Chapter 12: Multiple Regression

Section 12.1: How Can We Use Several Variables to Predict a Response?

Learning Objectives

1. Regression Models
2. The Number of Explanatory Variables
3. Plotting Relationships
4. Interpretation of Multiple Regression Coefficients
5. Summarizing the Effect While Controlling for a Variable
6. Slopes in Multiple Regression and in Bivariate Regression
7. Importance of Multiple Regression

Learning Objective 1: Regression Models

- **The model that contains only two variables, *x* and *y*, is called a *bivariate* model**

$$\mu_y = \alpha + \beta x$$

Learning Objective 1: Regression Models

- Suppose there are two predictors, denoted by x_1 and x_2
- This is called a *multiple regression* model

$$\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2$$

Learning Objective 1: Multiple Regression Model

- The multiple regression model relates the mean μ_y of a quantitative response variable y to a set of explanatory variables x_1, x_2,…

5

Learning Objective 1: Multiple Regression Model

- **Example: For three explanatory variables, the multiple regression equation is:**

$$\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

6

Learning Objective 1: Multiple Regression Model

- **Example: The sample prediction equation with three explanatory variables is:**

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

7

Learning Objective 1: Example: Predicting Selling Price Using House and Lot Size

- **The data set "house selling prices" contains observations on 100 home sales in Florida in November 2003**
- **A multiple regression analysis was done with *selling price* as the response variable and with *house size* and *lot size* as the explanatory variables**

8

Learning Objective 1:
Example: Predicting Selling Price Using House and Lot Size

- **Output from the analysis:**

TABLE 13.3: Regression of Selling Price on House Size and Lot Size

The regression equation is price = −10536 + 53.8 house_size + 2.84 lot_size

Predictor	Coef	SE Coef	T	P
Constant	−10536	9436	−1.12	0.267
House_size	53.779	6.529	8.24	0.000
Lot_size	2.8404	0.427	6.66	0.000

Learning Objective 1:
Example: Predicting Selling Price Using House and Lot Size

- **Prediction Equation:**

$$\hat{y} = -10{,}536 + 53.8x_1 + 2.84x_2$$

where y = selling price, x_1=house size and x_2 = lot size

Learning Objective 1:
Example: Predicting Selling Price Using House and Lot Size

- **One house listed in the data set had house size = 1240 square feet, lot size = 18,000 square feet and selling price = $145,000**
- **Find its predicted selling price:**

$$\hat{y} = -10{,}536 + 53.8(1240) + 2.84(18{,}000)$$
$$= 107{,}276$$

Learning Objective 1:
Example: Predicting Selling Price Using House and Lot Size

- **Find its residual:**

$$y - \hat{y} = 145{,}000 - 107{,}276 = 37{,}724$$

- **The residual tells us that the actual selling price was $37,724 higher than predicted**

Learning Objective 2: The Number of Explanatory Variables

- **You should not use many explanatory variables in a multiple regression model unless you have lots of data**
- **A rough guideline is that *the sample size n should be at least 10 times the number of explanatory variables***

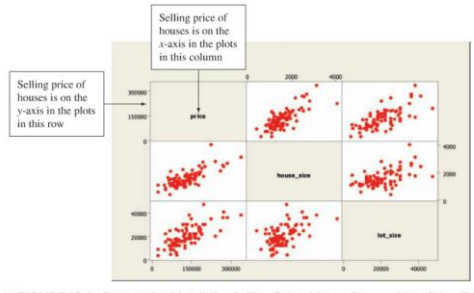
13

Learning Objective 3: Plotting Relationships

- **Always look at the data before doing a multiple regression**
- **Most software has the option of constructing scatterplots on a single graph for each pair of variables**
  - **This is called a scatterplot matrix**

14

Learning Objective 3: Plotting Relationships



▲ **FIGURE 13.1: Scatterplot Matrix for Selling Price, House Size, and Lot Size.** The middle plot in the top row has house size on the *x*-axis and selling price on the *y*-axis. The first plot in the second row reverses this, with selling price on the *x*-axis and house size on the *y*-axis. **Question:** Why are the plots of main interest the ones in the first row?

15

Learning Objective 4: Interpretation of Multiple Regression Coefficients

- **The simplest way to interpret a multiple regression equation looks at it in two dimensions as a function of a single explanatory variable**
- **We can look at it this way by fixing values for the other explanatory variable(s)**

16

Learning Objective 4: Interpretation of Multiple Regression Coefficients

Example using the housing data:

- **Suppose we fix x_1 = house size at 2000 square feet**
- **The prediction equation becomes:**

$$\hat{y} = -10,536 + 53.8(2000) + 2.84x_2$$
$$= 97,022 + 2.84x_2$$

17

Learning Objective 4: Interpretation of Multiple Regression Coefficients

- **Since the slope coefficient of x_2 is 2.84, the predicted selling price increases by $2.84 for every square foot increase in lot size when the house size is 2000 square feet**
- **For a 1000 square-foot increase in lot size, the predicted selling price increases by 1000(2.84) = $2840 when the house size is 2000 square feet**

18

Learning Objective 4: Interpretation of Multiple Regression Coefficients

Example using the housing data:

- **Suppose we fix x_2 = lot size at 30,000 square feet**
- **The prediction equation becomes:**

$$\hat{y} = -10,536 + 53.8x_1 + 2.84(30,000)$$
$$= 74,676 + 53.8x_1$$

19

Learning Objective 4: Interpretation of Multiple Regression Coefficients

- **Since the slope coefficient of x_1 is 53.8, for houses with a lot size of 30,000 square feet, the predicted selling price increases by $53.80 for every square foot increase in house size**

20

Learning Objective 4: Interpretation of Multiple Regression Coefficients

- **In summary, an increase of a square foot in house size has a larger impact on the selling price ($53.80) than an increase of a square foot in lot size ($2.84)**
- **We can compare slopes for these explanatory variables because their units of measurement are the same (square feet)**
- **Slopes cannot be compared when the units differ**

Learning Objective 5: Summarizing the Effect While Controlling for a Variable

- **The multiple regression model assumes that the slope for a particular explanatory variable is identical for all fixed values of the other explanatory variables**

Learning Objective 5: Summarizing the Effect While Controlling for a Variable

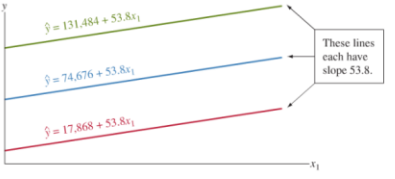
- **For example, the coefficient of x_1 in the prediction equation:**

$$\hat{y} = -10{,}536 + 53.8x_1 + 2.84x_2$$

is 53.8 regardless of whether we plug in x_2 = 10,000 or x_2 = 30,000 or x_2 = 50,000

Learning Objective 5: Summarizing the Effect While Controlling for a Variable

The slope effect of house size is 53.8 for each equation. Setting x_2 at a variety of values yields a collection of parallel lines, each having slope 53.8. See Figure 13.2.



▲ **FIGURE 13.2: The Relationship between $\hat{y}$ and x_1 for the Multiple Regression Equation $\hat{y} = -10{,}536 + 53.8x_1 + 2.84x_2$.** This shows how the equation simplifies when lot size x_2 = 10000, or x_2 = 30000, or x_2 = 50000. **Question:** The lines move upward (to higher $\hat{y}$-values) as x_2 increases. How would you interpret this fact?

Learning Objective 6:
Slopes in Multiple Regression and in Bivariate Regression

- **In multiple regression, a slope describes the effect of an explanatory variable while *controlling* effects of the other explanatory variables in the model**

Learning Objective 6:
Slopes in Multiple Regression and in Bivariate Regression

- **Bivariate regression has only a single explanatory variable**
- **A slope in bivariate regression describes the effect of that variable while ignoring all other possible explanatory variables**

Learning Objective 7:
Importance of Multiple Regression

- **One of the main uses of multiple regression is to identify potential lurking variables and control for them by including them as explanatory variables in the model**

Chapter 12:Multiple Regression

Section 12.2 Extending the Correlation and R-Squared for Multiple Regression

Learning Objectives

1. Multiple Correlation
2. R-squared
3. Properties of R^2

Learning Objective 1: Multiple Correlation

- **To summarize how well a multiple regression model predicts y, we analyze how well the observed y values correlate with the predicted $\hat{y}$ values**
- **The multiple correlation is the correlation between the observed y values and the predicted $\hat{y}$ values**
  - **It is denoted by R**

Learning Objective 1: Multiple Correlation

- **For each subject, the regression equation provides a predicted value**
- **Each subject has an observed y-value and a predicted y-value**

TABLE 13.4: Selling Prices and their Predicted Values

These values refer to the two home sales listed in Table 13.1. The predictors are x_1 = house size and x_2 = lot size.

Home	Selling Price	Predicted Selling Price
1	145,000	$\hat{y} = -10{,}536 + 53.8(1240) + 2.84(18{,}000) = 107{,}276$
2	69,000	$\hat{y} = -10{,}536 + 53.8(1120) + 2.84(17{,}000) = 97{,}983$

Learning Objective 1: Multiple Correlation

- **The correlation computed between all pairs of observed y-values and predicted y-values is the multiple correlation, R**
- **The larger the multiple correlation, the better are the predictions of y by the set of explanatory variables**

Learning Objective 1: Multiple Correlation

- **The R-value always falls between 0 and 1**
- **In this way, the multiple correlation 'R' differs from the bivariate correlation 'r' between y and a single variable x, which falls between -1 and +1**

Learning Objective 2: R-squared

- **For predicting y, the square of R describes the relative improvement from using the prediction equation instead of using the sample mean, $\bar{y}$**

Learning Objective 2: R-squared

- **The error in using the prediction equation to predict y is summarized by the *residual sum of squares:***

$$\Sigma(y-\hat{y})^2$$

Learning Objective 2: R-squared

- **The error in using $\bar{y}$ to predict y is summarized by the *total sum of squares*:**

$$\Sigma(y-\bar{y})^2$$

Learning Objective 2: R-squared

- **The proportional reduction in error is:**

$$R^2 = \frac{(y-\bar{y})^2 - (y-\hat{y})^2}{(y-\bar{y})^2}$$

Learning Objective 2: R-squared

- **The better the predictions are using the regression equation, the larger R^2 is**
- **For multiple regression, R^2 is the square of the multiple correlation, R**

Learning Objective 2: Example: How Well Can We Predict House Selling Prices?

- **For the 100 observations on y = selling price, x_1 = house size, and x_2 = lot size, a table, called the ANOVA (analysis of variance) table was created**
- **The table displays the sums of squares in the SS column**

TABLE 13.5: ANOVA Table and *R*-Squared for Predicting House Selling Price Using House Size and Lot Size

```
R-Sq = 71.1%
Analysis of Variance
Source           DF          SS
Regression        2      223676
Residual Error   97       90756
Total            99      314433
```


Learning Objective 2: Example: How Well Can We Predict House Selling Prices?

- **The R^2 value can be created from the sums of squares in the table**

$$R^2 = \frac{\sum(y-\bar{y})^2 - (y-\hat{y})^2}{(y-\bar{y})^2}$$

$$= \frac{314{,}433 - 90{,}756}{90{,}756} = 0.711$$

Learning Objective 2:
Example: How Well Can We Predict House Selling Prices?

- **Using house size and lot size together to predict selling price reduces the prediction error by 71%, relative to using $\bar{y}$ alone to predict selling price**

Learning Objective 2:
Example: How Well Can We Predict House Selling Prices?

- **Find and interpret the multiple correlation**

$$R = \sqrt{R^2} = \sqrt{0.711} = 0.84$$

- **There is a strong association between the observed and the predicted selling prices**
- **House size and lot size are very helpful in predicting selling prices**

Learning Objective 2:
Example: How Well Can We Predict House Selling Prices?

- **If we used a bivariate regression model to predict selling price with house size as the predictor, the r^2 value would be 0.58**
- **If we used a bivariate regression model to predict selling price with lot size as the predictor, the r^2 value would be 0.51**

Learning Objective 2:
Example: How Well Can We Predict House Selling Prices?

- **The multiple regression model has R^2 0.71, so it provides better predictions than either bivariate model**

Learning Objective 2:
Example: How Well Can We Predict House Selling Prices?

TABLE 12.6: R^2 Values for Multiple Regression Models for y = House Selling Price

NW is a geographic variable defined in Section 12.5.

Explanatory Variables in Model	R^2
Tax	0.679
Tax, House size	0.730
Tax, House size, Lot size	0.757
Tax, House size, Lot size, NW	0.769
Tax, House size, Lot size, NW, No. baths	0.773
Tax, House size, Lot size, NW, No. baths, No. bedrooms	0.775

Learning Objective 2:
Example: How Well Can We Predict House Selling Prices?

- **The single predictor in the data set that is most strongly associated with y is the house's real estate tax assessment**
  - **($r^2 = 0.679$)**
- **When we add house size as a second predictor, R^2 goes up from 0.679 to 0.730**
- **As other predictors are added, R^2 continues to go up, but not by much**

Learning Objective 2:
Example: How Well Can We Predict House Selling Prices?

- **R^2 does not increase much after a few predictors are in the model**
- **When there are many explanatory variables but the correlations among them are strong, once you have included a few of them in the model, R^2 usually doesn't increase much more when you add additional ones**

Learning Objective 2:
Example: How Well Can We Predict House Selling Prices?

- **This does not mean that the additional variables are uncorrelated with the response variable**
- **It merely means that they don't add much new power for predicting y, given the values of the predictors already in the model**

Learning Objective 3: Properties of R^2

- **The previous example showed that R^2 for the multiple regression model was larger than r^2 for a bivariate model using only one of the explanatory variables**
- **A key factor of R^2 is that it cannot decrease when predictors are added to a model**

49

Learning Objective 3: Properties of R^2

- **R^2 falls between 0 and 1**
- **The larger the value, the better the explanatory variables collectively predict y**
- **R^2 =1 only when all residuals are 0, that is, when all regression predictions are prefect**
- **R^2 = 0 when the correlation between y and each explanatory variable equals 0**

50

Learning Objective 3: Properties of R^2

- **R^2 gets larger, or at worst stays the same, whenever an explanatory variable is added to the multiple regression model**
- **The value of R^2 does not depend on the units of measurement**

51