Module 10 Discussion Notes

Module 10 Learning Objectives

1. Understand how regression analysis can be used to develop an equation that estimates mathematically how two variables are related.

2. Understand the improvement in the precision made possible by a regression approach over point estimation and interval estimation.

3. Know how to fit an estimated regression equation to a set of sample data based upon the least-squares method.

4. Be able to determine how good a fit is provided by the estimated regression equation.

5. Understand the assumptions necessary for statistical inference and be able to test for a significant relationship.

In most of our statistical procedures so far, we have been concerned with a single observation made on each element of the sample, that is, with a sample of values for a single variable x. We now consider the case where two measurements are made on each element of the sample: where the sample consists of pairs of values, one for each of the two variables x and y. For example, consider the heights and weights of individuals. If we take a sample of individuals, obtain from each his or her height and weight, and then let the height be represented by x and the weight by y, we obtain from the i^{th} person the pair of numbers (x_i, y_i). If there are n persons in the sample, we have a sample of size n which consists of the n number pairs: (x_1, y_1), (x_2, y_2), …,(x_n, y_n).

One way to study this relationship is by means of regression. *Regression analysis* is the process of obtaining a functional relationship, or equation, between the variables being investigated. For example, a financial analyst might be interested in the relationship between *the mean current yield on corporate bonds* and *the rate of transfers from bond funds to stock mutual funds*. If she had an equation showing the relationship between these two variables and if she had forecasts for future current yields on corporate bonds, she could use the equation to predict the movement between the two types of funds. Regression analysis is the tool or method that would be used to develop the prediction equation.

We begin at the very beginning by considering an actual real-world situation in which regression analysis might be useful.

A few years ago, a local civic organization located in San Diego, California became interested in understanding what variables might be associated with the attitude of San Diego's residents toward their city. Although they interviewed over one thousand residents, we will consider a small sub-sample (of size n = 12) of their data set. The question probing the attitude toward life in San Diego consisted of the following declarative statement accompanied by an agree/disagree response Likert Scale.

"On a scale from 1 to 11, where 1 represents 'very poor attitude toward life in San Diego' and 11 represents 'very positive attitude toward life in San Diego,' my attitude toward life in San Diego is best represented by _____."

The individuals conducting this research approached people in various locations in the city---in parks, on the waterfront, in malls---with the purpose of requesting a few minutes of their time for a brief interview about life in San Diego. After explaining who they represented as well as the purpose of the survey, the potential respondents were then "qualified," that is, they confirmed that they were indeed residents of San Diego (and not tourists or visitors). Several questions, including the one above, were asked, the information recorded, and the interviewees thanked for their time. Suppose the data for the first 12 respondents is represented in the table below.

Respondent No., n	Attitude Score
1	2
2	2
3	3
4	4
5	5
6	6
7	8
8	9
9	9
10	10
11	10
12	11

If this were all the data we had, could we make any statement about the typical attitude score of a randomly selected resident of the city? Of course we could. Clearly, we could calculate the best-known measure of central tendency, the mean: $\bar{x} = \frac{79}{12} = 6.5833$. This represents our best guess as to a randomly-selected resident's attitude toward life in San Diego. If this is all

the information we have, can we improve on it? Is it possible to make any stronger statements about the matter being investigated? The answer, as you no doubt know by now, is an unequivocal 'yes, we can develop an interval estimate of the population mean at, say, the 95% level of confidence.'

$$\bar{x} \pm t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t_{.025,11} \frac{3.3155}{\sqrt{12}}$$

$6.5833 \pm (2.201)\ (.9571)$

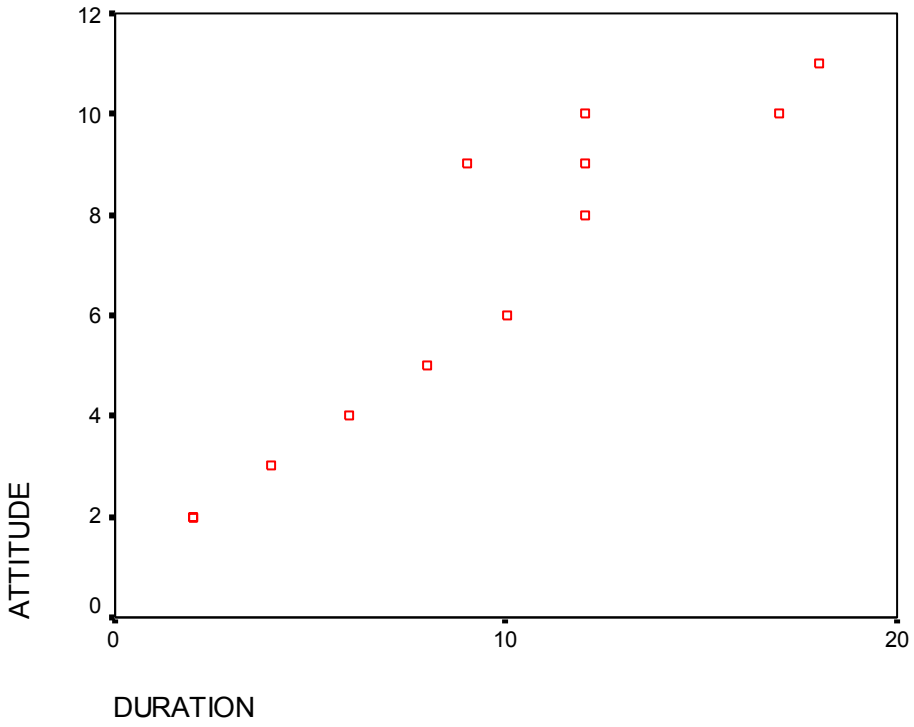
6.5833 ± 2.1066 or [4.4767, 8.6899]

Thus, we can be 95% confident that the mean attitude rating μ toward life in San Diego among its residents is, on an 11-point scale, between 4.4767 and 8.6899.

This is as far as we can go, given the analytic tools and data we have. What would regression do? The regression approach to this problem would involve our identifying some other variable which might be related to 'attitude' in some systematic way. Once we develop our regression model, we then make predictions about 'attitude' based on our knowledge of the new variable. Let us look at one possible way to do this. Suppose that during the interviews described above information on other variables was also collected, including a question about how long the interviewee had resided in San Diego. In other words, suppose that in addition to the question concerning the person's attitude toward life in San Diego, the data set also included each individual's duration of residency. We reproduce the table from above here with the additional information, the duration (in years) of residency.

Respondent No., n	Attitude Score	Duration (years)
1	2	2
2	2	2
3	3	4
4	4	6
5	5	8
6	6	10
7	8	12
8	9	12
9	9	9
10	10	12
11	10	17
12	11	18

Although we can see, by casually examining the data in the two right-hand columns of the above table, that the two variables seem to be positively related, we can understand the relationship better if we develop a scatter plot. When we do, we see that the two variables appear to be positively linearly related; that is, we see that there is a band of data points running from the lower left to the upper right.



When we perform simple linear regression analysis on two variables, such as 'attitude' and 'duration,' we are attempting to capture the relationship, if any, between them. In practical terms, this requires that we, first, place the "best" line through the data points, and, second, write out the equation of that line. The general form of the estimated simple linear regression equation is:

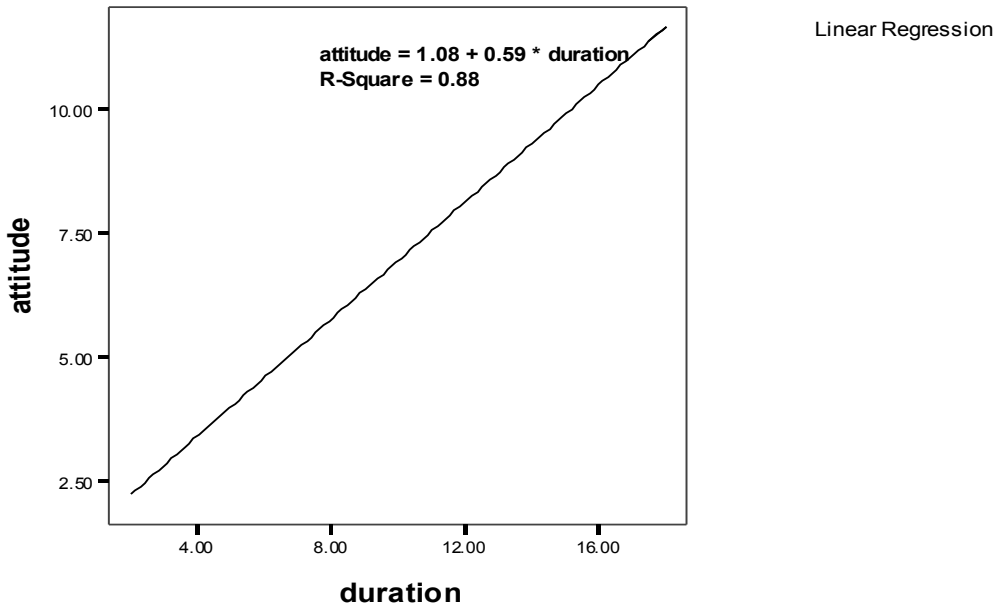
$\hat{y} = b_o + b_1 x$, where

$\hat{y}$ is the estimated value of the dependent variable (in this case, 'attitude')
b_o is the y-intercept
b_1 is the slope coefficient
x is the independent variable (in this case, 'duration')

The equation of the "best" line fitting the points in our scatterplot above is:

attitude = 1.079 + (.590)duration

When the "best" line is placed through the data points, we have the following result:



By "best" line, we mean that any other linear representation of this particular set of data points would not capture the relationship as well. The line depicted above seems to be the line "closest" to the 12 data points. Rotate the above line slightly, or move it up or down by the smallest amount, and the resulting line would lose a bit of its explanatory power.

The advantage of using the regression equation (rather than using the mean, 6.5833) to predict the average resident's *attitude toward life in San Diego* is that it incorporates additional information, the duration (in years) of residence in San Diego. Since the scatterplot clearly reveals that these two variables are positively linearly related, we will want to exploit our knowledge of 'duration' when trying to make predictions about 'attitude.'

The rationale behind the development of the estimated linear regression model is relatively intuitive. See Section 12.9 of Chapter 12 for the derivation of the values of the regression equation coefficients. The approach is variously known as the "ordinary least squares" or simply "least squares" method. You are encouraged to read Section 12.9 to learn how the regression equation is developed from a data set. We will here concern ourselves, however, with a more applied aspect of immediate importance: the strength of association.

The Strength of Association, r^2. The strength of association is measured by the coefficient of determination, or the 'famous' r^2. It signifies the proportion of total variation in the dependent variable, y, which is accounted for (or explained) by the variation in the independent variable, x. In simple linear regression, r^2 is simply the square of the simple correlation coefficient, r, obtained by correlating the two variables, x and y. As such, $0 \leq r^2 \leq 1$ (since $-1 \leq r \leq 1$).

In the case of our example, the $r^2 = .876$. What does this mean? It means that 87.6% of the variation in the dependent variable 'attitude' is explained, or accounted for, by the variation in the independent variable 'duration.' The section of the statistical package printout reporting 'strength of association' is provided in the center column of the Model Summary. (This printout below was developed using the SPSS statistical package, and is used here because it is a bit nicer to look at than the R printout.)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.936[a]	.876	.864	1.22329

a. Predictors: (Constant), DURATION

How is the r^2 calculated? To answer this question, we refer to the scatterplot with the regression line on the previous page. Upon examination, we note that our data set contains three people who report living in San Diego for 12 years, respondents 7, 8, and 10. We also note that, although each has lived in San Diego 12 years, each has provided a different attitude score: the attitude of respondent 7 is 8, the attitude of respondent 8 is 9, and the attitude of respondent 10 is 10. This reflects a simple fact of life: our regression includes only one independent variable, duration, and so the predictive or explanatory power of our model is not perfect. To be perfect, the regression model would have to incorporate enough information (that is, it would have to include additional, relevant independent variables) that there would be no error in its predictive ability. This, unfortunately, is rarely if ever possible in the real world of messy data analysis.

Let us focus on respondent 10 who reports having lived in San Diego for a period of 12 years as well as having an attitude score of 10 (out of a possible 11) toward life in the city. We notice that we would fail to predict accurately

this respondent's attitude using either the mean or the regression equation. That is, even though

y = attitude of respondent 10 = 10

$\bar{y} = \overline{attitude} = 6.5833$

$\hat{y} = 1.079 + (.590)$ duration $= 1.079 + (.590)\,(12) = 8.159$

We note also (perhaps with some satisfaction) that our regression-based prediction, 8.159, is "closer" to the actual value of 10 than is the mean-based prediction of 6.5833. (This will not necessarily always be the case: sometimes the "naïve" mean-based approach will actually predict better than the regression-based method; see, for example, respondent number 6.) These two types of error---one associated with using the mean, the other associated with using the regression equation-- have specific names, and they are related in a certain way. Here they are.

$(y_i - \bar{y})$ = *total variation* = the "error" associated with using the mean to predict = (10 – 6.5833) = 3.4167

$(y_i - \hat{y}_i)$ = *residual variation* = the "error" associated with using regression equation to predict = (10 – 8.159) = 1.841

$(\hat{y}_i - \bar{y})$ = *explained* or *regression variation* = (8.159 - 6.5833) = 1.5757

These three important terms are related in that the *total variation* equals the sum of the *residual variation* and the *explained* or *regression variation.* In the case of respondent number 10: (3.4167) = (1.841) + (1.5757). Since this is the case with all observations in the entire data set, not only for respondent number 10, we can represent this relationship in the following manner:

$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2$, where

$\sum(y_i - \bar{y})^2 = SS_y =$ sum of total squares = total variation in variable y

$\sum(y_i - \hat{y}_i)^2 = SS_{res} =$ sum of residual squares

$\sum(\hat{y}_i - \bar{y})^2 = SS_{reg} =$ sum of regression squares

Put another way, the total variation in any data set, SS_y, can be decomposed into the variation explained by the regression model, SS_{reg}, and the residual variation "left over," or unexplained by the regression model, SS_{res}. That is,

$$SS_y = SS_{res} + SS_{reg}$$

Since the coefficient of determination r^2 is the proportion of variation in the dependent variable y explained (or accounted for) by the variation in the independent variable x, we now have the following definition of r^2

$$r^2 = \frac{SS_{reg}}{SS_y} = \frac{SS_y - SS_{res}}{SS_y}$$

The section of the statistical printout reporting these sums of squares (called the ANOVA table) is provided below.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	105.952	1	105.952	70.803	.000[a]
	Residual	14.964	10	1.496		
	Total	120.917	11			

a. Predictors: (Constant), DURATION

b. Dependent Variable: ATTITUDE

Using our new expression for r^2 and the values in the 'sum of squares' column, we can now derive a value for the coefficient of determination. Note that it is exactly equal to the value in the center column of the Model Summary Table (top of next page).

$$r^2 = \frac{SS_{reg}}{SS_y} = \frac{SS_y - SS_{res}}{SS_y} = \frac{(120.917 - 14.964)}{120.917} = \frac{105.953}{120.917} = .876$$

Consider two other terms in the Model Summary Table. See next page.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.936[a]	.876	.864	1.22329

a. Predictors: (Constant), DURATION

First, the 'Adjusted R Square' entry, .864, is a term we will take up in a more meaningful way in the write up on multiple regression. Second, we see that R = .936. What is the meaning of this, and is it related to the next entry to the right, 'R Square'?

To answer this question, let us first perform a statistical procedure in which we correlate the original variables, 'attitude' and 'duration,' and provide the statistical package printout below.

Correlations

		ATTITUDE	DURATION
ATTITUDE	Pearson Correlation	1	.936**
	Sig. (2-tailed)	.	.000
	N	12	12
DURATION	Pearson Correlation	.936**	1
	Sig. (2-tailed)	.000	.
	N	12	12

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

Recall that in bivariate regression (that is, in those regression models where we have only one independent variable), r^2 is simply the square of the simple correlation coefficient obtained by correlating the two variables. Since r = .936, we can see that $r^2 = (.936)^2 = .876$. Thus, r and r^2 are related, at least in the bivariate case; in the instance where we have more than one independent variable, this relationship is not valid. An interpretation of r = .936 which is more generally valid, and one that is true in the case of multiple regression as well as in simple linear regression, is that it represents the correlation of the *actual* and *predicted dependent variables*. To understand this meaning, let us once again regress 'attitude' on 'duration;' this time, however, we will retain the unstandardized predicted value of the dependent variable.

Attitude	Duration	Predicted Attitude
2.00	2.00	2.25875
2.00	2.00	2.25875
3.00	4.00	3.43818
4.00	6.00	4.61761
5.00	8.00	5.79705
6.00	10.00	6.97648
8.00	12.00	8.15591
9.00	12.00	8.15591
9.00	9.00	6.38676
10.00	12.00	8.15591
10.00	17.00	11.10449
11.00	18.00	11.69420

How is 'Predicted Attitude' derived? We simply plug into the regression equation the value of 'duration' for each of the 12 subjects. For respondent 1, for example, 1.079 + (.590) (2) = 2.259 (adjusted for rounding errors); for respondent 12, 1.079 + (.590) (18) = 11.699.

Now suppose we correlate the *actual* and the *predicted dependent variables*.

Correlations

		Unstandardized Predicted Value	ATTITUDE
Unstandardized Predicted Value	Pearson Correlation	1	.936**
	Sig. (2-tailed)	.	.000
	N	12	12
ATTITUDE	Pearson Correlation	.936**	1
	Sig. (2-tailed)	.000	.
	N	12	12

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

Now we have a more general meaning for r. In the multivariate as well as the univariate case, r is simply the correlation between the *actual* and *predicted dependent variables*. Naturally, regression models that predict well will have higher values of r than will models that predict less well.

A well-known authority on statistics has written that "*as a practical matter, for typical data found in the social sciences (and, yes, most areas of business management and economics are considered social sciences), values of* r^2 *as low as .25 are often considered useful. For data in the physical sciences,* r^2

values of .60 or greater are often found…in business applications, r^2 values vary greatly, depending on the unique characteristics of each application."

Thus, we should not be tricked into thinking that the $r^2 = .876$ is normal: it is abnormally high. In fact, I cherry-picked the data (that is, n = 12 observations) for the purpose of illustrating some of the foundational ideas underlying regression analysis.

In the next module, we extend our discussion of regression analysis from the situation where we have only one independent variable (simple linear regression) to the case where we can have any number of independent variables. While it is essential to introduce the topic of regression in its most basic form, as we have done in this module with our focus on simple linear regression, we will find multiple regression analysis to be a much more useful and powerful method.