Chapter 14

Regression Analysis

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.

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:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{14.1}$$

14.1.1 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:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{14.2}$$

- The correlation coefficient has the same sign as the covariance; however, its value ranges between -1 and +1.

14.1.2 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:

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

14.1.3 The Test Statistic

- The test statistic is

$$t_{df} = \frac{r_{xy}}{s_r}, \tag{14.3}$$

where

$$s_r = \sqrt{\frac{1 - r_{xy}^2}{n - 2}} \tag{14.4}$$

The test statistic follows a t distribution with $df = n - 2$.

14.1.4 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 does not suggest that one variable causes the other.

- With regression analysis, we explicitly assume that one variance, 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 exploratory variables.
- The simple linear regression model is defined as

$$y = \beta_0 + \beta_1 x + \epsilon \tag{14.5}$$

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

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

14.2.1 Sample Regression Equation

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

$$\hat{y} = b_0 + b_1 x, \tag{14.6}$$

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 $e = y - \hat{y}$.

14.2.2 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

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{14.7}$$

- Then, compute the intercept:

$$b_0 = \bar{y} - b_1 \bar{x} \tag{14.8}$$

14.2.3 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.
- However 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.

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \tag{14.9}$$

  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.

14.3.1 The Estimated Equation

- The sample multiple regression equation is:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \tag{14.10}$$

- 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.

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

14.4.1 Mean Squared Error

- To compute the standard error of the estimate, we first compute the mean squared error (MSE).
- We first compute the error sum of squared:

$$\begin{aligned} SSE &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \end{aligned} \tag{14.11}$$

- Dividing SSE by the appropriate degrees of freedom, $n - k - 1$, yields the MSE:

$$MSE = \frac{SSE}{n - k - 1} \tag{14.12}$$

14.4.2 Standard Error of the Estimate

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

$$\begin{aligned} s_e &= \sqrt{MSE} \\ &= \sqrt{\frac{SSE}{n - k - 1}} \\ &= \sqrt{\frac{\sum e_i^2}{n - k - 1}} \\ &= \sqrt{\frac{\sum (y_i - \hat{y})^2}{n - k - 1}} \end{aligned} \tag{14.13}$$

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

14.4.3 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.

- The coefficient of determination can be computed as

$$\begin{aligned} R^2 &= 1 - \frac{SSE}{SST} \\ &= \frac{SSR}{SST} \end{aligned} \tag{14.14}$$

where SSE is (14.11) and SST is (14.15)

$$SST = \sum (y_i - \bar{y})^2. \tag{14.15}$$

- 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).

14.4.4 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 show not be in the model.
- The adjusted R^2 tries to balance the raw explanatory power against the desire to include only important predictors.
- The Adjusted R^2 is computed as:

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right) \tag{14.16}$$

- 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.