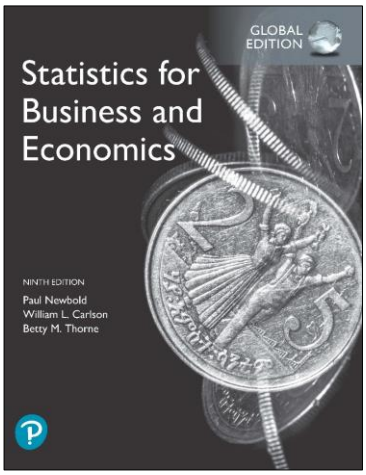
Statistics for Business and Economics

Ninth Edition, Global Edition



Chapter 13

Additional Topics in Regression Analysis

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

1

Chapter Goals

After completing this chapter, you should be able to:

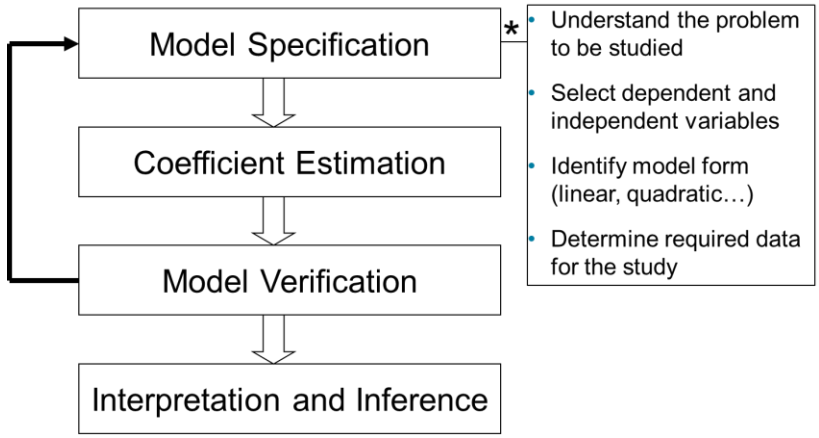
- Explain regression model-building methodology
- Apply dummy variables for categorical variables with more than two categories
- Explain how dummy variables can be used in experimental design models
- Incorporate lagged values of the dependent variable is regressors
- Describe specification bias and multicollinearity
- Examine residuals for heteroscedasticity and autocorrelation

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

2

Section 13.1 Model-Building Methodology

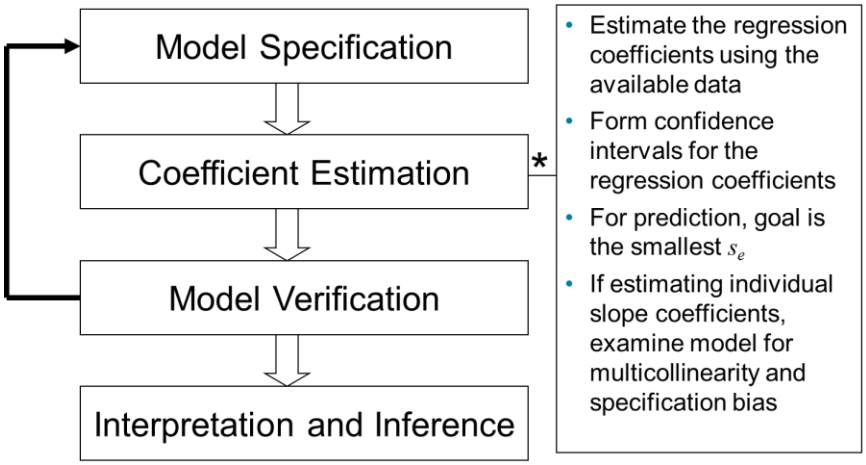
The Stages of Statistical Model Building



Pearson
Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

3

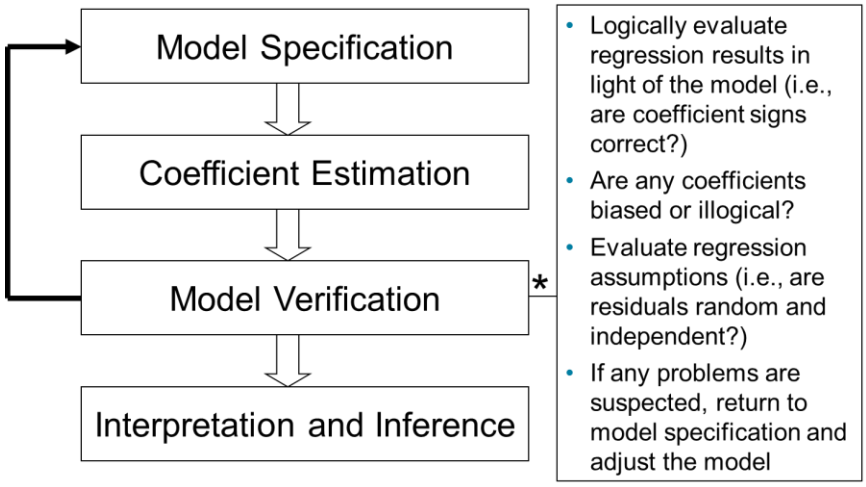
The Stages of Model Building (1 of 3)



Pearson
Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

4

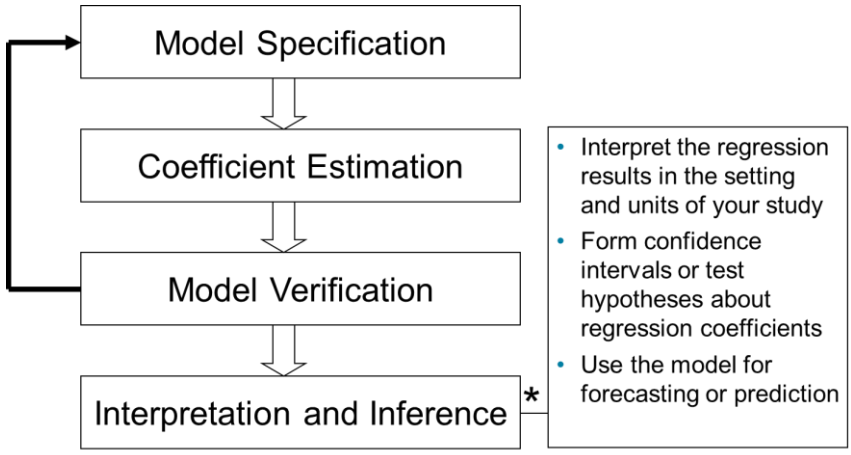
The Stages of Model Building (2 of 3)



Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

5

The Stages of Model Building (3 of 3)



Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

6

Section 13.2 Dummy Variables and Experimental Design

- Dummy variables can be used in situations in which the categorical variable of interest has more than two categories
- Dummy variables can also be useful in experimental design
  - Experimental design is used to identify possible causes of variation in the value of the dependent variable
  - *Y* outcomes are measured at specific combinations of levels for treatment and blocking variables
  - The goal is to determine how the different treatments influence the *Y* outcome

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

7

Dummy Variable Models (More Than 2 Levels) (1 of 2)

- Consider a categorical variable with *K* levels
- The number of dummy variables needed is **one less than the number of levels, *K* − 1**
- Example:

  y = house price ; x_1 = square feet
- If style of the house is also thought to matter:

Style = ranch, split level, condo

Three levels, so two dummy variables are needed

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

8

Dummy Variable Models (More Than 2 Levels) (2 of 2)

- Example: Let "condo" be the default category, and let x_2 and x_3 be used for the other two categories:

y = house price

x_1 = square feet

x_2 = 1 if ranch, 0 otherwise

x_3 = 1 if split level, 0 otherwise

The multiple regression equation is:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

9

Interpreting the Dummy Variable Coefficients (with 3 Levels)

Consider the regression equation:

$$\hat{y} = 20.43 + 0.045x_1 + 23.53x_2 + 18.84x_3$$

For a condo: $x_2 = x_3 = 0$

$\hat{y} = 20.43 + 0.045x_1$

For a ranch: $x_2 = 1; x_3 = 0$

$\hat{y} = 20.43 + 0.045x_1 + 23.53$

With the same square feet, a ranch will have an estimated average price of 23.53 thousand dollars more than a condo

For a split level: $x_2 = 0; x_3 = 1$

$\hat{y} = 20.43 + 0.045x_1 + 18.84$

With the same square feet, a split-level will have an estimated average price of 18.84 thousand dollars more than a condo.

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

10

Experimental Design (1 of 2)

- Consider an experiment in which
  - four treatments will be used, and
  - the outcome also depends on three environmental factors that cannot be controlled by the experimenter
- Let variable z_1 denote the treatment, where $z_1 = 1, 2, 3,$ or 4. Let z_2 denote the environment factor (the "blocking variable"), where $z_2 = 1, 2,$ or 3
- To model the four treatments, three dummy variables are needed
- To model the three environmental factors, two dummy variables are needed

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

11

Experimental Design (2 of 2)

- Define five dummy variables, $x_1, x_2, x_3, x_4,$ and x_5
- Let treatment level 1 be the default $(z_1 = 1)$
  - Define $x_1 = 1$ if $z_1 = 2$, $x_1 = 0$ otherwise
  - Define $x_2 = 1$ if $z_1 = 3$, $x_2 = 0$ otherwise
  - Define $x_3 = 1$ if $z_1 = 4$, $x_3 = 0$ otherwise
- Let environment level 1 be the default $(z_2 = 1)$
  - Define $x_4 = 1$ if $z_2 = 2$, $x_4 = 0$ otherwise
  - Define $x_5 = 1$ if $z_2 = 3$, $x_5 = 0$ otherwise

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

12

Experimental Design: Dummy Variable Tables

- The dummy variable values can be summarized in a table:

Z_1	X_1	X_2	X_3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Z_2	X_4	X_5
1	0	0
2	1	0
3	0	1

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Experimental Design Model

- The experimental design model can be estimated using the equation

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon$$

- The estimated value for β_2, for example, shows the amount by which the *y* value for treatment 3 exceeds the value for treatment 1

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Section13.3 Lagged Values of the Dependent Variable

- In time series models, data is collected over time (weekly, quarterly, etc…)
- The value of *y* in time period *t* is denoted y_t
- The value of y_t often depends on the value y_{t-1}, as well as other independent variables x_j :

$$y_i = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \gamma y_{t-1} + \varepsilon_t$$

A lagged value of the dependent variable is included as an explanatory variable

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Interpreting Results in Lagged Models (1 of 3)

- An increase of 1 unit in the independent variable x_j in time period *t* (all other variables held fixed), will lead to an expected increase in the dependent variable of
  - β_j in period t
  - $\beta_j\gamma$ in period $(t+1)$
  - $\beta_j\gamma^2$ in period $(t+2)$
  - $\beta_j\gamma^3$ in period $(t+3)$ and so on
- The total expected increase over all current and future time periods is $\dfrac{\beta_j}{(1-\gamma)}$
- The coefficients $\beta_0, \beta_1, \ldots, \beta_K, \gamma$ are estimated by least squares in the usual manner

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Interpreting Results in Lagged Models (2 of 3)

- Confidence intervals and hypothesis tests for the regression coefficients are computed the same as in ordinary multiple regression
  - (When the regression equation contains lagged variables, these procedures are only approximately valid. The approximation quality improves as the number of sample observations increases.)

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Interpreting Results in Lagged Models (3 of 3)

- Caution should be used when using confidence intervals and hypothesis tests with time-series data
  - There is a possibility that the equation errors $\mathcal{E}_i$ are no longer independent from one another.
  - When errors are correlated the coefficient estimates are unbiased, but not efficient. Thus confidence intervals and hypothesis tests are no longer valid.

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Section 13.4 Specification Bias

- Suppose an important independent variable *z* is omitted from a regression model
- If *z* is uncorrelated with all other included independent variables, the influence of *z* is left unexplained and is absorbed by the error term, ε
- But if there is any correlation between *z* and any of the included independent variables, some of the influence of *z* is captured in the coefficients of the included variables

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

19

Specification Bias

- If some of the influence of omitted variable *z* is captured in the coefficients of the included independent variables, then those coefficients are biased…
- …and the usual inferential statements from hypothesis test or confidence intervals can be seriously misleading
- In addition the estimated model error will include the effect of the missing variable(s) and will be larger

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

20

Section 13.5 Multicollinearity

- Collinearity: High correlation exists among two or more independent variables
- This means the correlated variables contribute redundant information to the multiple regression model

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

21

Multicollinearity

- Including two highly correlated explanatory variables can adversely affect the regression results
  - No new information provided
  - Can lead to unstable coefficients (large standard error and low *t*-values)
  - Coefficient signs may not match prior expectations

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

22

Indicators of Multicollinearity

- Coefficients differ from the values expected by theory or experience, or have incorrect signs
- Coefficients of variables believed to be a strong influence have small *t* statistics indicating that their values do not differ from 0
- All the coefficient student *t* statistics are small, indicating no individual effect, but the overall *F* statistic indicates a strong effect for the total regression model
- High correlations exist between individual independent variables or one or more of the independent variables have a strong linear regression relationship to the other independent variables or a combination of both

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Detecting Multicollinearity

- Examine the simple correlation matrix to determine if strong correlation exists between any of the model independent variables
- Look for a large change in the value of a previous coefficient when a new variable is added to the model
- Does a previously significant variable become insignificant when a new independent variable is added?
- Does the estimate of the standard deviation of the model increase when a variable is added to the model?

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Corrections for Multicollinearity

- Remove one or more of the highly correlated independent variables. But, as shown in Section 13.4, this might lead to a bias in coefficient estimation.
- Change the model specification, including possibly a new independent variable that is a function of several correlated independent variables.
- Obtain additional data that do not have the same strong correlations between the independent variables.

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Section 13.6 Heteroscedasticity

- Homoscedasticity
  - The probability distribution of the errors has constant variance
- Heteroscedasticity
  - The error terms do not all have the same variance
  - The size of the error variances may depend on the size of the dependent variable value, for example

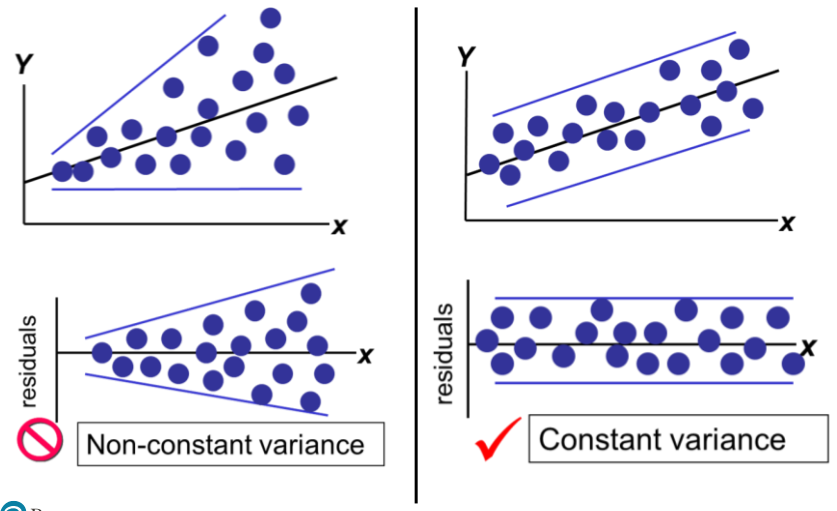
Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Heteroscedasticity (2 of 2)

- When heteroscedasticity is present:
  - least squares is not the most efficient procedure to estimate regression coefficients
  - The usual procedures for deriving confidence intervals and tests of hypotheses is not valid

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Residual Analysis for Homoscedasticity



Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Tests for Heteroscedasticity

- To test the null hypothesis that the error terms, ε_i, all have the same variance against the alternative that their variances depend on the expected values $\hat{y}_i$
- Estimate the simple regression $e_i^2 = a_0 + a_1\hat{y}_i$
- Let R^2 be the coefficient of determination of this new regression

  The null hypothesis is rejected if nR^2 is greater than $\chi^2_{1,\alpha}$

  – where $\chi^2_{1,\alpha}$ is the critical value of the chi-square random variable with 1 degree of freedom and probability of error α

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Section 13.7 Autocorrelated Errors

- Independence of Errors
  – Error values are statistically independent
- Autocorrelated Errors
  – Residuals in one time period are related to residuals in another period

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

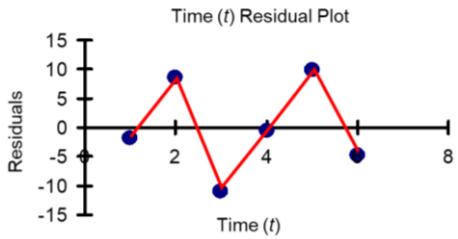
Autocorrelated Errors

- Autocorrelation violates a least squares regression assumption
  - Leads to s_b estimates that are too small (i.e., biased)
  - Thus *t*-values are too large and some variables may appear significant when they are not

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Autocorrelation

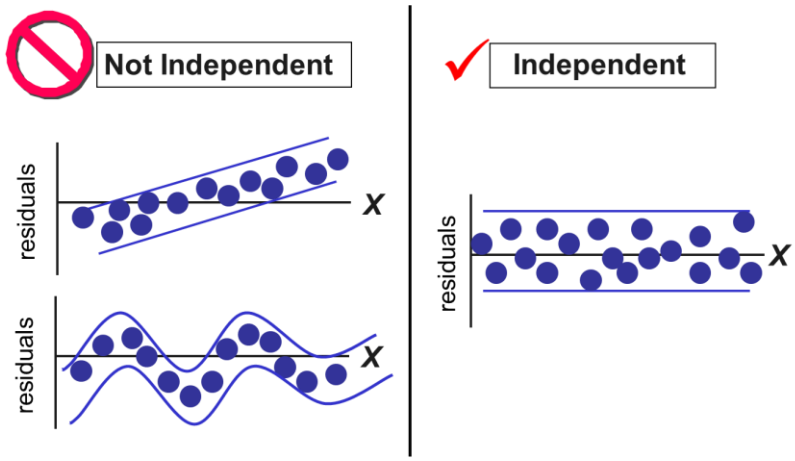
- Autocorrelation is correlation of the errors (residuals) over time
- Here, residuals show a cyclic pattern, not random



- Violates the regression assumption that residuals are random and independent

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Residual Analysis for Independence



Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

The Durbin-Watson Statistic (1 of 2)

- The Durbin-Watson statistic is used to test for autocorrelation

H_0 : successive residuals are not correlated
$\left(\text{i.e., } Corr\left(\varepsilon_t, \varepsilon_{t-1}\right) = 0\right)$
H_1 : autocorrelation is present

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

The Durbin-Watson Statistic (2 of 2)

H_0 : $\rho = 0$ (no autocorrelation)

H_1 : autocorrelation is present

- The Durbin-Watson test statistic (*d*):

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- The possible range is $0 \le d \le 4$
- *d* should be close to 2 if H_0 is true
- *d* less than 2 may signal positive autocorrelation,
- *d* greater than 2 may signal negative autocorrelation

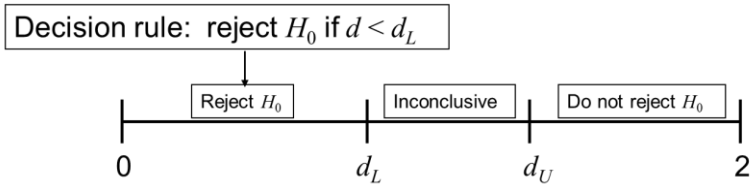
Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Testing for Positive Autocorrelation (1 of 3)

H_0 : positive autocorrelation does not exist

H_1 : positive autocorrelation is present

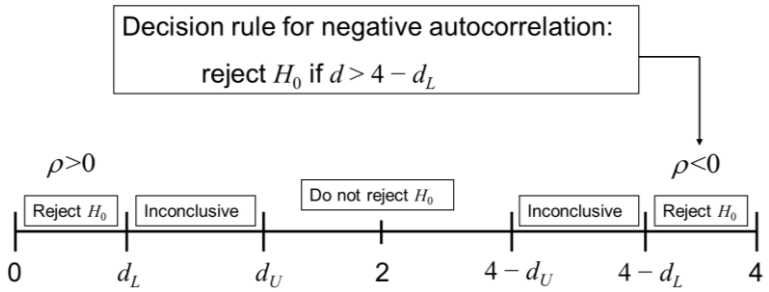
- Calculate the Durbin-Watson test statistic = *d*
  - *d* can be approximated by $d = 2(1-r)$, where *r* is the sample correlation of successive errors
- Find the values d_L and d_U from the Durbin-Watson table
  - (for sample size *n* and number of independent variables *K*)



Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Negative Autocorrelation

- Negative autocorrelation exists if successive errors are negatively correlated
  - This can occur if successive errors alternate in sign



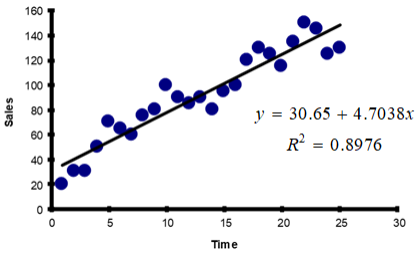
Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

37

Testing for Positive Autocorrelation (2 of 3)

- Example with n = 25:

Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	3296.18
Sum of Squared Residuals	3279.98
Durbin-Watson Statistic	**1.00494**



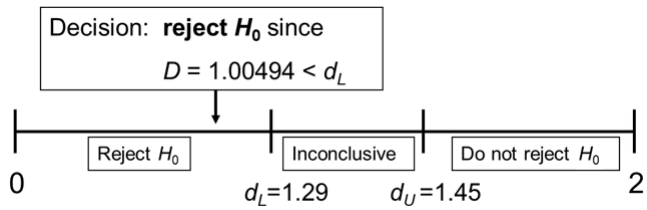
$$d = \frac{\sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^{\,2}} = \frac{3296.18}{3279.98} = 1.00494$$

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

38

Testing for Positive Autocorrelation (3 of 3)

- Here, n = 25 and there is K = 1 independent variable
- Using the Durbin-Watson table, $d_L = 1.29$ and $d_U = 1.45$
- $D = 1.00494 < d_L = 1.29$, so reject H_0 and conclude that significant positive autocorrelation exists
- Therefore the linear model is not the appropriate model to forecast sales



Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Estimation of Regression Models with Autocorrelated Errors (1 of 2)

- Suppose that we want to estimate the coefficients of the regression model

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

where the error term ε_t is autocorrelated

- Two steps:

(i) Estimate the model by least squares, obtaining the Durbin-Watson statistic, d, and then estimate the autocorrelation parameter using

$$r = 1 - \frac{d}{2}$$

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Estimation of Regression Models with Autocorrelated Errors (2 of 2)

(ii) Estimate by least squares a second regression with

- dependent variable $\left(y_t - ry_{t-1}\right)$
- independent variables $\left(x_{1t} - rx_{1,t-1}\right), \left(x_{2t} - rx_{2,t-1}\right), \ldots, \left(x_{Kt} - rx_{K,t-1}\right)$

- The parameters $\beta_1, \beta_2, \ldots, \beta_k$ are estimated regression coefficients from the second model
- An estimate of β_0 is obtained by dividing the estimated intercept for the second model by (1−*r*)
- Hypothesis tests and confidence intervals for the regression coefficients can be carried out using the output from the second model

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.

Chapter Summary

- Discussed regression model building
- Introduced dummy variables for more than two categories and for experimental design
- Used lagged values of the dependent variable as regressors
- Discussed specification bias and multicollinearity
- Described heteroscedasticity
- Defined autocorrelation and used the Durbin-Watson test to detect positive and negative autocorrelation

Pearson Copyright © 2020 Pearson Education Ltd. All Rights Reserved.