Chapter 15

Inference with Regression Models

LO 15.1: Conduct tests of individual significance.

LO 15.2: Conduct a test of joint significance.

LO 15.3: ~~Conduct a general test of linear restrictions.~~

LO 15.4: ~~Calculate and interpret interval estimates for predictions.~~

LO 15.5: Explain the role of the assumptions on the OLS estimators.

LO 15.6: Describe common violations of the assumptions and offer remedies.

15.1 Tests of Significance

LO 15.1 Conduct tests of individual significance.

- With two explanatory variables to choose from, we can formulate three linear models:

$$\begin{aligned}
&\text{Model 1: } \text{Win} = \beta_0 + \beta_1\text{BA} + \epsilon\\
&\text{Model 2: } \text{Win} = \beta_0 + \beta_1\text{ERA} + \epsilon\\
&\text{Model 3: } \text{Win} = \beta_0 + \beta_1\text{BA} + \beta_2\text{ERA} + \epsilon
\end{aligned}$$

15.1.1 Tests of Individual Significance

- Consider our standard multiple regression model: (**??**).
- In general, we can test whether β_j is equal to, greater than, or less than some hypothesized value β_{j0}.
- This test could have one of three forms:

Table 15.1: Three Forms of Individual Significance

Two-tailed Test	**Right-tailed Test**	**Left-tailed Test**
$H_0 : \beta_j = \beta_{j0}$	$H_0 : \beta_j \leq \beta_{j0}$	$H_0 : \beta_j \geq \beta_{j0}$
$H_A : \beta_j \neq \beta_{j0}$	$H_A : \beta_j > \beta_{j0}$	$H_A : \beta_j < \beta_{j0}$

15.1.2 The Test Statistic

- The appropriate test statistic is:

$$t_{df} = \frac{b_j - \beta_{j0}}{s_{b_j}} \tag{15.1}$$

- s_{b_j} is the standard error of the estimator b_j.
- The test statistic will follow a t-distribution with degrees of freedom $df = n - k - 1$.

15.1.3 Testing $\beta_j = 0$

- By far, the most common hypothesis test for an individual coefficient is to test whether its value differs from zero.
- To see why, consider our model: (**??**).
- If a coefficient is equal to zero, then it implies that the explanatory variable is not a significant predictor of the response variable.

15.1.4 Computer-Generated Output

- Virtually all statistical software will automatically report a test statistic and a p-value with each coefficient estimate.
- These values can be used whether the regression coefficient differs from zero.
- To perform a one-sided test where the hypothesized value is zero, divide the computer-reported p-value in half.
- If we wish to test whether the coefficient differs from a non-zero value, we need to compute a new test statistic.

15.1.5 Intervals for the Parameters

- A confidence interval for the β_j parameter can be constructed using the formula:

$$b_j \pm t_{\alpha/2,df} s_{b_j} \tag{15.2}$$

- This can also be used to perform the two-sided test to determine whether a coefficient differs from zero.

- For ERA, the interval of $[-0.15, -0.08]$ does not include 0, indicating ERA is a significant predictor.

15.1.6 Test for a Non-Zero Slope

- A capital asset pricing model follows the equation:

$$y = \alpha + \beta x + \epsilon \tag{15.3}$$

  where $y =$ the risk-adjusted return of an asset, $R - R_f$ and $x =$ the risk-adjusted return to the market $R_M - R_f$.

- The estimate of β is called the investment's beta value.
- A $\beta > 1$ implies the stock is "aggressive", while a $\beta < 1$ implies it is "conservative".

15.1.7 Test of Joint Significance

LO 15.2 **Conduct a test of joint significance.**

- In addition to conducting tests of individual significance, we may also want to test the joint significance of all k variables at once.
- The competing hypotheses for a test of joint significance are:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_A : \text{at least one } \beta_j \neq 0$$

15.1.8 The Test Statistic

- The test statistic for a test of joint significant is

$$\begin{aligned} F_{(df_1, df_2)} &= \frac{MSR}{MSE} \\ &= \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} \end{aligned} \tag{15.4}$$

  where MSR and MSE are the mean regression sum of squares and the mean error sum of squares, respectively.

- The numerator degrees of freedom, $df_1 = k$, while the denominator degrees of freedom, $df_2 = n - k - 1$.

15.2 A General Test of Linear Restrictions

LO 15.3 **Conduct a general test of linear restrictions.**

- The significance tests in the previous section can also be labeled tests of **linear restrictions**.
- For example, if we have $k = 3$ explanatory variables, testing whether $\beta_2 = \beta_3 = 0$ is equivalent to testing whether to restrict the model to only x_1.
- In this section, we apply the F-test for any number of linear restrictions; the resulting F-test is often referred to as the **partial F**-test.

15.2.1 Restricted and Unrestricted Models

- To conduct the partial F-test, we estimate the model with and without the restrictions.
- In the restricted model, we do not estimate the coefficients that are restricted under the null hypothesis.
- The unrestricted model is a complete model that imposes no restrictions on the coefficients.
- If restrictions are valid, then the restricted model's error sum of squares, SSE_R, will not be significantly larger than the unrestricted model's error sum of squares SSE_U.

15.2.2 The Test Statistic

- The test statistic for a partial F-test can be computed as

$$F_{(df_1,df_2)} = \frac{\frac{SSE_R - SSE_U}{df_1}}{\frac{SSE_U}{df_2}} \tag{15.5}$$

  where the numerator degrees of freedom, df_1, equals the number of restrictions on the model, and the denominator degrees of freedom, df_2, equals $n - k - 1$.

- If the test statistic is greater than the critical value, then we reject the null hypothesis and the restrictions are not valid.

15.2.3 Car Wash Example

- A manager at a car wash company wants to determine which promotions improve sales.
- He has information on sales, price discounts, and advertising expenditures on Radio and Newspaper in 40 Missouri counties. (Columns = [Country; Sales (in \$1,000s); Discount (in %); Radio (in \$1,000s); Newspaper (in \$1,000s)])

- More specifically, he would like to test whether either type of advertising impacts sales. To do so, we form the competing hypotheses as:

$$\begin{aligned} H_0 &: \ \beta_2 = \beta_3 = 0 \\ H_A &: \ \text{At least one of the coefficients is nonzero.} \end{aligned}$$

- To conduct the test, we need to estimate a restricted model (R) and an unrestricted model (U):

$$\begin{aligned} &\text{(R) Sales} = \beta_0 + \beta_1 \text{Discount} + \epsilon \\ &\text{(U) Sales} = \beta_0 + \beta_1 \text{Discount} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper} + \epsilon \end{aligned}$$

- From the tables (which I'm not adding to this PDF), we can see that $SSE_U = 1208.1348$ while the $SSE_R = 2182.5649$. We can now proceed to compute the value of the test statistic.
- The number of restrictions, df_1, equals 2. The unrestricted model has $df_2 = n - k - 1 = 40 - 3 - 1 = 36$ degrees of freedom.
- We compute the value of the test statistic as:

$$\begin{aligned} F_{2,36} &= \frac{\frac{2182.5649 - 1208.1348}{2}}{\frac{1208.1348}{36}} \\ &= \frac{487.2151}{33.5593} \\ &= 14.52 \end{aligned}$$

- Since $F_{(2,36)} = 14.52$ is greater than the critical value $F_{0.05,(2,36)} = 3.26$, we reject the null hypothesis and conclude the restrictions are not valid.

15.3 Interval Estimates for Predictions

LO 15.4 **Calculate and interpret interval estimates for predictions.**

- Once we have developed a regression model, we often want to use it to make predictions.
- From the introductory case, what would we predict for a team with an earned run average of 4.00 and a batting average of 0.250? Plugging these values into the estimated equation, we find:

$$\hat{\text{Win}} = 0.13 + 3.28(0.250) - 0.12(4.00) = 0.47$$

- But this is only a point estimate and ignores sampling error. We could also provide interval estimates.

15.3.1 Two Types of Predictions

- We will develop two types of interval estimates regarding y:
  1. A confidence interval for the expected value of y
  2. A prediction interval for an individual value of y
- It is common to refer to the first as a confidence interval and the second as a prediction interval.

15.3.2 The Confidence Interval

- The point estimate of $E(y^0)$ is just the $\hat{y}$ value:

$$\hat{y}^0 = b_0 + b_1 x_1^0 + b_2 x_2^0 + \cdots + b_k x_K^0 \tag{15.6}$$

- The confidence interval, as always, includes the point estimate, plus or minus the margin of error:

$$\hat{y}^0 \pm t_{\alpha/2,df} se(\hat{y}^0) \tag{15.7}$$

- The term $se(\hat{y}^0)$ is the standard error of the prediction. Though difficult to compute by hand if there is more than one explanatory variable in the model, we will develop a procedure to compute it with a statistical package.

15.4 Model Assumptions and Common Violations

LO 15.5 **Explain the role of the assumptions on the OLS estimators.**

- The statistical properties of OLS estimator, as well as the validity of the testing procedures, depend on a number of assumptions. We discuss those assumptions now.
  1. The model (**??**) is linear in the β parameters with an additive error ϵ.
  2. Conditional on the $x_1, \ldots, x_k$ values, the expected error is 0, thus:

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{15.8}$$

  3. There is no exact linear relationship among the $x_1, \ldots, x_k$ values (no perfect multicollinearity).
  4. The variance of the error term ϵ is the same for all $x_1, \ldots, x_k$ values. We call this homoskedasticity.
  5. The error term ϵ is uncorrelated across observations, conditional on the explanatory variables. There is no serial correlation or autocorrelation.
  6. The error term ϵ is not correlated with any of the predictors $x_1, \ldots, x_k$, In other words, there is no endogeneity.
  7. The error term ϵ is normally distributed. This assumption allows us to do hypothesis testing. If normality is not true, the tests may not be valid.

15.4.1 Checking the Assumptions

- The true error terms ϵ cannot be observed because they exist only in the population. We can, however, look at the residuals, $e = y - \hat{y}$, where $\hat{y} = b_0 + \sum_{i=1}^{k} b_i x_i$, for each observation.
- It is common to plot to residuals on the vertical axis and an explanatory variable on the horizontal axis.
- When estimating a regression in Excel, the dialog box that opens after choosing **Data** > **Data Analysis** > **Regression** allows us to select *Residuals* and Residual Plots options.

15.4.2 Common Violation 1: The Model Suffers from Multicollinearity

LO 15.6 **Describe common violations of the assumptions and offer remedies.**

- Perfect multicollinearity exists when two or more x variables have an exact linear relationship.
- For example, suppose the x data includes total cost, fixed cost and variable cost.
- Other data sets may have a great degree of multicollinearity that is not perfect.
- In these cases, we may see a high R^2 coupled with individually insignificant explanatory variables. Additionally, unintuitive result may be indicative.
- A sample correlation between explanatory variables that is > 0.80 or < -0.80 suggests severe multicollinearity.

Remedying Multicollinearity

- A good remedy may be simply drop one of the collinear variables if we can justify it as redundant.
- Alternatively, we could try to increase our sample size.
- Another option would be to try to transform our variables so that they are no longer collinear.
- Last, especially if we are interested only in maintaining a high predictive power, it may make sense to do nothing.

15.4.3 Common Violation 2: The Error Term is Heteroskedastic

- The variance of the error term changes for different values of at least one explanatory variable.
- Informal residual plots can gauge heteroskedasticity (display a marked pattern, such as increasing magnitude along an axis).

Remedying Heteroskedastic

- Heteroskedastic results in inefficient estimators and the hypothesis tests for significance are no longer valid.
- To get around the second problem, some researchers use OLS estimates along with corrected standard errors, called White's standard errors. Many statistical packages have this option available, unfortunately the current version of Excel does not.

15.4.4 Common Violation 3: The Error Term is Serially Correlated

- We assume that the error term is uncorrelated across observations when obtaining OLS estimates.
- But this often breaks down in time series data.
- Remedies are not easily accessibly using Excel.

15.4.5 Common Violation 4: The Explanatory Variable is Endogenous

- Endogeneity in the regression model refers to the error term being correlated with the explanatory variables.
- This commonly occurs due to an omitted explanatory variable.
- For example, a person's salary may be highly correlated with that person's innate ability. But since we cannot included it, ability gets incorporated in the error term. If we try to predict salary by years of education, which may also be correlated with innate ability, then we have an endogeneity problem.
- Endogeneity will result in biased estimators, and so is quite a serious problem.
- Unfortunately, endogeneity is difficult to fix. Most commonly, we would like to find an instrumental variable, one that is correlated with the endogenous explanatory variable but uncorrelated with the error term. But it may be difficult to find such a variable.