RELATIONSHIP ANALYSIS

This chapter focuses on how to use statistical analysis to test the hypothesis about relationship between concepts. In particular, there are two statistical analyses that are widely used for this task. They are (1) correlation analysis; and (2) regression analysis. Anyway, the objective of this chapter is not to dig down into the mathematics behind these techniques. The main objective is to understand how to use statistical software package to analyze the data, to know how the results are interpreted, and how to make a research conclusion based on the results interpretation.

Before we move into the detail of these two statistical analyses, let's come up with some research case that we can use as the example in this chapter. Here, the author was interested to study the factors that might predict grade point average (GPA) of students in a class. In particular, there are three main variables that the author aimed to explore. These three variables include:

(1) How many hours a student studies for the exam on average per day.
(2) How many time a student skips class.
(3) How many hours a student sleeps per night.

The first factor, hours of study per day, is quite obvious. The more a student spends on reviewing class materials and lectures, the more chance he/she will score better in the exam. So we expect a positive contribution of this factor on GPA that a student will get.

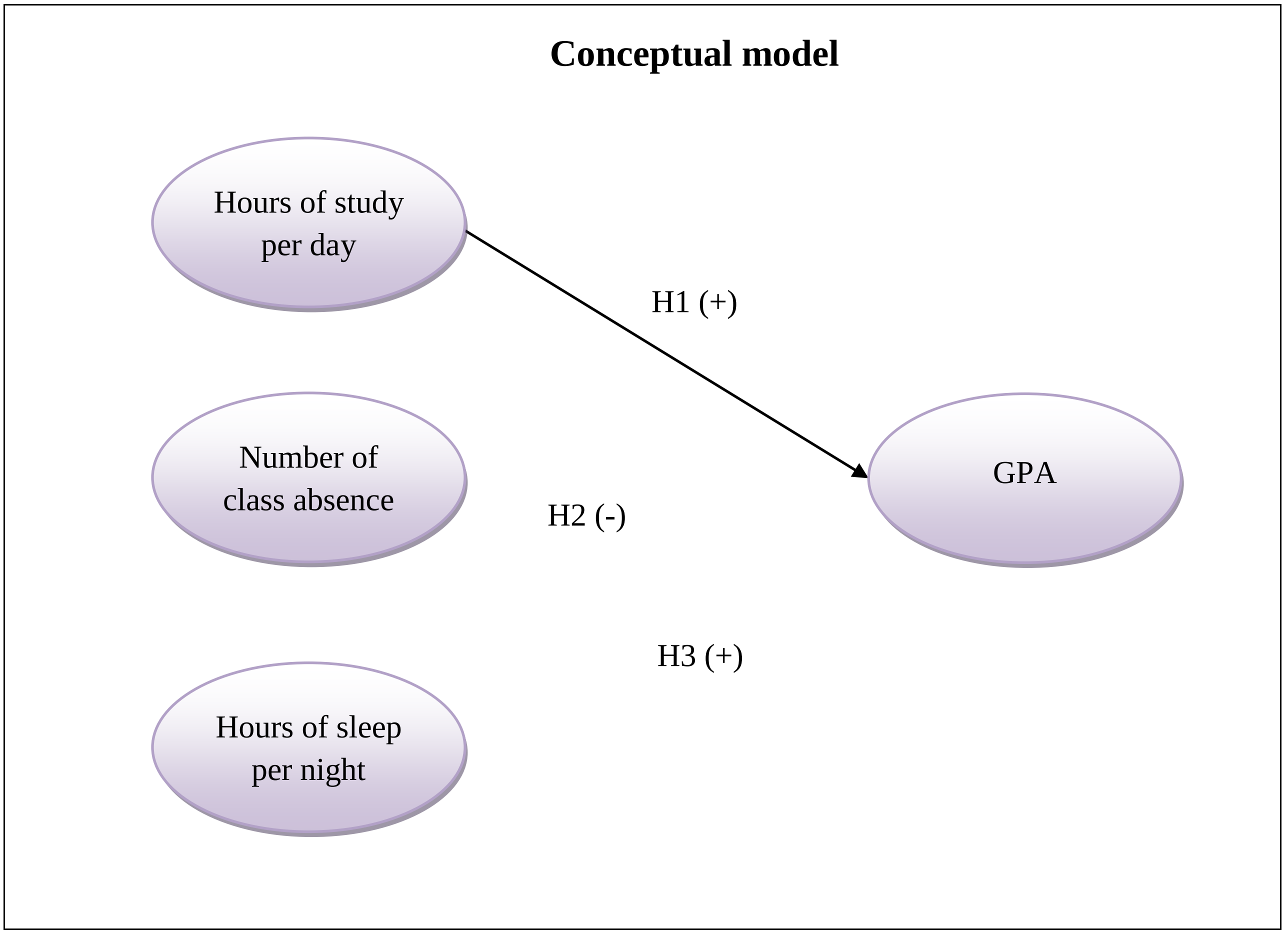
The second factor, number of class absence, can potentially affects class performance as well. If a student skips more classes, he/she will miss important lectures, thereby causing him/her to perform poorly in the exams. Also, some lecturer may count attendance as a part of grading, and that may affect the GPA of a student who skips more classes. In this case, we expect a negative relationship between number of class absence and GPA.

The third factor, hours of sleep per night, is the interesting factor that is worth exploring. However, it may be quite difficult to predict its effect on GPA. On the positive side, more sleep per night may benefit students because adequate and good night's sleep makes people become healthy and have clear mind. In this case, students who got more sleep may become more productive and be able to memorize what they had studied better than students who didn't have adequate sleep, thereby causing them to perform better in the exams. However, although we believe in the benefits of good night' sleep, we can't ignore the possible that too much sleep may cause negative impacts as well. For example, it may cause students to become lazy to study. Anyway, for now we may want to be optimistic about the benefit of sleep. So we predict that more sleep per night might help students achieve higher GPA.

Based on the logical reasons that we provided about the contributions of these three factors on GPA that students would achieve, we can declare the hypotheses as the following:

> Hypothesis 1: Hours of study will positively affect student GPA. In other words, students who spend more time studying will get higher GPA than students who spend less time studying.
>
> Hypothesis 2: Number of class absence will negatively affect student GPA. In other words, students who skip more class will get lower GPA than students who skip less class.
>
> Hypothesis 3: Hours of sleep will positively affect student GPA. In other words, students who get more sleep will get higher GPA than students who get less sleep.

If we draw a conceptual model of what we want to prove, we will get the conceptual model like the figure below. The predicted signs of the relationship are put in parentheses.



Now that we have the hypotheses in place, it is time that we can perform *empirical testing* to verify the hypotheses. The data were collected from a small group of undergraduate students. For hours of study and hours of sleep, students were asked to indicate the average hours they normally spent on reviewing lectures daily and the average hour they normally slept per night. The data about number of class absence and GPA were obtained from the lecturer's grading sheet after the class ended.

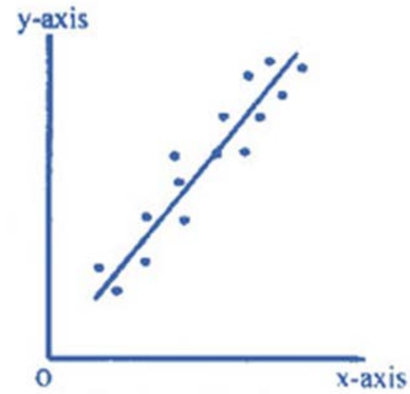
In addition to these data, other characteristics of a sample were also collected including gender (male/female), student classification (freshman, sophomore, junior, senior), and academic department (engineering, business, and art). We will use these data for relationship analyses throughout the chapter.

CORRELATION ANALYSIS

Correlation (or sometime is called *bi-variate correlation*) is a statistical technique that is used for estimating the relationship or the variation between two variables. In statistical analysis, the sign and strength of the correlation between variables can be represented by *correlation coefficient*, or a small "r". The correlation coefficient can be positive or negative, ranging from 1 to -1. The positive correlation (r is greater than 0, but less than 1) means that two variables move together in the same direction. For example, the positive correlation between X and Y means that as X increases, Y will also increase; if X decreases, Y will also decrease.

Positive correlation

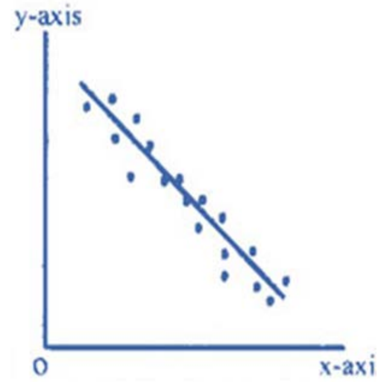
$(1 > r > 0)$



On the other hand, the negative correlation (r is less than 0, but still more than -1) means that two variables move together in the opposite direction. For example, the negative correlation between X and Y means that as X increases, Y will decrease; but if X decreases, Y will increase.

Negative correlation

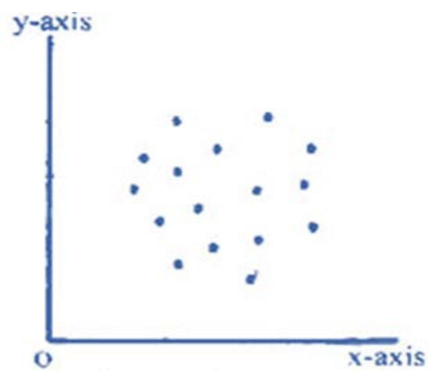
(0 > r > -1)



However, in some case we cannot identify whether the correlation between variables is positive or negative. In fact, two variables may not actually relate. In this case, we can imply that the correlation is equal to 0. It means that the relationship between two variables does not exist.

Correlation does not exist

(r = 0)



The more the value of correlation coefficient is close to either 1 or -1, the stronger the relationship between variables. Conversely, the more the value of correlation coefficient is close to 0, the weaker the relationship. For example, if the correlation coefficient between X and Y is equal to .95, and the correlation coefficient between X and Z is equal to .25, we can conclude that the relationship between X and Y is stronger than the relationship between X and Z. The same is for a negative correlation. If the correlation coefficient between X and Y is equal to -.75, and the correlation coefficient between X and Z is equal to -.35, we can conclude that the relationship between X and Y is stronger than the relationship between X and Z.

Calculating correlation coefficient

The value of the correlation coefficient can be calculated using the following equation:

$$r = \frac{\sum_{i=1}^{n} \left(\left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right) \right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}}$$

For example, if we want to estimate the relationship between average hours that students spent on study per day (X) and the score that students got from the test (Y), we can use the data that we collect, like what is shown in the table below, to calculate correlation coefficient.

Average hours of study per day (X)	Test score out of 100 (Y)
1	70
2	86
3	90
4	92
5	95

If we use the value of X and Y in the dataset and calculate them based on the formula shown above, we will get the correlation coefficient equal to 0.90, which represent the relationship between these two variables. A positive correlation coefficient means that these variables move together in the same direction. If average hours of study increases, test score will increase as well. Conversely, if average hours of study decrease, test score will also decrease. Because the value of the correlation coefficient is very close to 1, it suggests that the strength of relationship between these two variables is quite strong.

Hypothesis testing of correlation coefficient (r)

Just because we get positive or negative correlation coefficient, we can't confirm yet whether the relationship between two variables actually exists. In order to confirm whether the correlation between two variables can be supported or not, it is important to perform hypothesis testing to make sure that the correlation that we get does not happen by chance. Null hypothesis and alternative hypothesis for correlation can be declared as the following:

H_0: $r = 0$; there is no correlation between two variable
H_a: $r \neq 0$; there is a correlation between two variable
(r can be positive or negative)

Using the same rule, we can reject H_0 if a p-value of the correlation coefficient is less than or equals to .05 (or less than 5 percent). In this case, we can *reject null hypothesis* suggesting that correlation between variables is equal to zero. This means that alternative hypothesis is *statistically supported*. Note that correlation coefficient that is not equal to zero in this case can mean either it is greater than zero or less than zero. If the correlation coefficient is greater than zero, we conclude that two variables positively relate; if the correlation coefficient is less than zero, we conclude that two variables negatively relate.

On the other hand, if a p-value from correlation analysis is greater than zero, we *fail to reject the null hypothesis* suggesting that there is no correlation between two variables. In this case, the relationship between variables *is not statistically supported*. We cannot confirm the relationship between them.

Estimating correlation coefficient using SPSS

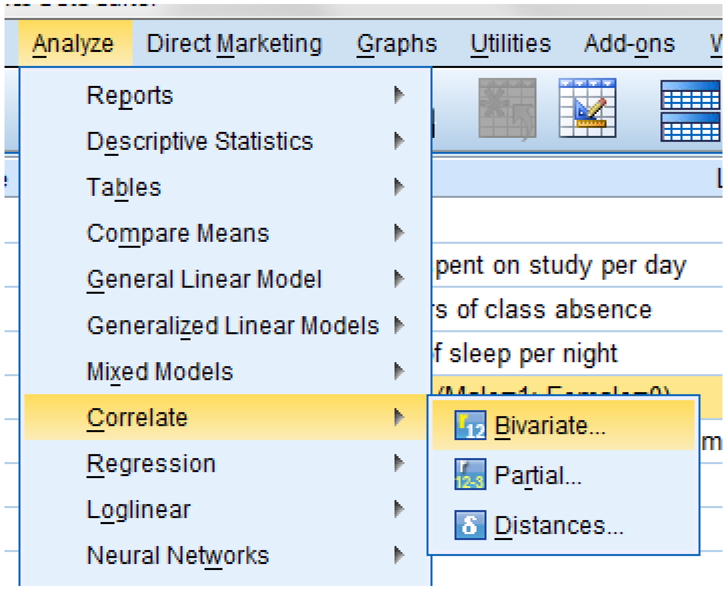
Now that we get some basic idea of what is correlation coefficient and how to interpret it, now let's make our life easier by using SPSS to do the calculation for us. We will use the research topic mentioned in the beginning of the chapter as the example case to see how to use correlation analysis to investigate the relationship between the variables of interest.

Open the file "GPA.sav". This file contains raw data of all variables that were collected from 14 student sample, including GPA, hours of study per day, number of class absence, and hours of sleep per night. The data also contain some demographic information of a sample including gender, student classification, and academic department. All variables and data are shown below:

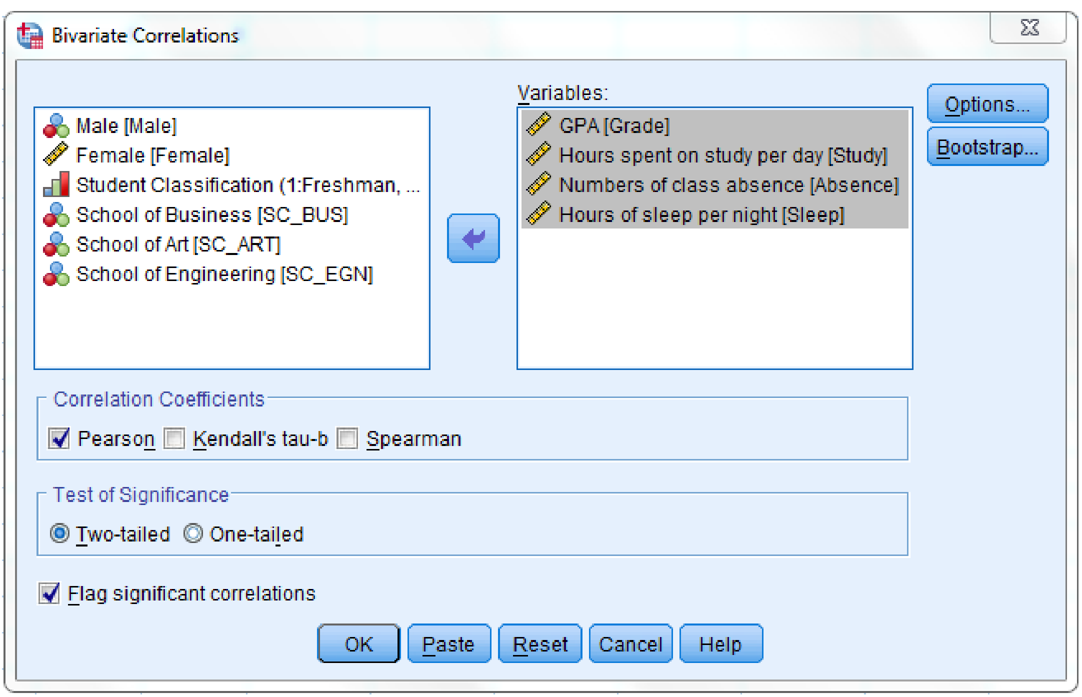
	Name	Type	Width	Decimals	Label
1	Grade	Numeric	8	2	GPA
2	Study	Numeric	8	0	Hours spent on study per day
3	Absence	Numeric	8	0	Numbers of class absence
4	Sleep	Numeric	8	0	Hours of sleep per night
5	Male	Numeric	8	0	Male
6	Female	Numeric	8	0	Female
7	Classification	Numeric	8	0	Student Classification (1:Freshman, 2:Sophomore, 3:Jonior, 4:Senior)
8	SC_BUS	Numeric	8	0	School of Business
9	SC_ART	Numeric	8	0	School of Art
10	SC_EGN	Numeric	8	0	School of Engineering

	Grade	Study	Absence	Sleep	Male	Female	Classification	SC_BUS	SC_ART	SC_EGN
1	2.10	3	2	6	0	1	1	0	1	0
2	2.44	3	2	6	1	0	2	0	1	0
3	2.50	3	1	6	0	1	2	0	0	0
4	2.60	3	1	6	1	0	2	0	1	0
5	2.60	4	1	6	0	1	3	1	0	0
6	2.70	3	1	6	0	1	3	1	0	0
7	2.80	5	2	6	0	1	3	0	1	0
8	3.20	6	0	7	0	1	3	0	1	0
9	3.27	6	0	8	0	1	4	0	1	0
10	3.92	8	0	8	0	1	4	1	0	0
11	3.90	8	1	7	1	0	4	0	1	0
12	1.78	2	3	6	1	0	1	0	0	1
13	1.90	3	2	5	1	0	1	0	0	1
14	2.01	3	3	5	1	0	1	0	0	1

From the menu bar, select “*Analyze*”, then “*Correlate*”, and “*Bivariate*”.



Now drag all variables that we want to estimate their correlations into the box on the right hand side. By default, the software automatically selects Pearson correlation coefficient and use two-tailed test for the analysis.



By using Pearson correlation coefficient, the software assumes that your data is normally distributed and the variables move together in linear fashion. However, if it appears that the variables tend to move together monotonically and your data is not normally distributed, Spearman or Kendall's tau-b correlation coefficients will be the more reliable methods in this case (Chen & Popovich, 2002; Kendall & Gibbons, 1990).

The following are correlation results that SPSS provides:

Correlations

		GPA	Hours spent on study	Numbers of class absence	Hours of sleep per night
GPA	Pearson Correlation	1	.944**	-.793**	.844**
	Sig. (2-tailed)		.000	.001	.000
	N	14	14	14	14
Hours spent on study	Pearson Correlation	.944**	1	-.672**	.802**
	Sig. (2-tailed)	.000		.008	.001
	N	14	14	14	14
Numbers of class absence	Pearson Correlation	-.793**	-.672**	1	-.787**
	Sig. (2-tailed)	.001	.008		.001
	N	14	14	14	14
Hours of sleep per night	Pearson Correlation	.844**	.802**	-.787**	1
	Sig. (2-tailed)	.000	.001	.001	
	N	14	14	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

The intercept between the variable in the row and the variable in column represent the correlation between them. For example, the variable *"GPA"* at the first row and the variable *"Hour spent on study"* at the second column represent the result from correlation analysis between these two variables. The number at the top .944 is the correlation coefficient that represents the relationship between them. In particular, this correlation coefficient is positive which means that they have a positive relationship. This correlation coefficient that is close to 1 suggests that the relationship between is very strong.

Correlations

		GPA	Hours spent on study
GPA	Pearson Correlation	1	.944**
	Sig. (2-tailed)		.000
	N	14	14

Correlation coefficient

p-value

Number of observation

Next, the value of .000 at the second line represents a p-value of the correlation coefficient. In particular, a p-value is significantly lower than 0.05. In fact, this means that a p-value is significant at less than .1 percent (p<.001). This suggests that the positive correlation coefficient between these two variables *can be supported statistically*. We can be confident that the positive relationship between GPA and hour spent on study does not just happen by chance.

		GPA	Hours spent on study	Numbers of class absence
GPA	Pearson Correlation	1	.944**	-.793**
	Sig. (2-tailed)		.000	.001
	N	14	14	14

If we want to assess the correlation between "*GPA*" and "*number of class absence*", we have to look at the intercept of these two variables in the correlation table, as highlighted above. It indicates that the correlation coefficient is -.793. Here, the correlation coefficient has negative sign. It means that there is a negative relationship between these two variables. Once again, negative correlation tells us that two variables tend to move in the opposite direction. In this case, as number of class absence increases, GPA will decrease; but if number of class absence decreases, GPA will increase. However, you cannot rush into making a final conclusion yet just by looking at the sign of the correlation coefficient,. We have to take a look at a p-value to confirm whether this negative correlation can be supported statistically or not. From the result, it shows that a p-value is equal to .001, which is lower than .05. In fact, it is significant at the 1 percent level. Thus, we can conclude that the negative relationship between GPA and number of class absence is *statistically supported*.

		GPA	Hours spent on study	Numbers of class absence	Hours of sleep per night
GPA	Pearson Correlation	1	.944**	-.793**	.844**
	Sig. (2-tailed)		.000	.001	.000
	N	14	14	14	14

Lastly, if we want to assess the correlation between "*GPA*" and "*hours of sleep per night*", we have to look at the intercept of these two variables in the table. From the correlation table, it indicates that the correlation coefficient is .844. Here, the correlation coefficient has a positive sign. It means that there is a positive relationship between these two variables. As hours of sleep per night increases, GPA also increases; but if of sleep per night decreases, GPA also decreases. When we consider a p-value of the correlation coefficient, it shows that the p-value is .000 which is a lot below .05. In fact, a p-value is significant at less than 0.1 percent ($p<.001$). Thus, we can conclude that the positive relationship between GPA and hours of sleep per night is *statistically supported*.

Limitations of correlation analysis

Although correlation analysis can be used to evaluate the relationship between variables, it has major limitations. First, correlation analysis only estimates one-on-one relationship between variables. In other words, it only allows us to estimate the relationship between one pair of variables at a time. This is why it is called a bivariate correlation. For example, the positive correlation between hours spent on study and GPA just takes into consideration the relationship between these two variables only. It does not consider the influence of other factors beyond hours spent on study that might affect GPA at the same time.

Second, correlation analysis only tells us whether two variables have positive, negative, or no relationship. It does not estimate the rate of change that two variables have with each other. For example, a positive correlation between hours spent on study and GPA only tells us that as hours spent on study increases, GPA also increases. However, it cannot tell what will happen to the GPA if a student spent 1 hour more on study; how much his/her GPA will increase? Nonetheless, these limitations can be overcome by the regression analysis that we will discuss in the next section.

SIMPLE REGRESSION

Regression analysis offers more benefits than correlation analysis when we use it to analyze the relationship between variables. It not only allows researchers to assess the relationship between variables, but it can also be used for prediction and explanation. Although there are many types of advanced regression analyses that are used in research, the basic type of regression that we need to know as a fundamental technique in this chapter is *ordinary least squares regression* (or *OLS regression*). OLS regression analysis is based on a linear function; it assumes that variables move together in a linear fashion. The basic form of OLS regression is known as a simple regression.

Like correlation analysis, *simple regression* is the statistical technique that we use to estimate the relationship between two variables. However, simple regression can give us more information than correlation analysis does. But before we go into detail of a simple regression analysis, it is important that we have to understand about linear equation. Linear equation can be expressed as the following:

$$\mathbf{Y} = \alpha + \beta\mathbf{X} + \boldsymbol{\varepsilon}$$

whereby:

Y = dependent variable
X = independent variable
α = alpha coefficient or intercept
β = beta coefficient
ε = error term

Generally, Y and X are variables that we aim to study their cause-and-effect relationship. Y is called dependent variable. It is the outcome variable that is affected by other variable. On the other hand, X is called independent variable. It is the variable that will affect the dependent variable.

α is called *alpha coefficient* or *intercept*. This is the indicator that tells us what will be the value of Y if X is equal to zero. In other words, it tells us what will be the value of Y if there is no X.

β is called *beta coefficient*. This indicator tells us about the rate of change that Y is caused by X. In other words, it indicates how many units Y will change if X increases or decreases by one unit.

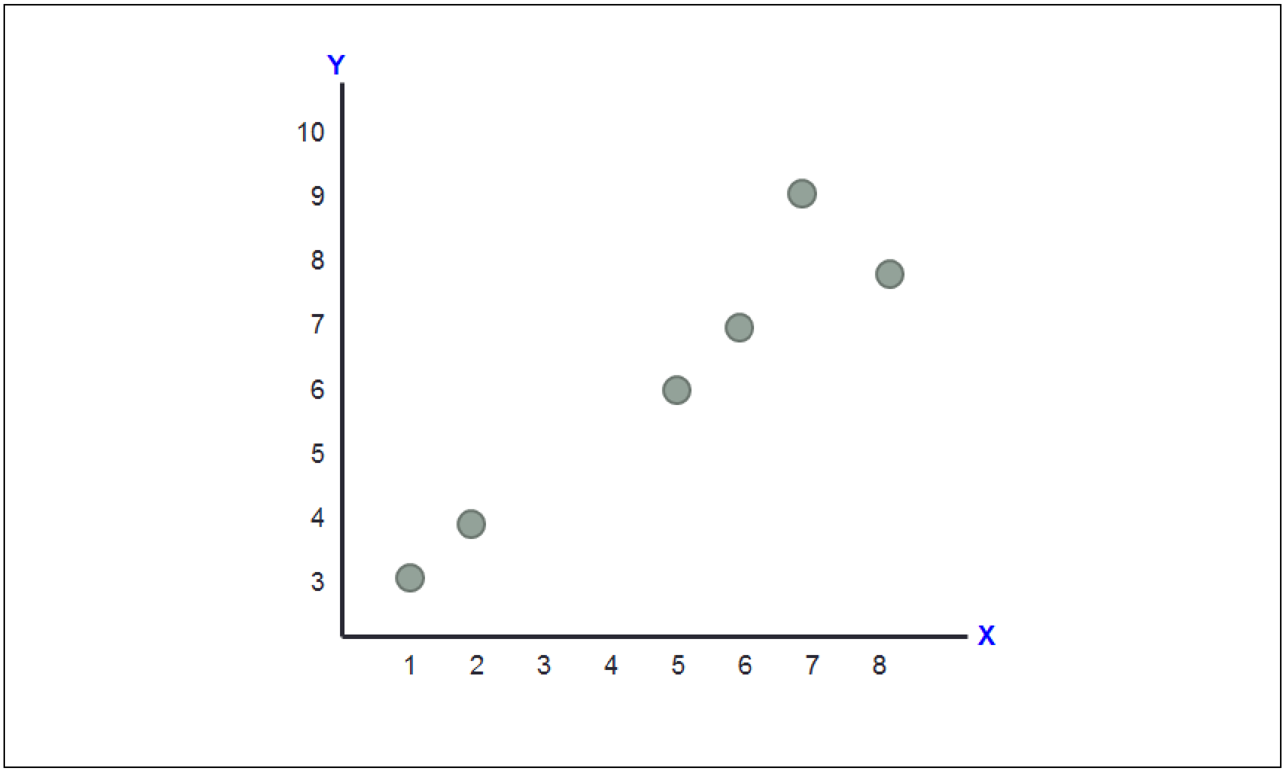
Finally, ε is called *error term* (or *residue*). As its name implies, the error term is the factor that make the predictive power of regression become less accurate or deviate from what it has to be. It also represents other key variables that affect Y but we fail to add them in a regression analysis.

In order to get a clearer understanding about each component in a regression equation, let's assume that we already collected the data of hours spent on study and test scores from eight students. The data are shown as the following.

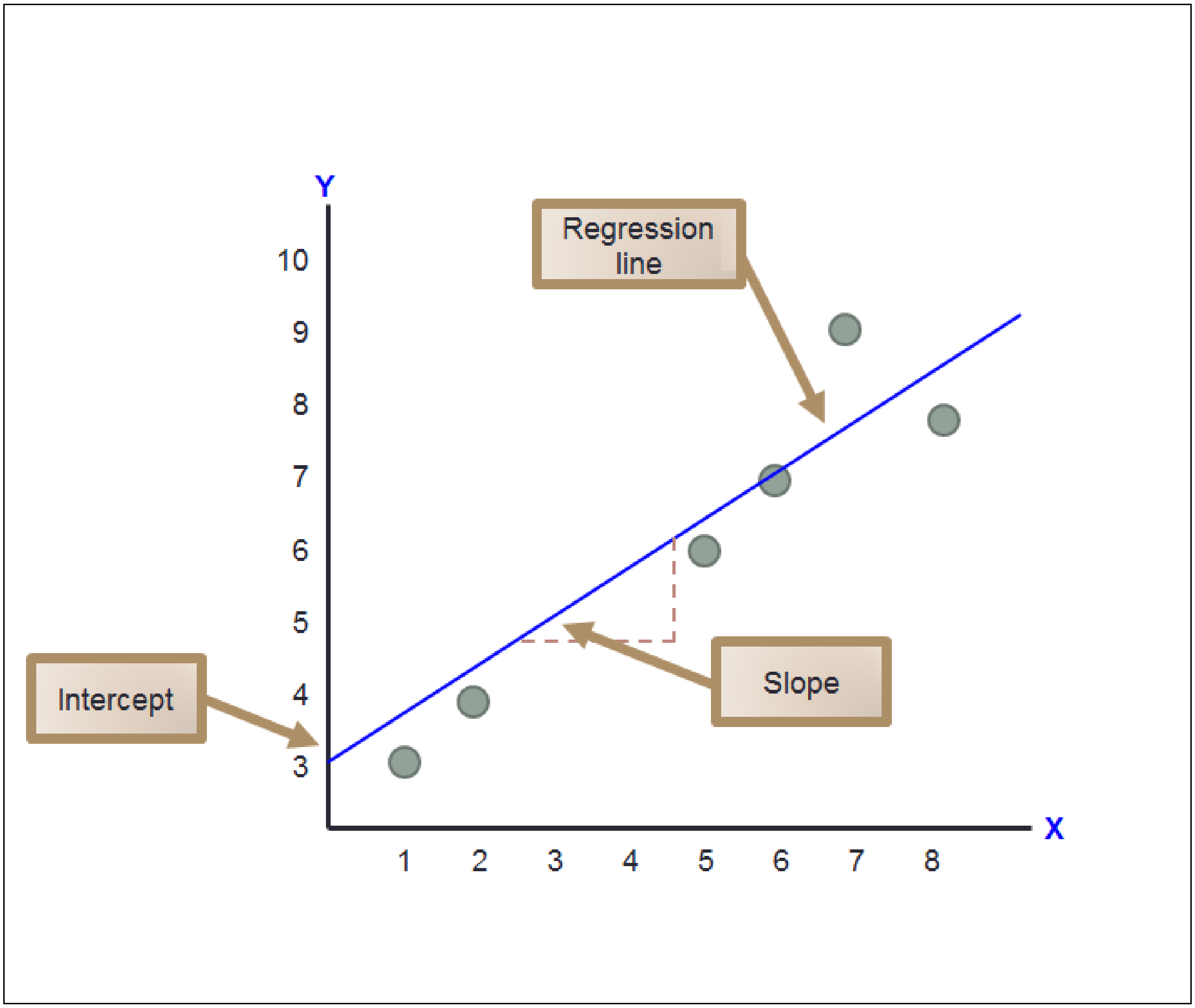
X Number of hours spent on study	Y Test score
1	3
2	4
5	6
6	7
7	9
8	8

From the data, the first student spent 1 hour on average and got 3 points in the exam; the second student spent 2 hour studying and got 4 points; the third student spent 5 hours studying and got 6 points..., the last student spent 8 hours studying and get 8 points in the test. In the case, we expected that hours spent on study is the factor that predicts test score of students. Because test score is the outcome, it is dependent variable (Y) in the regression equation. And because hour spent on study is the factor that will affect the outcome variable, it is independent variable (X) in the regression equation.

Now that we know X (independent variable) and Y (dependent variable), then we use these data to perform a scatter plot. If we plot the value of X and Y along the vertical and horizontal axis, we will get the data plot as the following:



From this data plot, if we draw a line across the middle of all data plots, we will obtain the line like in the figure below. The line that cut through the middle of all data points is a *regression line*. A regression line that we obtain will also tell us about alpha coefficient and beta coefficient of the regression equation.

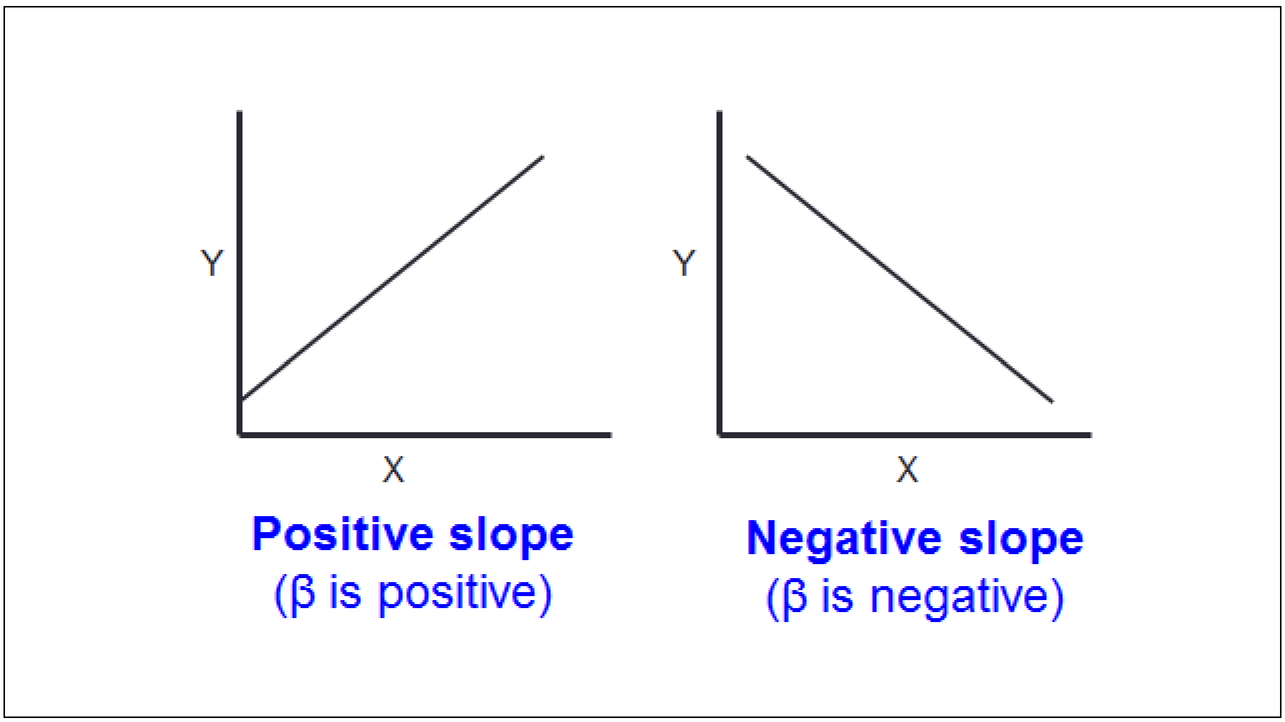


<u>Alpha coefficient</u>

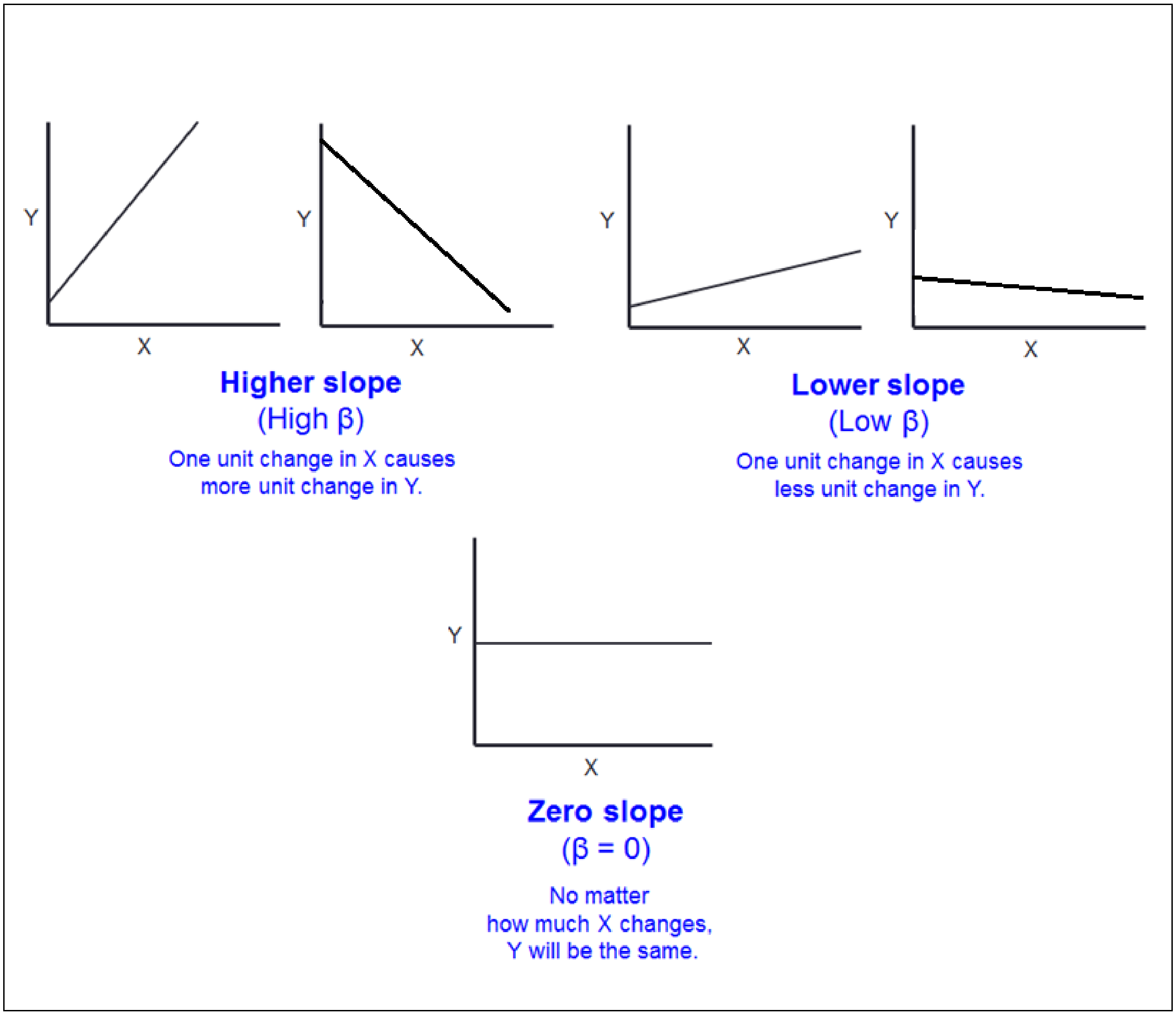
Alpha coefficient or *intercept* tells us *what will be the value of the dependent variable (Y) when the independent variable (X) is absent or equal to zero*. The value of the alpha coefficient can be represented by the point where the regression line touches the Y axis. If you take a look at the regression line in the figure above, you can see that the line touch the Y-axis at the area where Y is equal to 3. This point is also the area where X is equal to zero. Therefore, the value of alpha coefficient in this case means that Y will be equal to 3 when the value of X is equal to zero; or when X is absent, the value of Y will be equal to3.

Beta coefficient

A slope or a steepness of the regression line can be used to indicate the strength of relationship between dependent variable and independent variable. From the figure above we can see that the line has upward slope. Upward slope implies the positive relationship between variables. The slope of the regression line is represented by the *beta coefficient* (β). Beta coefficient has *sign* and *value*. The sign of beta coefficient is similar to the sign of correlation coefficient. If a regression line has upward slope, the sign of the beta coefficient will be positive. Positive beta coefficient means that the increase in the value of X will cause Y to increase (or the reduction in the value of X will cause Y to reduce). On the other hand, when a regression line has downward slope, the beta coefficient will have a negative sign. Negative beta coefficient means that the increase in the value of X will cause Y to reduce (or the reduction in the value of X will cause Y to increase).



Beta coefficient also has the value which indicates the strength of the effect that the independent variable causes to the dependent variable. The value of the beta coefficient reflects the steepness of the regression line. If beta coefficient is highly positive, the slope will be steeper in the upward direction; and this means that the effect of independent variable on dependent variable will be stronger in a positive sense (one unit increase in X will cause more unit increase in Y). If beta coefficient is highly negative, the slope will be steeper in the downward direction; and this means that the effect of independent variable on dependent variable will be stronger in a negative sense (one unit increase in X will cause more unit decrease in Y). But if beta coefficient is equal to zero, the regression line will be completely flat, and this means that the independent variable does not have any effect on the dependent variable.



Together, the sign and value of the beta coefficient indicate the "*rate of change*" that the independent variable influences the dependent variable. What can be interpreted by the sign and the value of the correlation coefficient is that *one unit* *changes in the independent variable will cause the dependent variable to change by how many unit*. In particular, the higher the value of the beta coefficient means the stronger the effect that the independent variable will cause to the dependent variable.

For example, if beta coefficient is equal to positive 5, it means that 1 unit increase in the value of the independent variable will cause the dependent variable to increase by 5 units; it can also mean that 1 unit decrease in the value of the independent variable will cause the dependent variable to reduce by 5 units. Conversely, beta coefficient equals to -5 means that 1 unit increases in the value of the independent variable will cause the dependent variable to reduce by 5 units; also, it can mean 1 unit decreases in the value of the independent variable will cause the dependent variable to increase by 5 units. Beta coefficient equals to zero suggests that no matter how much the independent variable increases or decreases, there will no effect on the dependent variable.

Making a prediction from the regression equation

From what we have learned about how to interpret alpha and beta coefficients in regression analysis, we will see how to make a prediction based on these indicators. If you still remember, the regression equation can be expressed as the following:

$$\mathbf{Y} = \alpha + \beta\mathbf{X} + \boldsymbol{\varepsilon}$$

For now, let's assume that there is no error term in the regression equation. In this case, $\varepsilon = 0$. We will discuss about what is error term in more detail later. At this point, let's assume that the value of alpha coefficient is equal to 10, and the value of beta coefficient is equal to 5. If we replace the values of alpha and beta coefficients into the equation above, we will obtain the equation as the following:

$$Y = 10 + 5X$$

What can be inferred from this equation? You know that alpha coefficient is the predicted value of Y when X is equal to zero. In this case, the alpha coefficient which equals to 10 means that if X is zero, Y will equal to 10. If you want to prove it, try to replace X with 0, then you will get:

$$
\begin{aligned}
Y &= 10 + 5\,(0) \\
  &= 10
\end{aligned}
$$

Next, the value of beta coefficient indicates what will be the predicted value of Y if the value of X changes by 1 unit. Here, the value of beta coefficient is equal to 5, which indicates that 1 unit increase in the value of X will cause Y to increase by 5 units. If you want to prove it, try to replace X with 1, then you will get:

$$
\begin{aligned}
Y &= 10 + 5\,(1) \\
  &= 15
\end{aligned}
$$

And what will be the value of Y if we increase the value of X by one more unit. Now, makes X equal to 2.

$$
\begin{aligned}
Y &= 10 + 5\,(2) \\
  &= 20
\end{aligned}
$$

Lastly, what will be the value of Y if we additionally increase the value of X by one more unit? Now, makes X equal to 3.

$$
\begin{aligned}
Y &= 10 + 5\,(3) \\
  &= 25
\end{aligned}
$$

As you can see, when X is equal to 0, the value of Y is 10, which is equal to the value of alpha coefficient. When the value of X increases from 0 to 1, the value of Y is 15. When the value of X increases from 1 to 2, the value of Y is 20. When the value of X increases from 2 to 3, the value of Y is 25. Overall, you can see that one unit incremented in the value of X appears to increase Y additionally by 5 units, which equal to the value of beta coefficient.

α = 10; **β** = 5		
X	Y	Y increases by
0	10	-
1	15	5
2	20	5
3	25	5

Calculating alpha and beta coefficients

Now that we already know how to interpret alpha and beta coefficients in the regression equation, we need to know briefly about how to estimate them. In order to calculate the alpha and beta coefficients in simple regression, we need to obtain the dataset that contain the values of the dependent variable (Y) and the independent variable (X).

X	Y
1	70
2	86
3	90
4	92
5	95

Then, we can calculate alpha and beta coefficients using the following equations:

$$\beta = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sum x^2 - \left(\sum x\right)^2/n}$$

$$\alpha = \bar{y} - b\bar{x}$$

Like other statistical analysis, in order to confirm whether alpha and beta coefficients that are obtained can be statistically supported, we have to perform hypothesis testing to justify them. Null hypothesis and alternative hypothesis for beta coefficient can be declared as the following:

H_0: $\alpha = 0$; alpha coefficient is equal to zero
H_a: $\alpha \neq 0$; alpha coefficient is unequal to zero
(α can be positive or negative)

H_0: $\beta = 0$; beta coefficient is equal to zero
H_a: $\beta \neq 0$; beta coefficient is unequal to zero
(β can be positive or negative)

In order to confirm that the alpha and beta coefficients are not statistically equal to zero, we have to reject the null hypothesis; therefore, we expect a p-value to be lower than or at least equal to .05.

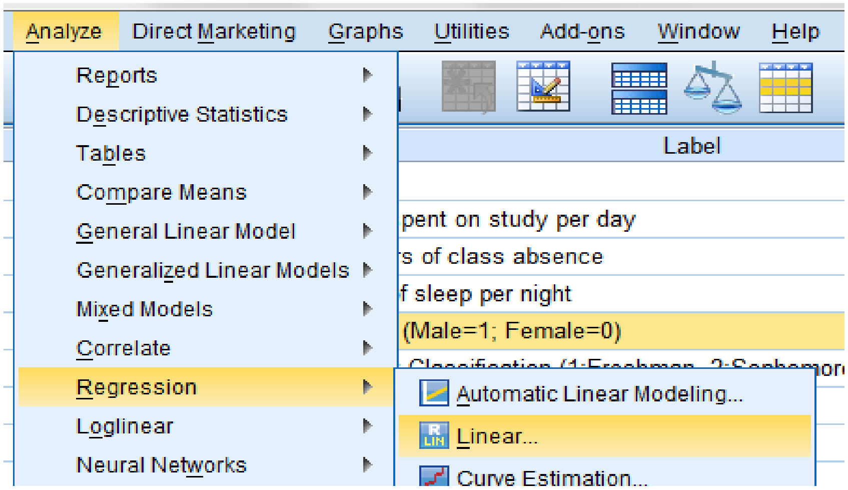
Performing simple regression analysis in SPSS

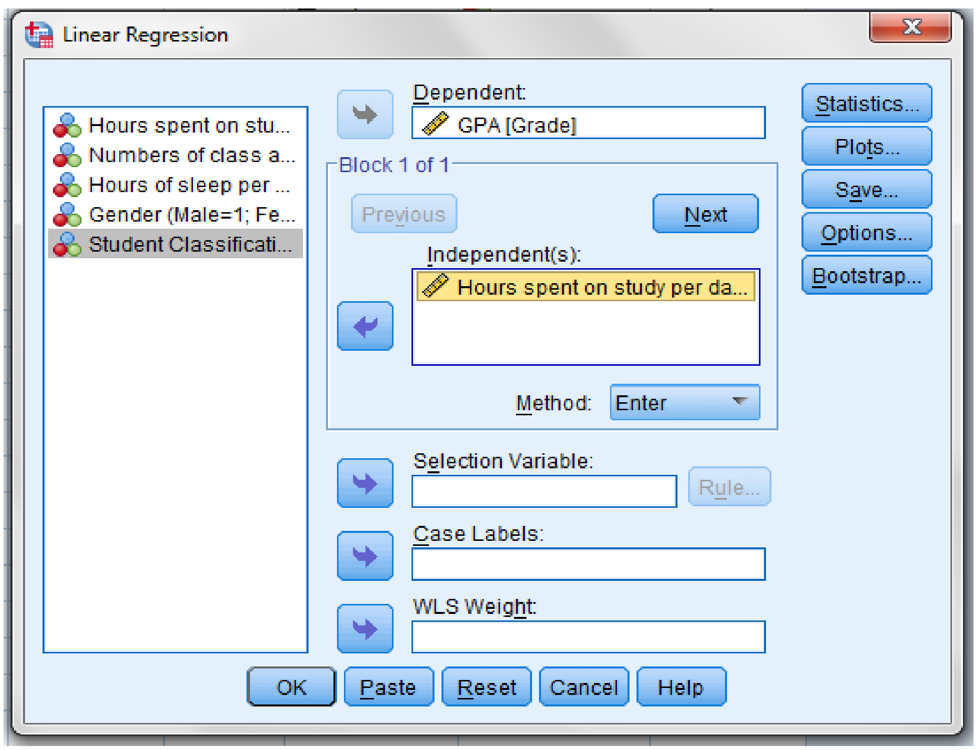
Making a manual calculation like what is shown earlier can be practical if you have a few data. Anyway, if you have big data that contain hundreds or thousands of samples, you can use SPSS to perform the calculation for us. To see how to perform simple regression in SPSS, let's try it again with the research case that we set for this chapter. Because a simple regression only consider one independent variable in the equation, we will try it one-by-one with each factor that is hypothesized to affect GPA.

<u>Estimate the effect of hours spent on study on GPA</u>

We will focus on the first variable, *hour spent on study per day*, to explore whether it can affect GPA or not.

From the menu bar, select Analyze → Regression → Linear





Because GPA is the dependent variable, you have to move it into "Dependent" field. And because hours spent on study per day is the independent variable, you move it to the "Independent(s)" field. Then, you can click Ok button.

If you click Ok button, you will obtain several tables in the Output window of SPSS. But at this point, I would like you to skip tables at the top part and move directly to the table "Coefficients". This is the table that report alpha and beta coefficients of the regression.

α

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1.309	.153		8.534	.000
	Hours spent on study per day	.323	.033	.944	9.877	.000

a. Dependent Variable: GPA

β

P-value

Alpha and beta coefficients of the regression are reported at the column "Unstandardized Coefficients". The first row "Constant" is the alpha coefficient in the regression equation, which equals to 1.309. The second row is the beta coefficient of hours spent on study per day, which equal to .323. This beta has a positive sign which suggests that it positively affects GPA. In particular, the increase in hours spent on study per day causes GPA to increase. However, in order to make sure that this positive effect is supported statistically, we have to look at a p-value of the beta coefficient to make a final confirmation.

A p-value of the beta coefficient can be found at the last column of the table that is labeled 'Sig'. It shows that a p-value of the beta coefficient is .000 which is statistically significant at the below .1 level. A p-value that is lower than .05 suggests that the positive beta is statistically supported. Therefore, we can be confident that hours spent on study per day is a significant variable that positively affects GPA.

<u>Making a prediction</u>

Now that we obtained the alpha and beta coefficients, we can make a prediction what will be the GPA that a student will get if he/she spend a certain hours on study per day. First of all, let's set up the linear regression equation by replacing Y and X with the variables of interest. Let's assume that there is no error term in the analysis for now.

$$\textbf{GPA} = \alpha + \beta \textbf{ Hours spent on study per day}$$

If we place the values of the alpha and beta coefficients into the equation, we will obtain the equation as the following:

$$\textbf{GPA} = 1.309 + .323 \textbf{ Hours spent on study per day}$$

From this regression equation, we can make some prediction about the GPA that a student will get based on how many hours he/she spend on study daily. For example, if a student does not study at all (in this case, hours spent on study is equal to zero), we can predict that he/she will get GPA = 1.309 + .323 (0) = 1.309.

If a student spent 1 hour studying (in this case, hours spent on study is equal to 1), we can predict that he/she will get GPA = 1.309 + .323 (1) = 1.309 + .323 = 1.632.

If a student spent 2 hour studying (in this case, hours spent on study is equal to 2), we can predict that he/she will get GPA = 1.309 + .323 (2) = 1.309 + .646 = 1.955.

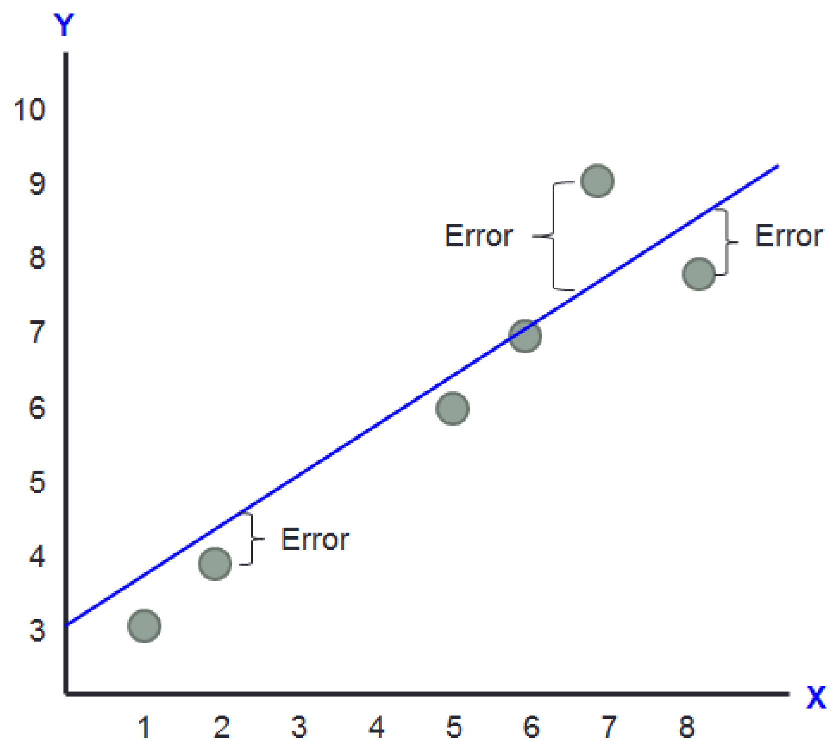
The interpretation can be summarized as the following:

α = 1.309; **β** = .323		
Hours spent on study per day	**GPA**	**GPA changes by**
0	1.309	
1	1.632	+.323
2	1.955	+.323

So what do these findings tell us? It is simple; GPA will increase when the number of hours spent on study per day increase. The rate of change is equal to the value of the beta coefficient. So from this regression analysis, if a student comes to ask you what he/she should do to get higher GPA, the recommendation is that you have to spend more time studying.

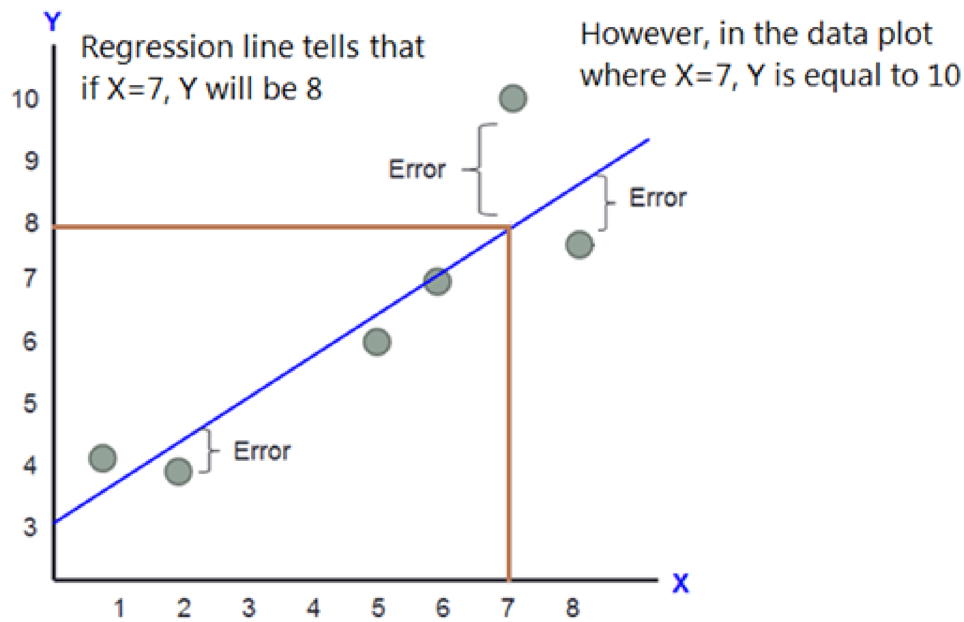
Error term and r-square

The last indicator in the linear regression equation that we have omitted so far is the *error term (ε)*. Now we will have to know what it is and what it tells us. In the figure below, if you take a look at the regression line that cuts through the midpoints of all data plot, you can see that some data points are quite far away from the regression line, some data points are pretty close to the regression line, and there are some data points that touch the regression line.



The distance of each data plot from the regression line is considered the *error* in the regression analysis. The greater the distance from the regression line, the higher the error it is. Errors in regression analysis will affect the predictive power of the regression because they represent the deviation in the predicted values that can be inferred from the regression line.

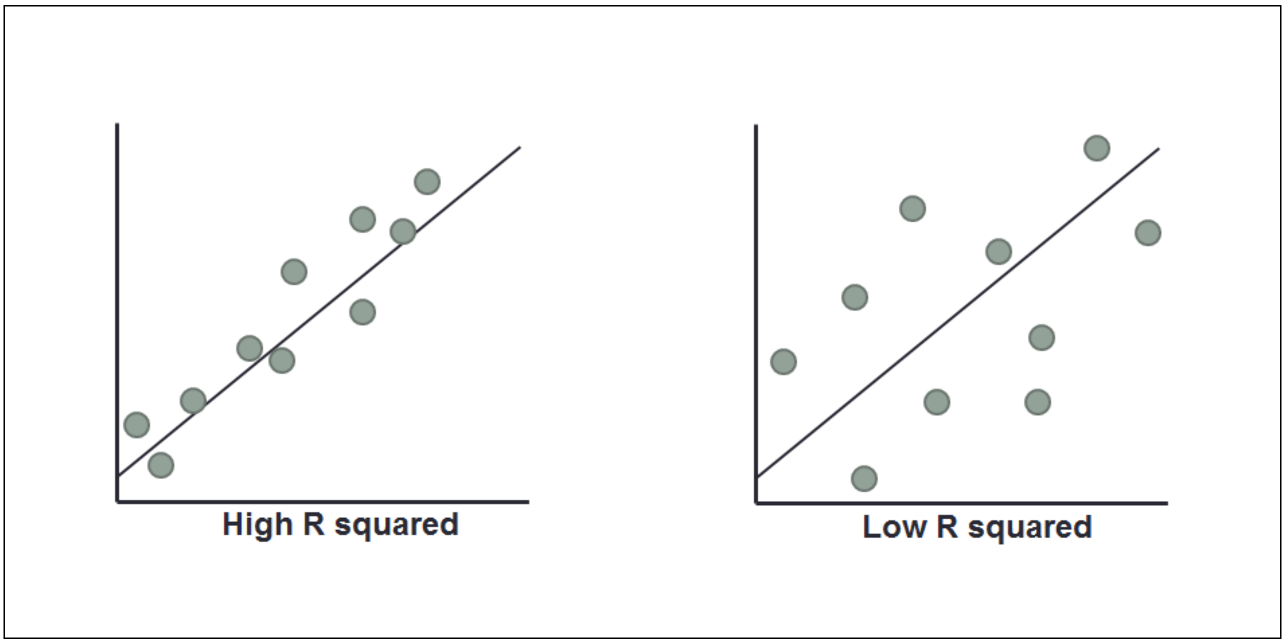
For example, the regression line below tells us that the point when X is equal to 7, Y is equal to 8. However, the data that we used to generate the regression line does not fit well with this. It shows that the point where X is equal to 7, Y is in fact equal to 10. There is a deviation in the actual value of Y in the data and the predicted value of Y as suggested by the regression line. The difference in the value of Y that the regression line predicts and the actual value of data that we observe is regarded as *error* in regression analysis.



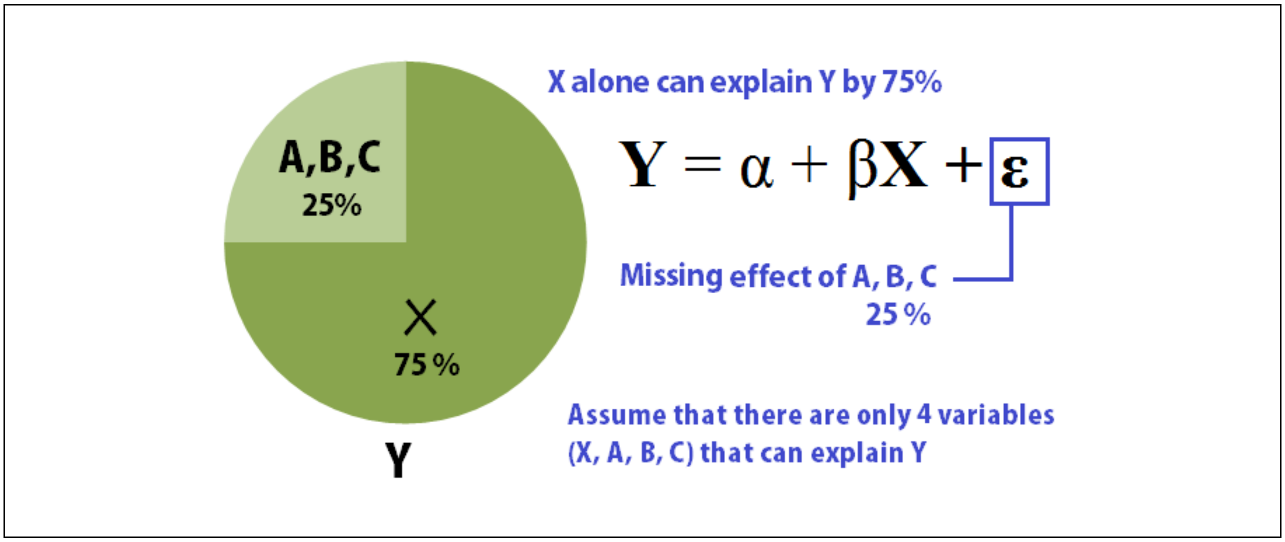
From the figure above you can see that most of the data do not fit well with the regression line. There are errors in almost every data plot. Anyway, although these errors will lower the predictive power of regression, it is ok to have errors in regression analysis. Nothing in this world is perfect; the same is true for a prediction. Generally speaking, it is a very rare case that a prediction can be 100 percent accurate. Thus, some level of error in regression analysis is not unusual. However, it is important that the error should not be too great; otherwise, we may not be able to tell anything from the regression analysis.

Error term in regression equation is also caused by the absence of key variables that might be relevant to predict the dependent variable which are not included in regression equation. In this case, it represents the important variables that we do not consider them in the regression analysis. For example, let's assume that in reality there are only 4 variables (X, A, B, and C) that can explain Y. But if we only think that X is a variable that affect Y and we fail to include the variable A, B, and C in the regression analysis, the error will be equal to the effects that A, B, and C have on Y.

In particular, one key indicator in regression analysis that inversely relates to the error term is called r-square (r^2). R-square or *coefficient of determination* is the indicator for the goodness of fit of the data in the regression analysis. The value of the r-square can range from 0 to 1; thus, it is normally expressed in percentage. When the majority of the data plots fit the regression line quite well (like in the left hand side of the figure), the value of the r-square will be high. However, when the majority of the data plots scatter far away from the regression line (like in the right hand side of the figure), the r-square will be low. Because r-square represents the how will the data fit the regression line, high r-square can signify smaller error in the regression model.



In particular, r-square is the indicator that tells you how many percent that the dependent variable can be explained or can be predicted by the independent variable(s) in the regression analysis. For example, the r-square of .75 means that the independent variable(s) that you put in regression analysis can explain or can predict the occurrence of the dependent variable by 75 percent. There are another 25 percent that might be explained by other variables that we do you put in the regression analysis yet. Referring to the previous example, if we think that X is the only independent variable that affects Y and we didn't realize that A, B, and C can affect Y as well, if we put only X in regression analysis and we obtain the r-square equals to 75 percent, this means that X can explain Y by 75 percent; another 25 percent of Y that can't be explain by X will be fulfilled if we also add the variable A, B, and C in the regression analysis.



Now come back to the simple regression analysis that we did lately about the effect of hours spent on study and GPA. Let's take a look at the model summary table that you recently obtained from SPSS. The table is located at the top part of the output window. The value of r-square is .89 or 89 percent. R-square in this case suggests that the independent variable that we have, that is hours spent on study, can predict or can explain 89 percent of GPA that students obtain. In fact 89 percent is considered a pretty high r-square; it is quite close to 100 percent. This r-square also implies that the remaining 11 percent (100 – 89) might be explained by other variables that we do not put in the regression analysis.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.944[a]	.890	.881	.23331

a. Predictors: (Constant), Hours spent on study per day

Another indicator next to r-square is *adjusted r-square*. Adjusted r-square is quite different from the interpretation of the r-square. For r-square, its value will keep increasing as you add more independent variables into the regression analysis. Even though the independent variable you add is not statistically significant, r-square will keep increasing. But for adjusted r-square, its value will only increase when the independent variable you add to the regression significantly affects the dependent variable. If the variable that you add to the regression is not statistically significant, adjusted r-square will drop in value. Thus, the objective of having adjusted r-square is to prevent researchers from pooling a lot of independent variables into the regression analysis without carefully consider that the variables will actually affect the dependent variable or not.

Using simple regression to estimate the effect of class absence on GPA

Ok, so now you know that number of hours spent on study positively affects GPA, let's see how other variables that we hypothesized will affect GPA. So we move to the next variable, *number of class absence*. In SPSS, we just simply replace the existing independent variable with the new one. In this case, you replace it with *number of class absence*. If you run the analysis, you will obtain the results from the coefficients table as the following:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.417	.197		17.341	.000
	Numbers of class absence	-.533	.118	-.793	-4.513	.001

a. Dependent Variable: GPA

From this table, the results show that the alpha coefficient or the constant is equal to 3.417. As a reminder, alpha coefficient tells us what will be the value of the dependent variable when the independent variable is equal to zero. In this case, it can be interpreted that a student will get 3.417 GPA if he/she has never skipped class. The beta coefficient of the variable *number of class absence* is -.533. You can see that the beta coefficient takes a negative sign, which means that this variable has a negative effect on the dependent variable. In this case, the more a student is absent from class, the lower GPA he/she will get. Again, the value of the beta coefficient is the rate of change that determines the change in dependent variable when the independent variable changes by 1 unit. Because the value is negative, it can be interpreted that one additional class absence will lower the GPA by .533.

But remember, before you use beta coefficient to make a final conclusion to confirm the effect of the independent variable, you have to consider whether it is statistically significant or not by assessing a p-value. In this case, a p-value of the beta coefficient is equal to .001, which is lower than .05. Thus, we can conclude that the negative effect of class absence on GPA is statistically significant at the below 1 percent level. Class absence is a significant variable that negatively affect GPA.

Making a prediction

If we set up the regression equation by replacing the Y and X with the variables of interest, we will get:

$$\textbf{GPA} = \alpha + \beta \textbf{ Number of class absence}$$

If we replace the values of alpha and beta coefficients with the values we got from SPSS, we will obtain the equation as the following:

$$\textbf{GPA} = 3.417 - .533 \textbf{ Number of class absence}$$

From this regression equation, we can make some prediction about the GPA that a student will get based on how many times he/she is absent. For example, if a student came to class every time (in this case, number of class absence is equal to zero), we can predict that he/she will get GPA = 3.417 - .533 (0) = 3.417.

If a student is absent 1 time (in this case, number of class absence is equal to 1), we can predict that he/she will get GPA = 3.417 - .533 (1) = 2.884.

If a student is absence 2 times (in this case, number of class absence is equal to 2), we can predict that he/she will get GPA = 3.417 - .533 (2) = 1.309 + 1.066 = 2.351.

The interpretations can be summarized as the following:

α = 3.417; β = -.533		
Number of class absence	**GPA**	**GPA changes by**
0	3.417	-
1	2.884	-.533
2	2.351	-.533

Next, let's take a look at the model summary table. The r-square of the regression is equal to .629 or 62.9 percent. This means that number of class absence alone can explain 62.9 percent of GPA that students obtain. There is 37.1 percent (100 - 62.9) that can be explained by other variables that are not included in the model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.793[a]	.629	.598	.42923

a. Predictors: (Constant), Numbers of class absence

<u>Using simple regression to estimate the effect of hours of sleep on GPA</u>

Lastly, let's quickly use simple regression to estimate the effect of sleep on GPA. If you perform the same analysis like what we did with the previous variables, you will get the beta coefficient of hours of sleep per night which is equal .626. A p-value is shown as .000, which is a lot lower than .05. This findings tell us that hours of sleep per night is a variable that positively affect GPA; its positive effect is also statistically significant at the .1 percent level ($p<.001$). Because the beta coefficient is statistically supported, we can conclude that students who got more sleep per night tend to get higher GPA than students who got less sleep per night.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.238	.728		-1.702	.115
	Hours of sleep per night	.626	.115	.844	5.457	.000

a. Dependent Variable: GPA

If you look at the model summary table, you will see that the r-square of the regression is equal to .713 or 71.3 percent. This r-square tells us that hours of sleep per night alone can explain GPA of students by 71.3 percent. There is another 28.7 percent (100 - 71.3) that can be explained by other variables that are not included in the regression model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.844[a]	.713	.689	.37782

a. Predictors: (Constant), Hours of sleep per night

Limitation of simple regression analysis

Although simple regression can be used to estimate the rate of change and the percentage that the dependent variable is explained by the independent variable, the major limitation of simple regression is that we only have one independent variable in the analysis. Due to this limitation, simple regression is not quite different from the correlation analysis that only estimates one-on-one relationship between variables. Therefore, simple regression is not normally used to draw a final conclusion in research.

Let's refer to the simple regression that we performed earlier. We studied the effect of three variables (hour spent on study, number of class absence, and hours of sleep per night) separately by performing three regression analyses. Each regression model contains only one independent variable. When there is only one independent variable in a regression, we ignore the possibility that more than one variable can affect the dependent variable at the same time. For example, referring to one regression analysis that estimated the effect of *number of class absence* on

GPA, we only have one independent variable, *number of class absence*, in the analysis. By having only *number of class absence* variable, we ignore the possibility that some students although they skipped several classes, they may spend more time reviewing class materials and also had enough sleep every night. In reality, *hours spent on study*, *number of class absence*, and *hours of sleep per night* altogether can determine student GPA. Therefore, using simple regression is not a practical way when there is more than one independent variable that can take part in affecting the dependent variable. This limitation can be overcome when using "multiple regression analysis" that we will discuss in the next section.

MULTIPLE REGRESSION

Multiple regression is similar to simple regression. The major difference is that more than one independent variable can be included in the analysis, thereby allowing us to assess the effect of multiple factors simultaneously. When multiple independent variables are analyzed together, we can also tell which independent variables have stronger effect on the dependent variable than others. In particular, multiple regression is a widely accepted technique that is used to perform relationship analysis in research.

Multiple regression can be expressed by the linear equation below:

$$\mathbf{Y} = \alpha + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \beta_3\mathbf{X_3} + \ldots + \beta_n\mathbf{X_n} + \boldsymbol{\varepsilon}$$

whereby:

Y = dependent variable
X_1 = independent variable 1
X_2 = independent variable 2
X_3 = independent variable 3
……
……
X_n = independent variable n

α = alpha coefficient or intercept

β_1 = beta coefficient of independent variable 1
β_2 = beta coefficient of independent variable 2
β_3 = beta coefficient of independent variable 3
……
……
β_n = beta coefficient of independent variable n

ε = error term

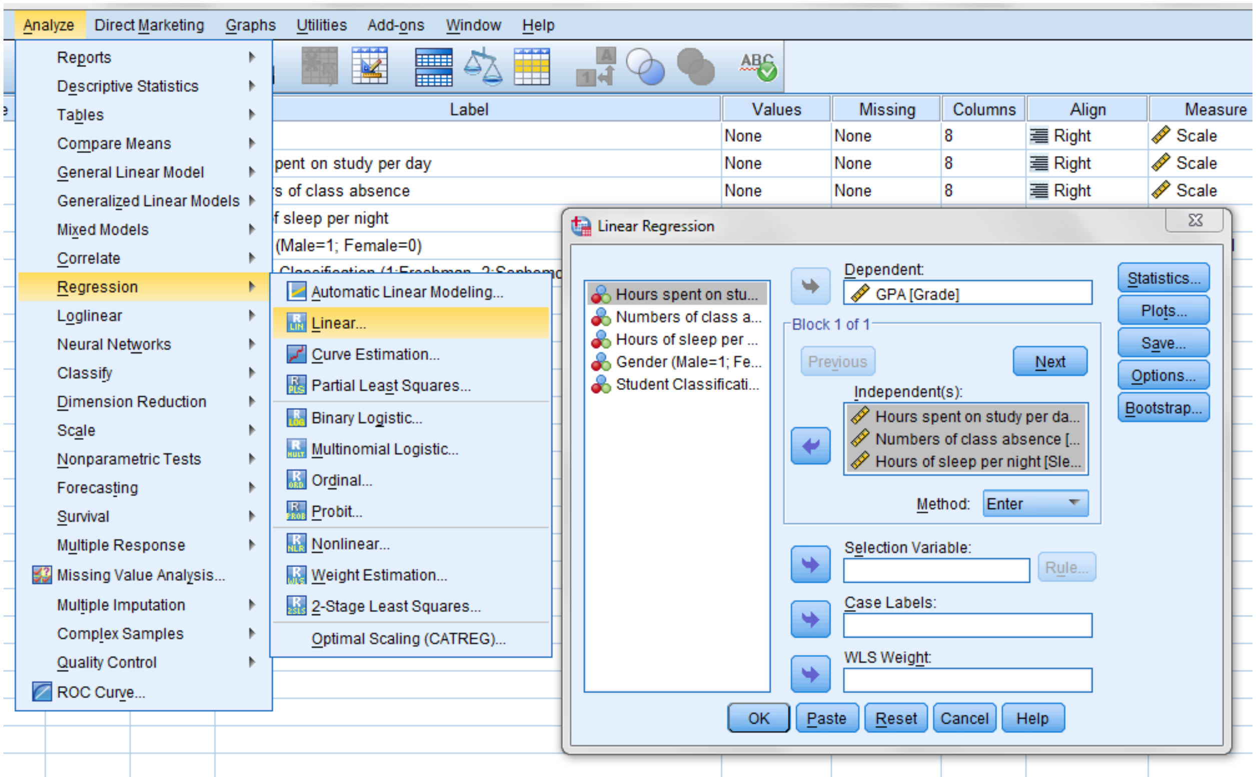
You can see that the linear equation of multiple regression is similar to simple regression. But instead of having one independent variable and one beta coefficient associated with it, there are more than one independent variable and each of them has beta coefficient associated with it as well.

The interpretation of the alpha and beta coefficients in multiple regression is similar to simple regression. The value of alpha coefficient in multiple regression indicates what is the value of the dependent variable when all independent variables are equal to zero. Each beta coefficient is also interpreted similar to simple regression. It is the rate of change that the independent variable causes to the dependent variable. For example, β_1 represents the rate of change that the independent variable X_1 affects the dependent variable; β_2 represents the rate of change that the independent variable X_2 affects the dependent variable, and so on.

Perform multiple regression in SPSS

As mentioned in the beginning, having more than one independent variable in multiple regression analysis allows researchers to compare the effect of all independent variables in the regression to assess which variables have stronger effect on the dependent variable than others. As an example, let's go back to the research case. When using multiple regression, we can include all three variables *hours spent on study*, *number of class absence*, and *hours of sleep per night* together in the analysis.

Now, go back to the place where we performed a simple regression. This time, you include all three variables in the field 'Independent(s)'.



After you put all three independent variables together and run the analysis, you will get the results from coefficients table as the following:

Significant at below 0.1%

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1.590	.735		2.164	.056
	Hours spent on study	.246	.046	.718	5.381	.000
	Numbers of class absence	-.176	.087	-.262	-2.028	.070
	Hours of sleep per night	.046	.119	.062	.389	.706

a. Dependent Variable: GPA

Statistically insignificant

Let's take a look at the beta coefficients and p-values of all three independent variables. Are the results similar to what we got earlier, or something has changed right now?

In particular, the signs of all beta coefficients are consistent with what we got earlier when we performed three simple regressions separately. *Hours spent on study* (β=.246) and *hours of sleep per night* (β=.046) are showed to affect *GPA* positively, while *number of class absence* (β =-.176) is shown to affect GPA negatively. However, when we look at their p-values, we will see that only a p-value of *hours spent on study* is statistically significant at the .1 percent level ($p<.001$). However, p-values of *number of class absence* and *hours of sleep per night* have now become higher than .05, which means that they are not statistically significant anymore. When multiple independent variables are included in regression, the variable that strongly explains the dependent variable can make the variables that weakly explain the dependent variable become less significant. In this case, although *number of class absence* and *hours of sleep per night* were found to significantly affect GPA when each of them were analyzed separately in simple regression, when they are considered together with *hours spent on study* in multiple regression, it appears that *hours spent on study* explain GPA a lot stronger than these two variables do.

So, what can be inferred from these findings? We may conclude that the number of hours that students spent on study tends to be the factor that strongly affects GPA that students get. Although the number of class absence and the number of hours of sleep per night can affect GPA, they may not be important as long as students spend more time studying. In this case, some student may skip classes and did not have a lot of sleep; but if he/she spent more time studying, he/she would get higher GPA as well. Therefore, multiple regression is used in the case like this to compare the effect of each independent variable on the dependent variable.

Anyway, when interpreting the rate of change from beta coefficient in multiple regression, we have to interpret them one-by-one while keeping other variables constant. For example, the beta coefficient of hours spent on study that is equal to .246 can be interpreted that 'all thing being equal (e.g. students have equal amount of class absence and have equal sleep per night), 1 hour spent more in study will cause GPA to increase by .246'. Next, the beta coefficient of class absence that is equal to -.176 can be interpreted that 'all thing being equal (e.g. students have equal hours of study and have equal sleep per night), 1 time increase in class absence will cause GPA to lower by .176'. Finally, the beta coefficient of hours of sleep that is equal to .046 can be interpreted that 'all thing being equal (e.g. students have equal hours of study and have equal amount of class absence), 1 hour increase in sleep per night will cause GPA to increase by .046'. Anyway, don't forget that only hours spent on study is the variable that is statistically significant.

Model Summary

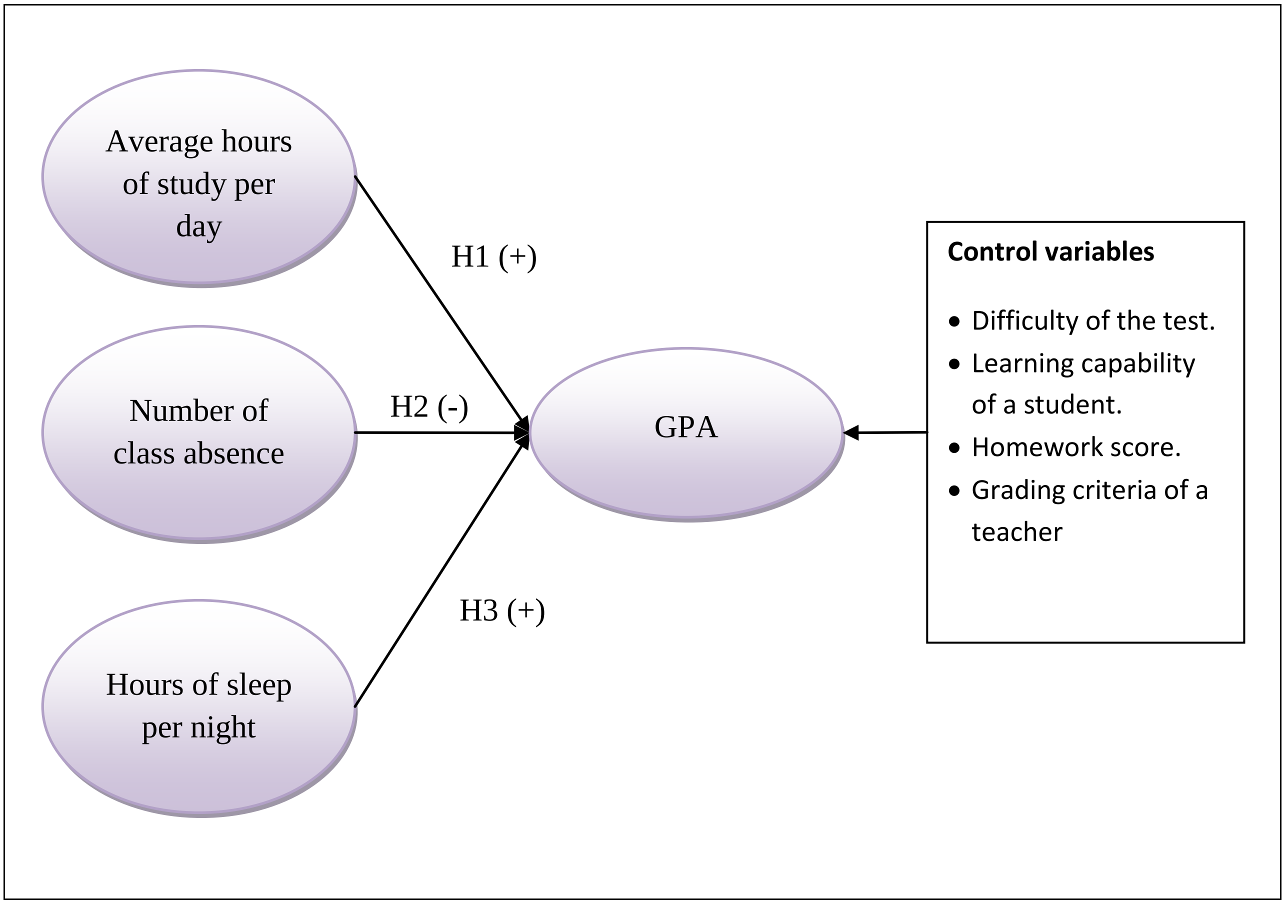
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.968[a]	.937	.919	.19314

a. Predictors: (Constant), Hours of sleep per night, Numbers of class absence, Hours spent on study

Lastly, let's take a look at the model summary table. The r-square is equal to .937 or 93.7 percent. For multiple regression, the r-square is interpreted similar to what we did in simple regression. In this case, it can be interpreted that all independent variables that are included in the regression (*hours spent on study, number of class absence,* and *hours of sleep per night)* can predict or explain GPA by 93.7 percent. Another 6.3 percent (100 – 93.7) may be explained by other variables that are not included in the regression.

Including control variables in regression analysis

In the example research about the determinants of GPA, we hypothesized that there are three variables that might affects GPA. Anyway, beyond these three variables that we analyzed in multiple regression previously, do you think they are enough to predict GPA of students? Do you think there are some other variables that might affect GPA rather than these three variables? In fact, there are uncounted numbers of factors that might influence GPA beyond these three variables. Some examples are the level of difficulty of the test, learning capability of a student, whether a student submitted homework regularly, how tough a teacher is, etc. In particular, other variables that might affect the dependent variables in addition to the independent variables that we declare in a hypothesis are called *control variables*. Indeed, control variables can be regarded as independent variables in regression because they are expected to have some effect on the dependent variable as well.



In regression analysis, it is necessary that we have to put control variables in addition to the main independent variables that we declared in the hypotheses. The main reason why we have to include control variables is to ensure that the effect of the independent variable is still significant even though other factors that might influence the dependent variables are included. When the effect of the main independent variable in a hypothesis loses its statistical significance after control variables are included, this means that control variables tend to explain the dependent variable more than the main variable in the hypothesis does. It may be possible that the significant effect of the main independent that we obtained without control variables is just spurious. In fact, we already discussed about this purpose of control variables when we discussed about criteria to justify causality in the early chapter. Thus, when the effect of the main independent that used to be statistically significant appears to lose its significance after control variables are included, the hypothesis cannot be supported. In order to support a hypothesis, it has to be statistically significant even though control variables are put in regression.

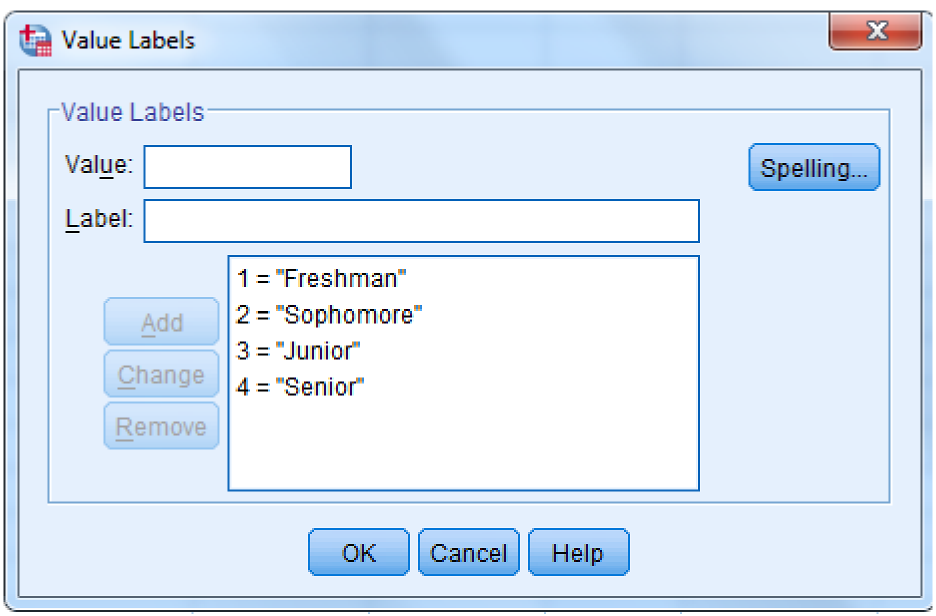
Anyway, someone may have questions regarding how many control variables should be put in regression analysis and what should be used as control variables. The first question is quite difficult to tell. You may use control variables as many as you can find them; but too much will overload the regression analysis and subsequently causes the estimation to be biased. Remember that control variables are regarded as additional independent variables in regression analysis. In order to have more independent variables, it is important that your sample size must be large enough to be proportionate to it. About the second question what should be used as control variables, the best way to justify what should be used is to look at existing literature to see what are factors that were used as control variables in those studies. In behaviors sciences research, for example, demographic characteristics such as age, gender, education, etc. are normally used as part of control variables because these demographic characteristics tend influence how people think, feel, and behave.

Using ranked variable and dummy variable in regression analysis

So far, the data that we used in regression analysis (*hours spent on study*, *number of class absence*, and *hours of sleep per night*) are ratio data. If you recall from the previous chapter, ratio data can take any value and we can calculate actual distance between them. For example, a student who skipped class 3 times has more class absence than a student who skipped class 1 time by 3-1 = 2 times. When ratio data are used in regression analysis, we can interpret the rate of change that the independent variable influences the dependent variable by using the actual unit of measurement like what we did earlier. For example, we can say that 1 hour increase in study increases GPA by how may point. However, when the data that are not measured on a ratio scale are used in regression, they will require different interpretation. If you remember from the previous chapter on scaling, we also have data that are measured on nominal scale and ordinal scale. In the next section, we will focus on how to interpret these two types of data in regression analysis. At this point, you need to know that a variable that is measured on nominal scale can also be called a dummy variable; a variable that is measured on ordinal scale can also be called a ranked variable. We will focus on them one-by-one. Let's start with ranked variable.

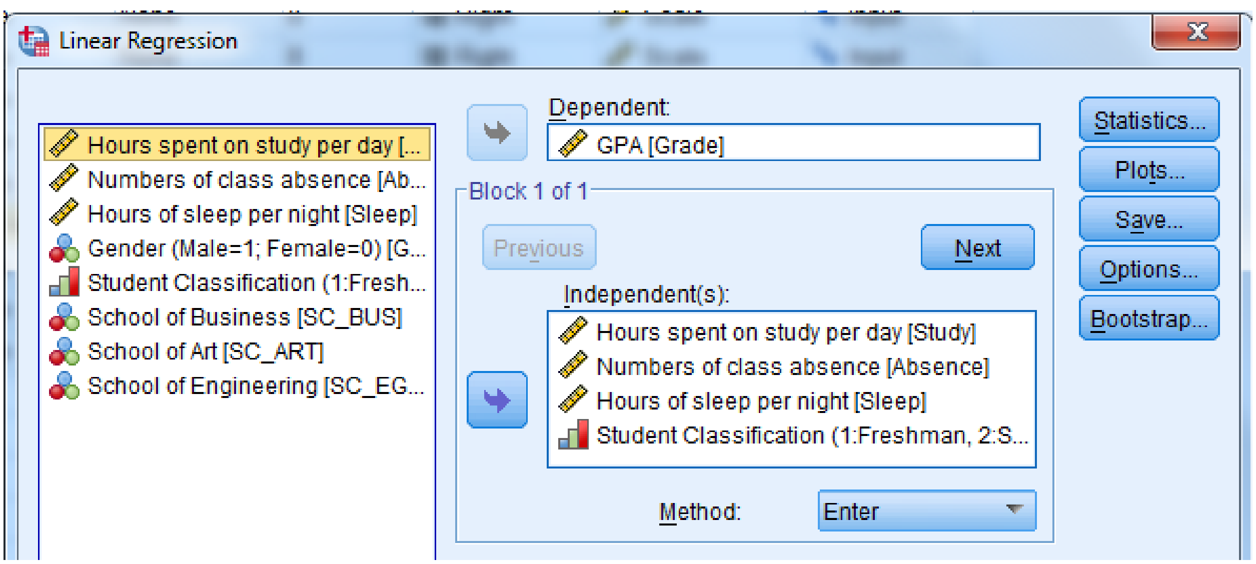
Interpreting ranked variable in regression analysis

A *ranked variable* (also called *ordinal-scale variable*) is a variable for which the individual observations can be put in order from smallest to largest. The example of ranked variable that we have in the dataset is the variable named "classification". This variable represents classification of undergraduate students who are freshmen (first-year students), sophomores (second-year students), juniors (third-year students), and seniors (last-year students). This variable is measured on ordinal scale because they can be ranked from low to high. Freshman is ranked lower than sophomore; sophomore is ranked lower than junior; junior is ranked lower than senior. Due to the clear order of ranking, this variable can be coded in number whereby the lowest-ranked students get the lowest score and the highest-ranked students get the highest score.



Now, in addition to the three main independent variables that we have in the hypotheses, we will add student classification into regression analysis as a control variable. But anyway, why student classification can be considered a control variable in this analysis. Do you think GPA of student can be explained by whether a student is a freshman, a sophomore, a junior, or a senior? Well, it could be possible. The effect of this variable is worth exploring. When students first went to college as freshmen, they may spend a majority of their free time enjoying college life, having fun, partying with new friends, attending college activities, etc. These students may not focus well on study during their first year in college and may do not care much about the GPA. However, as they spent more years in college and became junior or senior, they tended to be more mature and began to realize that they had to study hard to improve their GPA in order to get good jobs when they graduate.

Now let's try to add the variable that measure student classification in regression analysis to estimate whether it has some influence on GPA or not.



When you perform the analysis, you will get the results in the coefficients table as the following:

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1.467	.617		2.379	.041
	Hours spent on study per day	.184	.047	.539	3.960	.003
	Numbers of class absence	-.090	.081	-.134	-1.107	.297
	Hours of sleep per night	.007	.101	.010	.074	.943
	Student Classification (1: Freshman, 2:Sophomore, 3:Jonior, 4:Senior)	.211	.092	.361	2.303	.047

a. Dependent Variable: GPA

The beta coefficient of student classification is positive (β=.211). This coefficient is also *statistically supported* because its p-value of .047 is lower than .05; it is statistically significant at the 5 percent level. When using ranked variable in regression analysis, the positive beta coefficient means that the value of the dependent variable tend to be greater for subjects who are classified in the higher rank than subjects who are classified in the lower rank. Thus, the positive beta coefficient of student classification in this case means that 'students who are in higher classifications tend to have higher GPA than students who are in lower classifications.

Anyway, just in case that the beta coefficient turn to be negative, the interpretation has to be make oppositely. For example, if the beta coefficient of student classification was negative, the interpretation would be 'students who are in higher classifications tend to have lower GPA than students who are in lower classifications'.

Still, please note that when using a ranked variable in regression analysis, it can only tell whether the groups in the higher ranks tend to be more or less than the groups in the lower ranks in the factor that is measured by the dependent variable. From the result that we obtained earlier, it only tells that students who are in higher classifications tend to have higher GPA than students who are in lower classifications, but it can't tell specifically which group is the first, second, third, and fourth in terms of GPA. If you need specific detail about the comparison of each category in the one-on-one basis, you need to perform ANOVA test that we performs in the previous chapter to obtain this information.

Interpreting dummy variable in regression analysis

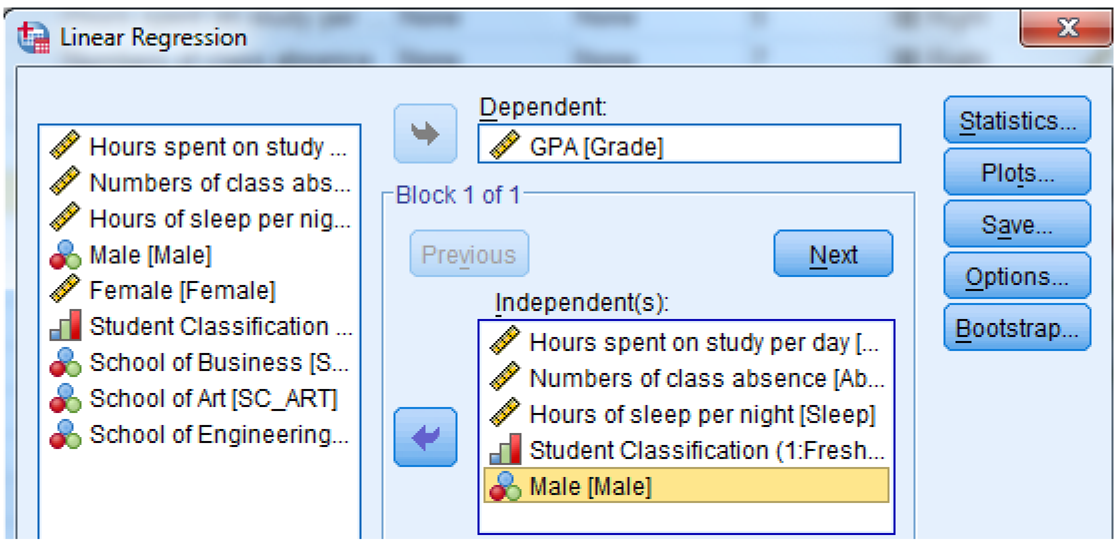
In the chapter on measurement and scaling concepts, we already mentioned about a dummy variable. Again, a *dummy variable* is a "*nominal-scale variable*" that takes the value "0" or "1" to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. The value of 0 means the absence of categorical effect, whereas the value of 1 means the presence of the categorical effect. For example, when measuring gender and we use male as a dummy variable, if a person is male we code 1 (yes: he is male); but if a person is female, we code 0 (no: she is not male). Because the main characteristic of a nominal scale is that we can't compare whether one category is better/worse or can be ranked higher/lower than another category, it is important to code the variable as 0 and 1 instead of ranking them numerically. Thus, each category must be represented by its own dummy variable. The examples of dummy variables that measure gender are coded as the following.

Gender	**Male dummy variable**	**Female dummy variable**
Male	1	0
Male	1	0
Female	0	1
Female	0	1
Female	0	1

There is one important rule that you have to follow when you put a dummy variable in regression analysis. The number of dummy variable that we put in regression is equal to the total number of category minus 1. For example, if you have two categories, we only put one category in regression. If we have three categories, we only put two categories in regression. The beta coefficient of the dummy variable that represent the category that you put in regression will be used to compare whether that category has higher or lower than the category that is not put in regression on the factor that is measured by the dependent variable.

Dummy variables of 2 categories

To help you understand more about how to analyze dummy variables in regression analysis, let's try it with some nominal-scale variable that we have in the dataset. We have 'gender' as a nominal-scale variable here. For gender, we know that by nature we have only two classifications which are male and female; so there are two categories. In the dataset, we have 2 dummy variables; one for male and one for female. In regression, the number of dummy variable that you can put in the analysis has to be the total number of category minus 1. In this case, there are 2 categories; so you have to decide whether you want to put male or female into regression. Here, let's put 'male' into regression.



Before we see the regression results, here is how to interpret the coefficient of the dummy variable. Positive coefficient of the dummy variable means that the average value of the dependent variable of the category that is included in regression is higher than the category that is not included in the regression. Conversely, negative coefficient of the dummy variable means that the average value of the dependent variable of the category that is included in regression is higher than the average value of the dependent variable of the category that is not included in the regression.

Approximated GPA of female

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1.412	.606		2.331	.048
	Hours spent on study per day	.170	.047	.496	3.597	.007
	Numbers of class absence	-.123	.085	-.183	-1.453	.184
	Hours of sleep per night	.017	.099	.024	.177	.864
	Student Classification (1: Freshman, 2:Sophomore, 3:Jonior, 4:Senior)	.229	.091	.391	2.512	.036
	Male	.127	.108	.096	1.173	.275

a. Dependent Variable: GPA

Positive beta means that male has higher GPA than female by .127

Not statistically significant

From the coefficients table, you can see that the coefficient of the male dummy variable is positive (β = .127). This can be interpreted that male tends to have higher GPA than female on average. In fact, the approximated value of GPA of female can be inferred from the alpha coefficient of the regression. Anyway, please note that the value of alpha coefficient is not the actual average GPA of female when there is other independent variable in regression; it is just an approximated value that we only use for comparison. If female has approximated GPA of 1.412, male will have approximated GPA of 1.412 + .127 = 1.539, which is higher than female. If you want to know the exact average GPA of male and female, you have to have only male dummy variable in regression analysis without other independent variables; in this case, the value of alpha coefficient will represent the real average GPA of female.

Now we know that the positive beta coefficient of male dummy variable tells us that male has higher GPA than female on average. Nonetheless, just looking at the coefficient is not enough to be confident in the result. You have to look at a p-value of the coefficient before you can make a final conclusion. From the table you can see that a p-value of this dummy variable is equal to .275 which is higher than .05; thus, it is not statistically significant. This suggests that although male tends to have higher GPA than female on average, this finding is not statistically supported. We can't conclude by statistics that male tends to have higher GPA than female.

You can also try whether the findings will be similar if we use female dummy variable instead of male dummy variable. You can try replacing male dummy variable with female dummy variable. From the coefficient table, we can see that the coefficient value of female dummy variable is similar to when we used male dummy variable, except for that it has a negative sign (β = -.127). When the coefficient of the dummy variable is negative, the interpretation is that the average value of the dependent variable of the category that is included in regression is higher than the average value of the dependent variable of the category that is not included in the regression. In this case, negative coefficient of female dummy variable can be interpreted that female tends to have lower GPA on average than male. However, because the p-value is equal to .275 which is higher than .05, we have to conclude that this finding is not statistically supported. In fact, when you have only 2 category-dummy variables (like male and female), it does not make any major different whether you will include or omit which category. The only difference is just the sign of the beta coefficient that you have to interpret it oppositely.

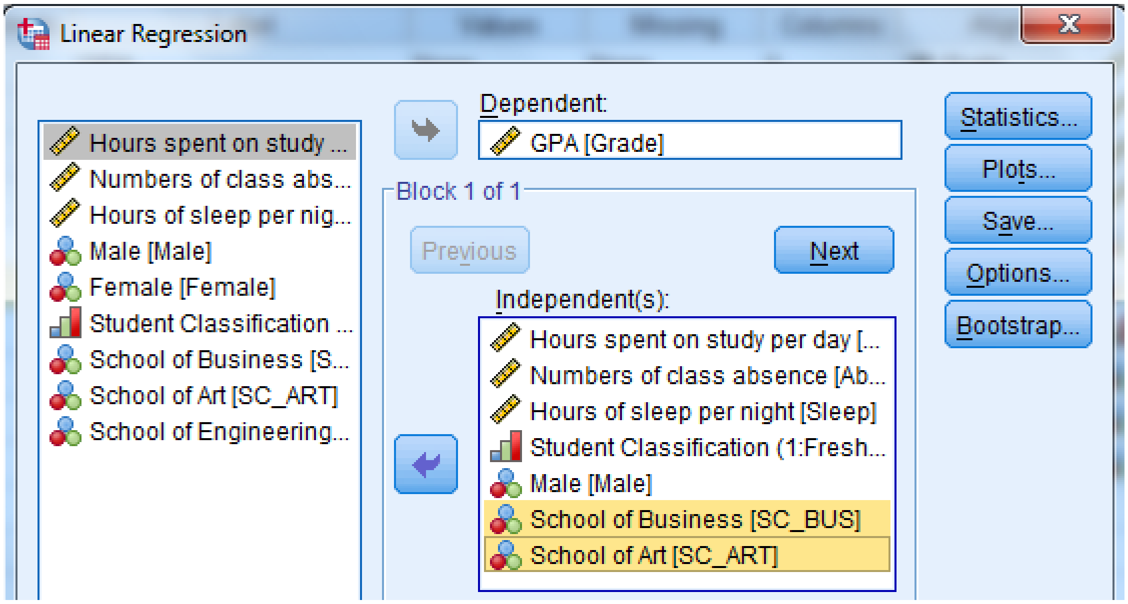
Dummy variables more than 2 categories

In the previous example we analyzed and interpreted dummy variable that contain two categories (male and female). Now let's have some example of how to analyze and interpret results when we work with dummy variables of more than two categories. In particular, the student sample we have came from three departments including school of engineer, school of business, and school of art. These are nominal data because we can't compare whether which department is better than another. Therefore, we need to have separated dummy variable for each of these three departments. For example:

Department	Engineering dummy variable	Business dummy variable	Art dummy variable
Engineering	**1**	0	0
Engineering	**1**	0	0
Engineering	**1**	0	0
Business	0	**1**	0
Business	0	**1**	0
Business	0	1	0
Art	0	0	**1**
Art	0	0	**1**
Art	0	0	**1**

But before we move to the analysis, let's ask yourself first do you think GPA that students achieve can be explained by the department that students are in? Do you think being in engineering, business, or art department can affect GPA of students? This may be quite difficult to justify. But the author thinks that being in the field of engineering may be more difficult to get good grade than being in the field of business and art, because their subjects are related to pure sciences which require extensive mathematical and analytical skills. Anyway, it is just a prediction. Let's see what the regression analysis will show us.

Remember the rule when we work with dummy variable. The number of dummy variable that you can put in the analysis has to be the total number of category minus 1. Because there are three departments in the dataset, we can only insert two of them in regression analysis. In this example, we will put only business dummy variable and art dummy variable in regression analysis. We will leave engineering dummy variable for comparison after we get the beta coefficients of business and art dummy variables from regression results.



After you perform regression analysis, you will get results from coefficients table as the following:

Appoximated GPA of students in engineering department

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1.385	.689		2.009	.091
	Hours spent on study per day	.176	.054	.514	3.244	.018
	Numbers of class absence	-.120	.095	-.179	-1.273	.250
	Hours of sleep per night	.018	.113	.024	.160	.878
	Student Classification (1: Freshman, 2:Sophomore, 3:Jonior, 4:Senior)	.196	.118	.336	1.665	.147
	Male	.140	.126	.106	1.107	.311
	School of Business	.114	.203	.072	.562	.595
	School of Art	.088	.144	.067	.611	.564

a. Dependent Variable: GPA

Positive beta coefficients

Not statistically significant

Their GPAs are higher than students in engineering department

The interpretation of dummy variable in this case is similar to what we did earlier. We use the sign and value of beta coefficient of the dummy variables in regression to assess whether GPAs of students in these two departments is higher or lower than GPA of students in engineering department that we did not include in regression. Anyway, the approximated GPA of the department that we omitted can be inferred from the alpha coefficient of the regression, which is 1.385. We can see that the beta coefficient of business dummy variable is .114 which is positive. This suggests that students in business department tended to have higher GPA than students in engineering department approximately by .114. Similarly, the beta coefficient of art dummy variable is .088 which is also positive. This suggests that students in art department tended to have higher GPA than students in engineering department approximately by .088. However, none of their p-values are lower than .05, which means that although the results we got showed that students in business and art department tended to have higher GPA than students in engineering department, these findings are not statistically supported. We cannot conclude that being in each department affect GPA of students.

Anyway, if we still want to rank the GPA of students in these three departments based on the beta coefficients that we obtained from regression analysis, we can rank them from the lowest to the highest by calculating from the alpha and beta coefficients, which yields the results as the following:

Department	Approximated GPA (for comparison only)
Engineering	1.385
Art	1.385 + .088 = 1.473
Business	1.385 + .114 = 1.499

Here, it is obvious that students who are in business department tend to get the highest GPA, following by students in art department, and students in engineering department. Again, these findings are just the approximated GPA, not the actual GPA that students in each department obtain on average. Still, it is consistent with what the author predicted that being in engineering department is more difficult to get good grade as comparing to being in business and art department. However, because the beta coefficients are not statistically significant, these ranking can't be used to make a final conclusion anyway.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.984[a]	.969	.932	.17681

a. Predictors: (Constant), School of Art, Male, Hours spent on study per day, School of Business, Numbers of class absence, Hours of sleep per night, Student Classification (1:Freshman, 2:Sophomore, 3:Jonior, 4:Senior)

Finally, if we take a look at the model summary table, now it shows that the r-square of the regression is equal to .969 or 96.9 percent, which is very high and almost close to 100 percent. This r-square can be interpreted that all variables that are included in the regression analysis can explain 96.9 percent of GPA.

Dummy variable regression

Earlier, we know the average GPA of students in each department is just the approximated values. But if we want to get the 'actual' average GPA of students in three departments, we have to put only the dummy variables in the regression analysis. This technique is call *dummy variable regression*. Using the same rule, we omitted engineering dummy variable from the regression and put only business and art dummy variables. The results from regression are shown as the following:

Actual average GPA of students in engineering department

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.048	.284		7.200	.000
	School of Business	1.026	.434	.645	2.362	.038
	School of Art	.854	.356	.654	2.395	.036

a. Dependent Variable: GPA

Actual differences in average GPA

Statistically supported

When only dummy variables are included in the regression, the alpha coefficient is the actual mean value of the category in which its dummy variable is not included. In this case, the alpha coefficient of **2.048** is the actual average GPA of students in engineering. The actual average GPA of students in business department can be determined by the beta coefficient of its dummy variable. Its value is positive and is equal to 1.026, which means that students in business department get higher GPA than students in engineering department by 2.048 + 1.026 = **3.074**. Similarly, the actual average GPA of students in art department can be determined by the beta coefficient of its dummy variable. Its value is positive and is equal to .854, which means that students in art department get higher GPA than students in engineering department by 2.048 + .854 = **2.902**. Their p-values are all statistically significant at the 5 percent level. These findings statistically confirm that GPA is affected by the department where students are in. In particular, business department got the highest GPA, following by art department, and lastly, engineering department.

Unstandardized beta coefficient VS standardized beta coefficient

From what we have interpreted from regression results so far, we use unstandardized beta coefficient for interpretation. When using unstandardized beta coefficient, the rate of change that the independent variable influences the dependent variable is expressed in actual unit. However, for standardized beta coefficient, the rate of change of every independent variable is standardized to transform them into the same unit of measurment. When using standardized beta coefficient, the rate of change should be interpreted in 'standard deviation' unit. For example, let's look at the standardized beta coefficient of *hours spent on study*. The standardized coefficient of .514 should be interpreted that '1 standard deviation increase in hour spent on study will cause GPA to increase by .514 standard deviation'.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.385	.689		2.009	.091
	Hours spent on study per day	.176	.054	.514	3.244	.018
	Numbers of class absence	-.120	.095	-.179	-1.273	.250
	Hours of sleep per night	.018	.113	.024	.160	.878
	Male	.140	.126	.106	1.107	.311
	Student Classification (1:Freshman, 2:Sophomore, 3:Jonior, 4:Senior)	.196	.118	.336	1.665	.147
	School of Business	.114	.203	.072	.562	.595
	School of Art	.088	.144	.067	.611	.564

a. Dependent Variable: GPA

Because unit of measurement of the independent variables are transformed into the same unit, standardized beta coefficient is more reliable when we compare the rate of change of the independent variables that are measured in different units. Using standardized beta coefficients is also particularly useful when we analyze the rate of change of the attitude variables that are measured on a Likert scale. For example, when we use questionnaire to collect data about work attitudes of employees that are measured from 1 (strongly disagree) to 5 (strongly agree), the unit of measurement tends to be unclear to expressed. By using standardized beta coefficient we can use standardized deviation as a unit of measurement to represent the rate of change between the attitude measures. For example, we can say that one standard deviation increase in one attitude measure will cause another attitude measure to change by how many standard deviations.

From the table above, if we want to compare the effect of all independent variables on GPA to see which variables tend to have the strongest effect, we should compare the standardized coefficients. Obviously, hours spent on study is the variable that has the strongest effect on GPA because its standardized beta coefficient of .514 is the highest among all independent variables'. Furthermore, it is also statistically significant, making it the most influential variable that affects GPA. Although student classification which has standardized beta coefficient of .336 and number of class absence which has standardized beta coefficient of -.179 can be regarded as the second and the third strongest variables that affect GPA, we cannot confirm their effect because both of them are not statistically significant.

Multicollinearity

In multiple regression, the independent variables are expected to have a strong relationship with the dependent variable. However, the relationship among independent variables in regression must not be highly significant; otherwise, *multicollinearity* problem will bias the regression results. Let's consider this regression equation that contains three independent variables:

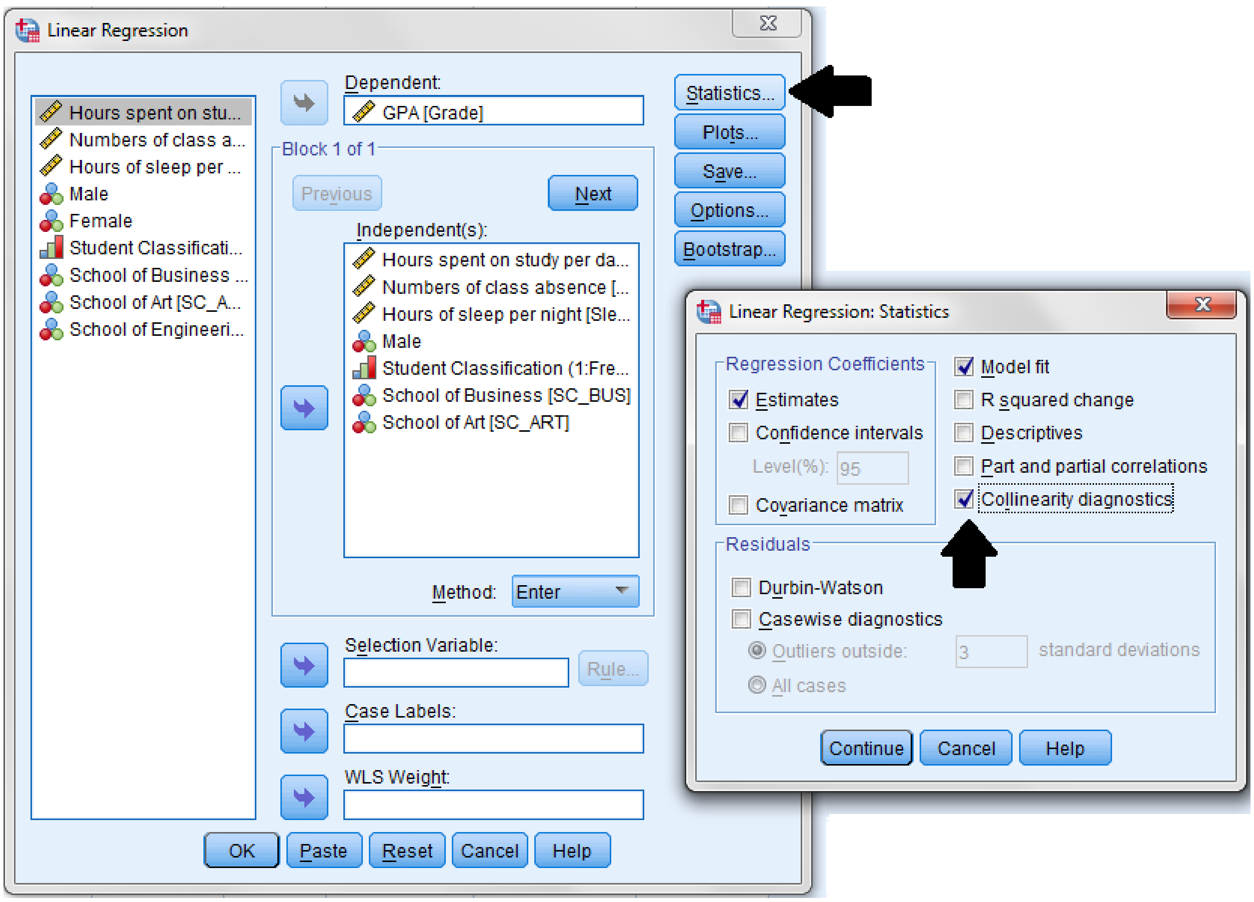
$$Y = \alpha + \beta_1 X1 + \beta_2 X2 + \beta_3 X3$$

We expect X1, X2, and X3 to have strong relationship with Y because all of them are predicted to affect Y. However, X1, X2, and X3 must not have strong correlation with one another; otherwise, multicollinearity problem will tamper with the estimation of beta coefficients that are obtained from regression analysis. In particular, multicollinearity can alter the significant level or even change the sign of the beta coefficients.

So how we can detect multicollinearity? In multiple regression, *the variance inflation factor (VIF)* is widely used as an indicator to detect multicollinearity in the analysis. Several scholars set the maximum value that VIF should not exceed. For example Hair et al. (2009) recommended that VIF should be lower than 10;

Rogerson (2001) argued that VIF should be lower than 5; Pan and Jackson (2008) suggested that it should be lower than 4. Anyway, the threshold of VIF that is used to justify the level of multicollinearity is up to the researchers. But despite different justifications, lower VIFs provide more confidence that the results from regression analysis are not susceptible to multicollinearity problem.

The estimation of multicollinearity using VIF in regression analysis can be performed in SPSS as the following:



VIF indicators can be seen at the last column of the coefficients table.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	1.385	.689		2.009	.091		
	Hours spent on study per day	.176	.054	.514	3.244	.018	.209	4.795
	Numbers of class absence	-.120	.095	-.179	-1.273	.250	.265	3.776
	Hours of sleep per night	.018	.113	.024	.160	.878	.226	4.433
	Male	.140	.126	.106	1.107	.311	.573	1.746
	Student Classification (1:Freshman, 2:Sophomore, 3:Jonior, 4:Senior)	.196	.118	.336	1.665	.147	.129	7.747
	School of Business	.114	.203	.072	.562	.595	.323	3.098
	School of Art	.088	.144	.067	.611	.564	.433	2.311

a. Dependent Variable: GPA

As you can see, the maximum VIF in the regression analysis is 7.747. Although it is higher than the maximum threshold of 5 as suggested by Rogerson (2001), it is still lower than 10 which is the still satisfactory as suggested by Hair et al. (2009). Therefore, we may conclude based on this finding that multicollinearity may not be a serious issue in the analysis.

Reporting and summarizing regression results

So far, the final regression that we analyzed not only covers the main independent variables, but it also contains control variables as well. The overall regression results can be used to summarize the research findings. As we can see, although more control variables are included in the regression, the positive effect of the independent variable *hours spent on study* is still statistically significant at the 5 percent level (p-value=.018). We can confirm that it is the variable that predicts GPA of students.

When reporting regression results in the paper, we can begin by repeating the hypothesis and then report the value and the significant level of the beta coefficient. After that, we conclude whether the hypothesis is supported or not. For example:

- Hypothesis 1 predicts a positive relationship between hours spent of study and GPA. The standardized beta coefficient is positive and statistically significant (β=.514; p=.018). Thus, hypothesis 1 is supported.

- Hypothesis 2 predicts a negative relationship between number of class absence and GPA. Although the standardized beta coefficient is negative as predicted, it is not statistically significant (β=-.179; p=.25). Thus, hypothesis 2 is not supported.

- Hypothesis 3 predicts a positive relationship between hours of sleep and GPA. Although the standardized beta coefficient is positive as predicted, it is not statistically significant (β=.024; p=.878). Thus, hypothesis 3 is not supported.

In addition to writing your results report, you may also summarize your statistical findings in table or in conceptual model. Anyway, there is no fixed rule about how you should summarize your findings. If you choose to summarize your results in table, you may follow the table below. In particular, you need to report beta coefficients of all independent variables, their significant level (you can use * to represent them), r-square and adjusted r-square, and number of observations. You may report VIFs to show that your results do not suffer significantly from multicollinearity problem. For beta coefficients, it is up to you whether you want to report standardized or unstandardized beta coefficients. However, whatever you choose you must mention it in the table or figure.

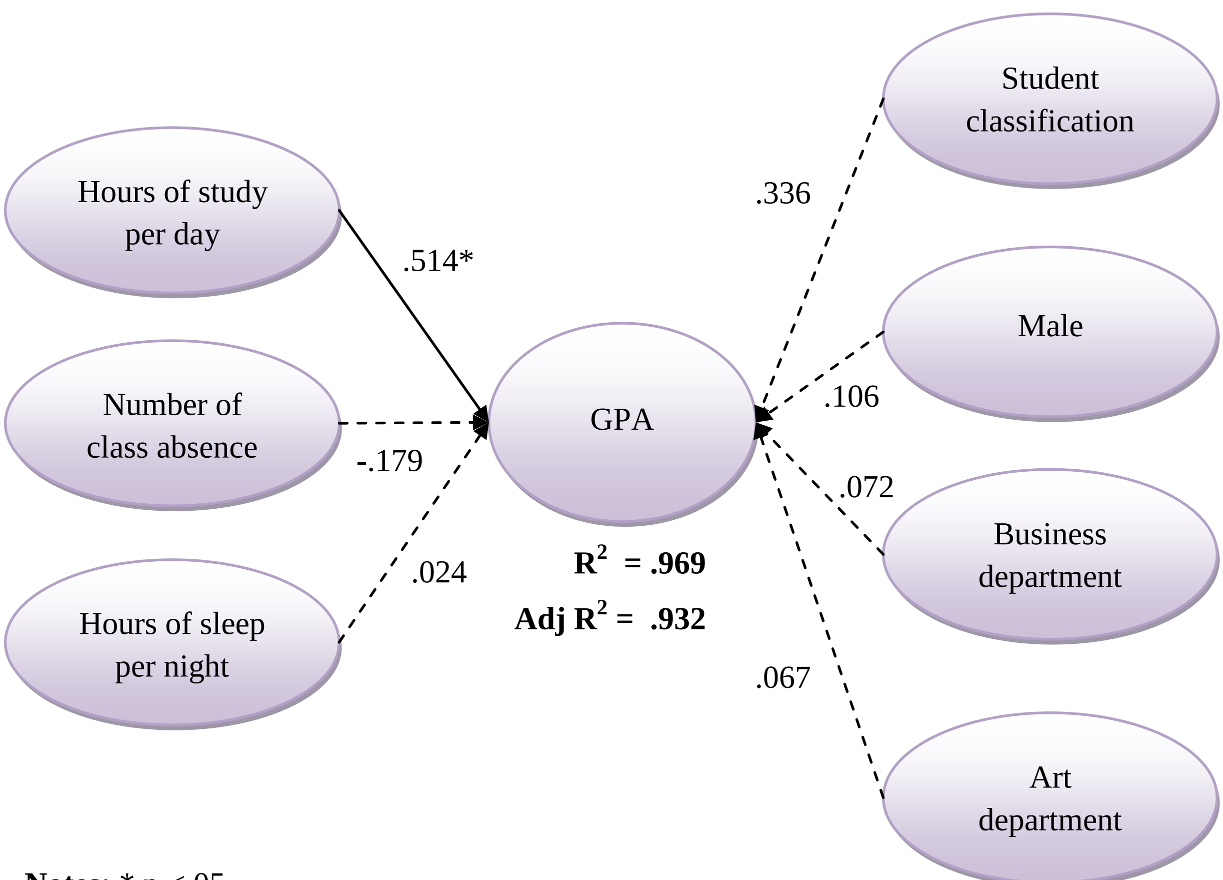
Regression results

Dependent variable is GPA	Standardized beta coefficient	VIF
Hypothesized variable		
Hours spent of study per day (H1)	.514*	4.795
Number of class absence (H2)	-.179	3.776
Hours of sleep per night (H3)	.024	4.433
Control variable		
Student classification	.336	1.746
Male dummy variable	.106	7.747
Business dummy variable	.072	3.098
Art dummy variable	.067	2.311
R-square	.969	
Adjusted r-square	.932	
Number of observations	14	

Note: * p <.05

In addition to summarizing your findings in table, you may summarize them in the conceptual model as the following. R-square and adjusted r-square should be reported close to the dependent variable. To make it easy for readers to recognize which variables in regression analysis is statistically supported, some authors also use dash lines to represent the relationship that is statistically significant and use the solid line to represent the relationship that is not statistically significant. Again, there is no fixed rule about how you should summarize your findings in conceptual model. It depends on your preferences also.

Regression results



Notes: * $p < .05$,
Standardized beta coefficients are reported,
Solid lines represent statistically significant results,
Number of observations = 14

<u>Example regression analysis: Factors affecting team performance</u>

Let's have some more example of how to interpret results from regression analysis. In this study, the researcher hypothesized that there are 6 main independent variables that might affect the dependent variable which is "<u>team performance</u>" of Master's students and doctoral students. Six independent variables, which were measured on Likert's scale, include the following:

(1) Teamwork
(2) Training
(3) Lack of communication
(4) Shared goal
(5) Team competition
(6) Trust among team members.

In addition to these 6 variables, the researcher also incorporated other '**control variables**' that might affect team performance as well. These control variables include:

- Number of team member (measured on ratio scale)
- Team age (measured on ratio scale)
- Whether a team has a leader (measured as a dummy variable)
    - yes = 1
    - no = 0
- Number of team meeting per week (measured on ratio scale)
- Student classification (measured on ordinal scale)
    - Master's degree was coded 1.
    - Doctoral degree was coded 2.
- Finally, the researcher also control for the academic major of students including accounting, international business, and marketing. Student major are coded as dummy variables.
    - Accounting major dummy variable (yes=1; no=0)
    - International business major dummy variable (yes=1; no=0)
    - Marketing major dummy variable (yes=1; no=0)

The followings are regression outputs from SPSS:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2.461	.417		5.905	.000		
	Teamwork	.174	.083	.184	2.087	.038	.429	2.332
	Training	.093	.087	.089	1.059	.291	.468	2.136
	Lack of communication	-.209	.064	-.218	-3.255	.001	.742	1.348
	Shared goal	.058	.069	.061	.836	.404	.615	1.625
	Competition	-.033	.052	-.040	-.640	.523	.846	1.182
	Trust	.020	.009	.164	2.326	.021	.672	1.489
	Number of members	.004	.024	.010	.163	.871	.847	1.181
	Team age	.137	.097	.092	1.415	.159	.787	1.271
	Does a team has a leader (1=Yes; 0=No)	.162	.131	.089	1.239	.217	.648	1.542
	Number of team meeting per week	-.042	.042	-.059	-.997	.320	.940	1.064
	Student classification	-.022	.088	-.015	-.249	.804	.926	1.080
	Major: International business	.530	.294	.111	1.801	.073	.884	1.131
	Major: Marketing	.743	.206	.217	3.599	.000	.918	1.090

a. Dependent Variable: Team performance

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.604[a]	.364	.321	.54826

a. Predictors: (Constant), Major: Marketing, Trust, Student classification, Training, Major: International business, Number of team meeting per week, Number of members, Competition, Team age, Lack of communication, Shared goal, Does a team has a leader (1=Yes; 0=No), Teamwork

What you need to answer from the regression results are the following:

1. Interpret the effect (rate of change) that each independent variable has on team performance. Also, you need to indicate whether the effect of each independent variable is statistically supported or not.
2. Team performance is significantly strong for students in which major(s)?
3. Among <u>6 main independent variables,</u> you have to rank which variables "<u>significantly</u>" affect team performance from the highest to the lowest (Note: Don't mention control variables).
4. How many percent that the regression can explain team performance?
5. Is multicollinearity a serious problem in the regression analysis?
6. Based on the regression results, what are practical recommendations that should be provided to help improving team performance?

Interpreting standardized beta coefficients

Because the majority of the variables are measured on Likert's scale and because we want to compare the effects of all independent variables to see which ones significantly affect the dependent variable and to rank them from the highest to the lowest, the standardized beta coefficients are appropriate in this case. By determining the standardized beta coefficients as well as their p-values, the interpretation can be made as the following:

- If *teamwork* score increases by 1 standard deviation, team performance will also increase by .184 standard deviations. This relationship is supported statistically.

- If *training* score decreases by 1 standard deviation, team performance will increase by .089 standard deviations. However, this relationship is not supported statistically.

- If *lack of communication* score increases by 1 standard deviation, team performance will decrease by .218 standard deviations. This relationship is also supported statistically.

- If *shared goal* score decreases by 1 standard deviation, team performance will also decrease by .061 standard deviations. However, this relationship is not supported statistically.

- If *team competition* score increases by 1 standard deviation, team performance will decrease by .04 standard deviations. However, this relationship is not supported statistically.

- If *trust* score decreases by 1 standard deviation, team performance will increase by .164 standard deviations. This relationship is supported statistically.

- Team performance tended to be higher in larger teams than in smaller teams. However, this finding is not supported statistically.

- Team performance tended to be higher in teams that have been established longer. However, this finding is not supported statistically.

- Team performance tended to be higher in teams that have a leader than in teams that don't have a leader. However, this finding is not supported statistically.

- Team performance tended to be lower in teams that have more meeting per week. However, this finding is not supported statistically.

- Teams whose members are Master's students tended to have higher team performance than teams whose members are Doctoral students. However, this finding is not supported statistically.

- To analyze whether team performance tend to vary significantly among three academic major (which is measured as dummy variables), the first major 'accounting' was omitted from regression analysis. The standardized beta of international business major dummy variable is positive (β=.111), which means that on average team performance of students in international business major tend to be higher than team performance of students in accounting major by .111 standard deviations. However, a p-value of international business major dummy variable is not statistically significant (p=.073). Thus, we *cannot* conclude that the difference in team performance is statistically supported.

  Next, the standardized beta of marketing major is positive (β=.217), which means that on average tem performance of students in marketing major tend to be higher than team performance of students in accounting major by .217 standard deviations. A p-value of marketing major dummy variable is also statistically significant (p<.001). Thus, we can conclude that the difference in team performance is statistically supported.

  If we want to use regression results to compare average team performance among students in three major, we can estimate their score based on the beta coefficients as the following:

Account major (omitted in regression):	X
International business major	X + .111
Marketing major	X + .217

This means that students in marketing major tend to have the highest team performance, following by students in international business major, and students in accounting. However, because only the beta of marketing major dummy variable is statistically significant, we can conclude that being in marketing major is actually matter for students to get higher team performance.

Comparing standardized beta coefficients

By considering all standardized beta coefficients and p-values in regression results, you can detect which variables significantly explain team performance by considering the value of the standardized beta coefficient of the variables that have a p-value lower than .5. From what we saw in the results, among 6 main independent variables, there are only 3 main independent variables that are statistically significant. If we rank them from the highest to the lowest based on the value of standardized beta coefficient, they can be ranked as the following:

Rank	Variable	Standardized beta coefficient	p-value
1	Lack of communication	-.218	.001
2	Teamwork	.184	.038
3	Trust	.164	.021

Suggesting practical implications

Practical implications that are made from regression analysis should only focus on the main independent variables that are statistically supported. We can see that lack of communication is the strongest variable that affects team performance, following by team work, and trust. From the finding we may suggest that in order to improve term performance, it is particularly important that communication among team members need to be improved first because lack of communication is the strongest factors that lowers team performance. The second issue that students need to is to promote teamwork. Lastly, they have to make sure that trust among team members is enhanced.

<u>Interpreting r-square</u>

The r-square that is equal to .364 or 36.4 percent tells us that all independent variables that we put in regression analysis altogether can explain 36.4 percent of team performance. There is another 63.6 percent that cannot be explained by the regression analysis.

<u>Assessing multicollinearity</u>

Multicollinearity can be assessed by VIF. Apparently the maximum VIF in the regression analysis is the VIF of the variable 'teamwork' which is equal to 2.332. Because all VIF is lower than 5 (it is still ok if they are lower than 10), we can claim that multicollinearity may not be a serious issue for the analysis.

References

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.

Hair, J. F., Anderson, R., Tatham, R. L., & Black, W. C. (2009). *Multivariate data analysis*. New York, NY: Macmillan.

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). London: Edward Arnold.

Pan, Y., & Jackson, R. T. (2008). Ethnic difference in the relationship between acute inflammation and and serum ferritin in us adult males. *Epidemiology and Infection, 136*(3), 421-431.

Rogerson, P. A. (2001). *Statistical methods for geography*. London: Sage.