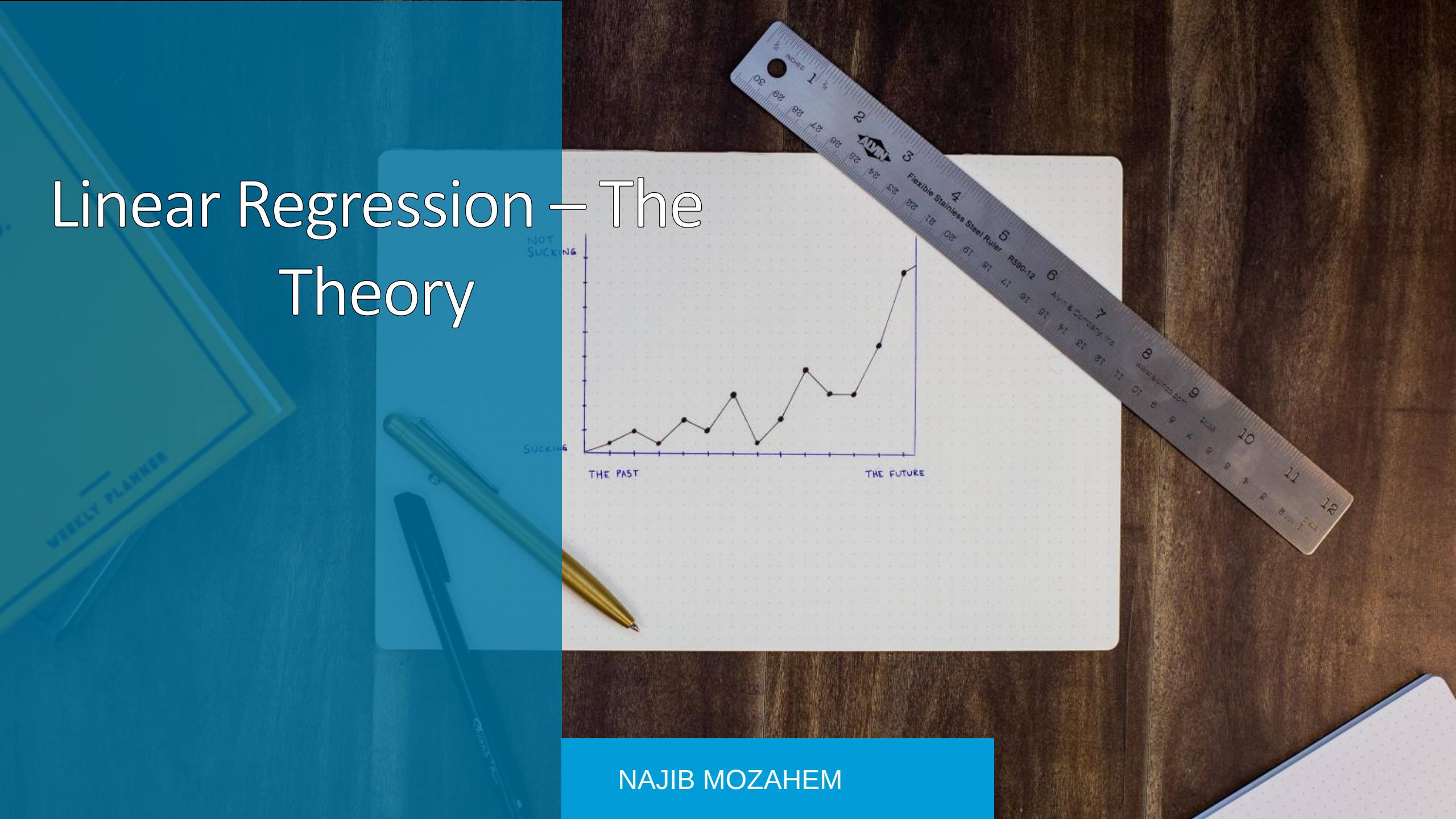
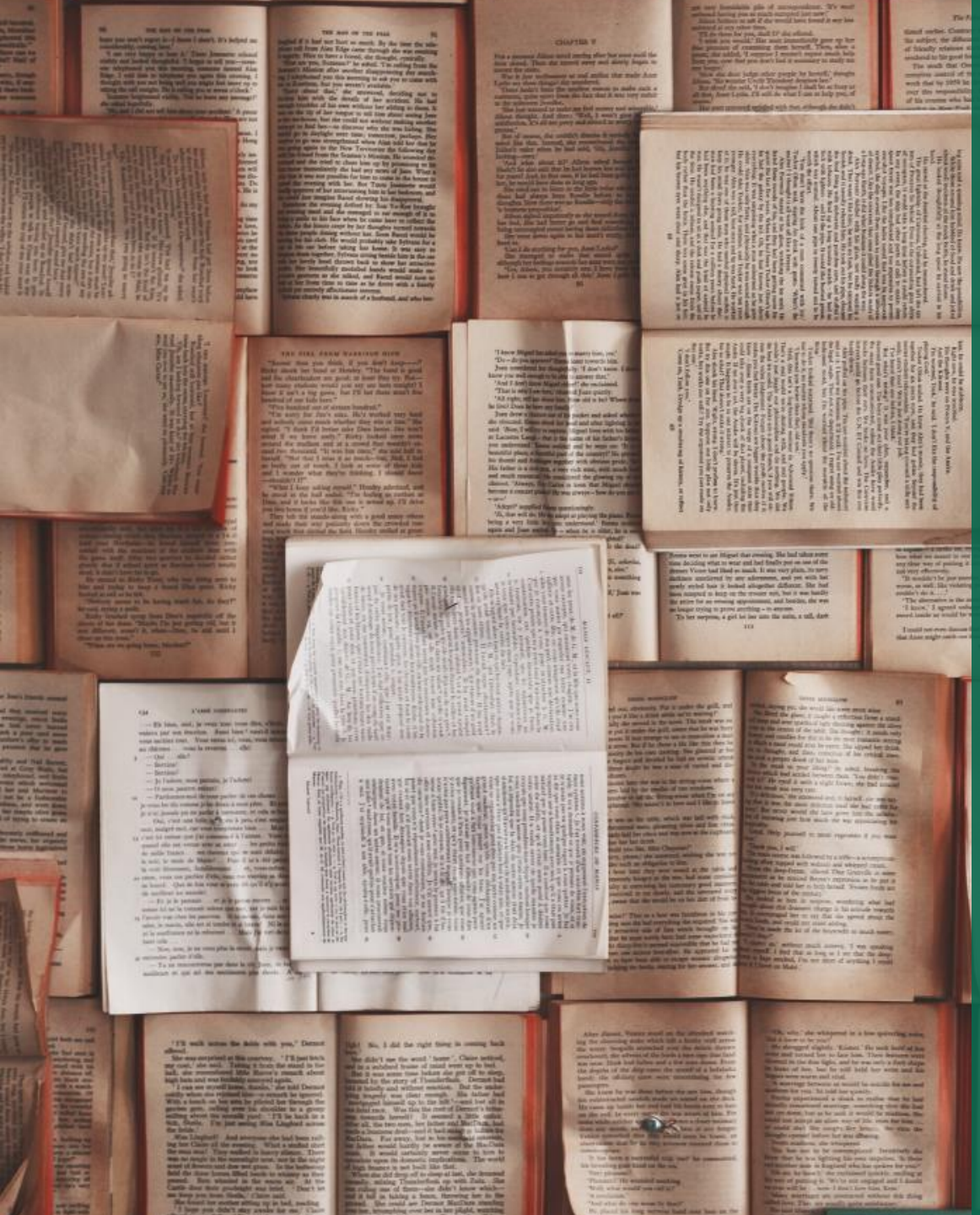
Linear Regression – The Theory



NAJIB MOZAHEM

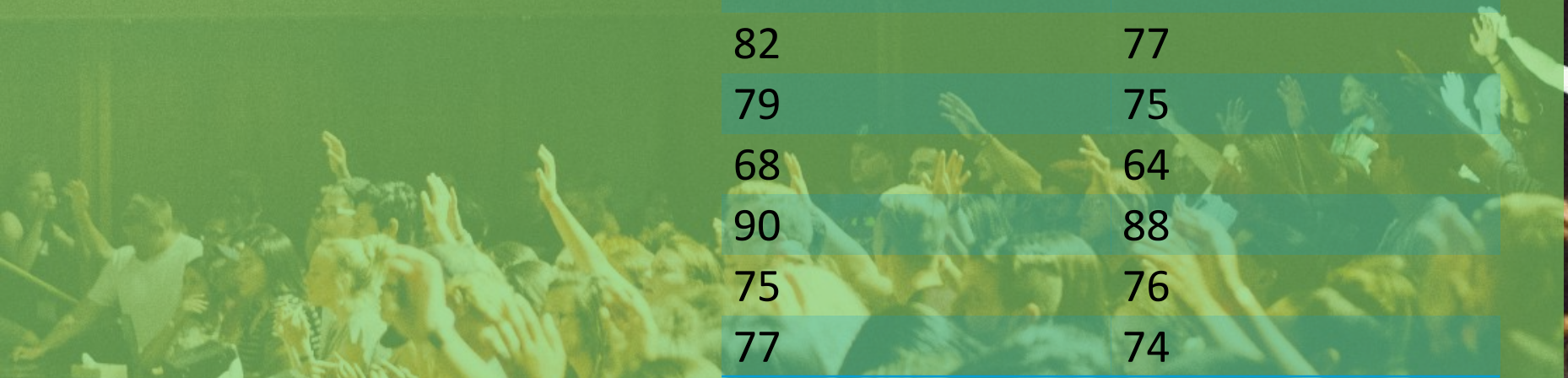


Part One: The Theory

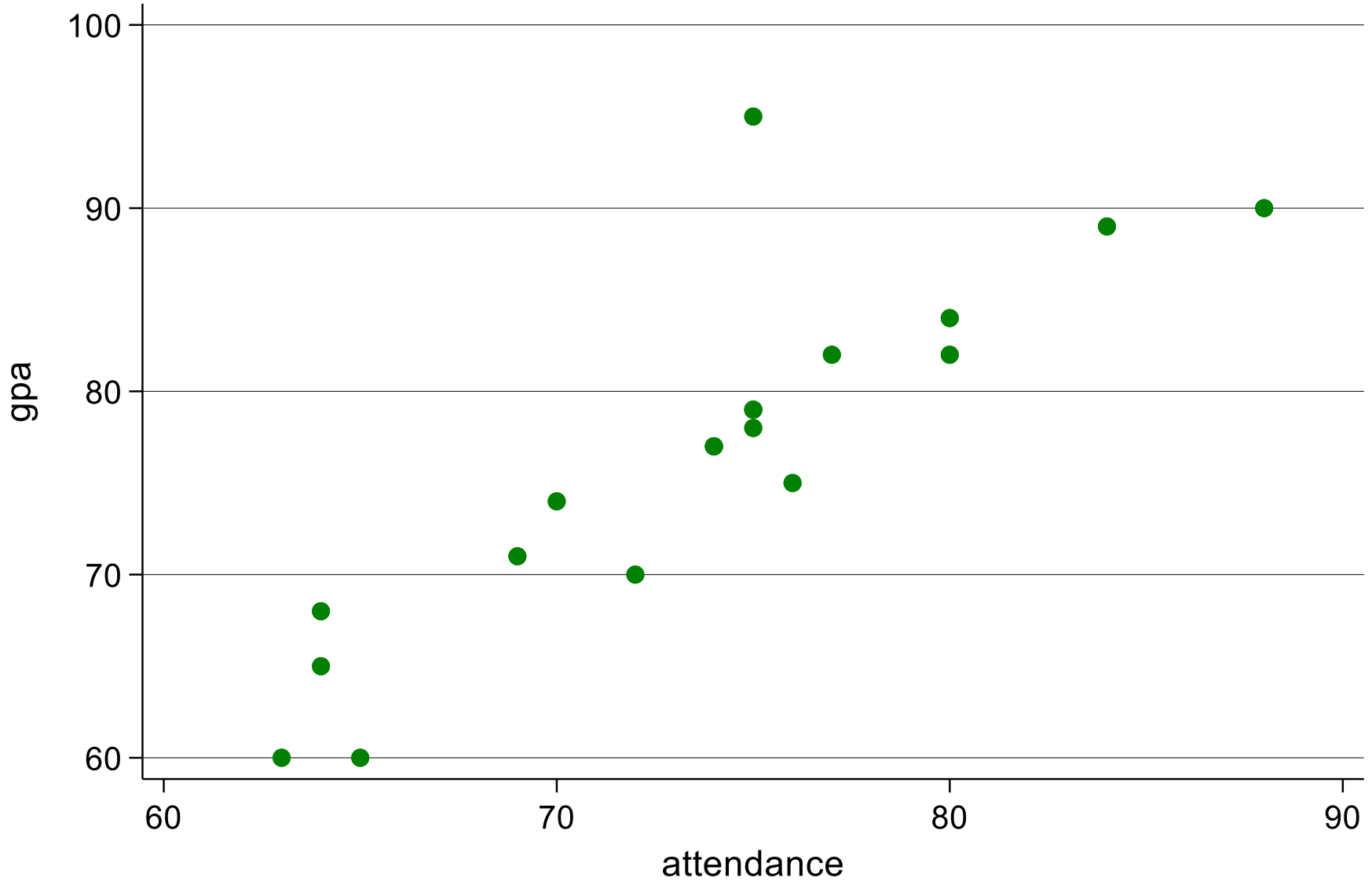
Simple Linear Regression

Grades and Attendance

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



Scatter Plot



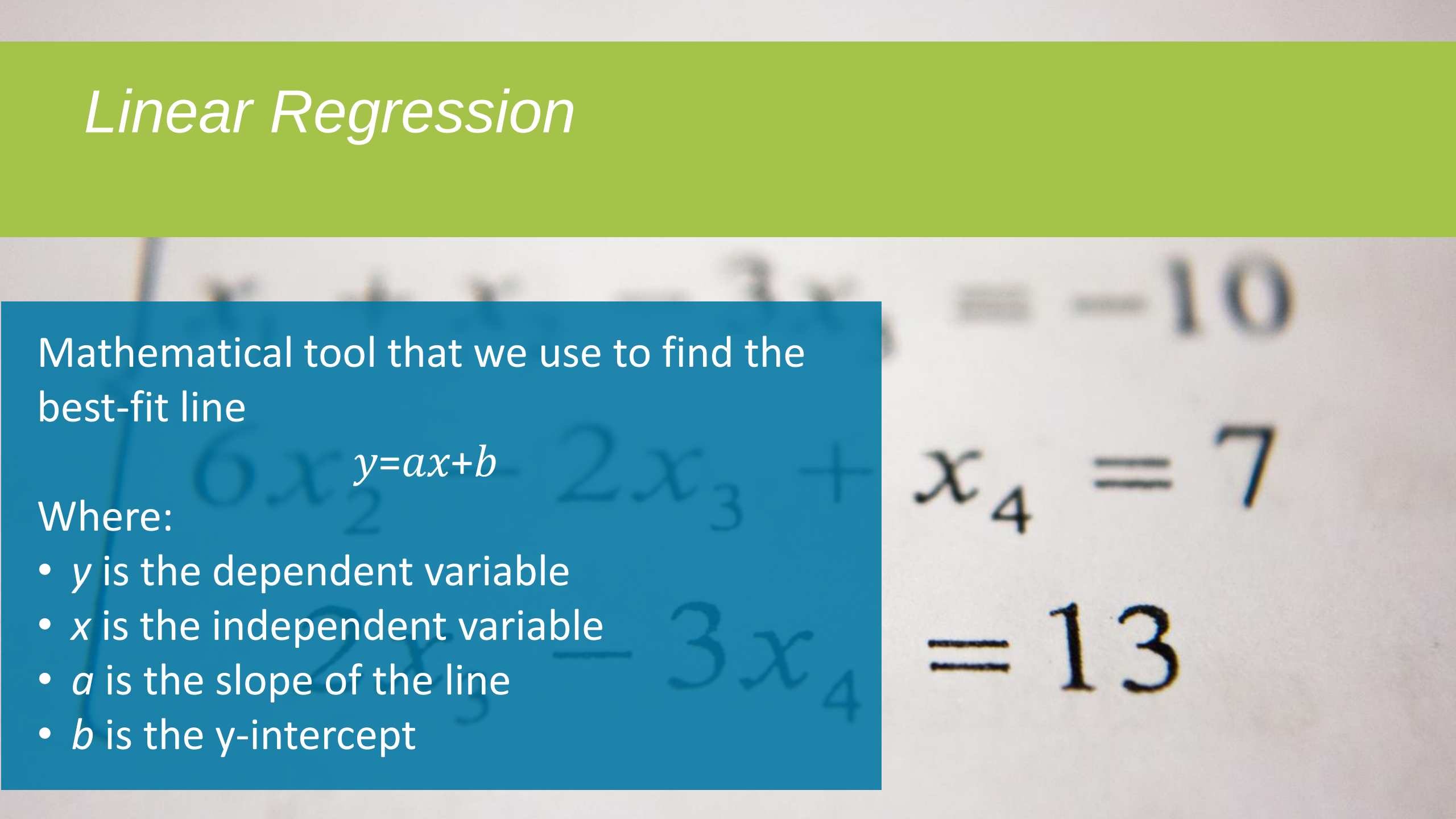
Linear Regression

Mathematical tool that we use to find the best-fit line

$$y=ax+b$$

Where:

- *y* is the dependent variable
- *x* is the independent variable
- *a* is the slope of the line
- *b* is the y-intercept



Linear Regression

- 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

The Slope

- 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
- For every one unit increase in *x*, the value of *y* increases by 3, which is the value of the slope
- The slope is the amount by which the dependent variable changes when the independent variable increases by one

The Slope

- Now let us look at a case where the slope is negative

$$y = -3x + 2$$

- 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
- For every one unit increase in *x*, the value of *y* decreases by 3, which is the value of the slope

Two Things to Look out for in the Slope

The sign 1

The magnitude 2

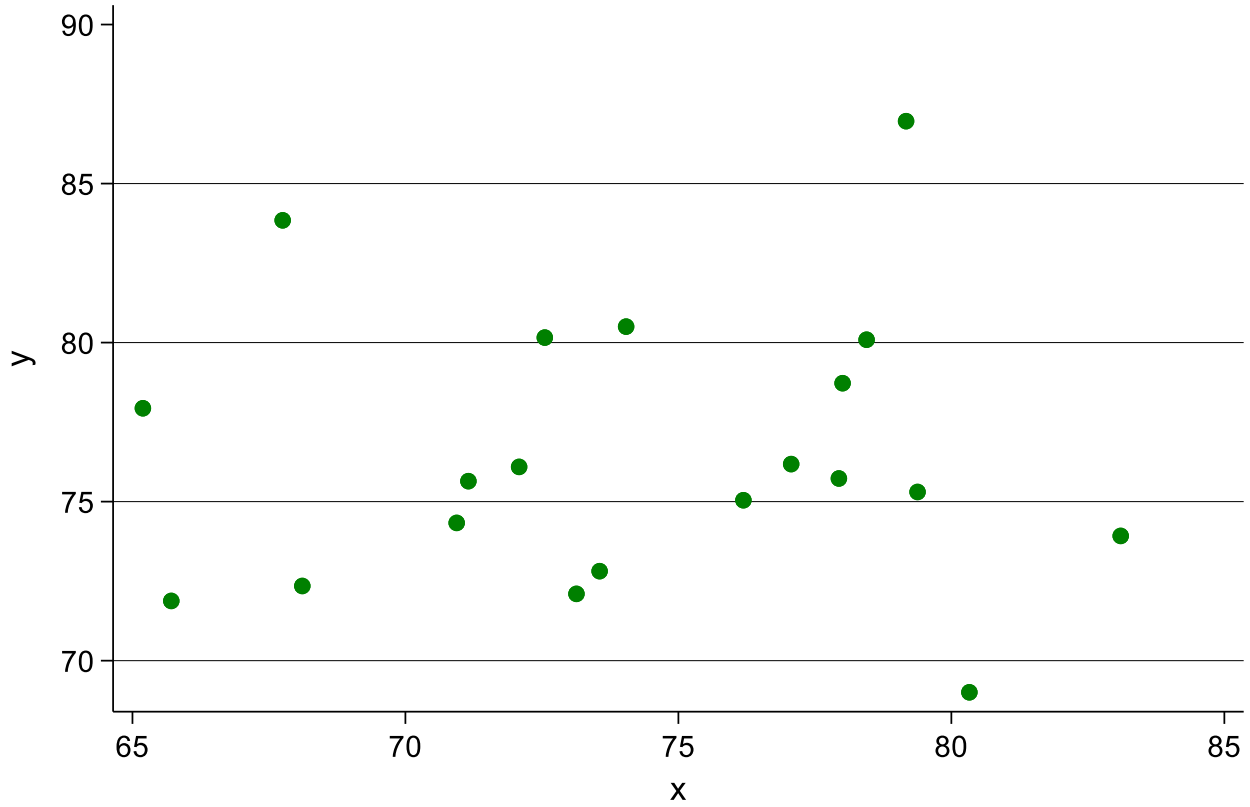
Linear Regression

- If we perform linear regression, the output will tell us that the following is the equation of the best-fit line:

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

R-Squared

- Linear regression gives us the best-fit line, but just because something is the best it doesn't mean that it is good



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

R-Squared

- R-squared calculates the proportion of the variation in the dependent variable that is explained by the line

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

- If the line explains most of the observed variation, the value of R-squared will be close to 1
- Otherwise, if the line fails to explain a large part of the variation, then the value of R-squared will be close to 0

The P-value

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

The P-value

- In order to test her, you conduct a small test
- You give her the first cup and she guesses correctly. Does this mean that she is right?
- No necessarily. Maybe it was a random lucky guess. After all, she has a probability of 0.5 to guess the correct answer

The P-value

- 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 doing that purely out of luck is 0.125

The P-value

- How many guesses must she make in order for her to prove that what we are observing is not due to luck, or randomness?
- 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

Linear Regression

- This brings us back to our line. The statistical software has told us that the best-fit line is:

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

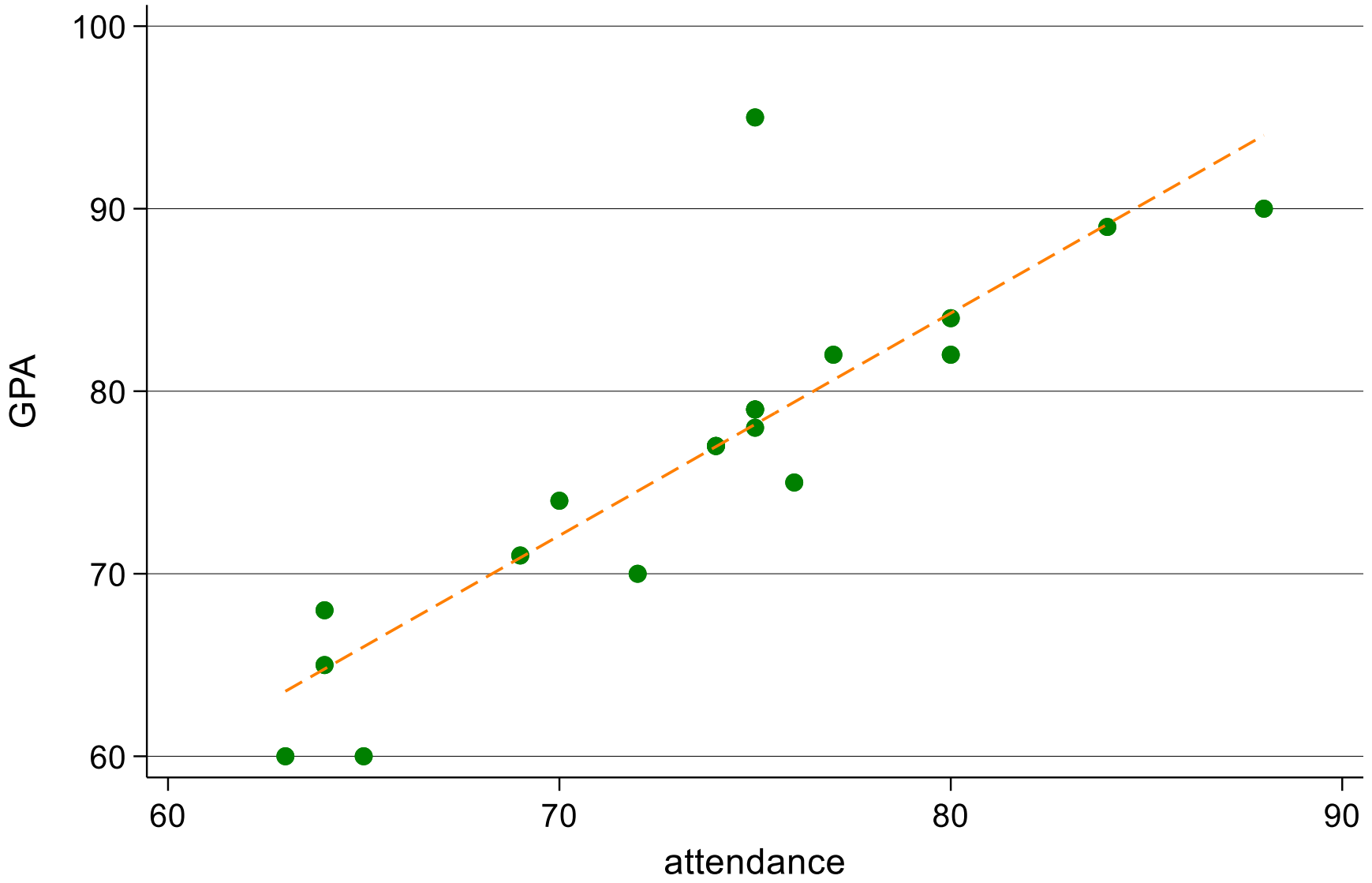
- The software has tells us that the value of R-squared is 0.75
- The statistical software tells us that the p-value of the slope is less than 0.05, therefore we reject the claim that the value that we obtained for the slope might have been due to randomness and nothing more. We therefore conclude that the value of the slope is significant

The Residuals

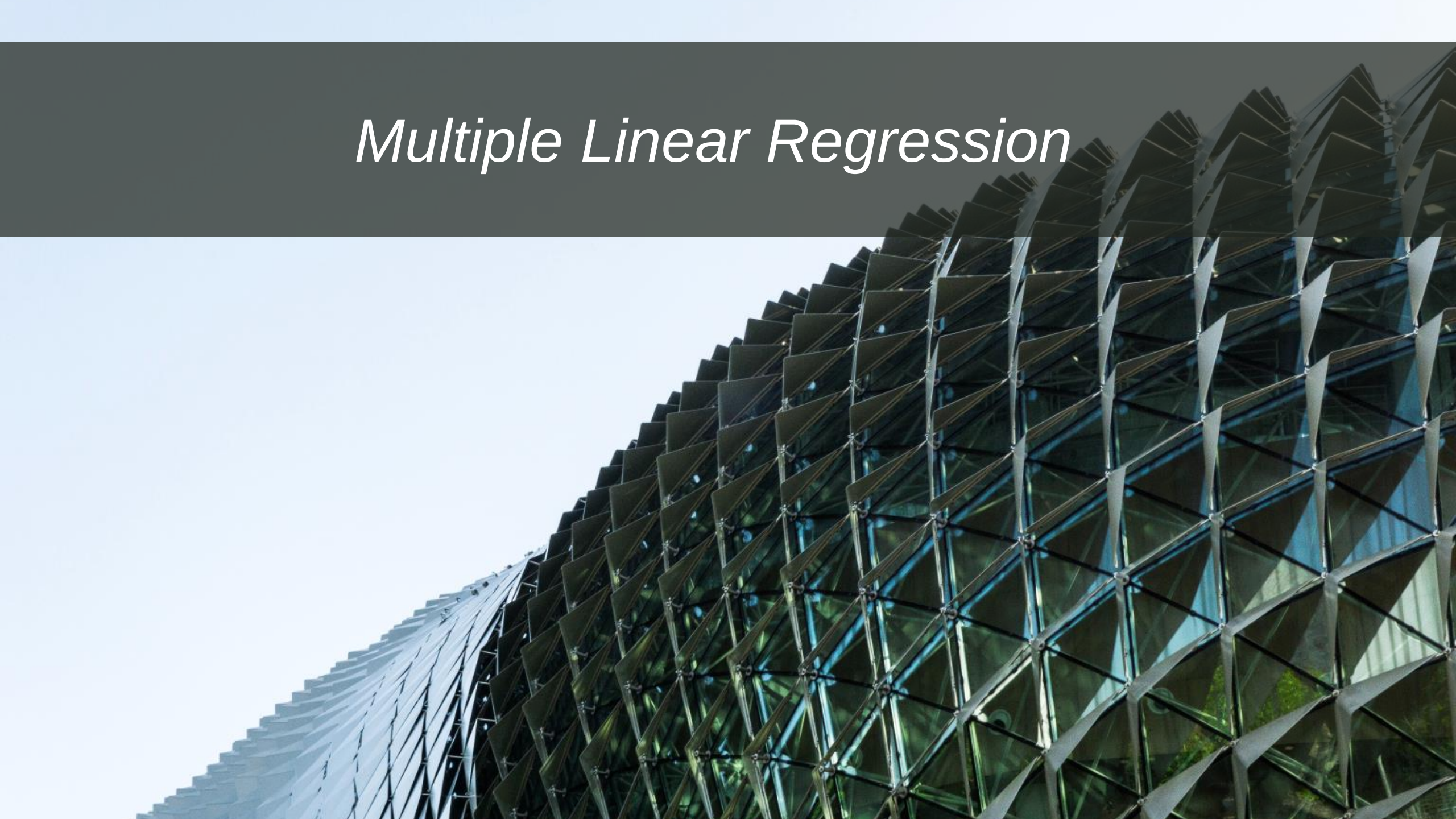
- Now that we have the equation for the best-fit line, we can calculate how accurate the line is
- This is done by predicting values
- What we want is for the predicted value to be as close to the observed value as possible

GPA	Attendance	Predicted GPA	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

Model Fit



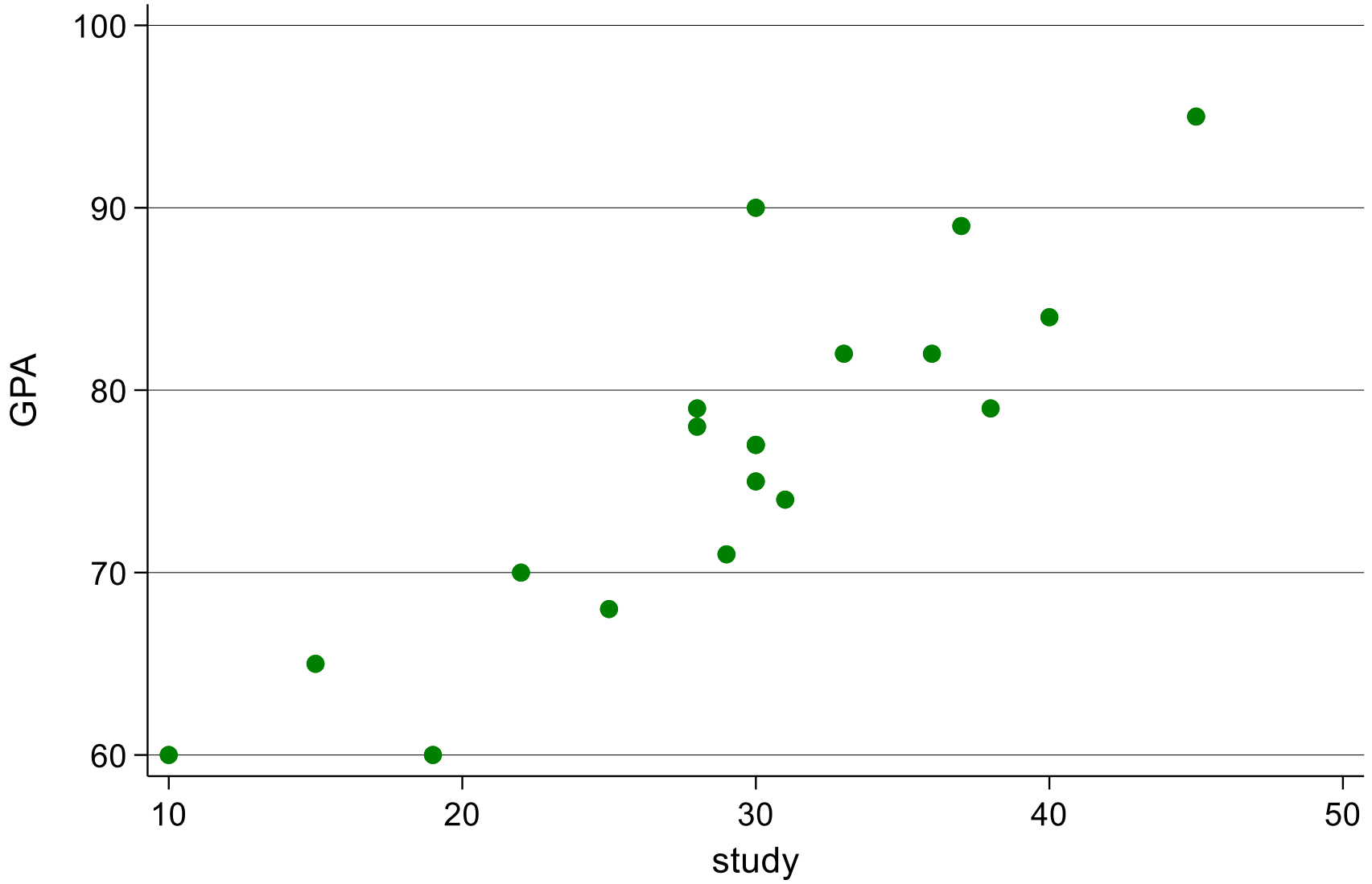
Multiple Linear Regression



Are grades only affected by attendance?

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

Scatter Plot



When the dependent variable is affected by more than one independent variable

$$y = a_1x_1 + a_2x_2 + \cdots + b$$

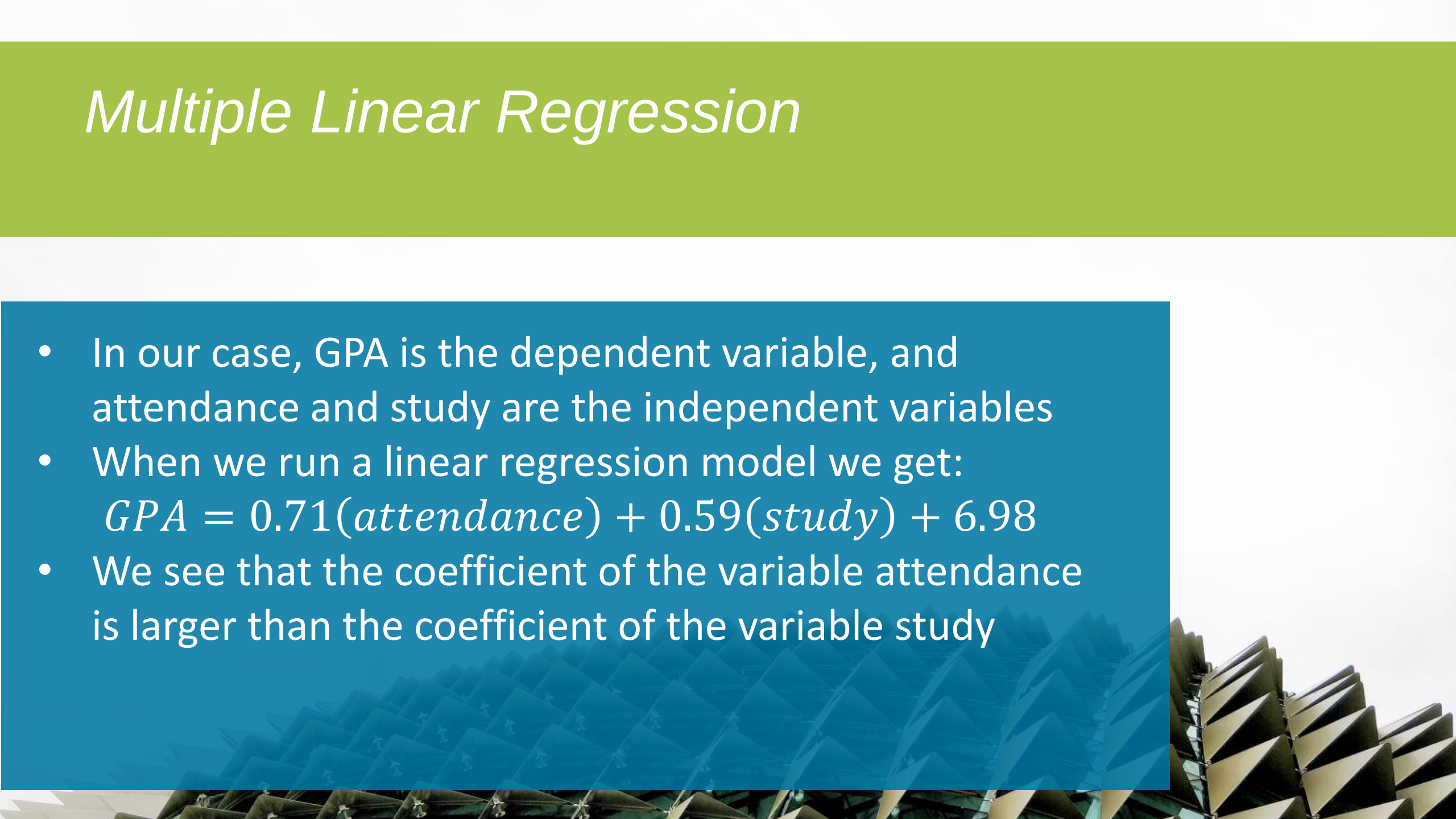
- In order to understand what each slope represents, consider the following case:

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

- 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

Multiple Linear Regression

- In our case, GPA is the dependent variable, and attendance and study are the independent variables
- When we run a linear regression model we get:
  $GPA = 0.71(attendance) + 0.59(study) + 6.98$
- We see that the coefficient of the variable attendance is larger than the coefficient of the variable study



R-Squared

- Which model is better? In the original model, we just had one independent variable. We now have two
- 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

The P-value

- The meaning of the p-values also remains the same
- The only difference is that there is one p-value associated with each independent variable
- In the output of the model that includes both attendance and study, both p-values are found to be less than 0.05

The Residuals

- Now that we have our equation, we can use it to predict the values of GPA
- 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

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

Binary Variables

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

Gender

- Both GPA and attendance levels are recorded as numbers
- Sometimes however, including variables that are not numeric in nature is necessary
- What if we wanted to investigate whether the variation in GPA could be explained by the gender of the students?
- Here, the variable gender is not numeric. It is categorical, in that it divides the observations into categories
- In such a case, we can create a binary variable to represent the two categories

Multiple Linear Regression with a Binary Variable

- In our case, GPA is the dependent variable, and attendance, study, and gender are the independent variables

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

- When we run a linear regression model we get:

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

Multiple Linear Regression with a Binary Variable

- What does it mean that the coefficient of gender is 4.29?
- 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 a female:

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

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

Multiple Linear Regression with a Binary Variable

- 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
- 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:

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

Categorical Variables

Major

- What if we had a categorical variable that divided the observations into more than two groups?
- Assume that the students included in our dataset were majoring in business, engineering, biology, or philosophy

	x_1	x_2	x_3
Business	0	0	0
Engineering	1	0	0
Biology	0	1	0
Philosophy	0	0	1

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

Multiple Linear Regression with a Categorical Variable

- 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 studying 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

Multiple Linear Regression with a Categorical Variable

- 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:

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

Engineering:

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

Biology:

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

Philosophy:

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

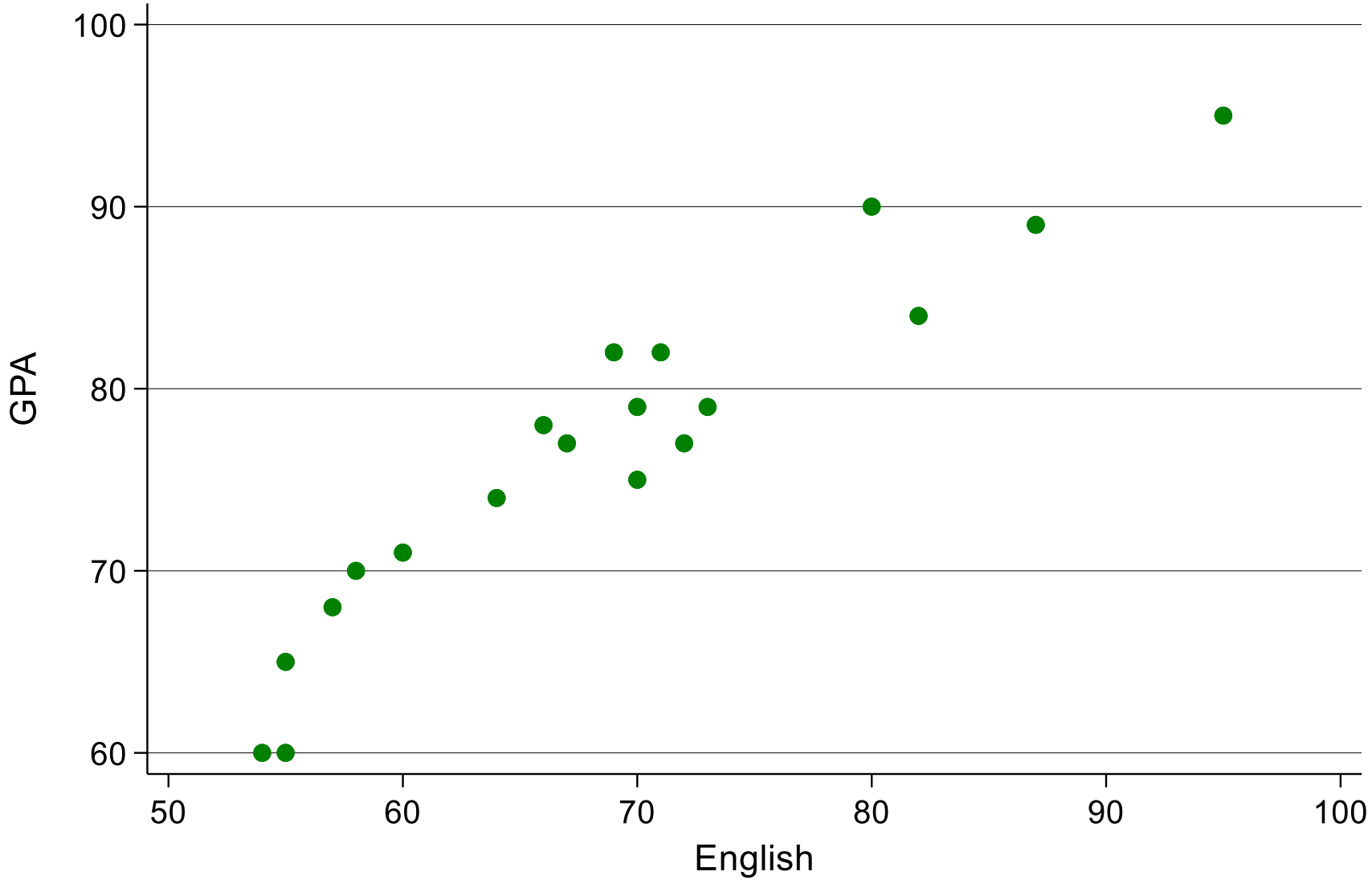
Quadratic Terms

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

English

- Assume that someone told you that a student's command of the English language also affects his or her GPA

Scatter Plot



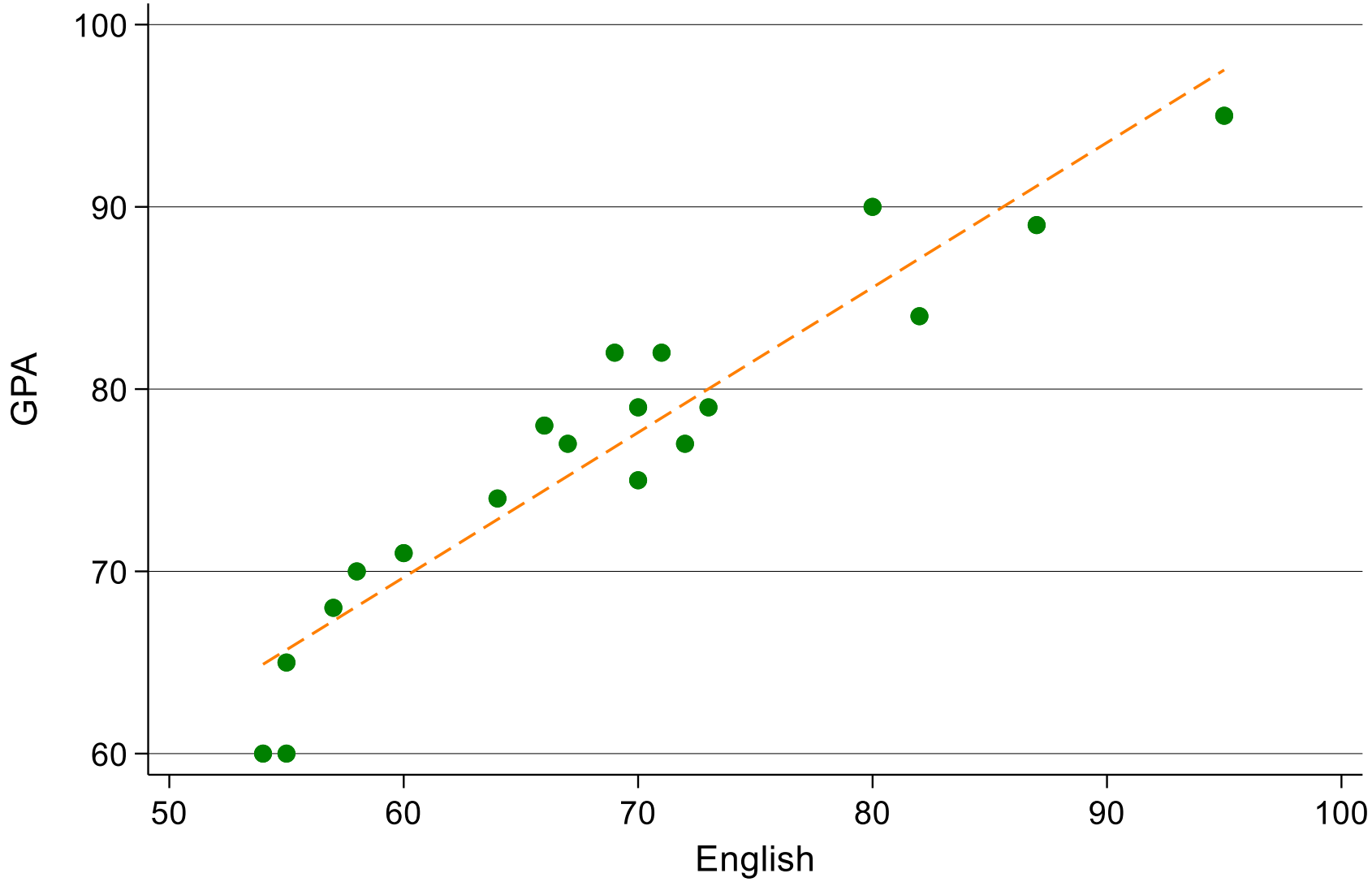
Linear Regression

- If we fit a simple linear regression model, we get the following equation:

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

- 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

Model Fit



Linear Regression

- If we take a closer look at the scatter plot, we will notice that the dots don't seem to fall on a line
- 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:

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

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

English-squared

- Instead of fitting this model:

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

- We can fit this model:

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

Linear Regression with a Quadratic Term

- The output of such a model will be

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

- 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

Model Fit (with quadratic term)

Checking Model Fit and Assumptions

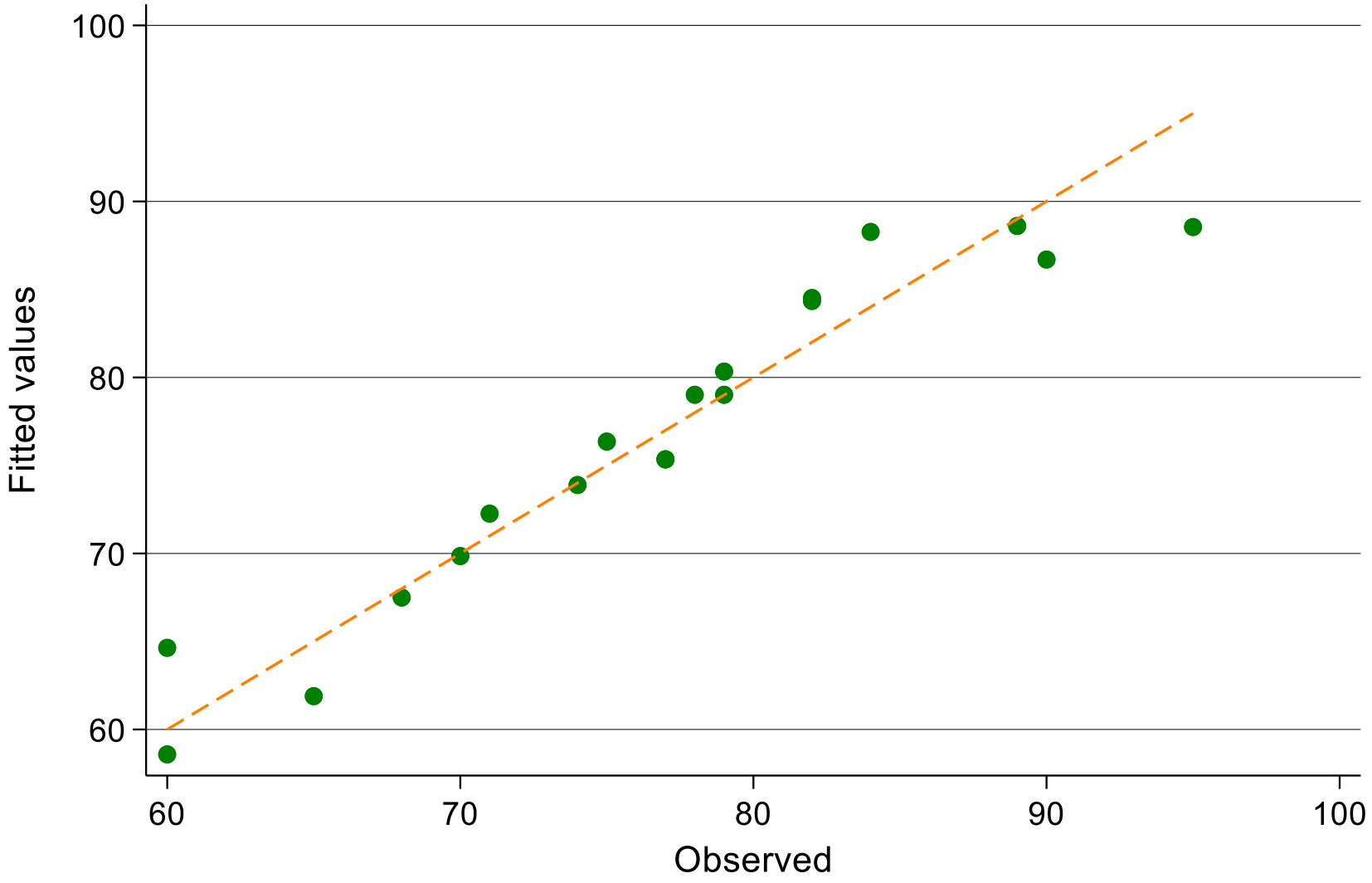
Model Fit

- 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

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

Checking Model Fit



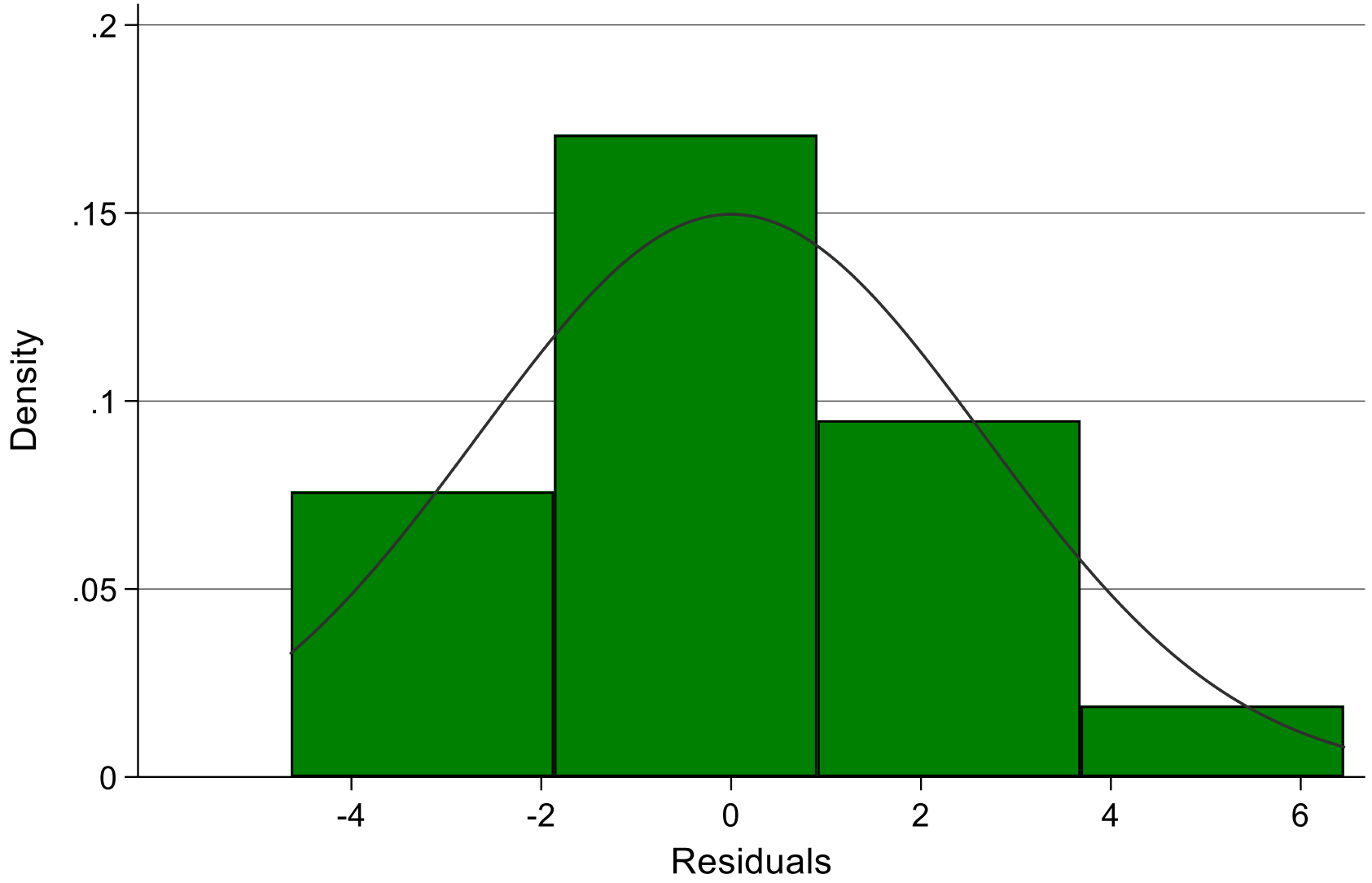
The Residuals

- Linear regression makes several assumptions about the distribution of the residuals

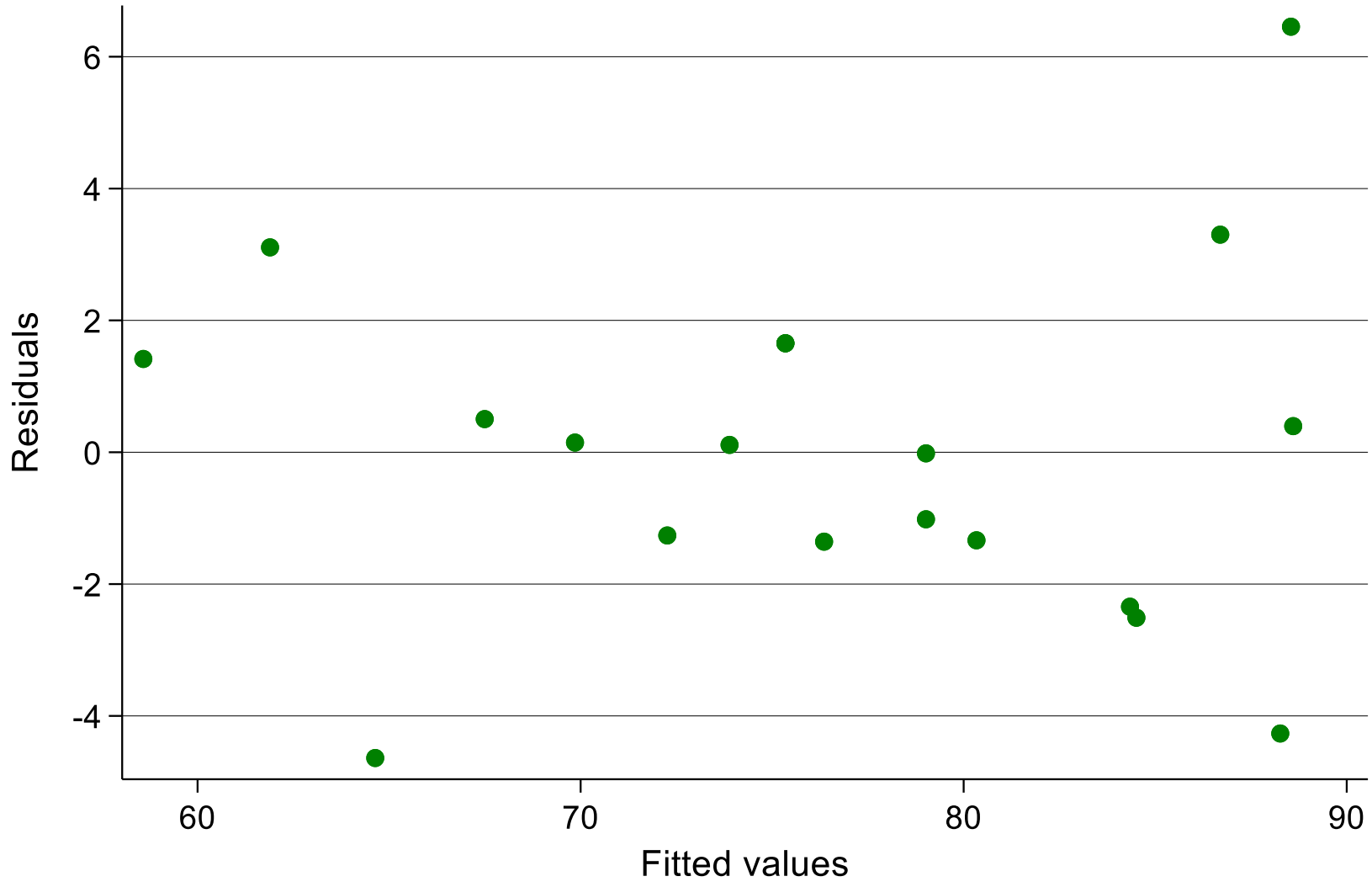
1. Normality
2. Independence
3. Constant variance

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

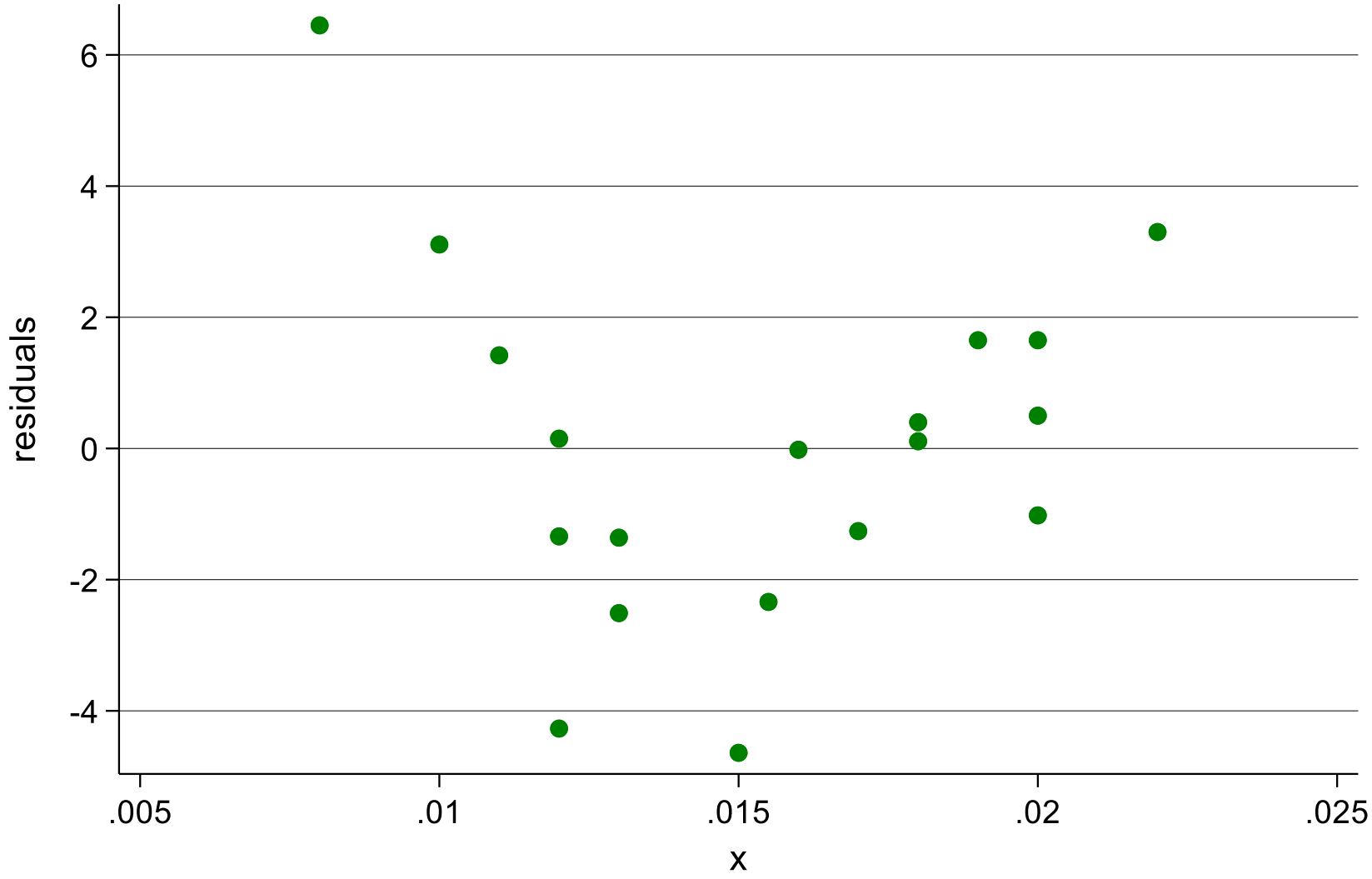
Residuals: Normality



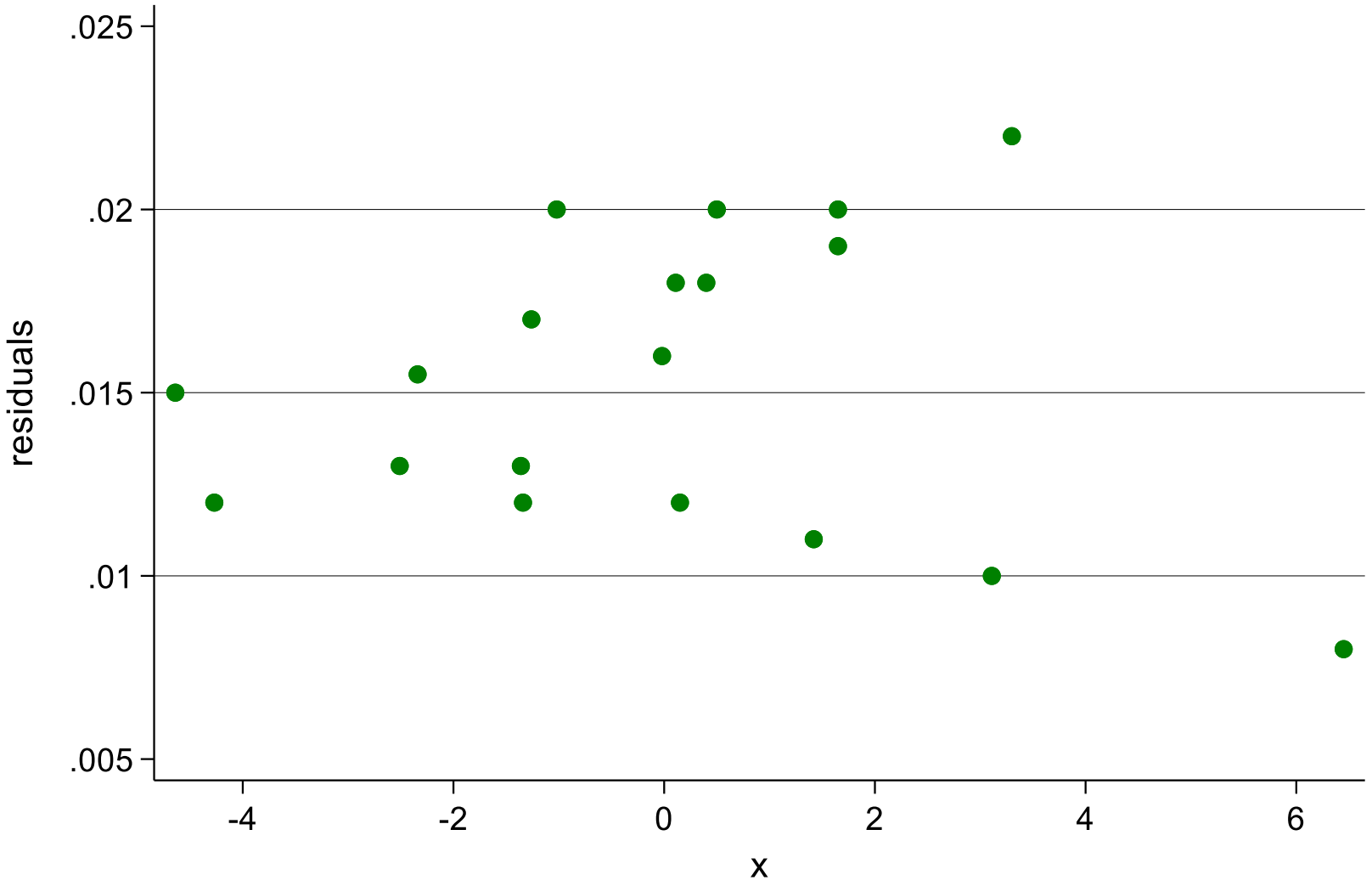
Residuals: Independence



Example of No Independence



Residuals: Example of No Constant Variance



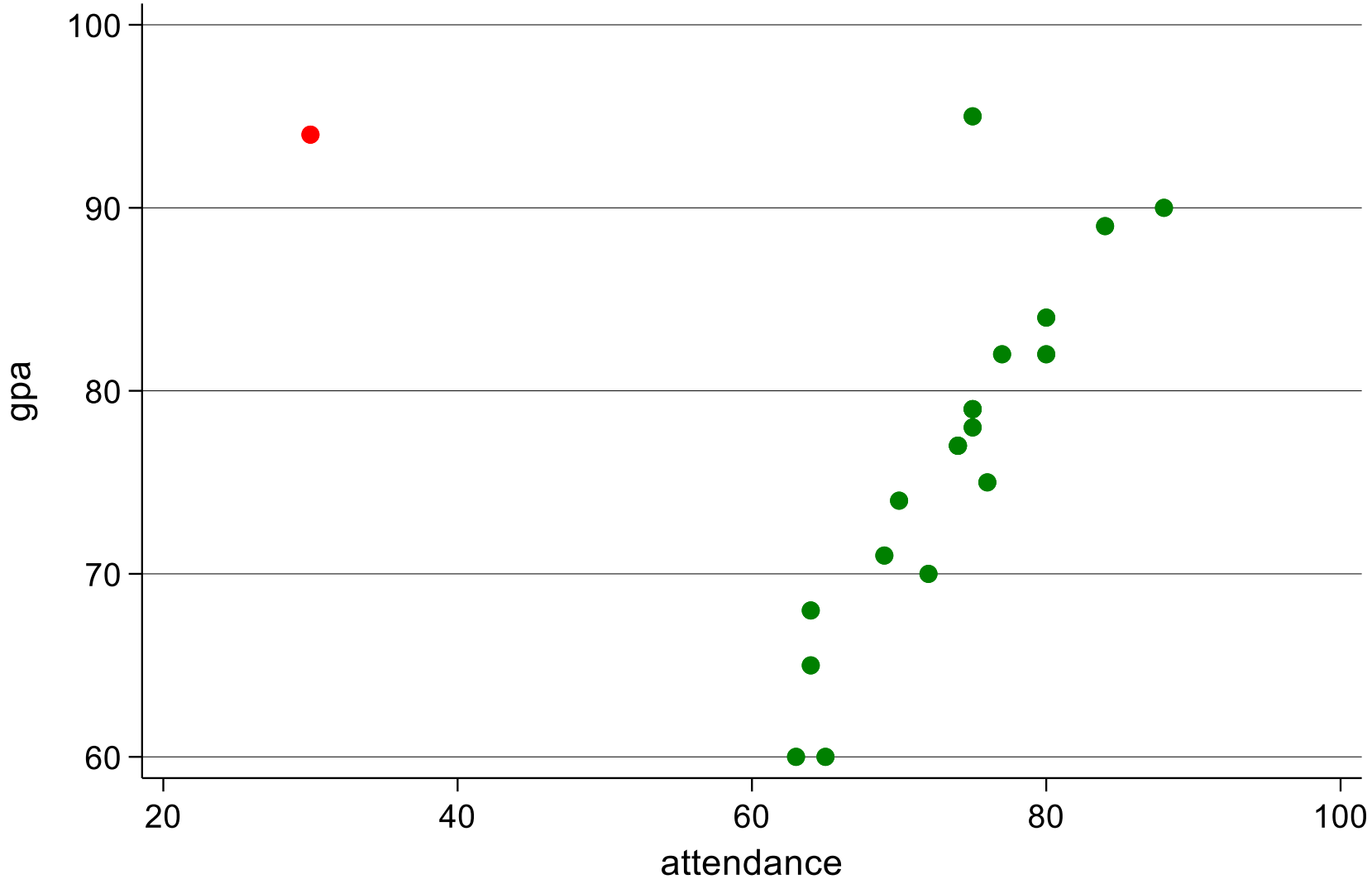
Multicolinearity

- 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
- 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

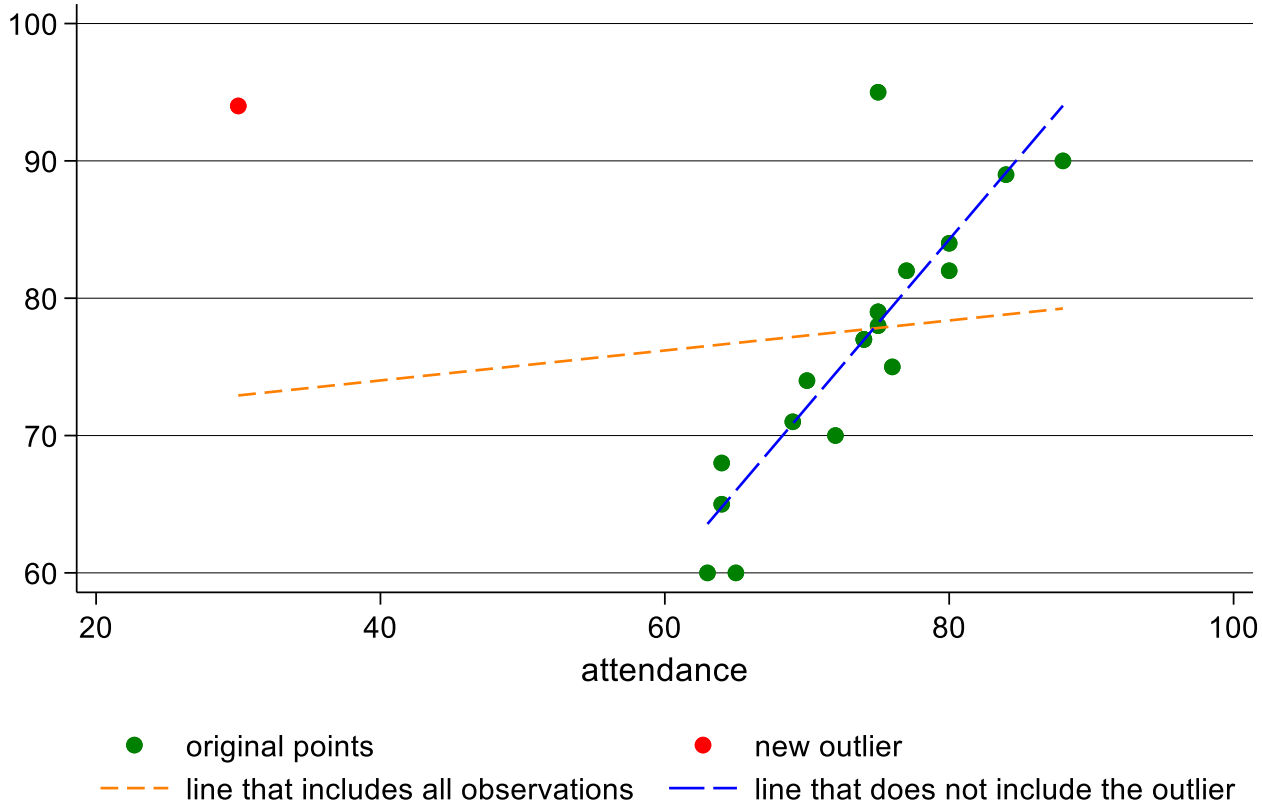
Diagnostics

- What if we checked the model fit and the assumptions of independence, normality, and homoscedasticity and found that some of these assumptions were violated?
- 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

Outliers



Outliers



Excluding outlier:

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

Including outlier:

$$GPA = 0.11(attendance) + 69.65$$

Influential Observations

- There are several statistics that are used in order to measure the influence that each observation has:

1. DFBETAS
2. DFFITS
3. Cook's D

Influential Observations

There are several statistics that are used in order to measure the influence that each observation has:

1. DFBETAS
2. DFFITS
3. Cook's D

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

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?
- Forward selection: this method adds independent variables one step at a time
- Backward elimination: this method removes one variable at a time
- Stepwise regression: this method is a combination of the previous two

Selection of Models

- 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