14 Regression Analysis

CHAPTER

Business Statistics: *Communicating with Numbers*

By Sanjiv Jaggia and Alison Kelly

Chapter 14 Learning Objectives (LOs)

LO 14.1: Conduct a hypothesis test for the population correlation coefficient.

LO 14.2: Discuss the limitations of correlation analysis.

LO 14.3: Estimate the simple linear regression model and interpret the coefficients.

LO 14.4: Estimate the multiple linear regression model and interpret the coefficients.

LO 14.5: Calculate and interpret the standard error of the estimate.

LO 14.6: Calculate and interpret the coefficient of determination R^2.

LO 14.7: Differentiate between R^2 and adjusted R^2.

How are debt payments and income related?

Metropolitan Area	Income (in $1,000s)	Unemployment	Debt
Washington, D.C.	$103.50	6.3%	$1,285
Seattle	81.70	8.5	1,135
⋮	⋮	⋮	⋮
Pittsburgh	63.00	8.3	763

- A study in 2010 showed that consumers in 26 cities made debt payments from $763 to $1,285 per month.
- Economist Madelyn Davis believes that income differences are the main reason for the disparity.
- She is less sure about the impact of unemployment.
- She uses **correlation** analysis and **regression** analysis to learn more.

14.1 Covariance and Correlation

LO 14.1 Conduct a hypothesis test for the population correlation coefficient.

- We examined covariance and correlation as exploratory tools in Chapters 2 and 3.
- Recall that covariance is a numerical measure that reveals the direction of the linear relationship between two variables.
- The sample covariance is computed as:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

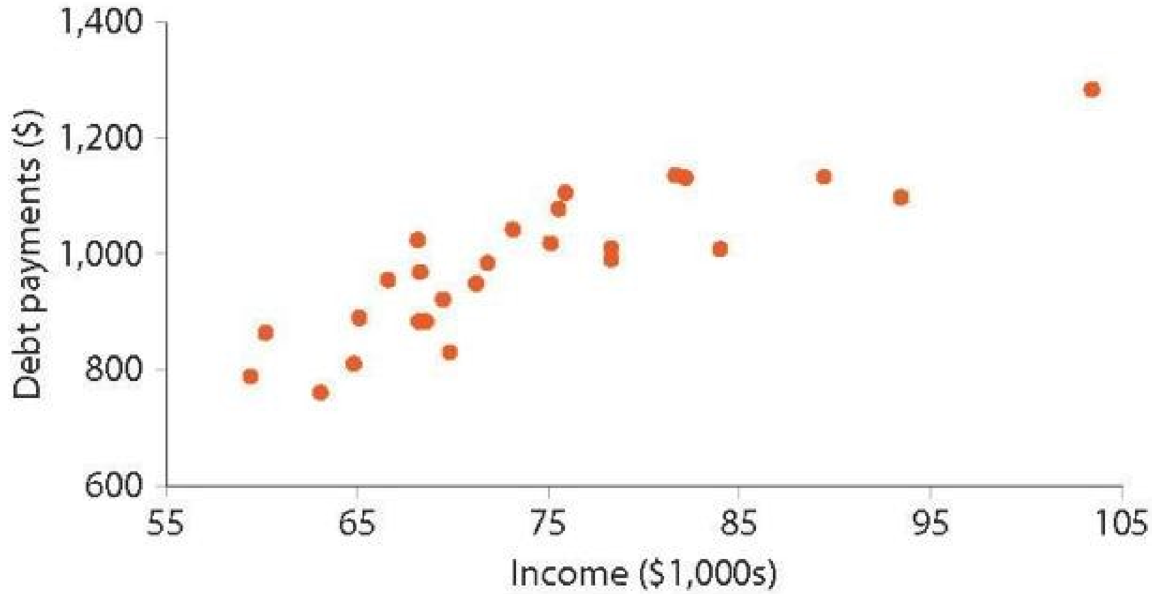
LO 14.1

Computing the Correlation

- The correlation coefficient indicates both the direction and the strength of the linear relationship.
- The sample correlation coefficient can be computed using: $r_{xy} = \frac{s_{xy}}{s_x s_y}$.
- The correlation coefficient has the same sign as the covariance; however, its value ranges between -1 and +1.

LO 14.1 Debt Payments and Income

Consider the introductory case. A scatterplot can graphically display the relationship between debt payments and income.



Here we see debt payments do indeed rise with incomes.

LO 14.1 Correlation in the Example

- For debt payments we have $\bar{y}$ = 983.5 and s_y = 124.61. For income we have $\bar{x}$ = 74.1 and s_x = 10.35.

- We compute the covariance as:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{27979.50}{26-1} = 1119.18$$

- The correlation coefficient is:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1119.18}{(10.35)(124.61)} = 0.87.$$

LO 14.1

Using Excel

- To compute the covariance, choose **Formulas > Insert Function > COVARIANCE.S** from the menu. Select the data for each variable as Array 1 and Array 2.
- To compute the correlation coefficient, choose **Formulas > Insert Function > CORREL**. Select the data just as was done for the covariance.

LO 14.1 Testing for Significant Correlation

- We need to be able to determine whether the relationship implied by the sample correlation coefficient is real or due to chance.

- In other words, we would like to test whether the population correlation coefficient is different from zero:

$$H_0: \rho_{xy} = 0$$
$$H_A: \rho_{xy} \neq 0$$

LO 14.1

The Test Statistic

- The test statistic is $t_{df} = \frac{r_{xy}}{s_r}$, where $s_r = \sqrt{(1 - r_{xy}^2)/(n-2)}$. The test statistic follows a *t* distribution with *df* = *n* - 2.
- Using the data from the introductory example, we first find $s_r = \sqrt{(1 - 0.87^2)(26-2)} = 0.1007$.
- Therefore, $t_{24} = \frac{0.87}{0.1007} = 8.64$. At the 5% significance level, 8.64 is greater than $t_{0.025,24}$ = 2.064, so we reject the null hypothesis. This implies that the correlation coefficient is significantly different from zero.

Limitations of Correlation Analysis

LO 14.2 Discuss the limitations of correlation analysis.

- The correlation coefficient captures only a linear relationship.
- The correlation coefficient may not be a reliable measure in the presence of outliers.
- Even if two variables are highly correlated, one does not necessarily cause the other.

14.2 The Simple Regression Model

LO 14.3 Estimate the simple linear regression model and interpret the coefficients.

- While the correlation coefficient may establish a linear relationship, it not suggest that one variable causes the other.
- With **regression analysis**, we explicitly assume that one variable, called the **response variable**, is influenced by other variables, called the **explanatory variables**.
- Using regression analysis, we may predict the response variable given values for our explanatory variables.

LO 14.3 Stochastic Relationships

- If the value of the response variable is uniquely determined by the values of the explanatory variables, we say that the relationship is **deterministic**.
- But if, as we find in most fields of research, that the relationship is inexact due to omission of relevant factors, we say that the relationship is **stochastic**.
- In regression analysis, we include a stochastic error term, that acknowledges that the actual relationship between the response and explanatory variables is not deterministic.

LO 14.3 The Simple Regression Model

The simple linear regression model is defined as

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where y and x are the response and explanatory variables, respectively and and ε is the random error term.

The coefficients β_0 and β_1 are the unknown parameters to be estimated.

LO 14.3 Sample Regression Equation

By fitting our data to the model, we obtain the equation

$$\hat{y} = b_0 + b_1 x,$$

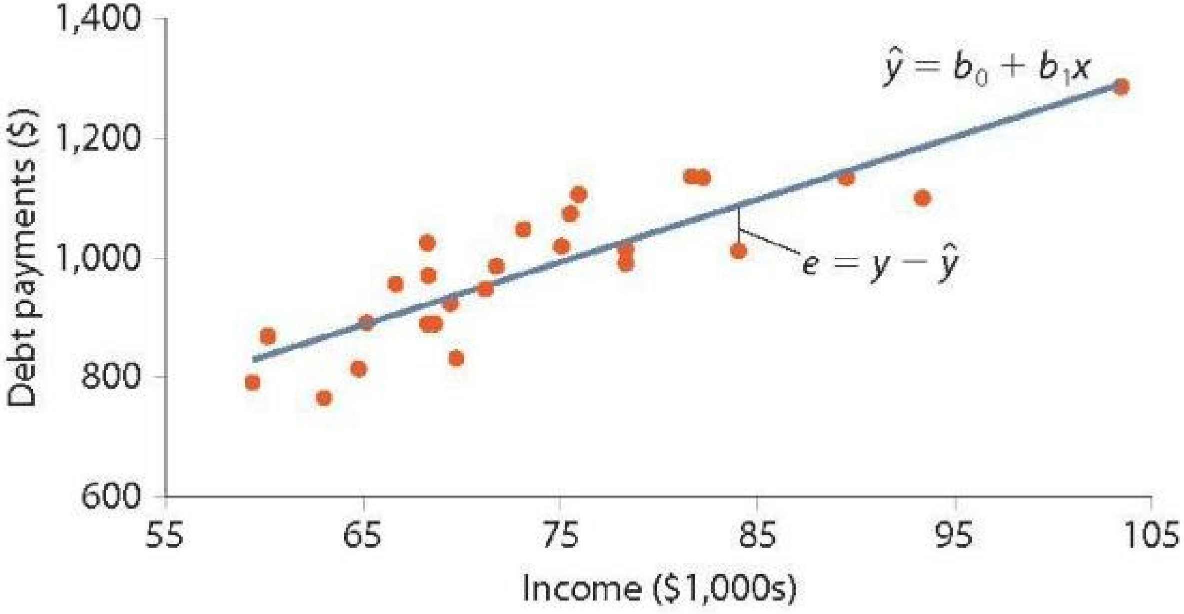
where $\hat{y}$ is the estimated response variable, b_0 is the estimate of β_0, and b_1 is the estimate of β_1.

Since the predictions cannot be totally accurate, the difference between the predicted and actual value represents the **residual** $e = y - \hat{y}$.

LO 14.3

Regression Illustration

This is a scatterplot of debt payments against income with a superimposed sample regression equation.



Debt payments rise with income. Vertical distance between y and $\hat{y}$ represents the residual, e.

LO 14.3 The Least Squares Estimates

- The two parameters β_0 and β_1 are estimated by minimizing the sum of squared residuals.

- The slope coefficient is estimated as

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}.$$

- Then compute the intercept: $b_0 = \bar{y} - b_1\bar{x}$.

LO 14.3

Debt Payments Example

- We denote Debt as y and Income as x. We have $\bar{y}$ = 983.46 and $\bar{x}$ = 74.05. In addition, we find:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 27979.50$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2679.75$$

- The slope: $$b_1 = \frac{27979.50}{2679.75} = 10.4411$$

- The intercept:
$b_0 = \bar{y} - b_1\bar{x} = 983.46 - 10.4411(74.05) = 210.30.$

LO 14.3 Interpreting the Coefficients

- The sample regression equation: $\hat{y} = 210.30 + 10.44x$
- The slope b_1 = 10.44 implies that in a city where the median household income increases by $1000, then average debt payments are expected to increase by $10.44.
- The intercept b_0 = 210.30 suggests that if income were 0, debt payments would still be $210.
- We could also use the sample regression equation to predict debt payments for other cities.

LO 14.3

Excel and Regression

- Open the data in an Excel spreadsheet and from the menu, choose **Data > Data Analysis > Regression** .

- After the dialog box opens, select the data for your response variable in the *Input Y Range* and the data for your explanatory variable(s) in the *Input X Range*.

- We can display the output on a new page, in the current worksheet, or even in a new workbook.

LO 14.3 Excel Output

- The Excel output will look like this:

Regression Statistics						
Multiple R	0.8675					
R Square	0.7526					
Adjusted R Square	0.7423					
Standard Error	63.26					
Observations	26					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	292136.91	292136.9	73.00	1E-08	
Residual	24	96045.55	4001.9			
Total	25	388182.46				
	Coefficients	*Standard Error*	*t Stat*	*p-value*	*Lower 95%*	*Upper 95%*
Intercept	**210.2977**	91.3387	2.3024	0.0303	21.78	398.81
Income	**10.4411**	1.2220	8.5440	0.0000	7.92	12.96

14.3 The Multiple Regression Model

LO 14.4 Estimate the multiple linear regression model and interpret the coefficients.

- If there is more than one explanatory variable available, we can use **multiple regression**.
- For example, we analyzed how debt payments are influenced by income, but ignored the possible effect of unemployment.
- Multiple regression allows us to explore how several variables influence the response variable.

LO 14.4

The Multiple Regression Model

Suppose there are *k* explanatory variables. The multiple linear regression model is defined as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

where $x_1, x_2, \ldots, x_k$ are the explanatory variables and the β_j values are the unknown parameters that we will estimate from the data.

As before, ε is the random error term.

LO 14.4 The Estimated Equation

- The sample multiple regression equation is:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k.$$

- In multiple regression, there is a slight modification in the interpretation of the slopes b_1 through b_k as they show "partial" influences.

- For example, if there are k = 3 explanatory variables, the value b_1 estimates how a change in x_1 will influence y assuming x_2 and x_3 are held constant.

LO 14.4

Adding a Second Predictor

- In addition to income, the unemployment rate may also influence an area's average debt payments.

- Utilizing Excel, we can easily add the additional explanatory variable by choosing **Data > Data Analysis > Regression** as before, but now select both income and the unemployment rate data for *Input X Range*.

LO 14.4 Expanded Computer Output

The output reflects the additional coefficient estimate for unemployment. The other coefficients also change slightly.

Regression Statistics						
Multiple R	0.8676					
R Square	0.7527					
Adjusted R Square	0.7312					
Standard Error	64.61					
Observations	26					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	292170.77	146085.39	35.00	1E-07	
Residual	23	96011.69	4174.42			
Total	25	388182.46				
	Coefficients	*Standard Error*	*t Stat*	*p-value*	*Lower 95%*	*Upper 95%*
Intercept	**198.9956**	156.3619	1.2727	0.2159	−124.46	522.45
Income	**10.5122**	1.4765	7.1195	0.0000	7.46	13.57
Unemployment	**0.6186**	6.8679	0.0901	0.9290	−13.59	14.83

LO 14.4

Interpretation of Slopes

- The estimated equation is
  $\hat{y} = 198.9956 + 10.5122x_1 + 0.6186x_2$.

- The coefficient of 10.51 on Income indicates that if Income increases by $1,000, then Debt is expected to increase by $10.51, assuming Unemployment does not change.

- The coefficient of 0.6186 on Unemployment indicates that an increase in Unemployment of 1% is expected to lead to an increase in Debt of 62 cents, assuming Income does not change.

LO 14.4 Predicting the Debt Level

- We can also use the estimated equation to predict debt payments given values for median income and the unemployment rate.

- Suppose we wish to predict debt payments that would occur in a city with a median income level of $80,000 and 7.5% unemployment.

- We simply plug those values into our estimated equation:

$$\hat{y} = 198.996 + 10.512(80) + 0.619(7.5) = 1{,}044.61$$

14.4: Goodness-of-Fit Measures

LO 14.5 Calculate and interpret the standard error of the estimate.

We will introduce three measures to judge how well the sample regression fits the data.

1. The Standard Error of the Estimate
2. The Coefficient of Determination
3. The Adjusted R^2

LO 14.5 Mean Squared Error

- To compute the standard error of the estimate, we first compute the mean squared error.

- We first compute the error sum of squares:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

- Dividing *SSE* by the appropriate degrees of freedom, *n – k – 1* , yields the mean squared error, *MSE*:

$$MSE = \frac{SSE}{n-k-1}$$

LO 14.5 Standard Error of the Estimate

- The square root of the *MSE* is the **standard error of the estimate**, s_e.

$$s_e = \sqrt{MSE} = \sqrt{\frac{\sum e_i^2}{n-k-1}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-k-1}}$$

- In general, the less dispersion around the regression line, the smaller the s_e, which implies a better fit to the model.

LO 14.5

s_e in Excel

- Here we show the standard error of the estimate for the simple linear regression (Model 1) and the multiple linear regression (Model 2):

	Model 1	Model 2
Multiple R	0.8675	0.8676
R Square	0.7526	0.7527
Adjusted R Square	0.7423	0.7312
Standard Error	63.26	64.61
Observations	26	26
Regression Equation	$\hat{y} = 210.30 + 10.44x$	$\hat{y} = 199 + 10.51x_1 + 0.62x_2$

- Notice that according to the standard error, adding the unemployment level as an explanatory variable did not help our goodness-of-fit.

The Coefficient of Determination

LO 14.6 Calculate and interpret the coefficient of determination R^2.

- The **coefficient of determination**, commonly referred to as the R^2, is another goodness-of-fit measure that is easier to interpret than the standard error.

- In particular, the R^2 quantifies the fraction of variation in the response variable that is explained by changes in the explanatory variables.

LO 14.5 Calculating R^2

- The coefficient of determination can be computed as $R^2 = 1 - \frac{SSE}{SST}$, where $SSE = \sum(y_i - \hat{y})^2$ and $SST = \sum(y_i - \bar{y})^2$.
- The *SST*, called the total sum of squares, denotes the total variation in the response variable.
- The *SST* can be broken down into two components: the variation explained by the regression equation (the regression sum of squares or *SSR*) and the unexplained variation (the error sum of squares or SSE).

LO 14.5 Excel Output and R^2

- The R^2 is also reported with the Regression Statistics in the Excel regression output. Here it is in second row from the top for each model:

	Model 1	Model 2
Multiple R	0.8675	0.8676
R Square	0.7526	0.7527
Adjusted R Square	0.7423	0.7312
Standard Error	63.26	64.61
Observations	26	26
Regression Equation	$\hat{y} = 210.30 + 10.44x$	$\hat{y} = 199 + 10.51x_1 + 0.62x_2$

The Adjusted R^2

LO 14.7 Differentiate between R^2 and adjusted R^2.

- More explanatory variables always result in a higher R^2.
- But some of these variables may be unimportant and should not be in the model.
- The **Adjusted R^2** tries to balance the raw explanatory power against the desire to include only important predictors.

LO 14.7 Computing Adjusted R^2

- The Adjusted R^2 is computed as

$$Adjusted\ R^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

- As you can see, the adjusted R^2 penalizes the R^2 for adding additional explanatory variables.

- As with our other goodness-of-fit measures, we typically allow the computer to compute the Adjusted R^2. It's shown directly below the R^2 in the Excel regression output.

LO 14.7 Model Comparison

Comparing the simple linear regression (Model 1) with the multiple linear regression model (Model 2):

	Model 1	Model 2
Multiple R	0.8675	0.8676
R Square	0.7526	0.7527
Adjusted R Square	0.7423	0.7312
Standard Error	63.26	64.61
Observations	26	26
Regression Equation	$\hat{y} = 210.30 + 10.44x$	$\hat{y} = 199 + 10.51x_1 + 0.62x_2$

Even though the R^2 is a bit higher in the multiple regression, the adjusted R^2 is lower and standard error higher, implying we are better off without the second predictor.