Linear Regression using Stata

Theory and Application

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

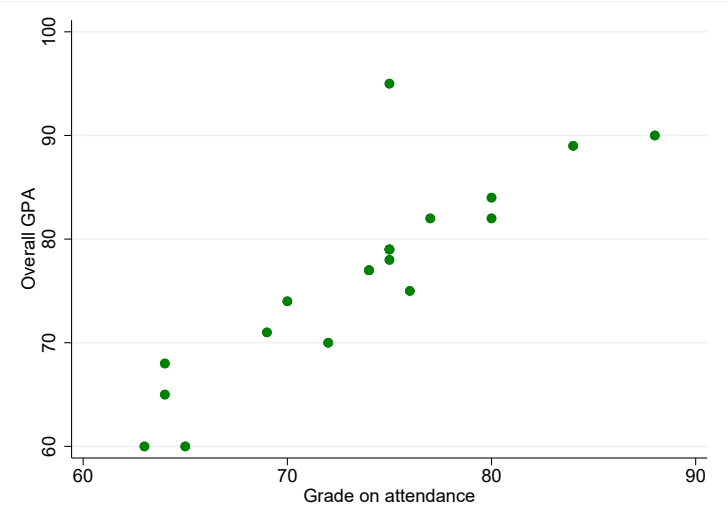

Figure 1.1: Scatter plot of the data points.

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 statis-

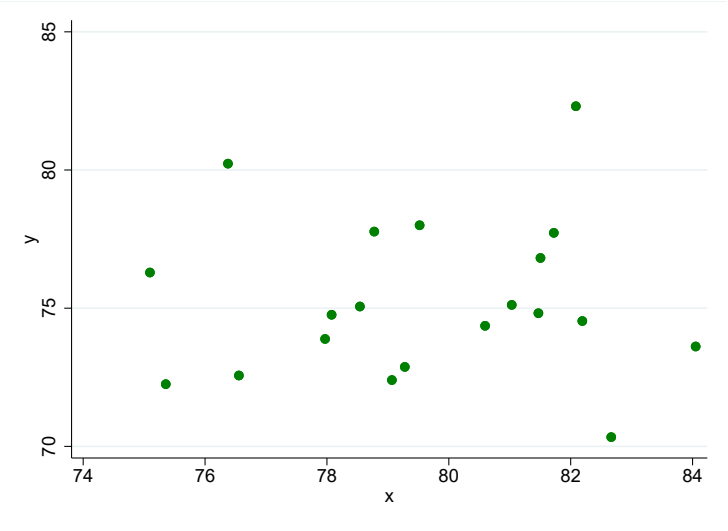

Figure 1.2: A scatter plot where there is no clear relationship between the variables y and x.

tical 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 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.

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

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

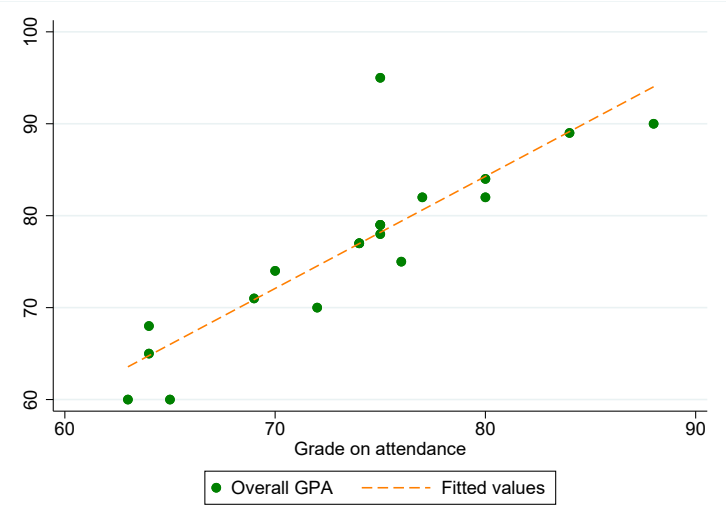

Figure 1.3: The best-fit line as computed by the simple regression model.

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 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.

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

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:

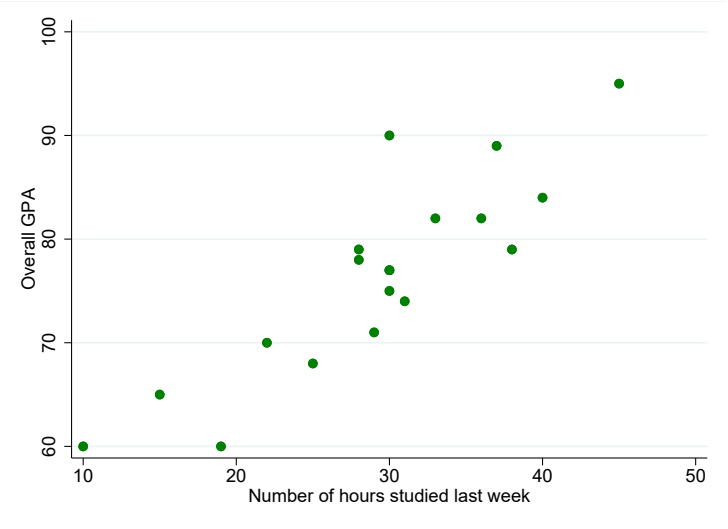

Figure 1.4: Scatter plot of GPA and study.

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

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

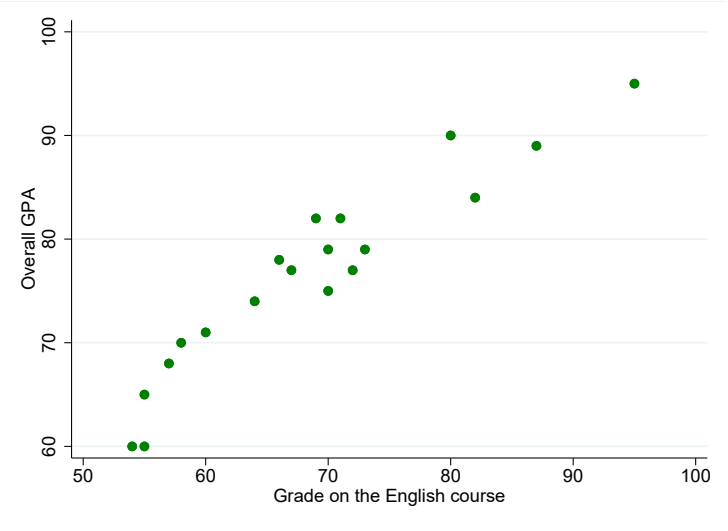

Figure 1.5: Scatter plot of the variables GPA and English.

drawing the 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:

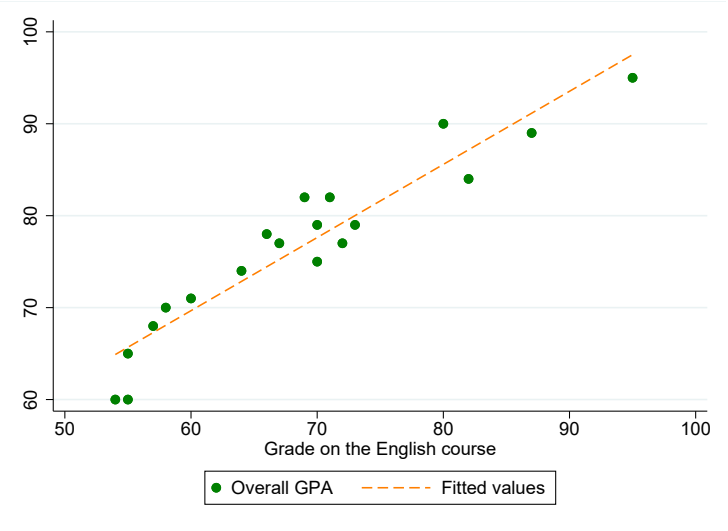

Figure 1.6: Drawing the best-fit line resulting from regressing GPA on English.

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

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.

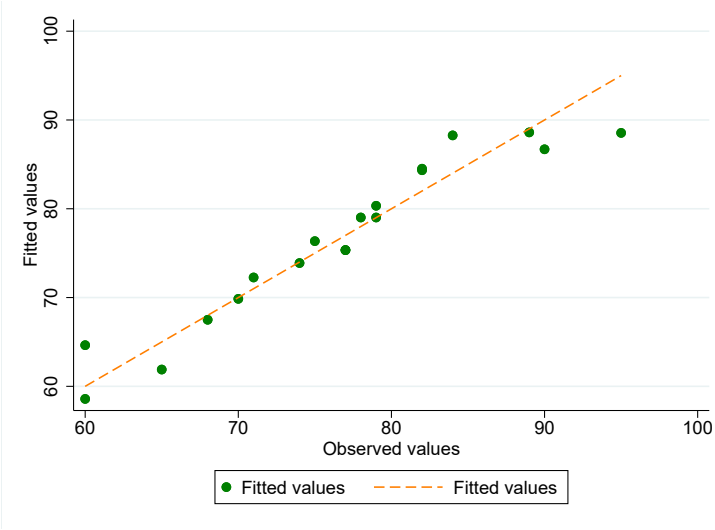

Figure 1.7: Checking model fit: Plotting the predicted values against the observed values (The dashed line represents the 45 degree diagonal along which the points should fall if the model was well fit).

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.

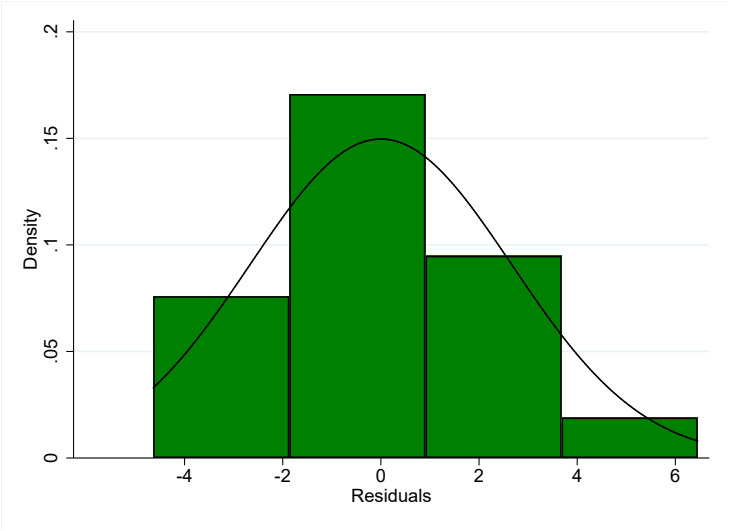

Figure 1.8: Checking the normality of the residuals: A histogram of the residuals overlaid with a normal curve.

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.

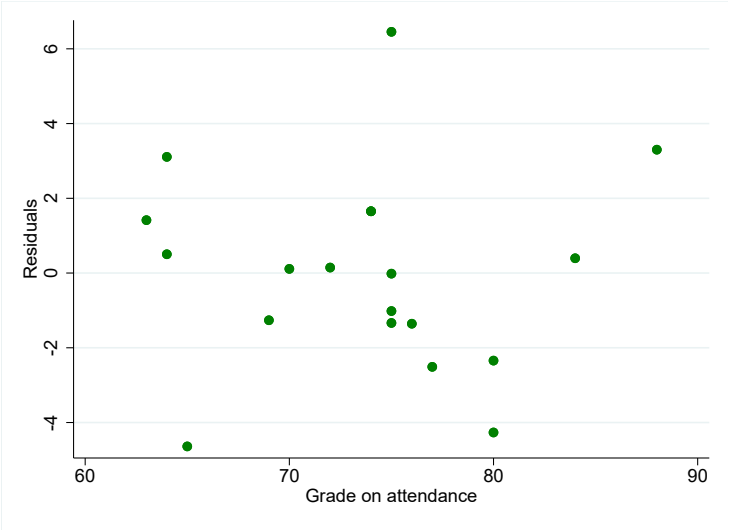

Figure 1.9: Testing for independence of the residuals: Plotting the residuals against the independent variable attendance.

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.

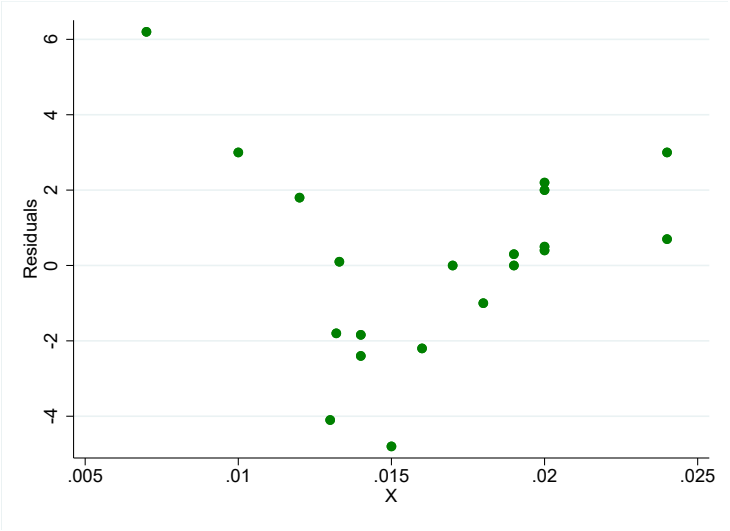

Figure 1.10: Testing for independence of the residuals: A case where the assumption is violated.

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

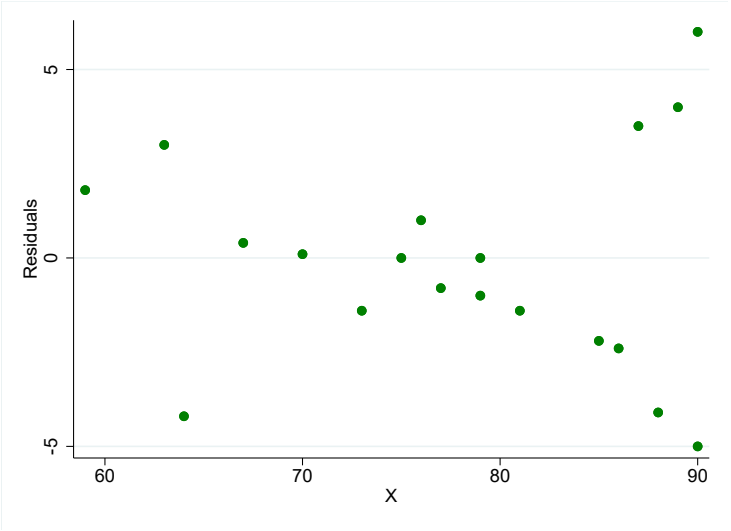

Figure 1.11: Testing for the independence of the residuals: Plotting the residuals against the predicted values of the dependent variable.

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.

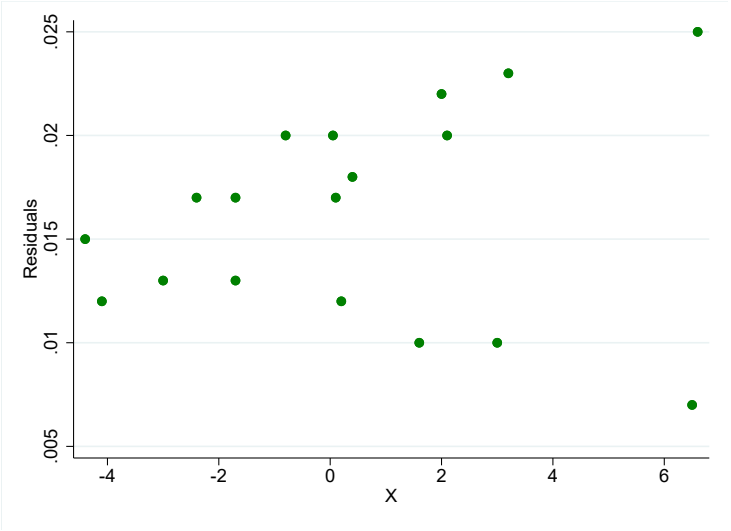

Figure 1.12: Testing for homoskedasticity: A case where the assumption is violated.

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.

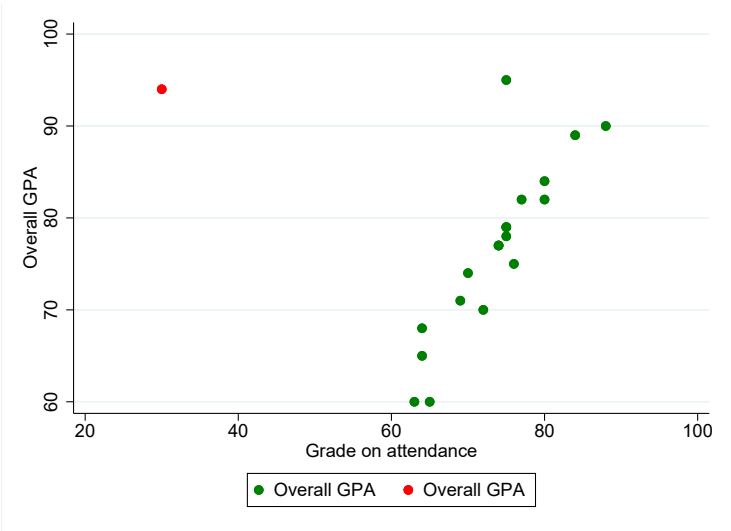

Figure 1.13: Outliers: The red dot is a new observation.

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 substan-

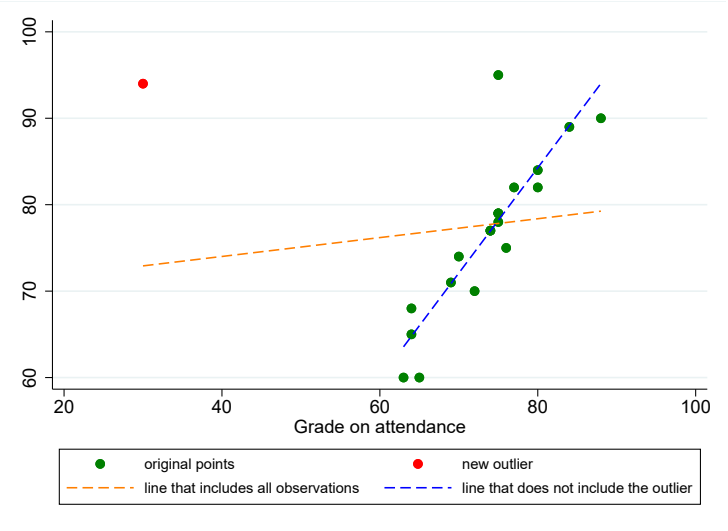

Figure 1.14: The effect that an outlier has: The short dashed line includes all observations, the long dashed line includes only the original observations (excludes the outlier).

tially, 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

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

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.

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 - Application

Now that the theory has been covered, it is time to see how to run a linear regression model in Stata. First you need to make sure that you have downloaded the file linear1.dta. This dataset contains the variables GPA, attendance, study, gender, english, and major.

2.1 Simple Linear Regression

2.1.1 The Model

As was previously mentioned, any good analysis starts with graphs. This is why the first step is to produce a scatter plot of GPA and attendance. This is done using the following command:

```
. scatter gpa attendance
```

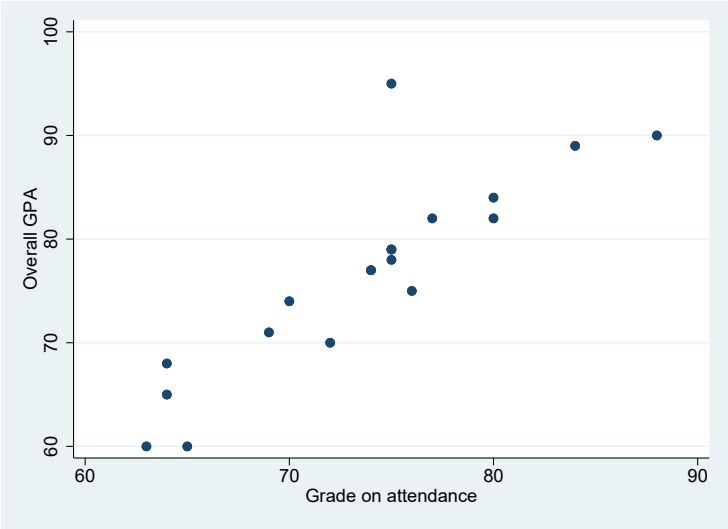

Figure 2.1: Scatter plot of GPA and attendance.

The result of running this command is shown in Figure 2.1. Notice that the figure has different colors from the one that was produced when the theory was being discussed. This is because I used options that allowed me to specify the exact colors to use. Figure 2.1 shows the graph as it appears by default.

Next we will run a simple linear regression model. To run a model that includes GPA as the dependent variable and attendance as the independent variable, we use the following command:

```
. regress gpa attendance
```

Source	SS	df	MS	Number of obs	=	19
				F(1, 17)	=	51.08
Model	1232.46901	1	1232.46901	Prob > F	=	0.0000
Residual	410.162567	17	24.1272098	R-squared	=	0.7503
				Adj R-squared	=	0.7356
Total	1642.63158	18	91.2573099	Root MSE	=	4.9119

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	1.218488	.1704853	7.15	0.000	.8587959	1.578181

```
       _cons |  -13.20441   12.61252    -1.05   0.310     -39.8145    13.40567
```

Notice that the dependent variable comes before the independent variable. From the output, we can see that the value of the coefficient of attendance is 1.22 and that the intercept is -13.20. We can also see that the p-value of the variable attendance is 0.000 which is less than the cut-off value of 0.05. This means that the result is significant. Finally, we see that the value of R-squared is 0.75. There is a lot more information in the output, but most of it is used to calculate the statistics that we just listed. For example, We can see that the total variation of the dependent variable is 1642.63, and that the variation explained by our model is 1232.47. If you divide 1232.47 by 1642.63 you will get 0.75, which is the value of R-squared. As you recall, R-squared is the proportion of the total variation that is explained by our model. Bases on the output, we can see that the linear equation is:

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

2.1.2 Model Fit

Now that we have the equation for the best-fit line, we can calculate how accurate the line is. In order to calculate the predicted values, we can use the following command directly after the regress command:

```
. predict gpa_predicted
(option xb assumed; fitted values)
```

The predict command tells Stata to calculate the predicted values. The

predicted values are stored and created in a new variable which I have chosen to name gpa_predicted. You can pick whatever name you want. Now that we have a variable that contains the predicted values, we can plot these predicted values against the actual values in order to see whether the predicted values are accurate. As you recall, the best-case scenario would be for the points on fall on the diagonal line that represents the equation y = x. To create this graph, we can use the following command:

```
. twoway (scatter gpa_predicted gpa) (line gpa gpa)
```

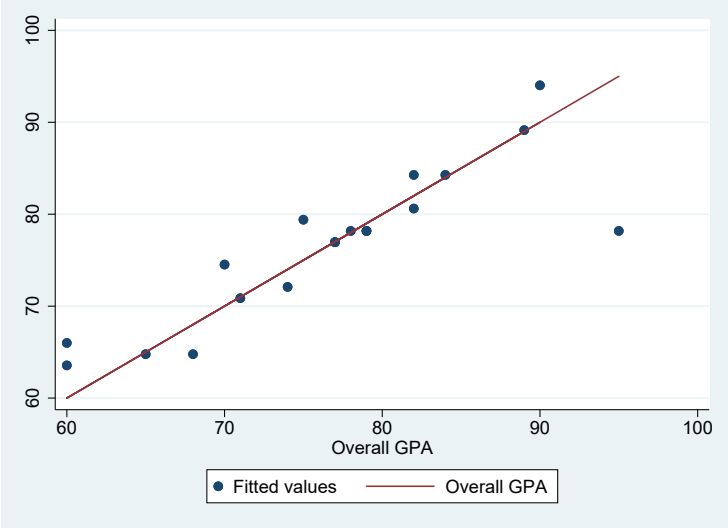

Figure 2.2: Checking model fit: Plotting the predicted values against the observed values (The line represents the 45 degree diagonal along which the points should fall if the model was well fit).

This command tells Stata that we want to draw two plots on the same graph. Each of these plots is enclosed in a set of parenthesis. In the first set of parenthesis, we tell Stata to plot a scatter plot of the predicted values against the actual values. In the second set of parenthesis, we tell Stata to draw the line y = x. Here we specified the variable GPA, so we are drawing the line

GPA = GPA. Figure 2.2 shows the output of the command. Looking at the graph, we see that in general the dots follow the diagonal line, with one exception.

We can also use the predict command in order to calculate the residuals. By default, the command calculates the predicted values. However, we can use the resid option:

```
. predict residuals, resid
```

The command creates a new variable names residuals. This variable will contain the residual for every observation. The residuals are used in order to test certain assumptions. For example, one of the assumptions is that the residuals have a normal distribution. To check this assumption, we can execute the following command:

```
. histogram residuals, normal
(bin=4, start=-5.9973369, width=5.7037786)
```

The output of the command is shown in Figure 2.3. The graph clearly shows that the normality assumption is violated. This is a serious red flag. Another useful way to check for normality is to plot a quantile-normal plot. This plot compares the distribution of any variable with the normal distribution:

```
. qnorm residuals
```

The output of the command is shown in Figure 2.4. If the variable, which is residuals in our case, follows a normal distribution, the dots would be located at or near the line. We can clearly see that this is not the case. Therefore, both Figure 18 and Figure 19 indicate that the normality assumption is

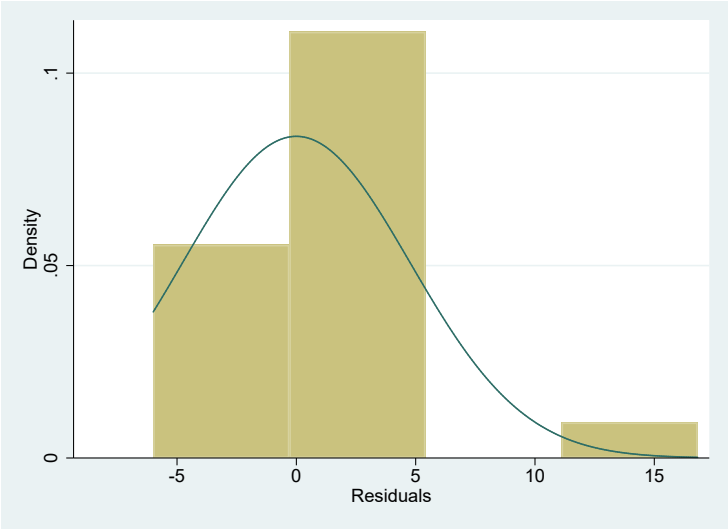

Figure 2.3: Checking the normality of the residuals: A histogram of the residuals overlaid with a normal curve.

violated. In general, we also need to test for the other assumptions. However, given that the first assumption is violated we already know that there is a serious problem with the model. We next try to fit a better model by including more than one independent variable. Perhaps this result in a more parsimonious model.

2.2 Multiple Linear Regression

We now want to consider the effect of another variable, which is study. As usual, we start with a scatter plot of the dependent variable and new the dependent variable:

```
. scatter gpa study
```

The output is shown in Figure 2.5. Once again, we see evidence that students

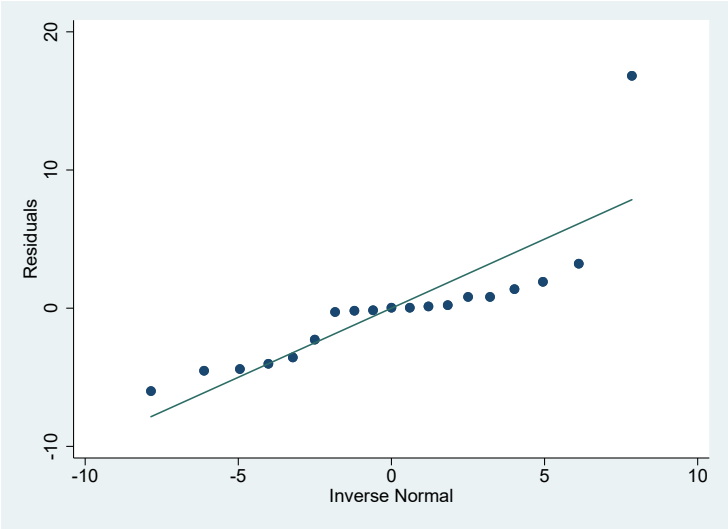

Figure 2.4: Checking the normality of the residuals: Using quantile-normal plots.

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. To run a multiple linear regression model, we use the same command that we used to run a simple model. The only difference is that we include more than one independent variable after the dependent variable:

```
. regress gpa attendance study
```

Source	SS	df	MS	Number of obs	=	19
				F(2, 16)	=	69.95
Model	1474.04889	2	737.024443	Prob > F	=	0.0000
Residual	168.582692	16	10.5364183	R-squared	=	0.8974
				Adj R-squared	=	0.8845
Total	1642.63158	18	91.2573099	Root MSE	=	3.246

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.7110424	.1546729	4.60	0.000	.3831505	1.038934
study	.5878398	.1227652	4.79	0.000	.3275893	.8480903
_cons	6.984299	9.340523	0.75	0.465	-12.81673	26.78532

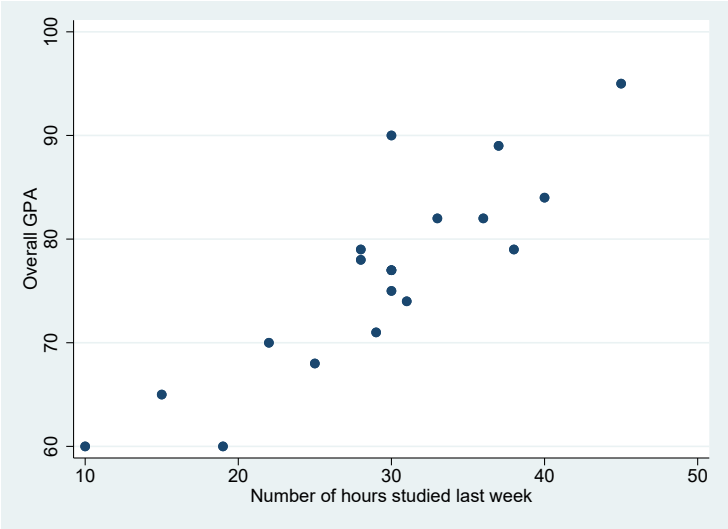

Figure 2.5: Scatter plot of GPA and study.

Based on the output, the best-fit line has the following form:

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

We can see that the p-values of both independent variables are less than 0.05. We can also see that R-squared is 0.90. Given that both independent variables are found to be significant, and that the R-squared value is higher than the R-squared obtained in the linear regression model, we can conclude that this model is better. Since both models contain a different number of independent variables, it would be better to look at the adjusted R-squared values. Again we see that the value of his statistics for the multiple linear regression model is larger than the value for the simple regression model.

2.2.1 Model Fit

In order to assess the accuracy of our model, we can calculate the predicted values and compare them to the actual values. To calculate the predicted values of the new model, we run the following command directly after running the regress command:

```
. predict gpa_predicted2
(option xb assumed; fitted values)
```

We name the variable gpa_predicted2 since there already exists a variable named gpa_predicted which we created in order to save the predicted values of the simple linear regression model. We next create a graph that will help us assess the fit of the model:

```
. twoway (scatter gpa_predicted2 gpa) (line gpa gpa)
```

Here we plot the predicted values of the multiple linear model against the observed values. We also draw a diagonal line in order to make it easier for us to visualize the output. The result of running the command is shown in Figure 2.6.

If you compare Figure 2.6 and Figure 2.2, you will notice that the multiple model that contains both attendance and study as independent variables does a better job. The dots are closer to the diagonal line. In addition, we had seen in Figure 2.2 that there is a point that is very far from the line. This is no longer the case in Figure 2.6.

We next look at the residuals. We can also use the predict command with the resid option in order to calculate the residuals:

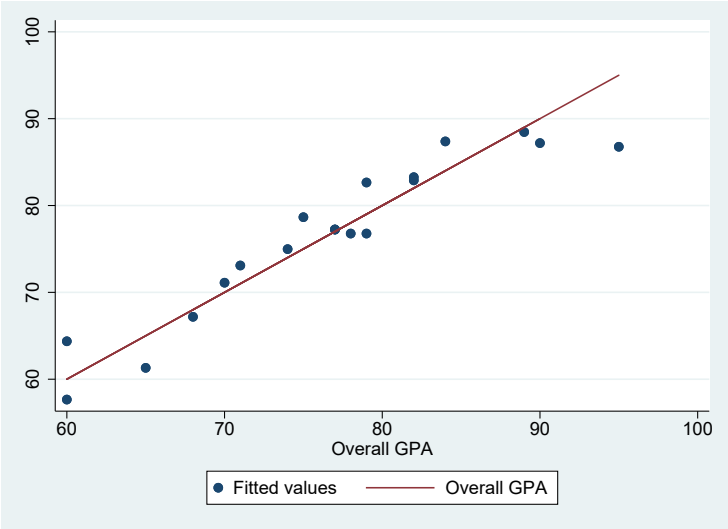

Figure 2.6: Checking model fit: Plotting the predicted values against the observed values (The line represents the 45 degree diagonal along which the points should fall if the model was well fit).

```
. predict residuals2, resid
```

The command creates a new variable names residuals2. This variable will contain the residual for every observation. To check the assumption of normality, we can execute the following command:

```
. histogram residuals2, normal
(bin=4, start=-4.3710127, width=3.1514351)
```

The output of the command is shown in Figure 2.7. Although the new residuals are more normal than the ones obtained from the previous simple model (the one that included only attendance as an independent vaiable), we still see that the normality assumption is violated. This is a serious red flag. We can confirm this by creating a quantile-normal plot:

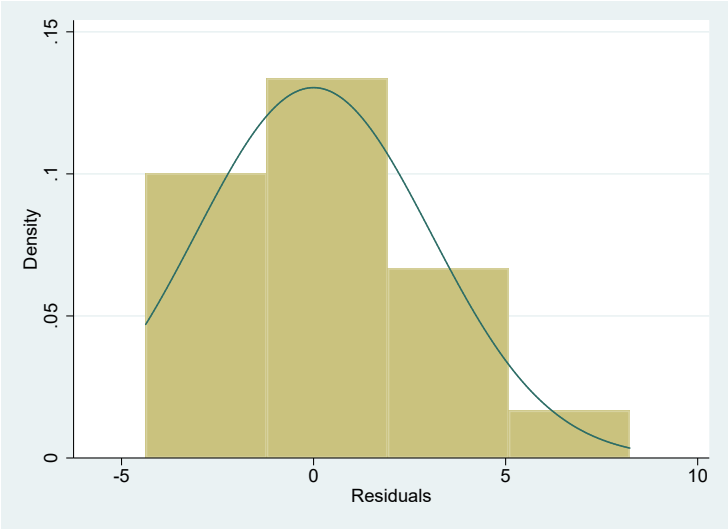

Figure 2.7: Checking the normality of the residuals: A histogram of the residuals overlaid with a normal curve.

```
. qnorm residuals2
```

The output of the command is shown in Figure 2.5. If we compare this figure to Figure **??**, we see that the new residuals are more normal than the ones obtained from the simple regression model. However, we still see that the normality assumption seems to be violated.

2.2.2 Binary Variables

We would now want to include gender as an independent variable. As we already discussed, this variable is not quantitative. Instead, it is categorical in that it divided people into categories. In this case there are two categories, male and female. The variable is already included in the dataset. To take a closer look at it, we can run the following command:

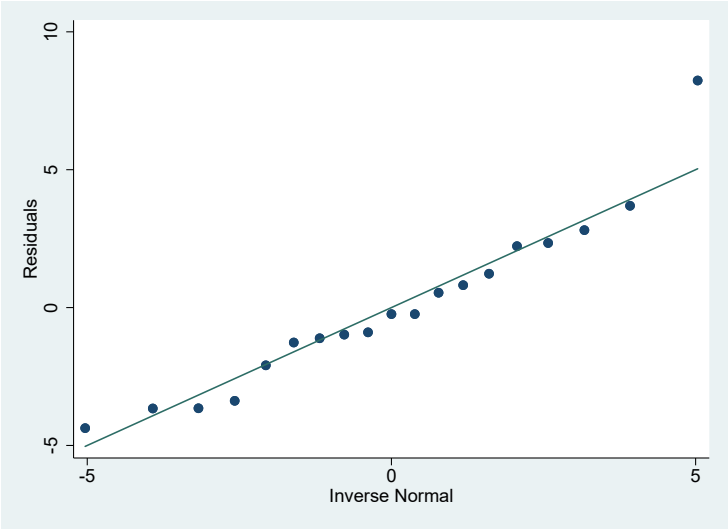

Figure 2.8: Checking the normality of the residuals: Using quantile-normal plots.

```
. codebook gender

-------------------------------------------------------------------------------
gender
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  gender

                 range:  [0,1]                        units:  1
         unique values:  2                    missing .:  0/19

            tabulation:  Freq.   Numeric  Label
                           11         0  male
                            8         1  female
```

From the output, we can see that the value of zero is associated with males and that females are assigned a value of one.

Once again, all we need is to include this variable in the regress command:

```
. regress gpa attendance study gender
```

Source	SS	df	MS	Number of obs	=	19
				F(3, 15)	=	59.22
Model	1514.7384	3	504.912801	Prob > F	=	0.0000
Residual	127.893177	15	8.5262118	R-squared	=	0.9221
				Adj R-squared	=	0.9066
Total	1642.63158	18	91.2573099	Root MSE	=	2.92

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.5047256	.1681634	3.00	0.009	.1462937	.8631574
study	.5605084	.1111414	5.04	0.000	.323616	.7974007
gender	4.286225	1.962058	2.18	0.045	.1041981	8.468253
_cons	21.18167	10.62247	1.99	0.065	-1.45958	43.82292

Although what we did works, there is a better way to include binary variables in Stata, and that is by including the prefix i before the variable name:

```
. regress gpa attendance study i.gender
```

Source	SS	df	MS	Number of obs	=	19
				F(3, 15)	=	59.22
Model	1514.7384	3	504.912801	Prob > F	=	0.0000
Residual	127.893177	15	8.5262118	R-squared	=	0.9221
				Adj R-squared	=	0.9066
Total	1642.63158	18	91.2573099	Root MSE	=	2.92

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.5047256	.1681634	3.00	0.009	.1462937	.8631574
study	.5605084	.1111414	5.04	0.000	.323616	.7974007
gender						
female	4.286225	1.962058	2.18	0.045	.1041981	8.468253
_cons	21.18167	10.62247	1.99	0.065	-1.45958	43.82292

This is the exact same command as before except that instead of specifying the name of the binary variable as it is, we included the i prefix with it.

Notice that the outputs of both commands are mostly identical except for how the name of the variable gender is included. In the first case, the output lists the variable like any other variable. The coefficient of the variable gender is 4.29. This means that when gender is equal to one, 4.29 is added to the prediction equation. In the second case, we see that the row of the variable gender contains no output. Instead, there is a new row below the variable that contains the name of one of the categories of the variable, which is female in this case. The output now makes more sense since Stata is telling is that when gender is female, 4.29 will be added. It is now clearer what the coefficient means. Females, on average, have a GPA that is 4.29 higher than males.

Why did Stata pick females instead of males? The reason is that females are assigned the value one. Stata always takes the lowest value as the referent, or base, value. Therefore, Stata is comparing females to the referent categories, which is males. This is the default behavior, and we can easily change it:

```
. regress gpa attendance study b1.gender
```

Source	SS	df	MS	Number of obs	=	19
				F(3, 15)	=	59.22
Model	1514.7384	3	504.912801	Prob > F	=	0.0000
Residual	127.893177	15	8.5262118	R-squared	=	0.9221
				Adj R-squared	=	0.9066
Total	1642.63158	18	91.2573099	Root MSE	=	2.92

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.5047256	.1681634	3.00	0.009	.1462937	.8631574
study	.5605084	.1111414	5.04	0.000	.323616	.7974007
gender						
male	-4.286225	1.962058	-2.18	0.045	-8.468253	-.1041981
_cons	25.4679	11.92431	2.14	0.050	.0518295	50.88396

Instead of using the i prefix, we used the b1 prefix. This is how to tell Stata to consider the value 1 as the referent, or base, category. So we want to compare other categories to this base category. We now see that the category included in the output is males, not females. We also see that while the magnitude of the coefficient is the same but that the sign is reversed. This is because we are now comparing males to females. If when we compared females to males we found that females had a GPA that was 4.29 points higher, when we compare males to females we should find that they have a GPA that is 4.29 points lower. This is exactly what is shown in the output. It doesn't matter what the base category is. What matters is the meaning of the coefficient, which remains the same.

We see that the model has an adjusted R-squared value that is greater than the model that did not include gender. We also see that the p-value of the variable gender is less than 0.05, which indicates that it is statistically significant. Therefore, the model seems to be a better fit.

It would be appropriate to calculated the predicted values and the residuals in order to check the model fit and the assumptions. However, since we started talking about binary variables, it would be better to continue directly to categorical variables with more than one category.

2.3 Categorical Variables with more than Two Categories

The variable major records the major of the students. To look at the values stored in this variable, we can run the following command:

```
. codebook major
```

```
major
```

```
                type:  numeric (float)
               label:  major

               range:  [0,3]                        units:  1
        unique values:  4                       missing .:  0/19

          tabulation:  Freq.   Numeric  Label
                           5         0  Business
                           5         1  Engineering
                           5         2  Biology
                           4         3  Philosophy
```

The output shows that there are four categories and that the category "Business" is assigned the lowest value, which is zero. Let us now include this variable in the regression model:

```
. regress gpa attendance study i.gender i.major
```

Source	SS	df	MS	Number of obs	=	19
				F(6, 12)	=	25.45
Model	1522.95561	6	253.825935	Prob > F	=	0.0000
Residual	119.67597	12	9.97299751	R-squared	=	0.9271
				Adj R-squared	=	0.8907
Total	1642.63158	18	91.2573099	Root MSE	=	3.158

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

attendance	.4822792	.1910708	2.52	0.027	.0659718	.8985866
study	.5331707	.1305492	4.08	0.002	.2487285	.8176129
gender						
female	4.785957	2.302437	2.08	0.060	-.2306215	9.802536
major						
Engineering	-.4114946	2.354197	-0.17	0.864	-5.540849	4.71786
Biology	-1.847053	2.219383	-0.83	0.422	-6.682672	2.988566
Philosophy	-.9144168	2.307835	-0.40	0.699	-5.942758	4.113925
_cons	24.21205	12.48119	1.94	0.076	-2.982127	51.40623

Notice that we use the i prefix with the variable major. We see that three majors are listed below the name of the variable major. The major business is missing because it is the referent category, since it was coded as zero. This means that the coefficients of each of the three other majors are comparing each major to the category business. The coefficient for engineering is -0.41 which means that, on average, engineering students have a GPA that is 0.41 lower than business students. Biology students on the other hand have a GPA that is 1.85 points lower. To change the base category to biology, we can run the following command:

```
. regress gpa attendance study i.gender b2.major
```

Source	SS	df	MS	Number of obs	=	19
				F(6, 12)	=	25.45
Model	1522.95561	6	253.825935	Prob > F	=	0.0000
Residual	119.67597	12	9.97299751	R-squared	=	0.9271
				Adj R-squared	=	0.8907
Total	1642.63158	18	91.2573099	Root MSE	=	3.158

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.4822792	.1910708	2.52	0.027	.0659718	.8985866
study	.5331707	.1305492	4.08	0.002	.2487285	.8176129

gender						
female	4.785957	2.302437	2.08	0.060	-.2306215	9.802536
major						
Business	1.847053	2.219383	0.83	0.422	-2.988566	6.682672
Engineering	1.435559	2.13894	0.67	0.515	-3.22479	6.095907
Philosophy	.9326363	2.319794	0.40	0.695	-4.121761	5.987034
_cons	22.365	11.92612	1.88	0.085	-3.619778	48.34978

Since biology was assigned the value 2, we use the prefix b2. We notice that the p-values of all categories in the variable major are greater than 0.05. This means that this variable is not significant. The adjusted R-squared value in the model that did not include the variable major is 0.91 while the adjusted R-squared value in Figure 31 is 0.89. Therefore, the addition of the variable major is resulting in a weaker mode. The p-values are not significant and the adjusted R-squared value has decreased. As such, it would make sense to eliminate this variable.

2.3.1 Quadratic Terms

There is another variable in our dataset that we have not used yet, and it is the variable English. We first produce a scatter plot using the command:

```
. scatter gpa english
```

We can see that students with a higher grade on English tend to have a higher GPA. Let us now fit a simple linear regression model:

```
. regress gpa english
```

Source	SS	df	MS	Number of obs	=	19
				F(1, 17)	=	142.85
Model	1467.93905	1	1467.93905	Prob > F	=	0.0000
Residual	174.692528	17	10.276031	R-squared	=	0.8937
				Adj R-squared	=	0.8874
Total	1642.63158	18	91.2573099	Root MSE	=	3.2056

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	.7954267	.0665517	11.95	0.000	.655015	.9358384
_cons	21.94569	4.62983	4.74	0.000	12.17761	31.71378

We see that the variable English is significant (p-value is less than 0.05), and the R-squared is quite high (0.89). However, the scatter plot (Figure 2.9) indicates that the relationship is not exactly linear.

```
. scatter gpa english
```

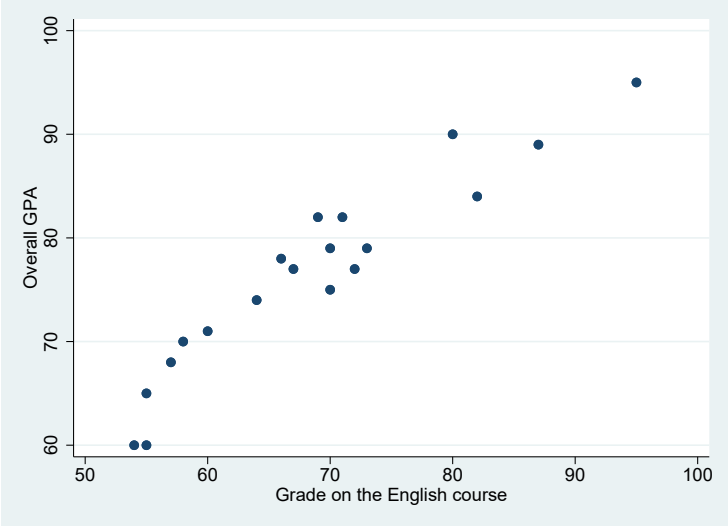

Figure 2.9: Scatter plot of GPA and English.

To investigate this further, we can ask Stata to overlay the scatter plot with

loess curve. A loess curve makes no assumption about the data. It lets the data speak for itself. This is done using the following command:

```
. twoway (scatter gpa english) (lowess gpa english)
```

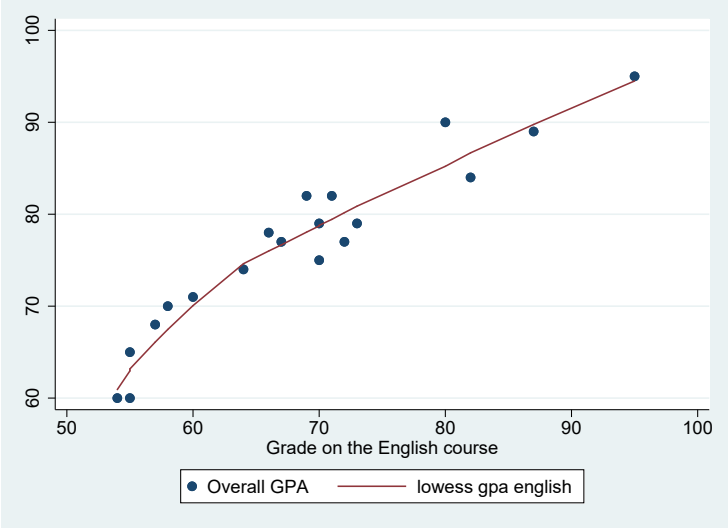

Figure 2.10: Checking for linearity: Scatter plot overlaid with a loess plot.

The output is shown in Figure 2.10. We can see that the curve bends along the way, thus indicating that a line might not be the best way to describe the relationship. Therefore, it would be prudent to fit a linear model that contains a quadratic term of the variable English. We can do this in two ways. First, we can create a new variable called English2 where this variable is the squared of English and then add this new variable to the model:

```
. gen english2 = english*english

. regress gpa english english2
```

Source	SS	df	MS	Number of obs	=	19
				F(2, 16)	=	94.05
Model	1513.86407	2	756.932035	Prob > F	=	0.0000
Residual	128.76751	16	8.04796937	R-squared	=	0.9216
				Adj R-squared	=	0.9118

Total	1642.63158	18	91.2573099	Root MSE	= 2.8369

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	2.352168	.6543368	3.59	0.002	.9650359	3.7393
english2	-.0107772	.0045115	-2.39	0.030	-.0203412	-.0012132
_cons	-32.82018	23.28926	-1.41	0.178	-82.1912	16.55085

We can see that the p-value of the squared term is less than 0.05, and is therefore significant. We can also see that the adjusted R-squared value of this model is higher than the one that does not include the quadratic term. This means that this model is a better model than the one that does not include a squared term. We can see that the coefficient of English is positive while the coefficient of the squared term is negative. This means that initially GPA increases with increasing values of English, but eventually this increase starts slowing down.

Another, and better, way to get the same output is to use the following command:

```
. regress gpa english c.english#c.english
```

Source	SS	df	MS	Number of obs	= 19
				F(2, 16)	= 94.05
Model	1513.86407	2	756.932035	Prob > F	= 0.0000
Residual	128.76751	16	8.04796937	R-squared	= 0.9216
				Adj R-squared	= 0.9118
Total	1642.63158	18	91.2573099	Root MSE	= 2.8369

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	2.352168	.6543368	3.59	0.002	.9650359	3.7393
c.english#c.english	-.0107772	.0045115	-2.39	0.030	-.0203412	-.0012132

```
       _cons |  -32.82018   23.28926    -1.41   0.178    -82.1912    16.55085
```

Notice that this command does not use any new variable. Instead, it includes the term c.english#c.english. This term is how we tell Stata to include a variable that is English x English. The prefix c tells Stata that the variable English is continuous. You are highly advised to use this terminology instead of creating a new variable. First, it minimizes the number of variables in the dataset. Second, it makes creating graphs that help us visualize the results easier (this will be covered later when we cover the margins command).

2.4 Checking Model Fit and Assumptions

2.4.1 Model Fit

Let us now see how we can check the model fit and the assumptions. We will do so for the following model:

```
. regress gpa attendance study i.gender
```

Source	SS	df	MS	Number of obs	=	19
				F(3, 15)	=	59.22
Model	1514.7384	3	504.912801	Prob > F	=	0.0000
Residual	127.893177	15	8.5262118	R-squared	=	0.9221
				Adj R-squared	=	0.9066
Total	1642.63158	18	91.2573099	Root MSE	=	2.92

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	.5047256	.1681634	3.00	0.009	.1462937	.8631574
study	.5605084	.1111414	5.04	0.000	.323616	.7974007

gender						
female	4.286225	1.962058	2.18	0.045	.1041981	8.468253
_cons	21.18167	10.62247	1.99	0.065	-1.45958	43.82292

We start with model fit. In order to test model fit, we can calculate the predicted values and compare them to the actual values. We had previously done this:

```
. drop gpa_predicted

. predict gpa_predicted
(option xb assumed; fitted values)

. twoway (scatter gpa_predicted gpa) (line gpa gpa)
```

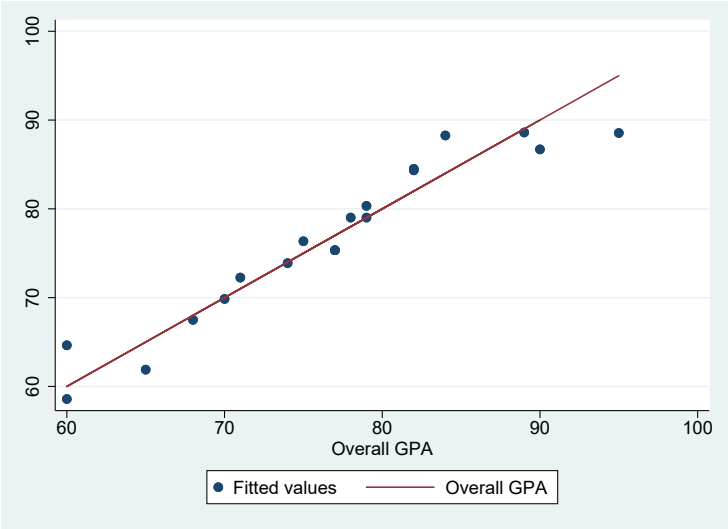

Figure 2.11: Checking model fit: Plotting the predicted values against the observed values (The line represents the 45 degree diagonal along which the points should fall if the model was well fit).

Note that because there was a previously defined variable called gpa_predicted, I dropped it before generating the newly predicted values. The result of run-

ning these commands is shown in Figure 2.11. As can be seen from the figure, it seems that the model is doing a good job.

2.4.2 Checking Model Assumptions

We now go on to check the assumptions. We have four assumptions to check: (1) normality of the residuals, (2) the randomness of the residuals, (3) homoscedasticity (the constant variance of the residuals), and (4) that multicolinearity does not exist. So far in this course we have discussed using graphical methods. Stata however offers us non-graphical ways to check these assumptions. For each of the assumptions, we will use both the graphical and the non-graphical tools.

Normality of the Residuals

With regards to the normality of the residuals, we first need to calculate the values of the residuals:

```
. drop residuals
. predict residuals, resid
```

Note that because there was a previously defined variable called residuals, I dropped it before generating the newly predicted residual values. Next we can use a quantile-normal plot to visualize whether the residuals are normal or not:

```
. qnorm residuals
```

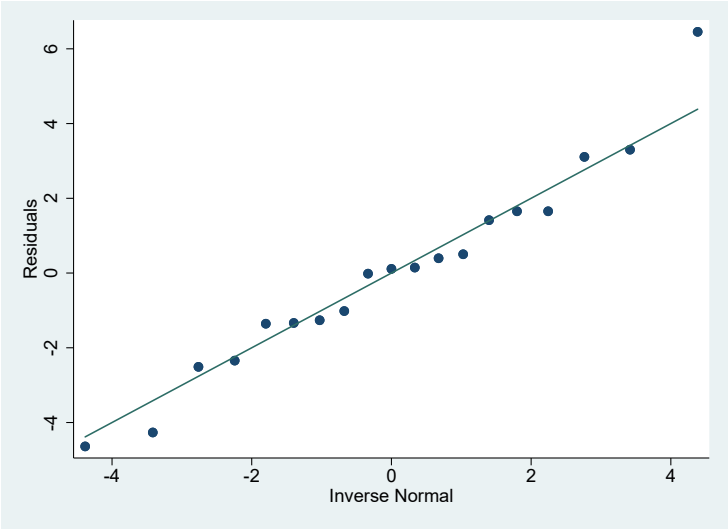

Figure 2.12: Checking the normality of the residuals: the quantile-normal plot.

The quantile-normal plot is shown in Figure 2.12. We can see that the residuals seem to follow a normal distribution. We next use a non-graphical test. There are several tests from which we can choose. Here, I will show you two of these tests. The first test is called a skewness-kurtosis test and it evaluates the null hypothesis that the distribution is in fact normal:

```
. sktest residuals

                    Skewness/Kurtosis tests for Normality
                                                      ------- joint ------
    Variable |        Obs  Pr(Skewness)  Pr(Kurtosis) adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
   residuals |         19      0.3837        0.3426        1.86          0.3952
```

As you can see, the result is not statistically significant (the p-value is greater than 0.05). This means that we do not reject the null hypothesis which is that the distribution is normal.

Another test is the Shapiro-Wilk test:

```
. swilk residuals

                    Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+-------------------------------------------------------
   residuals |         19    0.97134       0.654    -0.852    0.80289
```

This test all evaluates the null hypothesis that the distribution is normal. Once again, we see that the result is not statistically significant, so we cannot reject the null hypothesis. As such, both these tests, as well as the quantile-normal plot, indicate that the errors are normal.

Independence of the Residuals and Homoscedasticity

We list both the independence and the homoscedasticity assumptions together because they can be tested using the same tools. In order to check these assumptions, we can plot the residuals against the independent variables. Since this is a multiple linear regression model in which we have more than one independent variable, it would be better if we just plotted the residuals against the predicted values. We have already calculated both the residuals and the predicted values, so we can just use the twoway command in order to plot them. Stata however makes our life easy by providing a command that will do the work for us, without us having to calculate the predicted values or the residuals. This is accomplished using the following command:

```
. rvfplot, yline(0)
```

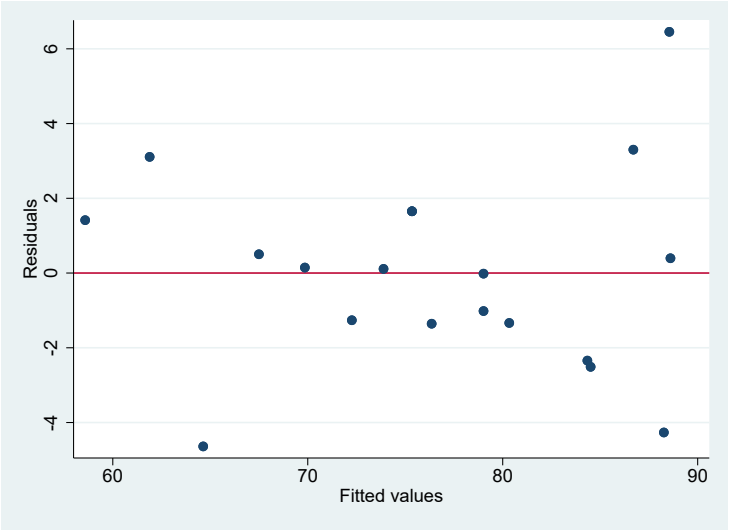

Figure 2.13: Checking for randomness of the residuals: using the rvfplot command.

We specify the yline(0) in order to tell Stata to draw a horizontal line at the y equal to zero point. This will make it easier for us to visualize the result. The result of executing this command is shown in Figure 2.13.

The figure shows that the errors seem to be random, since there is no clear pattern. However, we suspect that there might be a problem with the assumption of homoscedasticity. Remember, homoscedasticity means that the variance of the residuals is constant. In Figure 2.13 we see that the residuals near the edges tend to drift away more than the ones near the middle. We can use a non-graphical test in order to investigate this further. One such test is the Breusch-Pagan test, which tests the null hypothesis that that the residuals are homoscedastic:

```
. estat hettest

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

```
chi2(1)      =     1.64
Prob > chi2  =   0.2009
```

Since the result is statistically not significant, we cannot reject the null hypothesis that the residuals are homoscedastic.

What if the test had revealed that there was homoscedasticity? In fact, Figure 42 looks suspicious. There is reason to suspect that the homoscedasticity assumption might be violated. However, the Breusch-Pagan test failed to reject the null hypothesis. In such a case, it would be prudent to play it safe. If our analysis of the assumptions of normality, randomness, and homoscedasticity reveal that there is reason to suspect that any of these assumptions is violated, we can tell Stata to perform the regression while relaxing these assumptions. We tell Stata that we do not believe that these assumptions hold for our data. Stata will then relax the assumptions for us and this would produce more reliable results. Telling Stata to relax these assumptions is actually easy. It is accomplished by including the vce(robust) option with the regress command:

```
. regress gpa attendance study i.gender, vce(robust)

Linear regression                               Number of obs     =         19
                                                F(3, 15)          =      45.32
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9221
                                                Root MSE          =       2.92

------------------------------------------------------------------------------
             |               Robust
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  attendance |   .5047256   .2355552     2.14   0.049     .0026516    1.0068
       study |   .5605084   .1671962     3.35   0.004     .2041382    .9168785
             |
```

gender						
female	4.286225	1.772065	2.42	0.029	.5091587	8.063292
_cons	21.18167	13.22529	1.60	0.130	-7.007361	49.3707

Notice that what changes when we use the vce(robust) option are the values for the standard errors. This will result in a change in the p-values. The values of the coefficients of each variable remains unchanged. Given that all of the variables are still significant, this indicates that our initial results (before specifying the vce(robust) option) were acceptable. It seems that our data does not violate the assumptions.

Multicollinearity

Multicolinearity exists when the independent variables are correlated with each other, or when one of the independent variables is a linear combination of other variables. In order to test for this assumption, 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. The command to do this in Stata is the following:

```
. vif
```

Variable	VIF	1/VIF
attendance	2.75	0.363212
study	1.91	0.523833
1.gender	2.09	0.478193
Mean VIF	2.25	

Since the VIF for all variables is less than 10, we have no problem.

2.5 Diagnostics

Once we have fit a model and tested our assumptions, we should always look for outliers and for influential observations. In order to illustrate how this is done in Stata, you will need to use the file named linear2.dta. This file contains the exact same dataset as the file linear1.dta except for one extra observation. Now we run a regression model on this new dataset:

```
. use linear2, clear
(Data used in the course Linear Regression)

. regress gpa attendance study i.gender
```

Source	SS	df	MS	Number of obs	=	20
				F(3, 16)	=	12.90
Model	1366.20301	3	455.401002	Prob > F	=	0.0002
Residual	564.746995	16	35.2966872	R-squared	=	0.7075
				Adj R-squared	=	0.6527
Total	1930.95	19	101.628947	Root MSE	=	5.9411

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	-.5601817	.159506	-3.51	0.003	-.8983192	-.2220442
study	.8126303	.2144771	3.79	0.002	.3579592	1.267301
gender						
female	11.3828	3.444959	3.30	0.004	4.079812	18.68579
_cons	89.74928	9.340737	9.61	0.000	69.9478	109.5508

In order to make it easier for us to understand the change that took place once this new observation has been included, I have produced Table 2.1 which displays the values of the coefficients before and after the outlier has been

included. The table also includes the R-squared value for both models. It is

Table 2.1: Changes in the coefficients of the independent variables when the outlier is included.

	No outlier	**With outlier**
Attendance	0.51	-0.56
Study	0.56	0.81
Gender (female)	4.29	11.38
R-squared	0.92	0.71

clear just by looking at Table 2.1 that this single observation has resulted in a dramatic change in the results. Attendance now has a negative effect on the GPA. Females have a GPA that is on average 11.38 points higher than that of males. This example illustrates why we need to check for outliers, and to see whether these outliers are influential or not. How can we visualize the problem? In simple linear regression, when there is only one independent variable, we can easily produce a scatter plot of the dependent variable and the independent variable in order for us to see whether there is an outlier or not. However, in multiple linear regression, we cannot produce a scatter plot because we have more than one independent variable. Fortunately, Stata has a command that allows us to visualize the relationship between each independent variable and the dependent variable separately. These plots are called added-variable plots (avplots in Stata). The best way to understand the power of these graphs is to see them. The command to produce the added-variable plots is the following:

```
. avplots
```

The output of this command in our case is shown in Figure 2.14. Stata

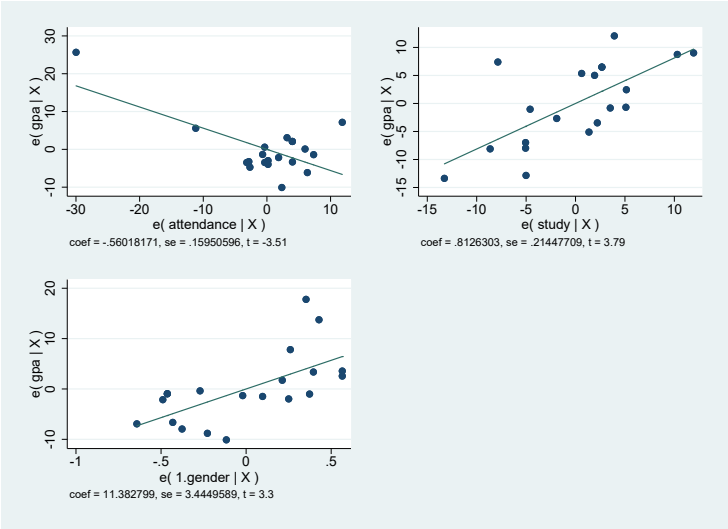

Figure 2.14: Checking for outliers: Added-variable plots.

produces three graphs because there are three independent variables. The x-axis label tells us which independent variable does each graph belong to. Looking at the first graph, which is for the independent variable attendance, we see that the relationship between GPA and attendance is represented by a line with a negative slope (the slope is -0.56 as indicates below the graph). Looking at the points, we see that one of them is particularly distant from the rest. This is our outlier. Looking at the other two graphs, we see that the points are more or less close to each other. Therefore, we conclude that the outlier is an outlier due to the value that it has for the variable attendance.

2.5.1 Influential Observations

Now that we have located our outlier, we need to investigate its level of influence (We already know that it is influential because we saw the change in the regression results once this observation was included. However, in

normal cases you will have a dataset that you haven't worked with before and won't be aware that a certain observation is influencing the results). As discussed in the theory section, there are three statistics that are useful when we want to calculate the influence of each observation, and they are DFBETAS, DFFITS, and Cook's D statistics. The following commands in Stata calculate these three statistics in order:

```
. dfbeta
                        _dfbeta_1: dfbeta(attendance)
                        _dfbeta_2: dfbeta(study)
                        _dfbeta_3: dfbeta(1.gender)

. predict dfits if e(sample), dfits

. predict cook, cooksd
```

The first command calculates the DFBETA statistic for each independent variable. This is why when you run it, you will notice that Stata creates three new variables (our regression model has three independent variables). The second command calculates the DFFITS statistic for each observation, and the third command calculates Cook's D for each observation as well.

Once we execute the above commands, we can see that Stata generates the variables that store the results. We now need to look at these results. There are two ways to do this, graphically and non-graphically. We will start with the non-graphical way. We can sort the observations using any of the three statistics. Let us start with Cook's D:

```
. sort cook

. list gpa attendance study gender major cook in -3/l, noobs
```

gpa	attend~e	study	gender	major	cook

60	65	19	male	Engineering	.0818629
90	88	30	female	Philosophy	.5864815
94	30	15	male	Engineering	14.09888

First we sort the observations, and then we tell Stata to display the last three results. Since the sort command sorts the observations in ascending order of the variable, observations with large values end up at the end of the list. I used the noobs option in order to tell Stata not to display the observation numbers We notice that the very last observation has a very large Cook's D. The observation belongs to a male student who has a GPA of 94, despite the fact that he has a very bad attendance record (30). We also see that the student is majoring in Engineering.

We next repeat the steps, but this time for the statistic DFFITS:

```
. sort dfits

. list gpa attendance study gender major dfits in -3/1, noobs
```

gpa	attend~e	study	gender	major	dfits
89	84	37	female	Business	.3491984
90	88	30	female	Philosophy	1.994035
94	30	15	male	Engineering	15.27957

We see from the output that the exact same student ends up at the bottom of the list. Although the above is useful, visualizing data would make things easier for us. To generate useful graphs, what we can do is to plot these statistics against each other. Observations with high values of either, or both, are influential. For example, execute the following command:

```
. scatter dfits cook, mlabel(attendance)
```

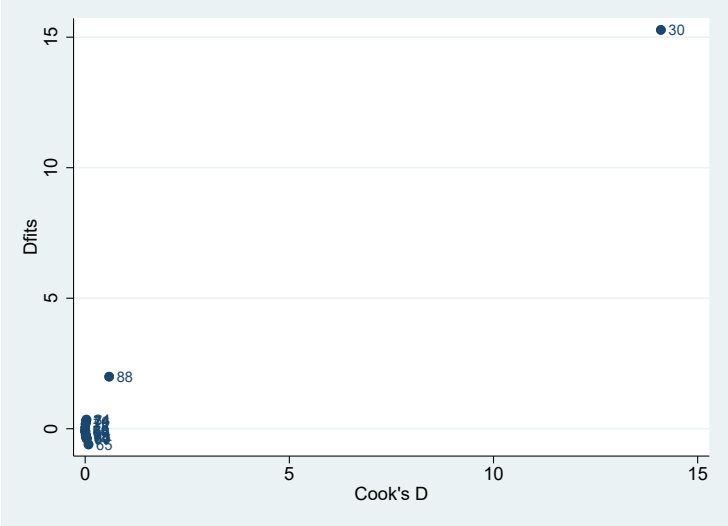

Figure 2.15: Visualizing influential points: A scatter plot of DFFITS and Cooks' D.

We ask Stata to label the points so that we can identify the influential observation. The output is shown in Figure 2.15. We see that the student who has an attendance grade of 30 is very influential.

2.6 Selection of Independent Variables

This final section will illustrate how to use the three selection methods: forward selection, backward elimination, and stepwise regression. As discussed in the theory section, you should never rely on an algorithm to do the selection for you. However, you can use these algorithms as a starting pointing to know more about the data.

The Stata command that allows users to choose any of these methods is the

stepwise command. The options that we specify in the command determine which type of algorithm is selected. This section will illustrate how to use the forward selection and the backward elimination methods in Stata.

2.6.1 Forward Selection

As you recall, the forward selection algorithm starts with no independent variable and then adds these variables one at a time. In order to use the forward selection algorithm in Stata, the pe(#)option needs to be specified, where # is the p-value for the significance level. In our case, this value is 0.05:

```
. stepwise, pe(0.05): regress gpa attendance study gender major english
                      begin with empty model
p = 0.0000 <  0.0500  adding  english
```

Source	SS	df	MS	Number of obs	=	20
				F(1, 18)	=	163.40
Model	1739.34716	1	1739.34716	Prob > F	=	0.0000
Residual	191.602837	18	10.6446021	R-squared	=	0.9008
				Adj R-squared	=	0.8953
Total	1930.95	19	101.628947	Root MSE	=	3.2626

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	.822192	.0643198	12.78	0.000	.6870611	.9573229
_cons	20.30766	4.529367	4.48	0.000	10.79181	29.8235

Notice that we included also the variables major and English in our model. The output tells us that Stata begins with an empty model. The first variable that is added is English, followed by attendance. These are the only two variables that are included in the final model. The value of R-squared is 0.93

which is pretty high.

2.6.2 Backward Elimination

As you recall, the backward elimination algorithm starts with all independent variable and then eliminates these variables one at a time. In order to use the backward elimination algorithm in Stata, the pr(#)option needs to be specified, where # is the p-value for the significance level. In our case, this value is 0.05:

```
. stepwise, pr(0.05): regress gpa attendance study gender major english
                      begin with full model
p = 0.5337 >= 0.0500  removing study
p = 0.3605 >= 0.0500  removing gender
p = 0.5699 >= 0.0500  removing attendance
p = 0.0642 >= 0.0500  removing major
```

Source	SS	df	MS	Number of obs	=	20
				F(1, 18)	=	163.40
Model	1739.34716	1	1739.34716	Prob > F	=	0.0000
Residual	191.602837	18	10.6446021	R-squared	=	0.9008
				Adj R-squared	=	0.8953
Total	1930.95	19	101.628947	Root MSE	=	3.2626

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	.822192	.0643198	12.78	0.000	.6870611	.9573229
_cons	20.30766	4.529367	4.48	0.000	10.79181	29.8235

Stata tells us that it starts with the full model. It then eliminates major, followed by gender, and finally followed by study. The resulting model is the same as the model that we got when we used forward selection.

2.7 Visualizing the Model

In a simple linear regression model, where there is only one independent variable, visualizing the result is very easy. All we have to do is to plot the line y = ax + b. In multiple linear regression, where there are more than one independent variable, things get more complicated. How can we plot the graph when there is only one x-axis but two independent variables? Luckily for us, Stata once again makes things easy for us.

2.7.1 Two Independent Variables

In order to illustrate this, let us take an example:

```
. regress gpa attendance i.gender
```

Source	SS	df	MS	Number of obs	=	20
				F(2, 17)	=	6.82
Model	859.494558	2	429.747279	Prob > F	=	0.0067
Residual	1071.45544	17	63.0267907	R-squared	=	0.4451
				Adj R-squared	=	0.3798
Total	1930.95	19	101.628947	Root MSE	=	7.9389

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	-.2598842	.1849687	-1.41	0.178	-.6501339	.1303656
gender						
female	15.73184	4.340363	3.62	0.002	6.574471	24.8892
_cons	89.73898	12.48179	7.19	0.000	63.40472	116.0733

Assume that we now want to plot a graph that shows how GPA changes with attendance. Since the value of GPA depends on gender, it is not enough to

just plot a two-dimensional graph with GPA on the y-axis and attendance on the x-axis. We need to take gender into consideration. What we can do is to tell Stata to use our model in order to predict the value of GPA for certain values of each independent variable. Let's say that we want to see how GPA varies for students as the attendance grade varies from 40 to 100. Since our model contains the variable gender, we want visualize how GPA is affected by gender. This can be done using the following command:

```
. margins, at(attendance=(40(1)100) gender=(0 1)) noatlegend

Adjusted predictions                            Number of obs     =         20
Model VCE    : OLS

Expression   : Linear prediction, predict()
```

	Delta-method					
	Margin	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at						
1	79.34362	5.383059	14.74	0.000	67.98635	90.70088
2	95.07545	7.783716	12.21	0.000	78.65325	111.4977
3	79.08373	5.216285	15.16	0.000	68.07833	90.08913
4	94.81557	7.611484	12.46	0.000	78.75674	110.8744
5	78.82385	5.050779	15.61	0.000	68.16764	89.48006
6	94.55569	7.439865	12.71	0.000	78.85894	110.2524
7	78.56396	4.886669	16.08	0.000	68.254	88.87393
8	94.2958	7.268899	12.97	0.000	78.95976	109.6318
9	78.30408	4.7241	16.58	0.000	68.3371	88.27106
10	94.03592	7.098637	13.25	0.000	79.0591	109.0127
11	78.0442	4.563237	17.10	0.000	68.41661	87.67178
12	93.77603	6.929128	13.53	0.000	79.15685	108.3952
13	77.78431	4.404268	17.66	0.000	68.49212	87.0765
14	93.51615	6.760429	13.83	0.000	79.25289	107.7794
15	77.52443	4.247404	18.25	0.000	68.56319	86.48567
16	93.25627	6.592604	14.15	0.000	79.34709	107.1654
17	77.26454	4.092889	18.88	0.000	68.6293	85.89978
18	92.99638	6.42572	14.47	0.000	79.4393	106.5535
19	77.00466	3.940997	19.54	0.000	68.68988	85.31944
20	92.7365	6.259852	14.81	0.000	79.52936	105.9436
21	76.74478	3.792045	20.24	0.000	68.74426	84.74529

22	92.47661	6.095084	15.17	0.000	79.61711	105.3361
23	76.48489	3.646393	20.98	0.000	68.79167	84.17811
24	92.21673	5.931508	15.55	0.000	79.70234	104.7311
25	76.22501	3.504452	21.75	0.000	68.83126	83.61875
26	91.95684	5.769223	15.94	0.000	79.78485	104.1288
27	75.96512	3.366692	22.56	0.000	68.86202	83.06822
28	91.69696	5.608344	16.35	0.000	79.86439	103.5295
29	75.70524	3.233647	23.41	0.000	68.88284	82.52764
30	91.43708	5.448994	16.78	0.000	79.9407	102.9334
31	75.44535	3.105922	24.29	0.000	68.89243	81.99828
32	91.17719	5.291311	17.23	0.000	80.0135	102.3409
33	75.18547	2.984202	25.19	0.000	68.88935	81.48159
34	90.91731	5.135449	17.70	0.000	80.08246	101.7522
35	74.92559	2.869251	26.11	0.000	68.872	80.97918
36	90.65742	4.981578	18.20	0.000	80.14721	101.1676
37	74.6657	2.761913	27.03	0.000	68.83857	80.49283
38	90.39754	4.82989	18.72	0.000	80.20736	100.5877
39	74.40582	2.66311	27.94	0.000	68.78715	80.02449
40	90.13765	4.680596	19.26	0.000	80.26246	100.0129
41	74.14593	2.573824	28.81	0.000	68.71564	79.57623
42	89.87777	4.533934	19.82	0.000	80.31201	99.44353
43	73.88605	2.495078	29.61	0.000	68.62189	79.1502
44	89.61789	4.390165	20.41	0.000	80.35545	98.88033
45	73.62616	2.427897	30.33	0.000	68.50375	78.74858
46	89.358	4.249585	21.03	0.000	80.39216	98.32384
47	73.36628	2.373264	30.91	0.000	68.35913	78.37343
48	89.09812	4.112521	21.67	0.000	80.42146	97.77478
49	73.1064	2.33206	31.35	0.000	68.18618	78.02661
50	88.83823	3.979335	22.32	0.000	80.44257	97.2339
51	72.84651	2.305007	31.60	0.000	67.98337	77.70965
52	88.57835	3.85043	23.00	0.000	80.45465	96.70205
53	72.58663	2.292604	31.66	0.000	67.74966	77.4236
54	88.31847	3.72625	23.70	0.000	80.45677	96.18017
55	72.32674	2.29509	31.51	0.000	67.48453	77.16896
56	88.05858	3.607284	24.41	0.000	80.44788	95.66929
57	72.06686	2.312416	31.17	0.000	67.18809	76.94563
58	87.7987	3.494064	25.13	0.000	80.42687	95.17053
59	71.80698	2.344254	30.63	0.000	66.86103	76.75292
60	87.53881	3.387166	25.84	0.000	80.39252	94.68511
61	71.54709	2.390024	29.94	0.000	66.50458	76.5896
62	87.27893	3.287207	26.55	0.000	80.34353	94.21433
63	71.28721	2.448944	29.11	0.000	66.12039	76.45403
64	87.01904	3.194839	27.24	0.000	80.27852	93.75957

65	71.02732	2.520092	28.18	0.000	65.71039	76.34425
66	86.75916	3.110738	27.89	0.000	80.19608	93.32224
67	70.76744	2.602466	27.19	0.000	65.27672	76.25816
68	86.49928	3.035591	28.50	0.000	80.09474	92.90381
69	70.50755	2.695036	26.16	0.000	64.82153	76.19358
70	86.23939	2.970077	29.04	0.000	79.97308	92.50571
71	70.24767	2.796791	25.12	0.000	64.34696	76.14838
72	85.97951	2.914846	29.50	0.000	79.82972	92.1293
73	69.98779	2.906765	24.08	0.000	63.85505	76.12053
74	85.71962	2.870493	29.86	0.000	79.66341	91.77583
75	69.7279	3.024063	23.06	0.000	63.34769	76.10812
76	85.45974	2.837526	30.12	0.000	79.47308	91.4464
77	69.46802	3.147865	22.07	0.000	62.8266	76.10943
78	85.19986	2.816346	30.25	0.000	79.25788	91.14183
79	69.20813	3.277435	21.12	0.000	62.29335	76.12292
80	84.93997	2.807221	30.26	0.000	79.01725	90.86269
81	68.94825	3.412116	20.21	0.000	61.74931	76.14718
82	84.68009	2.810266	30.13	0.000	78.75094	90.60923
83	68.68837	3.551326	19.34	0.000	61.19572	76.18101
84	84.4202	2.825443	29.88	0.000	78.45904	90.38137
85	68.42848	3.694553	18.52	0.000	60.63366	76.22331
86	84.16032	2.852558	29.50	0.000	78.14195	90.17869
87	68.1686	3.841348	17.75	0.000	60.06406	76.27313
88	83.90043	2.891276	29.02	0.000	77.80038	90.00049
89	67.90871	3.991318	17.01	0.000	59.48777	76.32966
90	83.64055	2.941137	28.44	0.000	77.43529	89.84581
91	67.64883	4.144117	16.32	0.000	58.90551	76.39215
92	83.38067	3.001588	27.78	0.000	77.04787	89.71346
93	67.38894	4.299444	15.67	0.000	58.31791	76.45998
94	83.12078	3.072002	27.06	0.000	76.63942	89.60214
95	67.12906	4.457035	15.06	0.000	57.72554	76.53258
96	82.8609	3.151712	26.29	0.000	76.21137	89.51043
97	66.86918	4.616657	14.48	0.000	57.12888	76.60947
98	82.60101	3.240033	25.49	0.000	75.76514	89.43688
99	66.60929	4.778108	13.94	0.000	56.52836	76.69022
100	82.34113	3.336279	24.68	0.000	75.30219	89.38006
101	66.34941	4.941208	13.43	0.000	55.92437	76.77445
102	82.08124	3.439787	23.86	0.000	74.82393	89.33856
103	66.08952	5.105799	12.94	0.000	55.31723	76.86182
104	81.82136	3.549921	23.05	0.000	74.33168	89.31104
105	65.82964	5.271742	12.49	0.000	54.70724	76.95204
106	81.56148	3.666083	22.25	0.000	73.82672	89.29624
107	65.56975	5.438912	12.06	0.000	54.09465	77.04486

108	81.30159	3.78772	21.46	0.000	73.3102	89.29298
109	65.30987	5.6072	11.65	0.000	53.47971	77.14003
110	81.04171	3.91432	20.70	0.000	72.78321	89.3002
111	65.04999	5.776508	11.26	0.000	52.86262	77.23735
112	80.78182	4.045419	19.97	0.000	72.24674	89.31691
113	64.7901	5.946749	10.90	0.000	52.24356	77.33665
114	80.52194	4.180592	19.26	0.000	71.70166	89.34222
115	64.53022	6.117846	10.55	0.000	51.62269	77.43774
116	80.26206	4.319457	18.58	0.000	71.1488	89.37531
117	64.27033	6.289728	10.22	0.000	51.00017	77.5405
118	80.00217	4.46167	17.93	0.000	70.58887	89.41547
119	64.01045	6.462332	9.91	0.000	50.37612	77.64478
120	79.74229	4.60692	17.31	0.000	70.02254	89.46204
121	63.75057	6.635603	9.61	0.000	49.75067	77.75046
122	79.4824	4.75493	16.72	0.000	69.45038	89.51443

The margins command uses the last model that we ran to calculate the values of the dependent variable given some values of the independent variables that we specify. In the above command, we are telling Stata to use the last run model to calculate (or predict) the values of the dependent variable (GPA in this case) when attendance ranges from 40 to 100 in one-unit increments, and when gender is zero and when it is one. In other words, the above command will calculate the following:

Attendance = 40, gender = 0: $GPA = 0.93(40) + 5.40(0) + 5.87$

Attendance = 40, gender = 1: $GPA = 0.93(40) + 5.40(1) + 5.87$

Attendance = 41, gender = 0: $GPA = 0.93(41) + 5.40(0) + 5.87$

Attendance = 41, gender = 1: $GPA = 0.93(41) + 5.40(1) + 5.87$

Attendance = 42, gender = 0: $GPA = 0.93(42) + 5.40(0) + 5.87$

Attendance = 42, gender = 1: $GPA = 0.93(42) + 5.40(1) + 5.87$

.

.

Attendance = 100, gender = 0: $GPA = 0.93(100) + 5.40(0) + 5.87$

Attendance = 100, gender = 1: $GPA = 0.93(100) + 5.40(1) + 5.87$

If you take a look at the first part of the output from the command, which I have suppressed by using the noatlegend option, you will see that this is what Stata is doing. Once Stata has calculated all of the values, we can ask it to produce the graph using the following command:

```
. marginsplot, noci

  Variables that uniquely identify margins: attendance gender
```

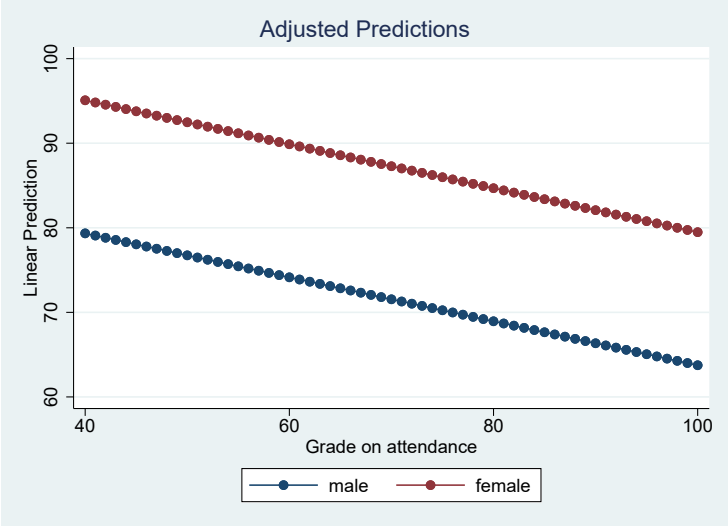

Figure 2.16: The command marginsplot: Visualizing the model with attendance and gender as independent variables.

The marginsplot command is used directly after the margins command. The noci option is how we tell Stata not to include the confidence intervals (you

can go ahead and see what happens when you run the command without including this option). The graph produced by this command in shown in Figure 2.16. The graph shows that as attendance increase, so does GPA. The graph also shows that female students have a higher GPA for each level of attendance.

2.7.2 Three Independent Variables

Let us now try to visualize the result for a more complex model:

```
. regress gpa attendance study i.gender
```

Source	SS	df	MS	Number of obs	=	20
				F(3, 16)	=	12.90
Model	1366.20301	3	455.401002	Prob > F	=	0.0002
Residual	564.746995	16	35.2966872	R-squared	=	0.7075
				Adj R-squared	=	0.6527
Total	1930.95	19	101.628947	Root MSE	=	5.9411

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attendance	-.5601817	.159506	-3.51	0.003	-.8983192	-.2220442
study	.8126303	.2144771	3.79	0.002	.3579592	1.267301
gender						
female	11.3828	3.444959	3.30	0.004	4.079812	18.68579
_cons	89.74928	9.340737	9.61	0.000	69.9478	109.5508

This model has three independent variables. After executing the command, we need to use the margins command in order to tell Stata to calculate the values of the dependent variable:

```
. margins, at(attendance=(40(1)100) study=(15 25 35) gender=(0 1)) noatlegend

Adjusted predictions                            Number of obs     =         20
```

Model VCE : OLS

Expression : Linear prediction, predict()

	Delta-method					
	Margin	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at						
1	79.53146	4.028714	19.74	0.000	70.99097	88.07196
2	90.91426	5.92757	15.34	0.000	78.34838	103.4801
3	87.65777	4.587292	19.11	0.000	77.93314	97.38239
4	99.04057	5.9182	16.73	0.000	86.49454	111.5866
5	95.78407	5.920814	16.18	0.000	83.2325	108.3356
6	107.1669	6.641851	16.14	0.000	93.08677	121.247
7	78.97128	3.903717	20.23	0.000	70.69577	87.24679
8	90.35408	5.816488	15.53	0.000	78.02368	102.6845
9	87.09759	4.43979	19.62	0.000	77.68565	96.50952
10	98.48038	5.777591	17.05	0.000	86.23244	110.7283
11	95.22389	5.777937	16.48	0.000	82.97521	107.4726
12	106.6067	6.490737	16.42	0.000	92.84694	120.3664
13	78.4111	3.781317	20.74	0.000	70.39507	86.42713
14	89.7939	5.707701	15.73	0.000	77.69411	101.8937
15	86.5374	4.293147	20.16	0.000	77.43634	95.63847
16	97.9202	5.637987	17.37	0.000	85.9682	109.8722
17	94.66371	5.635952	16.80	0.000	82.71602	106.6114
18	106.0465	6.340034	16.73	0.000	92.60623	119.4868
19	77.85092	3.661774	21.26	0.000	70.0883	85.61353
20	89.23372	5.601345	15.93	0.000	77.3594	101.108
21	85.97722	4.147453	20.73	0.000	77.18501	94.76943
22	97.36002	5.499467	17.70	0.000	85.70167	109.0184
23	94.10352	5.494928	17.13	0.000	82.4548	105.7523
24	105.4863	6.189773	17.04	0.000	92.36459	118.6081
25	77.29074	3.545379	21.80	0.000	69.77487	84.8066
26	88.67354	5.497559	16.13	0.000	77.01923	100.3278
27	85.41704	4.002813	21.34	0.000	76.93146	93.90262
28	96.79984	5.362112	18.05	0.000	85.43267	108.167
29	93.54334	5.354941	17.47	0.000	82.19137	104.8953
30	104.9261	6.039986	17.37	0.000	92.12194	117.7303
31	76.73056	3.432451	22.35	0.000	69.45409	84.00703
32	88.11335	5.396492	16.33	0.000	76.6733	99.55341
33	84.85686	3.859344	21.99	0.000	76.67541	93.0383
34	96.23966	5.226016	18.42	0.000	85.161	107.3183
35	92.98316	5.216076	17.83	0.000	81.92557	104.0407

36	104.366	5.890708	17.72	0.000	91.87822	116.8537
37	76.17037	3.323343	22.92	0.000	69.1252	83.21555
38	87.55317	5.298301	16.52	0.000	76.32128	98.78507
39	84.29668	3.717183	22.68	0.000	76.4166	92.17675
40	95.67948	5.09128	18.79	0.000	84.88644	106.4725
41	92.42298	5.078423	18.20	0.000	81.6572	103.1888
42	103.8058	5.741982	18.08	0.000	91.63332	115.9782
43	75.61019	3.218444	23.49	0.000	68.7874	82.43299
44	86.99299	5.203147	16.72	0.000	75.96281	98.02317
45	83.73649	3.576485	23.41	0.000	76.15469	91.3183
46	95.11929	4.958013	19.18	0.000	84.60877	105.6298
47	91.8628	4.942084	18.59	0.000	81.38605	102.3395
48	103.2456	5.593849	18.46	0.000	91.38717	115.104
49	75.05001	3.118179	24.07	0.000	68.43977	81.66025
50	86.43281	5.1112	16.91	0.000	75.59755	97.26807
51	83.17631	3.43743	24.20	0.000	75.88929	90.46334
52	94.55911	4.826339	19.59	0.000	84.32773	104.7905
53	91.30262	4.807172	18.99	0.000	81.11187	101.4934
54	102.6854	5.446358	18.85	0.000	91.13965	114.2312
55	74.48983	3.023009	24.64	0.000	68.08134	80.89832
56	85.87263	5.022637	17.10	0.000	75.22511	96.52014
57	82.61613	3.300226	25.03	0.000	75.61997	89.6123
58	93.99893	4.696391	20.02	0.000	84.04303	103.9548
59	90.74243	4.673809	19.42	0.000	80.8344	100.6505
60	102.1252	5.299564	19.27	0.000	90.89066	113.3598
61	73.92965	2.933429	25.20	0.000	67.71105	80.14824
62	85.31245	4.937639	17.28	0.000	74.84512	95.77977
63	82.05595	3.165113	25.93	0.000	75.34621	88.76569
64	93.43875	4.568315	20.45	0.000	83.75435	103.1231
65	90.18225	4.542131	19.85	0.000	80.55336	99.81114
66	101.5651	5.153525	19.71	0.000	90.64007	112.49
67	73.36947	2.849968	25.74	0.000	67.3278	79.41113
68	84.75226	4.856394	17.45	0.000	74.45717	95.04736
69	81.49577	3.032371	26.88	0.000	75.06743	87.92411
70	92.87857	4.442275	20.91	0.000	83.46136	102.2958
71	89.62207	4.412291	20.31	0.000	80.26843	98.97571
72	101.0049	5.008308	20.17	0.000	90.38773	111.622
73	72.80928	2.773177	26.25	0.000	66.93041	78.68816
74	84.19208	4.779093	17.62	0.000	74.06086	94.32331
75	80.93559	2.902325	27.89	0.000	74.78293	87.08824
76	92.31839	4.318449	21.38	0.000	83.16368	101.4731
77	89.06189	4.284455	20.79	0.000	79.97925	98.14453
78	100.4447	4.863986	20.65	0.000	90.1335	110.7559

79	72.2491	2.703626	26.72	0.000	66.51767	77.98053
80	83.6319	4.70593	17.77	0.000	73.65577	93.60803
81	80.3754	2.775354	28.96	0.000	74.49192	86.25889
82	91.7582	4.197032	21.86	0.000	82.86089	100.6555
83	88.50171	4.158807	21.28	0.000	79.68543	97.31798
84	99.88451	4.720641	21.16	0.000	89.87719	109.8918
85	71.68892	2.641885	27.14	0.000	66.08837	77.28947
86	83.07172	4.637102	17.91	0.000	73.2415	92.90194
87	79.81522	2.651901	30.10	0.000	74.19344	85.437
88	91.19802	4.078239	22.36	0.000	82.55254	99.8435
89	87.94153	4.035552	21.79	0.000	79.38654	96.49652
90	99.32432	4.578366	21.69	0.000	89.61862	109.03
91	71.12874	2.588514	27.48	0.000	65.64133	76.61614
92	82.51154	4.572804	18.04	0.000	72.81763	92.20545
93	79.25504	2.532479	31.30	0.000	73.88643	84.62366
94	90.63784	3.962307	22.88	0.000	82.23812	99.03755
95	87.38134	3.914917	22.32	0.000	79.08209	95.6806
96	98.76414	4.437262	22.26	0.000	89.35757	108.1707
97	70.56856	2.544039	27.74	0.000	65.17543	75.96168
98	81.95136	4.51323	18.16	0.000	72.38374	91.51897
99	78.69486	2.417685	32.55	0.000	73.5696	83.82012
100	90.07766	3.849494	23.40	0.000	81.9171	98.23822
101	86.82116	3.797151	22.86	0.000	78.77156	94.87076
102	98.20396	4.297446	22.85	0.000	89.09378	107.3141
103	70.00837	2.508935	27.90	0.000	64.68967	75.32708
104	81.39117	4.458568	18.26	0.000	71.93943	90.84292
105	78.13468	2.308212	33.85	0.000	73.24149	83.02787
106	89.51748	3.740082	23.93	0.000	81.58886	97.4461
107	86.26098	3.682528	23.42	0.000	78.45437	94.06759
108	97.64378	4.159048	23.48	0.000	88.82699	106.4606
109	69.44819	2.483597	27.96	0.000	64.1832	74.71318
110	80.83099	4.409002	18.33	0.000	71.48432	90.17766
111	77.5745	2.20485	35.18	0.000	72.90042	82.24857
112	88.95729	3.634378	24.48	0.000	81.25276	96.66183
113	85.7008	3.571353	24.00	0.000	78.12987	93.27173
114	97.0836	4.022213	24.14	0.000	88.55689	105.6103
115	68.88801	2.468327	27.91	0.000	63.65539	74.12063
116	80.27081	4.364706	18.39	0.000	71.01805	89.52357
117	77.01431	2.1085	36.53	0.000	72.54449	81.48413
118	88.39711	3.532717	25.02	0.000	80.90809	95.88614
119	85.14062	3.463956	24.58	0.000	77.79736	92.48388
120	96.52342	3.887107	24.83	0.000	88.28312	104.7637
121	68.32783	2.463313	27.74	0.000	63.10584	73.54982

122	79.71063	4.325841	18.43	0.000	70.54025	88.881
123	76.45413	2.020164	37.85	0.000	72.17158	80.73669
124	87.83693	3.435455	25.57	0.000	80.55409	95.11977
125	84.58044	3.3607	25.17	0.000	77.45607	91.7048
126	95.96323	3.753916	25.56	0.000	88.00529	103.9212
127	67.76765	2.468615	27.45	0.000	62.53442	73.00088
128	79.15045	4.292555	18.44	0.000	70.05064	88.25026
129	75.89395	1.940938	39.10	0.000	71.77935	80.00856
130	87.27675	3.342977	26.11	0.000	80.18995	94.36355
131	84.02025	3.261978	25.76	0.000	77.10517	90.93534
132	95.40305	3.622852	26.33	0.000	87.72295	103.0832
133	67.20747	2.48417	27.05	0.000	61.94126	72.47367
134	78.59026	4.264978	18.43	0.000	69.54892	87.63161
135	75.33377	1.871977	40.24	0.000	71.36535	79.30218
136	86.71657	3.255692	26.64	0.000	79.81481	93.61833
137	83.46007	3.168215	26.34	0.000	76.74376	90.17639
138	94.84287	3.494153	27.14	0.000	87.4356	102.2501
139	66.64728	2.509785	26.55	0.000	61.32678	71.96779
140	78.03008	4.243222	18.39	0.000	69.03485	87.02531
141	74.77359	1.814454	41.21	0.000	70.92712	78.62006
142	86.15639	3.174027	27.14	0.000	79.42775	92.88502
143	82.89989	3.079862	26.92	0.000	76.37087	89.42891
144	94.28269	3.368092	27.99	0.000	87.14265	101.4227
145	66.0871	2.545158	25.97	0.000	60.69161	71.4826
146	77.4699	4.227376	18.33	0.000	68.50826	86.43154
147	74.21341	1.769484	41.94	0.000	70.46227	77.96454
148	85.5962	3.098427	27.63	0.000	79.02783	92.16458
149	82.33971	2.997399	27.47	0.000	75.98551	88.69391
150	93.72251	3.244976	28.88	0.000	86.84347	100.6015
151	65.52692	2.589888	25.30	0.000	60.0366	71.01724
152	76.90972	4.217508	18.24	0.000	67.969	85.85044
153	73.65322	1.738041	42.38	0.000	69.96874	77.33771
154	85.03602	3.029345	28.07	0.000	78.6141	91.45795
155	81.77953	2.921324	27.99	0.000	75.5866	87.97246
156	93.16233	3.125152	29.81	0.000	86.5373	99.78735
157	64.96674	2.643501	24.58	0.000	59.36277	70.57071
158	76.34954	4.213659	18.12	0.000	67.41698	85.2821
159	73.09304	1.720867	42.47	0.000	69.44497	76.74112
160	84.47584	2.967238	28.47	0.000	78.18558	90.7661
161	81.21935	2.852149	28.48	0.000	75.17306	87.26563
162	92.60214	3.009015	30.77	0.000	86.22332	98.98097
163	64.40656	2.705468	23.81	0.000	58.67122	70.14189
164	75.78936	4.215845	17.98	0.000	66.85216	84.72655

165	72.53286	1.71839	42.21	0.000	68.89004	76.17569
166	83.91566	2.912551	28.81	0.000	77.74133	90.08999
167	80.65916	2.790387	28.91	0.000	74.74381	86.57452
168	92.04196	2.897006	31.77	0.000	85.90058	98.18334
169	63.84638	2.77523	23.01	0.000	57.96315	69.7296
170	75.22917	4.224058	17.81	0.000	66.27457	84.18378
171	71.97268	1.730674	41.59	0.000	68.30381	75.64154
172	83.35548	2.86571	29.09	0.000	77.28044	89.43051
173	80.09898	2.73654	29.27	0.000	74.29778	85.90019
174	91.48178	2.789625	32.79	0.000	85.56804	97.39552
175	63.28619	2.852215	22.19	0.000	57.23977	69.33262
176	74.66899	4.238262	17.62	0.000	65.68428	83.65371
177	71.4125	1.757408	40.64	0.000	67.68696	75.13804
178	82.7953	2.827104	29.29	0.000	76.8021	88.78849
179	79.5388	2.691082	29.56	0.000	73.83396	85.24364
180	90.9216	2.687424	33.83	0.000	85.22451	96.61868
181	62.72601	2.935854	21.37	0.000	56.50228	68.94975
182	74.10881	4.258398	17.40	0.000	65.08141	83.13621
183	70.85232	1.797948	39.41	0.000	67.04084	74.6638
184	82.23511	2.797074	29.40	0.000	76.30558	88.16465
185	78.97862	2.654445	29.75	0.000	73.35145	84.60579
186	90.36142	2.591019	34.87	0.000	84.8687	95.85413
187	62.16583	3.025597	20.55	0.000	55.75185	68.57981
188	73.54863	4.284381	17.17	0.000	64.46615	82.63111
189	70.29213	1.851388	37.97	0.000	66.36737	74.2169
190	81.67493	2.775899	29.42	0.000	75.79029	87.55957
191	78.41844	2.626998	29.85	0.000	72.84945	83.98742
192	89.80124	2.501078	35.91	0.000	84.49919	95.10328
193	61.60565	3.120916	19.74	0.000	54.9896	68.2217
194	72.98845	4.316106	16.91	0.000	63.83871	82.13818
195	69.73195	1.916648	36.38	0.000	65.66884	73.79506
196	81.11475	2.763782	29.35	0.000	75.2558	86.97371
197	77.85826	2.609031	29.84	0.000	72.32736	83.38915
198	89.24105	2.418324	36.90	0.000	84.11444	94.36767
199	61.04547	3.221317	18.95	0.000	54.21658	67.87435
200	72.42827	4.353448	16.64	0.000	63.19937	81.65716
201	69.17177	1.992568	34.71	0.000	64.94772	73.39583
202	80.55457	2.760842	29.18	0.000	74.70185	86.40729
203	77.29807	2.600741	29.72	0.000	71.78475	82.8114
204	88.68087	2.343517	37.84	0.000	83.71284	93.64891
205	60.48529	3.326339	18.18	0.000	53.43376	67.53681
206	71.86808	4.396264	16.35	0.000	62.54842	81.18775
207	68.61159	2.07798	33.02	0.000	64.20647	73.01671

208	79.99439	2.767109	28.91	0.000	74.12838	85.8604
209	76.73789	2.60222	29.49	0.000	71.22143	82.25435
210	88.12069	2.277441	38.69	0.000	83.29273	92.94865
211	59.9251	3.435559	17.44	0.000	52.64204	67.20816
212	71.3079	4.444394	16.04	0.000	61.88621	80.7296
213	68.05141	2.171763	31.33	0.000	63.44747	72.65534
214	79.43421	2.782521	28.55	0.000	73.53553	85.33289
215	76.17771	2.613451	29.15	0.000	70.63744	81.71798
216	87.56051	2.220875	39.43	0.000	82.85246	92.26855
217	59.36492	3.548589	16.73	0.000	51.84225	66.8876
218	70.74772	4.497669	15.73	0.000	61.21309	80.28235
219	67.49123	2.272883	29.69	0.000	62.67293	72.30952
220	78.87402	2.806926	28.10	0.000	72.92361	84.82444
221	75.61753	2.63431	28.70	0.000	70.03304	81.20202
222	87.00033	2.174563	40.01	0.000	82.39046	91.61019
223	58.80474	3.665077	16.04	0.000	51.03512	66.57436
224	70.18754	4.555909	15.41	0.000	60.52944	79.84563
225	66.93104	2.380403	28.12	0.000	61.88481	71.97727
226	78.31384	2.840093	27.57	0.000	72.29311	84.33457
227	75.05735	2.664571	28.17	0.000	69.40871	80.70598
228	86.44015	2.139168	40.41	0.000	81.90531	90.97498
229	58.24456	3.784703	15.39	0.000	50.22135	66.26777
230	69.62736	4.618924	15.07	0.000	59.83568	79.41904
231	66.37086	2.493497	26.62	0.000	61.08488	71.65684
232	77.75366	2.88172	26.98	0.000	71.64469	83.86263
233	74.49716	2.703917	27.55	0.000	68.76512	80.22921
234	85.87996	2.115241	40.60	0.000	81.39585	90.36407
235	57.68438	3.907179	14.76	0.000	49.40153	65.96723
236	69.06718	4.686524	14.74	0.000	59.13219	79.00216
237	65.81068	2.611441	25.20	0.000	60.27467	71.34669
238	77.19348	2.931446	26.33	0.000	70.97909	83.40787
239	73.93698	2.75196	26.87	0.000	68.10309	79.77088
240	85.31978	2.103172	40.57	0.000	80.86126	89.77831
241	57.1242	4.032245	14.17	0.000	48.57622	65.67217
242	68.50699	4.758512	14.40	0.000	58.4194	78.59459
243	65.2505	2.733606	23.87	0.000	59.45551	71.04548
244	76.6333	2.988867	25.64	0.000	70.29718	82.96941
245	73.3768	2.808253	26.13	0.000	67.42357	79.33003
246	84.7596	2.103165	40.30	0.000	80.30109	89.21811
247	56.56401	4.159669	13.60	0.000	47.74591	65.38212
248	67.94681	4.834692	14.05	0.000	57.69772	78.1959
249	64.69032	2.859452	22.62	0.000	58.62855	70.75208
250	76.07312	3.053549	24.91	0.000	69.59988	82.54635

251	72.81662	2.872311	25.35	0.000	66.72759	78.90565
252	84.19942	2.115221	39.81	0.000	79.71535	88.68349
253	56.00383	4.289238	13.06	0.000	46.91105	65.09661
254	67.38663	4.914869	13.71	0.000	56.96757	77.80569
255	64.13013	2.988513	21.46	0.000	57.79477	70.4655
256	75.51293	3.125041	24.16	0.000	68.88814	82.13772
257	72.25644	2.943627	24.55	0.000	66.01623	78.49665
258	83.63924	2.139136	39.10	0.000	79.10447	88.174
259	55.44365	4.420766	12.54	0.000	46.07204	64.81526
260	66.82645	4.998852	13.37	0.000	56.22936	77.42354
261	63.56995	3.120391	20.37	0.000	56.95502	70.18489
262	74.95275	3.202886	23.40	0.000	68.16294	81.74257
263	71.69626	3.021687	23.73	0.000	65.29057	78.10195
264	83.07905	2.174517	38.21	0.000	78.46928	87.68883
265	54.88347	4.554082	12.05	0.000	45.22924	64.53769
266	66.26627	5.086451	13.03	0.000	55.48347	77.04906
267	63.00977	3.254743	19.36	0.000	56.11002	69.90952
268	74.39257	3.286635	22.63	0.000	67.42522	81.35992
269	71.13607	3.105982	22.90	0.000	64.55169	77.72046
270	82.51887	2.220818	37.16	0.000	77.81095	87.2268
271	54.32329	4.689034	11.59	0.000	44.38298	64.2636
272	65.70609	5.177484	12.69	0.000	54.73031	76.68186
273	62.44959	3.391276	18.41	0.000	55.26041	69.63877
274	73.83239	3.375847	21.87	0.000	66.67591	80.98886
275	70.57589	3.196021	22.08	0.000	63.80063	77.35115
276	81.95869	2.277373	35.99	0.000	77.13088	86.78651
277	53.76311	4.825485	11.14	0.000	43.53353	63.99268
278	65.1459	5.271771	12.36	0.000	53.97025	76.32156
279	61.88941	3.529736	17.53	0.000	54.4067	69.37211
280	73.27221	3.4701	21.12	0.000	65.91592	80.62849
281	70.01571	3.29133	21.27	0.000	63.0384	76.99302
282	81.39851	2.343439	34.73	0.000	76.43064	86.36638
283	53.20292	4.963311	10.72	0.000	42.68118	63.72467
284	64.58572	5.369142	12.03	0.000	53.20365	75.96779
285	61.32923	3.669905	16.71	0.000	53.54937	69.10908
286	72.71203	3.568996	20.37	0.000	65.14609	80.27796
287	69.45553	3.391466	20.48	0.000	62.26594	76.64512
288	80.83833	2.418237	33.43	0.000	75.7119	85.96476
289	52.64274	5.1024	10.32	0.000	41.82614	63.45935
290	64.02554	5.469432	11.71	0.000	52.43086	75.62022
291	60.76904	3.811595	15.94	0.000	52.68882	68.84927
292	72.15184	3.67216	19.65	0.000	64.36721	79.93647
293	68.89535	3.496015	19.71	0.000	61.48413	76.30657

294	80.27815	2.500983	32.10	0.000	74.9763	85.57999
295	52.08256	5.242652	9.93	0.000	40.96864	63.19648
296	63.46536	5.572484	11.39	0.000	51.65222	75.2785
297	60.20886	3.954642	15.22	0.000	51.8254	68.59233
298	71.59166	3.779242	18.94	0.000	63.58003	79.6033
299	68.33517	3.604591	18.96	0.000	60.69377	75.97656
300	79.71796	2.590916	30.77	0.000	74.22547	85.21046
301	51.52238	5.383976	9.57	0.000	40.10886	62.9359
302	62.90518	5.678146	11.08	0.000	50.86804	74.94231
303	59.64868	4.098904	14.55	0.000	50.95939	68.33797
304	71.03148	3.889918	18.26	0.000	62.78522	79.27774
305	67.77498	3.716843	18.23	0.000	59.89563	75.65434
306	79.15778	2.687315	29.46	0.000	73.46093	84.85464
307	50.9622	5.52629	9.22	0.000	39.24699	62.67741
308	62.345	5.786277	10.77	0.000	50.07864	74.61135
309	59.0885	4.244257	13.92	0.000	50.09108	68.08592
310	70.4713	4.00389	17.60	0.000	61.98343	78.95917
311	67.2148	3.832447	17.54	0.000	59.09038	75.33923
312	78.5976	2.789509	28.18	0.000	72.68411	84.5111
313	50.40201	5.669519	8.89	0.000	38.38317	62.42086
314	61.78481	5.89674	10.48	0.000	49.28428	74.28534
315	58.52832	4.390593	13.33	0.000	49.22068	67.83596
316	69.91112	4.120886	16.97	0.000	61.17523	78.647
317	66.65462	3.951109	16.87	0.000	58.27864	75.0306
318	78.03742	2.896885	26.94	0.000	71.8963	84.17854
319	49.84183	5.813596	8.57	0.000	37.51756	62.16611
320	61.22463	6.009406	10.19	0.000	48.48526	73.964
321	57.96814	4.537817	12.77	0.000	48.34839	67.58788
322	69.35093	4.240654	16.35	0.000	60.36115	78.34072
323	66.09444	4.072561	16.23	0.000	57.461	74.72788
324	77.47724	3.008889	25.75	0.000	71.09868	83.8558
325	49.28165	5.958459	8.27	0.000	36.65028	61.91302
326	60.66445	6.124155	9.91	0.000	47.68182	73.64708
327	57.40795	4.685845	12.25	0.000	47.47441	67.3415
328	68.79075	4.362966	15.77	0.000	59.54168	78.03983
329	65.53426	4.196562	15.62	0.000	56.63794	74.43057
330	76.91706	3.125022	24.61	0.000	70.29231	83.54181
331	48.72147	6.104052	7.98	0.000	35.78146	61.66148
332	60.10427	6.24087	9.63	0.000	46.87421	73.33432
333	56.84777	4.834603	11.76	0.000	46.59887	67.09667
334	68.23057	4.487615	15.20	0.000	58.71725	77.74389
335	64.97408	4.322892	15.03	0.000	55.80995	74.1382
336	76.35687	3.244842	23.53	0.000	69.47812	83.23563

337	48.16129	6.250324	7.71	0.000	34.91119	61.41138
338	59.54409	6.359445	9.36	0.000	46.06267	73.02551
339	56.28759	4.984025	11.29	0.000	45.72193	66.85325
340	67.67039	4.614411	14.67	0.000	57.88828	77.4525
341	64.41389	4.451354	14.47	0.000	54.97745	73.85034
342	75.79669	3.367955	22.51	0.000	68.65695	82.93644
343	47.60111	6.397229	7.44	0.000	34.03959	61.16263
344	58.9839	6.479776	9.10	0.000	45.24739	72.72042
345	55.72741	5.134055	10.85	0.000	44.8437	66.61112
346	67.11021	4.743181	14.15	0.000	57.05511	77.1653
347	63.85371	4.581766	13.94	0.000	54.1408	73.56662
348	75.23651	3.494013	21.53	0.000	67.82953	82.64349
349	47.04092	6.544724	7.19	0.000	33.16673	60.91512
350	58.42372	6.601768	8.85	0.000	44.4286	72.41885
351	55.16723	5.28464	10.44	0.000	43.96429	66.37016
352	66.55003	4.87377	13.65	0.000	56.21809	76.88196
353	63.29353	4.713968	13.43	0.000	53.30036	73.2867
354	74.67633	3.622708	20.61	0.000	66.99653	82.35613
355	46.48074	6.69277	6.94	0.000	32.2927	60.66878
356	57.86354	6.725331	8.60	0.000	43.60648	72.12061
357	54.60705	5.435733	10.05	0.000	43.08381	66.13029
358	65.98984	5.006035	13.18	0.000	55.37752	76.60216
359	62.73335	4.847814	12.94	0.000	52.45644	73.01026
360	74.11615	3.75377	19.74	0.000	66.15851	82.07378
361	45.92056	6.841331	6.71	0.000	31.41759	60.42354
362	57.30336	6.850379	8.36	0.000	42.78121	71.82551
363	54.04686	5.587294	9.67	0.000	42.20233	65.8914
364	65.42966	5.139847	12.73	0.000	54.53367	76.32565
365	62.17317	4.98317	12.48	0.000	51.60932	72.73702
366	73.55597	3.886958	18.92	0.000	65.31598	81.79595

This command tells Stata to calculate the values of GPA when attendance ranges from 40 to 100 in one-unit increments, when study takes on the three values 15, 25, and 35, and when gender takes on the values zero and one. When we execute the command, Stata will produce a long list of calculated values. We next use the marginsplot command in order to tell Stata to produce a visual representation of the output:

```
. marginsplot, noci

  Variables that uniquely identify margins: attendance study gender
```

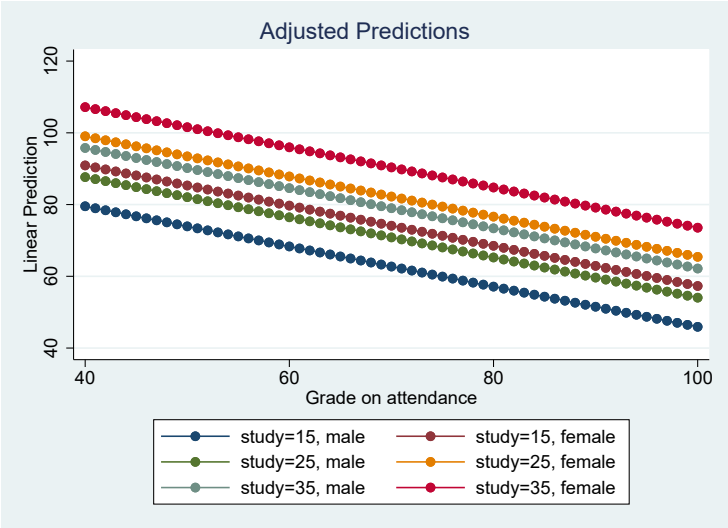

Figure 2.17: The command marginsplot: Visualizing the model with attendance, study, and gender as independent variables.

The result of running this command is shown in Figure 2.17. We can see that the more a student studies, the higher the GPA.

2.7.3 Quadratic Variable

We will use one last example in order to illustrate the importance of using interaction terms with quadratic variables. If you recall, when we were looking at the relationship between GPA and the grade on English, we found that including a squared term of English produced a better result. This was done two ways. In the first method, we created a new variable that we named English2, and we included this variable in the model:

```
. gen english2 = english*english

. regress gpa english english2
```

Source	SS	df	MS	Number of obs	=	20
				F(2, 17)	=	102.35
Model	1782.87955	2	891.439775	Prob > F	=	0.0000
Residual	148.07045	17	8.71002644	R-squared	=	0.9233
				Adj R-squared	=	0.9143
Total	1930.95	19	101.628947	Root MSE	=	2.9513

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	2.338304	.6806553	3.44	0.003	.9022468	3.774361
english2	-.0104834	.0046893	-2.24	0.039	-.020377	-.0005899
_cons	-33.07587	24.22765	-1.37	0.190	-84.19175	18.04001

In the second method, instead of creating a new variable, we included the interaction term c.english#c.english:

```
. regress gpa english c.english#c.english
```

Source	SS	df	MS	Number of obs	=	20
				F(2, 17)	=	102.35
Model	1782.87955	2	891.439775	Prob > F	=	0.0000
Residual	148.07045	17	8.71002644	R-squared	=	0.9233
				Adj R-squared	=	0.9143
Total	1930.95	19	101.628947	Root MSE	=	2.9513

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	2.338304	.6806553	3.44	0.003	.9022468	3.774361
c.english#c.english	-.0104834	.0046893	-2.24	0.039	-.020377	-.0005899
_cons	-33.07587	24.22765	-1.37	0.190	-84.19175	18.04001

At that point, I noted that the user should always use the second method,

even though both produced the same output. Now that we have covered the use of the margins command, it will be possible to see exactly why this is the case. First, run the following commands:

```
. regress gpa english english2
```

Source	SS	df	MS	Number of obs	=	20
				F(2, 17)	=	102.35
Model	1782.87955	2	891.439775	Prob > F	=	0.0000
Residual	148.07045	17	8.71002644	R-squared	=	0.9233
				Adj R-squared	=	0.9143
Total	1930.95	19	101.628947	Root MSE	=	2.9513

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	2.338304	.6806553	3.44	0.003	.9022468	3.774361
english2	-.0104834	.0046893	-2.24	0.039	-.020377	-.0005899
_cons	-33.07587	24.22765	-1.37	0.190	-84.19175	18.04001

```
. margins, at(english=(40(1)100)) noatlegend

Predictive margins                              Number of obs     =         20
Model VCE    : OLS

Expression   : Linear prediction, predict()
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at						
1	8.470031	20.09017	0.42	0.679	-33.91653	50.85659
2	10.80834	19.4099	0.56	0.585	-30.14297	51.75964
3	13.14664	18.72965	0.70	0.492	-26.36947	52.66275
4	15.48494	18.04943	0.86	0.403	-22.59604	53.56592
5	17.82325	17.36925	1.03	0.319	-18.82267	54.46917
6	20.16155	16.68911	1.21	0.244	-15.04939	55.37249
7	22.49986	16.00901	1.41	0.178	-11.2762	56.27591
8	24.83816	15.32896	1.62	0.124	-7.503113	57.17943
9	27.17646	14.64896	1.86	0.081	-3.730145	58.08307
10	29.51477	13.96903	2.11	0.050	.0426876	58.98685
11	31.85307	13.28917	2.40	0.028	3.815363	59.89078
12	34.19138	12.6094	2.71	0.015	7.587857	60.79489

13	36.52968	11.92974	3.06	0.007	11.36014	61.69922
14	38.86798	11.25019	3.45	0.003	15.13217	62.6038
15	41.20629	10.57078	3.90	0.001	18.9039	63.50868
16	43.54459	9.891541	4.40	0.000	22.67526	64.41392
17	45.8829	9.212514	4.98	0.000	26.44619	65.3196
18	48.2212	8.533746	5.65	0.000	30.21657	66.22583
19	50.5595	7.855306	6.44	0.000	33.98626	67.13275
20	52.89781	7.177284	7.37	0.000	37.75506	68.04055
21	55.23611	6.499814	8.50	0.000	41.5227	68.94952
22	57.57442	5.823086	9.89	0.000	45.28878	69.86005
23	59.91272	5.147393	11.64	0.000	49.05267	70.77277
24	62.25102	4.473206	13.92	0.000	52.81338	71.68866
25	64.58933	3.801325	16.99	0.000	56.56923	72.60942
26	66.92763	3.133234	21.36	0.000	60.31709	73.53818
27	69.26594	2.472008	28.02	0.000	64.05045	74.48142
28	71.60424	1.825123	39.23	0.000	67.75357	75.45491
29	73.94254	1.215692	60.82	0.000	71.37766	76.50743
30	76.28085	.7425121	102.73	0.000	74.71428	77.84741
31	78.61915	.7425121	105.88	0.000	77.05259	80.18572
32	80.95746	1.215692	66.59	0.000	78.39257	83.52234
33	83.29576	1.825123	45.64	0.000	79.44509	87.14643
34	85.63406	2.472008	34.64	0.000	80.41858	90.84955
35	87.97237	3.133234	28.08	0.000	81.36182	94.58291
36	90.31067	3.801325	23.76	0.000	82.29058	98.33077
37	92.64898	4.473206	20.71	0.000	83.21134	102.0866
38	94.98728	5.147393	18.45	0.000	84.12723	105.8473
39	97.32558	5.823086	16.71	0.000	85.03995	109.6112
40	99.66389	6.499814	15.33	0.000	85.95048	113.3773
41	102.0022	7.177284	14.21	0.000	86.85945	117.1449
42	104.3405	7.855306	13.28	0.000	87.76725	120.9137
43	106.6788	8.533746	12.50	0.000	88.67417	124.6834
44	109.0171	9.212514	11.83	0.000	89.5804	128.4538
45	111.3554	9.891541	11.26	0.000	90.48608	132.2247
46	113.6937	10.57078	10.76	0.000	91.39132	135.9961
47	116.032	11.25019	10.31	0.000	92.2962	139.7678
48	118.3703	11.92974	9.92	0.000	93.20078	143.5399
49	120.7086	12.6094	9.57	0.000	94.10511	147.3121
50	123.0469	13.28917	9.26	0.000	95.00922	151.0846
51	125.3852	13.96903	8.98	0.000	95.91315	154.8573
52	127.7235	14.64896	8.72	0.000	96.81693	158.6301
53	130.0618	15.32896	8.48	0.000	97.72057	162.4031
54	132.4001	16.00901	8.27	0.000	98.62409	166.1762
55	134.7384	16.68911	8.07	0.000	99.52751	169.9494

```
  56 |   137.0768   17.36925     7.89   0.000     100.4308    173.7227
  57 |   139.4151   18.04943     7.72   0.000     101.3341     177.496
  58 |   141.7534   18.72965     7.57   0.000     102.2373    181.2695
  59 |   144.0917    19.4099     7.42   0.000     103.1404     185.043
  60 |     146.43   20.09017     7.29   0.000     104.0434    188.8165
  61 |   148.7683   20.77047     7.16   0.000     104.9464    192.5901
------------------------------------------------------------------------------

. marginsplot, noci

  Variables that uniquely identify margins: english
```

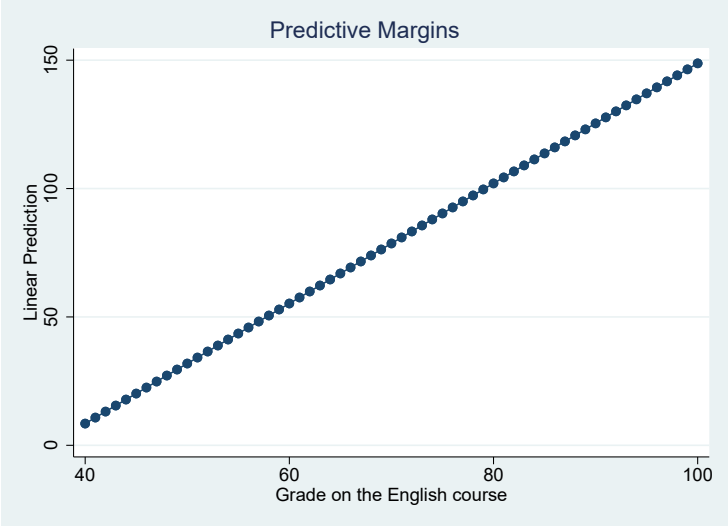

Figure 2.18: Including quadratic terms using generated variables.

The output is shown in Figure 2.18. The output is strange. Our linear model includes a squared term, yet the graph is a straight line. Why? The reason is simply because Stata does not know that the variable English2 is the square of the variable English. Stata includes the variable just like any other independent variable. When an independent variable that is included in the model is not included in the at() option of the margins command, Stata sets the value of the variable to the mean. Since the mean of the variable English2 is 4839.632, Stata is calculating the following:

English = 40: $GPA = 2.352168(40) - 0.0107772(4839.632) - 32.82018$

English = 41: $GPA = 2.352168(41) - 0.0107772(4839.632) - 32.82018$

English = 42: $GPA = 2.352168(42) - 0.0107772(4839.632) - 32.82018$

.

.

English = 100: $GPA = 2.352168(100) - 0.0107772(4839.632) - 32.82018$

Notice that the value of English2 is not changing. The variable is fixed at the mean. Stata does not know that when English is 40 English2 is 40*40, even though this is the formula that we used to generate the variable English2.

Now let us use the interaction term:

```
. regress gpa english c.english#c.english
```

Source	SS	df	MS	Number of obs	=	20
				F(2, 17)	=	102.35
Model	1782.87955	2	891.439775	Prob > F	=	0.0000
Residual	148.07045	17	8.71002644	R-squared	=	0.9233
				Adj R-squared	=	0.9143
Total	1930.95	19	101.628947	Root MSE	=	2.9513

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
english	2.338304	.6806553	3.44	0.003	.9022468	3.774361
c.english#c.english	-.0104834	.0046893	-2.24	0.039	-.020377	-.0005899
_cons	-33.07587	24.22765	-1.37	0.190	-84.19175	18.04001

```
. margins, at(english=(40(1)100)) noatlegend

Adjusted predictions                            Number of obs     =         20
Model VCE    : OLS
```

Expression : Linear prediction, predict()

	Delta-method					
	Margin	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at						
1	43.68281	4.635347	9.42	0.000	33.90308	53.46254
2	45.17196	4.340542	10.41	0.000	36.01422	54.3297
3	46.64014	4.055569	11.50	0.000	38.08363	55.19664
4	48.08735	3.780504	12.72	0.000	40.11118	56.06352
5	49.5136	3.515432	14.08	0.000	42.09668	56.93051
6	50.91888	3.26046	15.62	0.000	44.03991	57.79785
7	52.30319	3.015717	17.34	0.000	45.94058	58.66579
8	53.66653	2.781359	19.30	0.000	47.79838	59.53469
9	55.00891	2.557578	21.51	0.000	49.61289	60.40493
10	56.33032	2.344611	24.03	0.000	51.38363	61.27702
11	57.63077	2.14275	26.90	0.000	53.10996	62.15158
12	58.91025	1.952356	30.17	0.000	54.79113	63.02936
13	60.16876	1.773878	33.92	0.000	56.4262	63.91131
14	61.4063	1.607869	38.19	0.000	58.01399	64.79861
15	62.62288	1.455005	43.04	0.000	59.55309	65.69267
16	63.81849	1.316094	48.49	0.000	61.04177	66.59521
17	64.99313	1.192072	54.52	0.000	62.47808	67.50819
18	66.14681	1.083959	61.02	0.000	63.85986	68.43376
19	67.27952	.9927506	67.77	0.000	65.185	69.37404
20	68.39126	.9192269	74.40	0.000	66.45186	70.33066
21	69.48204	.8636856	80.45	0.000	67.65982	71.30426
22	70.55185	.8256618	85.45	0.000	68.80985	72.29384
23	71.60069	.8037658	89.08	0.000	69.90489	73.29649
24	72.62857	.7957474	91.27	0.000	70.94969	74.30745
25	73.63548	.7987889	92.18	0.000	71.95018	75.32077
26	74.62142	.8099008	92.14	0.000	72.91268	76.33016
27	75.58639	.8262682	91.48	0.000	73.84312	77.32967
28	76.5304	.845464	90.52	0.000	74.74663	78.31417
29	77.45344	.8655308	89.49	0.000	75.62733	79.27955
30	78.35552	.8849765	88.54	0.000	76.48838	80.22266
31	79.23663	.9027283	87.77	0.000	77.33204	81.14122
32	80.09677	.9180769	87.24	0.000	78.15979	82.03374
33	80.93594	.9306283	86.97	0.000	78.97249	82.89939
34	81.75415	.9402671	86.95	0.000	79.77036	83.73794
35	82.55139	.9471353	87.16	0.000	80.55311	84.54967
36	83.32766	.9516228	87.56	0.000	81.31991	85.33541

37	84.08297	.9543689	88.10	0.000	82.06943	86.09651
38	84.81731	.9562695	88.70	0.000	82.79976	86.83486
39	85.53068	.9584837	89.24	0.000	83.50846	87.55291
40	86.22309	.9624334	89.59	0.000	84.19253	88.25364
41	86.89453	.9697831	89.60	0.000	84.84846	88.94059
42	87.545	.9823895	89.11	0.000	85.47234	89.61766
43	88.1745	1.002213	87.98	0.000	86.06002	90.28899
44	88.78304	1.03119	86.10	0.000	86.60742	90.95866
45	89.37062	1.071089	83.44	0.000	87.11081	91.63042
46	89.93722	1.123377	80.06	0.000	87.5671	92.30734
47	90.48286	1.189125	76.09	0.000	87.97402	92.99169
48	91.00753	1.268984	71.72	0.000	88.33021	93.68485
49	91.51123	1.36322	67.13	0.000	88.63509	94.38738
50	91.99397	1.471794	62.50	0.000	88.88876	95.09919
51	92.45574	1.594459	57.99	0.000	89.09173	95.81976
52	92.89655	1.730837	53.67	0.000	89.2448	96.54829
53	93.31638	1.880499	49.62	0.000	89.34888	97.28389
54	93.71525	2.043003	45.87	0.000	89.40489	98.02561
55	94.09316	2.217929	42.42	0.000	89.41374	98.77258
56	94.45009	2.404892	39.27	0.000	89.37622	99.52397
57	94.78606	2.603546	36.41	0.000	89.29306	100.2791
58	95.10107	2.813591	33.80	0.000	89.16491	101.0372
59	95.3951	3.034765	31.43	0.000	88.99231	101.7979
60	95.66817	3.266843	29.28	0.000	88.77574	102.5606
61	95.92027	3.509629	27.33	0.000	88.5156	103.3249

```
. marginsplot, noci

  Variables that uniquely identify margins: english
```

The output is shown in Figure 2.19. This is the graph that we expect to see, where the relationship between GPA and English is nonlinear.

What happened here? By using the c.english#c.english notation, we have explicitly told Stata that the second term is the squared of English. Stata is now calculating the following:

English = 40: $GPA = 2.352168(40) - 0.0107772(40^2) - 32.82018$

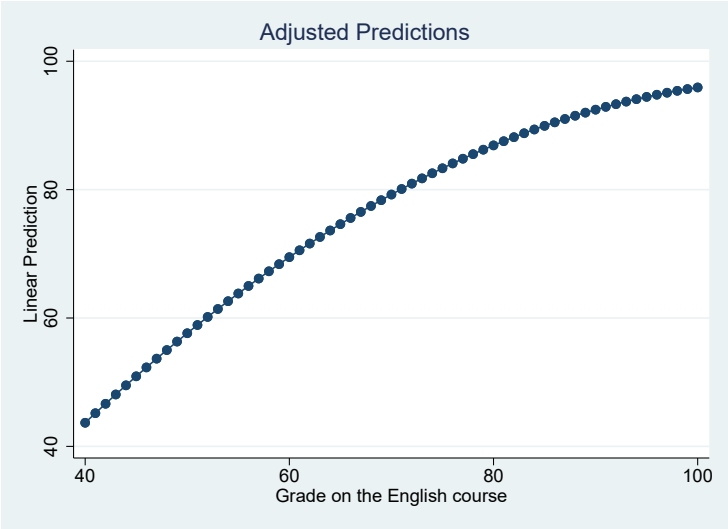

Figure 2.19: Including quadratic terms using interaction terms.

English = 41: $GPA = 2.352168(41) - 0.0107772(41^2) - 32.82018$

English = 42: $GPA = 2.352168(42) - 0.0107772(42^2) - 32.82018$

.

.

English = 100: $GPA = 2.352168(100) - 0.0107772(100^2 2) - 32.82018$

This is why you should always make it a habit to use the interaction terms whenever possible.

Chapter 3

References

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

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