Statistical Analysis in Fin Mkts

MSF 502

Li Cai

15 Inference with Regression Models

CHAPTER

Chapter 15 Learning Objectives (LOs)

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.

Analyzing the Winning Percentage in Baseball

Team	League	Win	BA	ERA
Baltimore Orioles	AL	0.407	0.259	4.59
Boston Red Sox	AL	0.549	0.268	4.20
⋮	⋮	⋮	⋮	⋮
Washington Nationals	NL	0.426	0.250	4.13

- Sports analysts frequently quarrel over what statistics separate winning teams from the losers.
- Is a high batting average (BA) the best predictor, or is it a low earned run average (ERA)? Or both?
- We will fit three regression models and use the statistical significance of the predictors to help decide.

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:

$$\text{Model 1: } \text{Win} = \beta_0 + \beta_1 \text{BA} + \varepsilon$$

$$\text{Model 2: } \text{Win} = \beta_0 + \beta_1 \text{ERA} + \varepsilon$$

$$\text{Model 3: } \text{Win} = \beta_0 + \beta_1 \text{BA} + \beta_2 \text{ERA} + \varepsilon$$

	Model 1	Model 2	Model 3
Multiple *R*	0.4596	0.6823	0.8459
R Square	0.2112	0.4656	0.7156
Adjusted *R* Square	0.1830	0.4465	0.6945
Standard Error	0.0614	0.0505	0.0375
Observations	30	30	30

LO

Tests of Individual Significance

- Consider our standard multiple regression model:

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

- 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:

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

LO

The Test Statistic

- The appropriate test statistic is $t_{df} = \frac{b_j - \beta_{j0}}{s_{b_j}}$.
- 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*.

LO

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:

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

- If a coefficient is equal to zero, then it implies that the explanatory variable is not a significant predictor of the response variable.

LO 15.1

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 to test 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 nonzero value, we need to compute a new test statistic.

LO

Example 15.1

- To test whether batting average influences winning percentage, we set up the following hypotheses:

$$H_0: \beta_1 = 0; H_A: \beta_1 \neq 0$$

- Then, examine the regression output.

	Coefficients	Standard Error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	0.1269	0.1822	0.6964	0.4921	−0.25	0.50
BA	3.2754	0.6723	4.8719	0.0000	1.90	4.65
ERA	−0.1153	0.0167	−6.9197	0.0000	−0.15	−0.08

- The value of the test statistic is $t_{27} = 4.817$ and its *p*-value is very close to zero. We reject the null hypothesis and conclude that batting average is a significant predictor.

LO

Intervals for the Parameters

- A confidence interval for a β_j parameter can be constructed using the formula: $b_j \pm t_{\alpha/2,df} s_{b_j}$
- 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.

	Coefficients	Standard Error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	0.1269	0.1822	0.6964	0.4921	−0.25	0.50
BA	3.2754	0.6723	4.8719	0.0000	1.90	4.65
ERA	−0.1153	0.0167	−6.9197	0.0000	−0.15	−0.08

LO

A Test for a Non-Zero Slope

- The capital asset pricing model follows the equation:

$$y = \alpha + \beta x + \varepsilon$$

y = the risk-adjusted return of an asset, $R - R_f$.

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 value < 1 implies it is "conservative."

LO

Example 15.3

- We use 60 months of data to estimate the beta for Johnson & Johnson (J&J). The data are available on the text website, labeled ***Johnson & Johnson***.

- We want to test whether J&J is a conservative stock, so we set up our competing hypotheses:

H_0: $\beta \geq 1$ (J&J stock is not conservative)

H_A: $\beta < 1$ (J&J stock is conservative)

LO 15.1

Example 15.3

- A simple regression yields b_1 = 0.5844 with a standard error s_{b_1} = 0.0803.

	Coefficients	Standard Error	*t* Stat	*p*-value
Intercept	0.2666	0.4051	0.6580	0.5131
$R_M - R_f$	0.5844	0.0803	7.2759	0.0000

- The test statistic for our claim is then calculated as

$$t_{58} = (0.5844 - 1)/0.0803 = -7.276.$$

- Since the value of the test statistic is less than the critical value $-t_{.05,58} = -1.672$, we reject the null hypothesis and conclude that the stock's "beta" is less than 1, making it a conservative investment.

Test of Joint Significance

LO 15.2 Conduct a test of joint significance.

- In addition to conducting tests of individual significance, we also may want to test the joint significance of all *k* variables at once.

- The competing hypotheses for a test of joint significance are:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_A: \text{at least one } \beta_j \neq 0$$

LO 15.2

The Test Statistic

- The test statistic for a test of joint significance is

$$F_{(df_1,\, df_2)} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE},$$

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, equal *k*, while the denominator degrees of freedom, df_2, are $n - k - 1$.

- Fortunately, statistical software will generally report the value of $F_{(df_1,\, df_2)}$ and its *p*-value as standard output, making computation by hand rarely necessary.

LO

Example 15.4

- We want to conduct a joint test of significance for the model: $Win = \beta_0 + \beta_1 BA + \beta_2 ERA + \varepsilon$
- So we set up the following hypotheses:

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_A: \text{at least one } \beta_j \neq 0$$

- From the ANOVA portion of the regression results, we see that $F_{(2,27)}$=33.9963 and its *p*-value is quite small (see value under *Significance F*), so we reject the null hypothesis, and conclude that the explanatory variables (regression) are jointly significant.

Synopsis of the Introductory Case

- Goodness-of-fit measures indicated that including both batting average and ERA is most appropriate.

Variable	Model 1	Model 2	Model 3
Intercept	−0.2731 (0.3421)	0.9504* (0.0000)	0.1269 (0.4921)
Batting Average	3.0054* (0.0106)	NA	3.2754* (0.0000)
Earned Run Average	NA	−0.1105* (0.0000)	−0.1153* (0.0000)
s_e	0.0614	0.0505	0.0375
R^2	0.2112	0.4656	0.7156
Adjusted R^2	0.1830	0.4465	0.6945
F-test (p-value)			33.9663*(0.0000)

- In Model 3, explanatory variables are individually significant and the regression is jointly significant.
- We can conclude that both batting average and earned run average are good predictors of overall winning percentage.

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.

LO

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.

LO

The Test Statistic

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

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

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

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

LO

Example 15.5

- 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.

County	Sales (in $1,000s)	Discount (in %)	Radio (in $1,000s)	Newspaper (in $1,000s)
1	62.72	40	2.27	3.00
2	49.65	20	3.78	1.78
⋮	⋮	⋮	⋮	⋮
40	49.95	40	3.57	1.57

LO Testing the Effects of Advertising

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

  H_0: $\beta_2 = \beta_3 = 0$

  H_A: At least one of the coefficients is nonzero.

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

$$\text{(R)}\quad \text{Sales} = \beta_0 + \beta_1 \text{Discount} + \varepsilon$$

$$\text{(U)}\quad \text{Sales} = \beta_0 + \beta_1 \text{Discount} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper} + \varepsilon.$$

LO

Estimate the Two Models

The table on the right displays the estimates and *SSE* for each model. The *p*-values are in parentheses.

Variable	Restricted	Unrestricted
Intercept	43.4541* (0.0000)	6.7025 (0.3559)
Discount	0.4016* (0.0001)	0.3417* (0.0000)
Radio	NA	6.0624* (0.0007)
Newspaper	NA	9.3968* (0.0001)
SSE	2182.5649	1208.1348

We can see from the table that the $SSE_U = 1208.1348$, while the $SSE_R = 2182.5649$. We can now proceed to computing the value of the test statistic.

LO

Example 15.5

- 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:

$$F_{(2,36)} = \frac{(2182.5649 - 1208.1348)/2}{1208.1348/36} = \frac{487.2151}{33.5593} = 14.52$$

- 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.

LO

Are Advertising Returns Equal?

- The same car wash company manager does not think that advertising expenditures in newspapers and on radio influence sales in the same way.

- To test this claim, we develop these hypotheses:

  H_0: $\beta_2 = \beta_3$ (Advertising returns are equal)

  H_A: $\beta_2 \neq \beta_3$ (One type is more effective than the other)

- But, the restriction is of a different type in this case, so we must develop our restricted model in a slightly different way.

LO

Example 15.6

- The unrestricted model is as before:

$$\text{(U)}\ \ \text{Sales} = \beta_0 + \beta_1 \text{Discount} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper} + \varepsilon.$$

- If we apply our restriction, the model changes to:

$$\text{Sales} = \beta_0 + \beta_1 \text{Discount} + \beta_2 \text{Radio} + \beta_2 \text{Newspaper} + \varepsilon.$$

- But, we can rewrite this as:

$$\text{(R)}\ \ \text{Sales} = \beta_0 + \beta_1 \text{Discount} + \beta_2 (\text{Radio} + \text{Newspaper}) + \varepsilon.$$

- So when we estimate the model, the second explanatory

LO 15.3

Conducting the Test

We first estimate both regression models:

Variable	Restricted	Unrestricted
Intercept	7.9524 (0.2740)	6.7025 (0.3559)
Discount	0.3517* (0.0000)	0.3417* (0.0000)
Radio	7.1831* (0.0000)	6.0624* (0.0007)
Newspaper	Same as for Radio	9.3968* (0.0001)
SSE	1263.6243	1208.1348

- We can then compute the value of the test statistic as:

$$F_{(1,36)} = \frac{(1263.6243 - 1208.1348)/1}{1208.1348/36} = 1.65$$

- We fail to reject the null hypothesis with a *p*-value of 0.2072 (Excel function F.DIST.RT(1.65,1,36)); we cannot conclude that differences exist in type of advertising.

15.3: Interval Estimates for Predictions

LO 15.4 Calculate and interpret interval estimates for

- 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:

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

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

LO

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.

LO

The Confidence Interval

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

$$\hat{y}^0 = b_0 + b_1 x_1^0 + b_2 x_2^0 + \ldots + b_k x_k^0$$

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

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

- 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.

LO 15.4

Modified Regression

- Many statistics programs will compute confidence intervals, but Excel's Data Analysis Tools do not.
- However, here is a neat trick you can use. Shift the values of the explanatory variables in your data set by the value of interest for each one:

$$x_1^* = x_1 - x_1^0,\ x_2^* = x_2 - x_2^0,\ \ldots,\ x_k^* = x_k - x_k^0$$

- When we estimate this modified regression, the resulting estimate of the intercept and its standard error equal $\hat{y}^0$ and $se(\hat{y}^0)$, respectively.
- The 95% confidence interval is given in the same row.

LO 15.4

Confidence Interval for Winning Percentage

- In the baseball example, we first shift the data by our hypothesized values:

y	x_1	x_2	$x_1^* = x_1 - 0.25$	$x_2^* = x_2 - 4$
0.407	0.259	4.59	$0.259 - 0.25 = 0.009$	$4.59 - 4 = 0.59$
0.549	0.268	4.20	$0.268 - 0.25 = 0.018$	$4.20 - 4 = 0.20$
⋮	⋮	⋮	⋮	⋮
0.426	0.250	4.13	$0.250 - 0.25 = 0.000$	$4.13 - 4 = 0.13$

- Estimating the modified regression now reveals the confidence interval:

	Coefficients	Standard Error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	**0.4847**	**0.0085**	57.2582	0.0000	**0.4673**	**0.5021**
x_1^*	3.2754	0.6723	4.8719	0.0000	1.8960	4.6549
x_2^*	−0.1153	0.0167	−6.9197	0.0000	−0.1494	−0.0811

LO

Confidence Interval for Winning Percentage

- To summarize, after shifting the explanatory variables, the intercept row in the regression output gives us all the information we need.

- For a 95% confidence interval when the BA is 0.250 and the ERA is 4.00, $\hat{y}^0 \pm t_{\alpha/2,df} \ se(\hat{y}^0)$ = [0.4673, 0.5021].

- If we want to compute an interval with a different confidence level, we simply need to find the correct $t_{\alpha/2,df}$ statistic and plug in the intercept and standard error of the intercept from that same regression, or alternatively, invoke a different level with Excel.

LO 15.4

The Prediction Interval

- The formula for the prediction interval:

$$\hat{y}^0 \pm t_{\alpha/2,df} \sqrt{\left(se(\hat{y}^0) \right)^2 + {s_e}^2}$$

- The point estimate and the standard error of the prediction are computed using the same regression technique as for the confidence interval.

- But now we need to include the standard error of the estimate into the margin of error calculation.

LO

Prediction Interval for Winning Percentage

- To compute the prediction interval for a team with a batting average of 0.250 and an ERA of 4.00, we simply plug in the appropriate values from the previous example, plus the standard error of the estimate, which is 0.375:

$$0.4847 \pm 2.052\sqrt{0.0085^2 + 0.0375^2} = 0.4058 \text{ to } 0.5636$$

- Remember that the prediction interval is an interval estimate for *one* team with these characteristics while the confidence interval pertains to the average of all teams with these characteristics.

15.4: Model Assumptions and Common Violations

LO 15.5 Explain the role of the assumptions on the OLS

The statistical properties of the 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 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$ 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 $\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$

3. There is no exact linear relationship among the $x_1, \ldots, x_k$ values (no perfect **multicollinearity**).

LO

Assumptions (continued)

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.

LO 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 $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$, for each observation.
- It is common to plot the residuals on the vertical axis and an explanatory 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.

Common Violation 1: The Model Suffers from Multicollinearity

LO 15.6 Describe common violations of the assumptions and offer

- 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 results may be indicative.
- A sample correlation between explanatory variables that is greater than 0.80 or less than -0.80 suggests severe multicollinearity.

LO

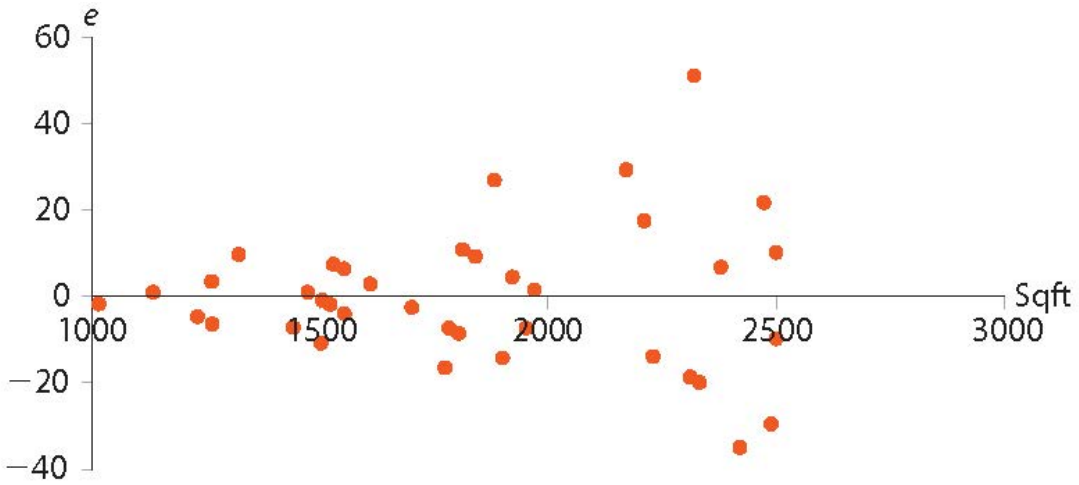
Remedying Multicollinearity

- A good remedy may be to 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.

LO

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. In Example 15.10, we try to predict sales by the square footage of convenience stores. The residuals display a marked pattern:



LO

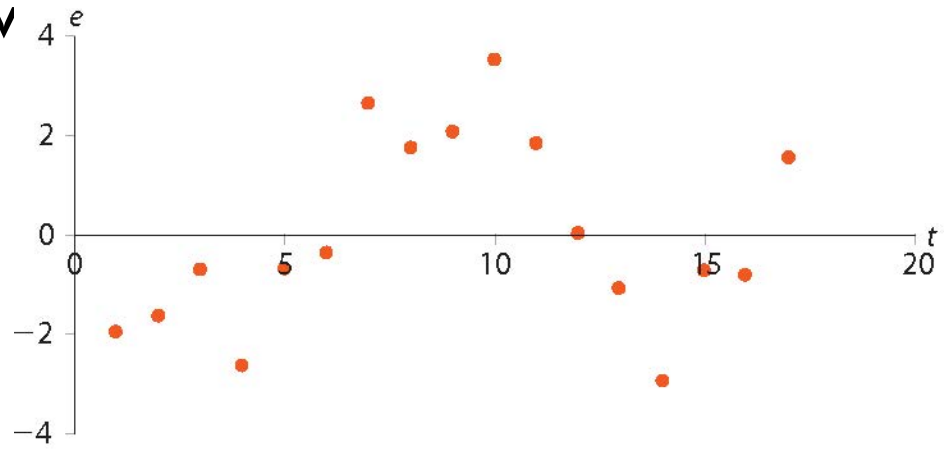
Remedying Heteroskedasticity

- Heteroskedasticity 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.

LO

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. In Example 15.11, we predict sales at a sushi restaurant over a period of time. A plot of the residuals against time show



- Remedies are not easily accessible using Excel.

LO

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 include 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.

LO

Common Violation 4: The Explanatory Variable is Endogenous

- 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.

- Further discussion of the instrumental variable approach is beyond the scope of the text.

End of Chapter