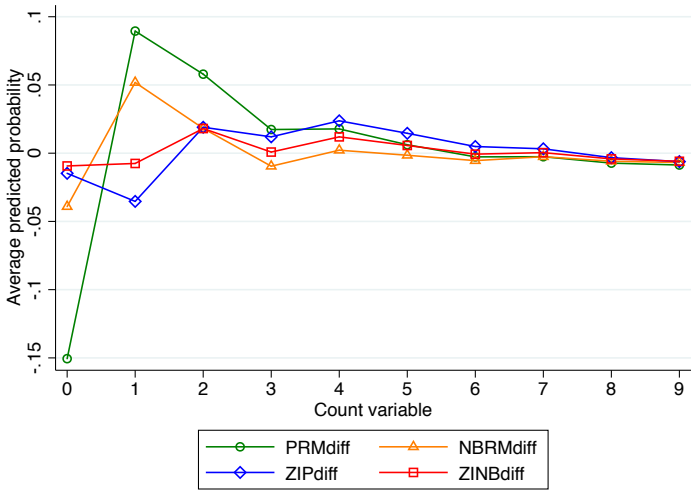

Figure 6.5: Difference between observed and predicted values in each of the four models.

We see in the figure that the Poisson regression model (PRM) produces the largest differences between observed and predicted values, followed by the zero-inflated Poisson model (ZIP). The smallest deviations are produced

by the negative binomial regression model (NBRM) and the zero-inflated negative binomial (ZINB) model, with the ZINB doing a slightly better job.

Another way to compare the four models is to look at the AIC and BIC statistics for each. As you recall from the theory part, we favor the model that produced the smallest values of these statistics. The following output displays side-by-side the output from each model, along with the AIC and BIC statistics of each:

Variable	PRM	NBRM	ZIP	ZINB
total_fail				
Student overall GPA	0.928	0.859	0.966	0.941
	-17.66	-11.68	-6.74	-4.75
Student average grade on Eng~h	0.967	0.971	0.985	0.975
	-6.92	-2.84	-2.95	-2.65
College				
Engineering	1.218	1.450	1.177	1.291
	3.28	3.04	2.53	2.25
Gender				
male	1.553	1.426	1.519	1.459
	5.11	2.56	4.54	2.77
Constant	2409.199	5.01e+05	66.849	689.760
	25.64	14.77	11.20	7.17
lnalpha				
Constant		1.222		0.731
		2.05		-2.26
inflate				
Student overall GPA			1.318	1.485
			11.59	7.91
Constant			0.000	0.000
			-11.61	-7.83
Statistics				

alpha		1.222		
N	755	755	755	755
ll	-1427.307	-1098.999	-1195.208	-1064.134
bic	2887.747	2237.759	2436.802	2181.282
aic	2864.614	2209.999	2404.415	2144.268

legend: b/t

In the first part, we see the exponentaited parameters for the four models. We note that the values are very similar across the models. Most importantly, the direction of the variables is the same. By direction I mean that the models agree whether a variable increases or decreases the counts. Since the values that are displayed are exponentiated, these are incidence-rate ratios. This means that values that are greater than one indicate that an increase in the variable will lead to an increase in the count, and a value that is less than one indicates that an increase in the variable will lead to a decrease in the count. We also see that the output includes an "inflate" section for the zero-inflated models. The output also contains the AIC and BIC goodness of fit statistics at the bottom. Models that produce lower values of these statistics are considered a better fit. We see that the zero-inflated negative binomial model produces the lowest AIC and BIC statistics.

Therefore, it is safe to deduce that the zero-inflated negative binomial model is the best suited model to be used with this dataset.

6.6 Visualizing the Results

In my opinion, the best way to understand models is to summarize them using meaningful graphs. As an example, take the case of the independent

GPA. We would like to know how the expedted number of failed courses changes with respect to changes in the value of student GPA. Let us start by plotting the relationship between the dependent variable and GPA. I would like to see how the expected number of failed courses changes when GPA varies from 60 to 100. The resulting graph is shown in Figure 6.6.

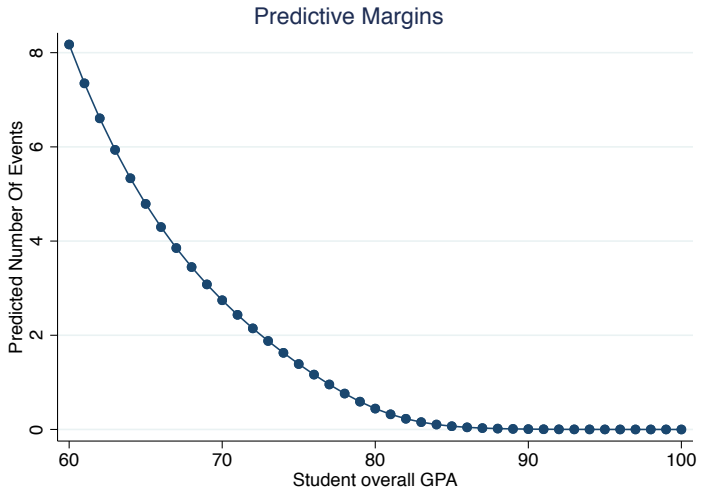

Figure 6.6: Using marginsplot to visualize the relationship between the expected number of events and GPA.

We see that there is a drop in the expected number of failed courses as GPA increases, and that this drop tends to level off when the GPA is above 80 since students with a GPA that is higher than 80 are expected to fail in zero courses. There is no difference between a student who has a GPA of 87 and one who has a GPA of 94.

We can go a step further and plot group differences across colleges for example. The result is shown in Figure 6.7.

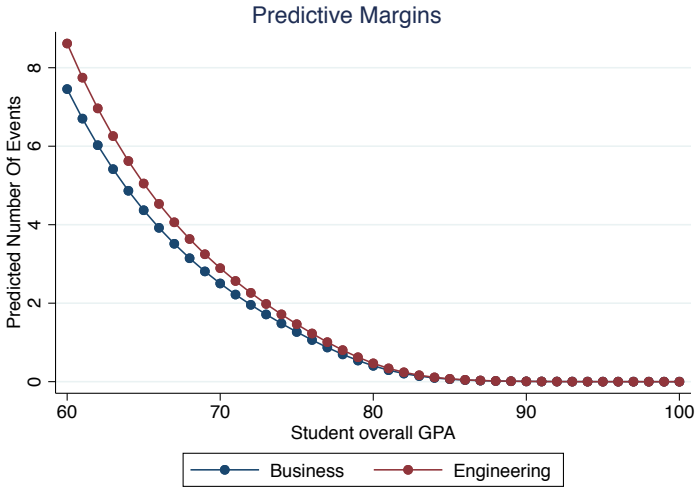

Figure 6.7: Visualizing the effect that two variables have on the expected number of outcomes.

We see that the differences between engineering students and business students is actually small (as you recall, the result is not significant), and that these differences vanish for students with high GPAs.

Not only can we visualize the expected number of events, but we can also visualize the probability that the event will occur a certain number of times. For example, assume that we want to calculate the probability that a student will fail in four courses for different values of the variable gpa and for each of the genders. Such a graph is shown in Figure 6.8. If you look at the title of the y-axis, Stata is clearly telling us that the predicted values are the probabilities that the total failed courses is equal to four.

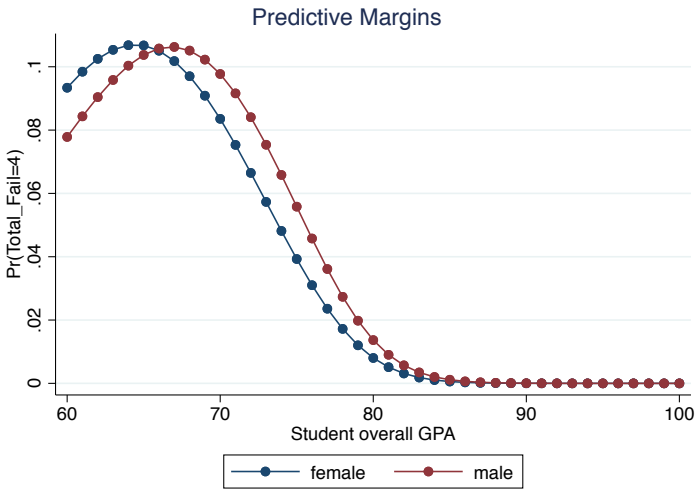

Figure 6.8: Plotting the probability that the event will occur exactly four times.

Chapter 7

References

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

Hilbe, J.M. (2011). Negative Binomial Regression. 2nd edition. Cambridge University 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.

Ryan, T.P. (2009). Modern Regression Methods. 2nd edition. Wiley.