# Lecture 11

# 11.1. Simple Linear Regression

## *Summary*

In this lecture we showed how regression analysis can be used to determine how a dependent variable *y* is related to an independent variable *x*. In simple linear regression, the regression model is $y = \beta_0 + \beta_1 x + \varepsilon$. The simple linear regression equation $E(y) = \beta_0 + \beta_1 x$ describes how the mean or expected value of *y* is related to *x*. We used sample data and the least squares method to develop the estimated regression equation $y = b_0 + b_1 x$. In effect, $b_0$ and $b_1$ are the sample statistics used to estimate the unknown model parameters $\beta_0$ and $\beta_1$.

The coefficient of determination was presented as a measure of the goodness of fit for the estimated regression equation; it can be interpreted as the proportion of the variation in the dependent variable *y* that can be explained by the estimated regression equation. We reviewed correlation as a descriptive measure of the strength of a linear relationship between two variables.

The assumptions about the regression model and its associated error term ε were discussed, and *t* and *F* tests, based on those assumptions, were presented as a means for determining whether the relationship between two variables is statistically significant. We showed how to use the estimated regression equation to develop confidence interval estimates of the mean value of *y* and prediction interval estimates of individual values of *y*.

The lecture concluded with the use of residual analysis to validate the model assumptions and to identify outliers and influential observations.

## *Glossary*

**Dependent variable** The variable that is being predicted or explained. It is denoted by *y*.

**Independent variable** The variable that is doing the predicting or explaining. It is denoted by *x*.

**Simple linear regression** Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line.

**Regression model** The equation describing how *y* is related to *x* and an error term; in simple linear regression, the regression model is $y = \beta_0 + \beta_1 x + \varepsilon$.

**Regression equation** The equation that describes how the mean or expected value of the dependent variable is related to the independent variable; in simple linear regression, $E(y) = \beta_0 + \beta_1 x$

**Estimated regression equation** The estimate of the regression equation developed from sample data by using the least squares method. For simple linear regression, the estimated regression equation is $y = b_0 + b_1 x$.

**Least squares method** A procedure used to develop the estimated regression equation. The objective is to minimize $\sum (y_i - \hat{y}_i)^2$

**Scatter diagram** A graph of bivariate data in which the independent variable is on the horizontal axis and the dependent variable is on the vertical axis.

**Coefficient of determination** A measure of the goodness of fit of the estimated regression equation. It can be interpreted as the proportion of the variability in the dependent variable *y* that is explained by the estimated regression equation.

***i*-th residual** The difference between the observed value of the dependent variable and the value predicted using the estimated regression equation; for the *i-th* observation the *i*-th residual is $y_i - \hat{y}_i$.

**Correlation coefficient** A measure of the strength of the linear relationship between two variables (previously discussed in Chapter 3).

**Mean square error** The unbiased estimate of the variance of the error term $\sigma^2$. It is denoted by MSE or $s^2$.

**Standard error of the estimate** The square root of the mean square error, denoted by *s*. It is the estimate of $\sigma$, the standard deviation of the error term ε.

**ANOVA table** The analysis of variance table used to summarize the computations associated with the *F* test for significance.

**Confidence interval** The interval estimate of the mean value of *y* for a given value of *x*.

**Prediction interval** The interval estimate of an individual value of *y* for a given value of *x*.

**Residual analysis** The analysis of the residuals used to determine whether the assumptions made about the regression model appear to be valid. Residual analysis is also used to identify outliers and influential observations.

**Residual plot** Graphical representation of the residuals that can be used to determine whether the assumptions made about the regression model appear to be valid.

**Standardized residual** The value obtained by dividing a residual by its standard deviation.

**Normal probability plot** A graph of the standardized residuals plotted against values of the normal scores. This plot helps determine whether the assumption that the error term has a normal probability distribution appears to be valid.

**Outlier** A data point or observation that does not fit the trend shown by the remaining data.

**Influential observation** An observation that has a strong influence or effect on the regression results.

**High leverage points** Observations with extreme values for the independent variables.

## *Key formulas*

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{11.1}$$

Simple Linear Regression Equation

$$E(y) = \beta_0 + \beta_1 x \tag{11.2}$$

Estimated Simple Linear Regression Equation

$$y = \beta_0 + \beta_1 x \tag{11.3}$$

Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2 \tag{11.4}$$

Slope and v-Intercept for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(x_1 - \bar{x})^2} \tag{11.5}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{11.6}$$

Sum of Squares Due to Error

$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{11.7}$$

Total Sum of Squares

$$SST = \sum (y_i - \bar{y})^2 \tag{11.8}$$

Sum of Squares Due to Regression

$$SSR = \sum (\hat{y}_i - \bar{y})^2 \tag{11.9}$$

Relationship Among SST, SSR, and SSE

$$SST = SSR + SSE \tag{11.10}$$

Coefficient of Determination

$$r^2 = \frac{SSR}{SST} \tag{11.11}$$

Sample Correlation Coefficient

$$r_{xy} = sign(b_1)\sqrt{Coefficient\ of\ \ determination} = sign(b_1)\sqrt{r^2} \tag{11.12}$$

Mean Square Error (Estimate of $\sigma^2$)

$$s^2 = MSE = \frac{SSE}{n-2} \tag{11.13}$$

Standard Error of the Estimate

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} \tag{11.14}$$

Standard Deviation of $b_1$

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{11.15}$$

Estimated Standard Deviation of $b_1$

$$s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{11.16}$$

$t$ Test Statistic ($df = n - 2$)

$$t = \frac{b_1}{s_{b_1}} \tag{11.17}$$

Mean Square Regression

$$MSR = \frac{SSR}{Number\ of\ \ independent\ variables} \tag{11.18}$$

$F$ Test Statistic

$$F = \frac{MSR}{MSE} \tag{11.19}$$

Estimated Standard Deviation of $\hat{y}_p$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{11.20}$$

Confidence Interval for $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \tag{11.21}$$

Estimated Standard Deviation of an Individual Value

$$s_{ind} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{11.22}$$

Prediction Interval for $y_p$

$$\hat{y}_p \pm t_{\alpha/2} s_{ind} \tag{11.23}$$

Residual for Observation $i$

$$y_i - \hat{y}_i \tag{11.24}$$

Standard Deviation of the $i$-th Residual

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{11.25}$$

Standardized Residual for Observation $i$

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{11.26}$$

Leverage of Observation $i$

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2} \tag{11.27}$$

## Lecture supplementary material

**FIGURE 14.2** THE ESTIMATION PROCESS IN SIMPLE LINEAR REGRESSION

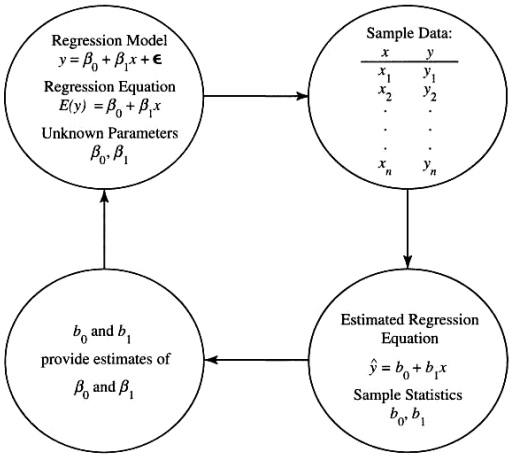
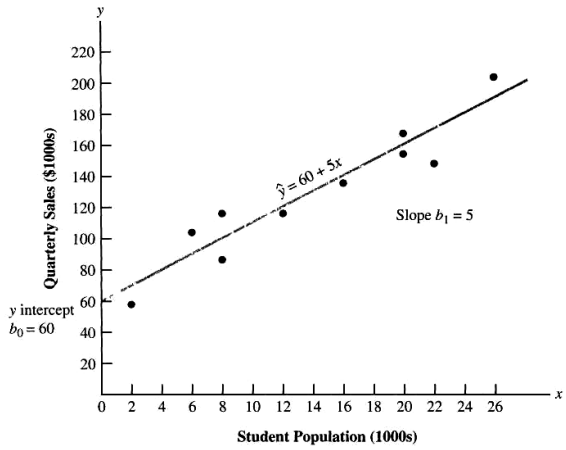



**FIGURE 14.4** GRAPH OF THE ESTIMATED REGRESSION EQUATION FOR ARMAND'S PIZZA PARLORS: $\hat{y} = 60 + 5x$




# 11.2. Multiple Regression

## Summary

In this lecture, we introduced multiple regression analysis as an extension of simple linear regression analysis presented in Lecture 14. Multiple regression analysis enables us to understand how a dependent variable is related to two or more independent variables. The regression equation $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ shows that the expected value or mean value of the dependent variable $y$ is related to the values of the independent variables $x_1, x_2, \ldots, x_p$. Sample data and the least squares method are used to develop the estimated regression equation $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$. In effect $b_0, b_1, b_2, \ldots, b_p$ are sample statistics used to estimate the unknown model parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Computer printouts were used throughout the chapter to emphasize the fact that statistical software packages are the only realistic means of performing the numerous computations required in multiple regression analysis.

The multiple coefficient of determination was presented as a measure of the goodness of fit of the estimated regression equation. It determines the proportion of the variation of $y$ that can be explained by the estimated regression equation. The adjusted multiple coefficient of determination is a similar measure of goodness of fit that adjusts for the number of independent variables and thus avoids overestimating the impact of adding more independent variables.

An $F$ test and a $t$ test were presented as ways to determine statistically whether the relationship among the variables is significant. The $F$ test is used to determine whether there is a significant overall relationship between the dependent variable and the set of all independent variables. The $t$ test is used to determine whether there is a significant relationship between the dependent variable and an individual independent variable given the other independent variables in the regression model. Correlation among the independent variables, known as multicollinearity, was discussed.

The section on qualitative independent variables showed how dummy variables can be used to incorporate qualitative data into multiple regression analysis. The section on residual analysis showed how residual analysis can be used to validate the model assumptions, detect outliers, and identify influential observations. Standardized residuals, leverage, studentized deleted residuals, and Cook's distance measure were discussed. The chapter concluded with a section on how logistic regression can be used to model situations in which the dependent variable may only assume two values.

## Glossary

**Multiple regression analysis** Regression analysis involving two or more independent variables.

**Multiple regression model** The mathematical equation that describes how the dependent variable $y$ is related to the independent variables $x_1, x_2, \ldots, x_p$ and an error term $\epsilon$.

**Multiple regression equation** The mathematical equation relating the expected value or mean value of the dependent variable to the values of the independent variables; that is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

**Estimated multiple regression equation** The estimate of the multiple regression equation based on sample data and the least squares method; it is $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$.

**Least squares method** The method used to develop the estimated regression equation. It minimizes the sum of squared residuals (the deviations between the observed values of the dependent variable, $y_i$ and the estimated values of the

dependent variable, $\hat{y}_i$).

**Multiple coefficient of determination** A measure of the goodness of fit of the estimated multiple regression equation. It can be interpreted as the proportion of the variability in the dependent variable that is explained by the estimated regression equation.

**Adjusted multiple coefficient of determination** A measure of the goodness of fit of the estimated multiple regression equation that adjusts for the number of independent variables in the model and thus avoids overestimating the impact of adding more independent variables.

**Multicollinearity** The term used to describe the correlation among the independent variables.

**Qualitative independent variable** An independent variable with qualitative data.

**Dummy variable** A variable used to model the effect of qualitative independent variables. A dummy variable may take only the value zero or one.

**Leverage** A measure of how far the values of the independent variables are from their mean values.

**Outlier** An observation that does not fit the pattern of the other data.

**Studentized deleted residuals** Standardized residuals that are based on a revised standard error of the estimate obtained by deleting observation *i* from the data set and then performing the regression analysis and computations.

**Influential observation** An observation that has a strong influence on the regression results.

**Cook's distance measure** A measure of the influence of an observation based on both the leverage of observation *i* and the residual for observation *i*.

**Logistic regression equation** The mathematical equation relating $E(y)$, the probability that $y = 1$, to the values of the independent variables; that is, $E(y) = P(y = 1 \mid x_1, x_2, ..., x_p) = \dfrac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}$

**Estimated logistic regression equation** The estimate of the logistic regression equation based on sample data; that is $\hat{y}$ =estimate of $P(y = 1 \mid x_1, x_2, ..., x_p) = \dfrac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}$

**Odds in favor of an event occurring** The probability the event will occur divided by the probability the event will not occur.

**Odds ratio** The odds that $y = 1$ given that one of the independent variables increased by one unit ($odd_1$) divided by the odds that $y = 1$ given no change in the values for the independent variables ($odds_0$); that is, Odds ratio = $odds_1/odds_2$.

**Logit** The natural logarithm of the odds in favor of $y = 1$; that is, $g(x_1,x_2,\ldots,x_p) = \beta_0 + \beta_l x_l + \beta_2 x_2 + \ldots + \beta_p x_p$

**Estimated logit** An estimate of the logit based on sample data; that is, $g\hat{}(x_1,x_2,\ldots,x_p) = \beta_0 + \beta_l x_l + \beta_2 x_2 + \ldots + \beta_p x_p$

## *Key formulas*

Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{11.28}$$

Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{11.29}$$

Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \tag{11.30}$$

Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2 \tag{11.31}$$

Relationship Among SST, SSR. and SSE

$$SST = SSR + SSE \tag{11.32}$$

Multiple Coefficient of Determination

$$R^2 = \frac{SSR}{SST} \tag{11.33}$$

Adjusted Multiple Coefficient of Determination

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{11.34}$$

Mean Square Regression

$$MSR = \frac{SSR}{p} \tag{11.35}$$

Mean Square Error

$$MSE = \frac{SSE}{n-p-1} \tag{11.36}$$

*F* Test Statistic

$$F = \frac{MSR}{MSE} \tag{11.19}$$

*t* Test Statistic

$$t = \frac{b_i}{s_{b_i}} \tag{11.17}$$

Standardized Residual for Observation *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{11.26}$$

Standard Deviation of Residual *i*

$$s_{y_i - \hat{y}_i} = s\sqrt{1-h_i} \tag{11.25}$$

Cook's Distance Measure

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p-1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right] \tag{11.37}$$

Logistic Regression Equation

$$E(y) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}} \tag{11.38}$$

Estimated Logistic Regression Equation

$$\hat{y} = estimate\ of\ \ P(y=1 \mid x_1, x_2, \ldots, x_p) = \frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1+e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}} \tag{11.39}$$

Odds Ratio

$$Odds\ raio = \frac{odds_1}{odds_0} \tag{11.40}$$

Logit

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p \tag{11.41}$$

Estimated Logit

$$\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p \tag{11.42}$$

# 11.3. Regression Analysis: Model Building

## *Summary*

In this chapter we discussed several concepts used by model builders in identifying the best estimated regression equation. First, we introduced the concept of a general linear model to show how the methods discussed in Chapters 14 and 15 could be extended to handle curvilinear relationships and interaction effects. Then we discussed how transformations involving the dependent variable could be used to account for problems such as nonconstant variance in the error term.

In many applications of regression analysis, a large number of independent variables are considered. We presented a general approach based on an *F* statistic for adding or deleting variables from a regression model. We then introduced a

larger problem involving 25 observations and eight independent variables. We saw that one issue encountered in solving larger problems is finding the best subset of the independent variables. To help in that task, we discussed several variable selection procedures: stepwise regression, forward selection, backward elimination, and best-subsets regression.

We extended the applications of residual analysis to show the Durbin-Watson test for autocorrelation. The chapter concluded with a discussion of how multiple regression models could be developed to provide another approach for solving analysis of variance and experimental design problems.

## Glossary

**General linear model** A model of the form $y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_p z_p + \varepsilon$, where each of the independent variables $z_j (j = 1,2,..., p)$ are functions of $x_1, x_2, \ldots, x_k$, the variables for which data have been collected.

**Interaction** The effect of two independent variables acting together.

**Variable selection procedures** Methods for selecting a subset of the independent variables for a regression model.

**Autocorrelation** Correlation in the errors that arises when the error terms at successive points in time are related.

**Serial correlation** Same as autocorrelation.

**Durbin-Watson test** A test to determine whether first-order autocorrelation is present.

## Key formulas

General Linear Model

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \varepsilon \tag{11.43}$$

*F* Test Statistic for Adding or Deleting $p - q$ Variables

$$F = \frac{\dfrac{SSE(x_1, x_2, \ldots, x_q) - SSE(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{p - q}}{\dfrac{SSE(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{n - p - 1}} \tag{11.44}$$

First-Order Autocorrelation

$$\varepsilon_i = \rho \varepsilon_{i-1} + z_i \tag{11.45}$$

Durbin-Watson Test Statistic

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^{\,2}} \tag{11.46}$$

## Lecture supplementary material

FIGURE 16.1 SCATTER DIAGRAM FOR THE REYNOLDS EXAMPLE