Statistics 1
Chapter 14
Multiple Regression and Correlation Analysis

Ankara University, Faculty of Political Science, Department of Economics, Onur Özsoy

Chapter 14

Multiple Regression and Correlation Analysis

GOALS

When you have completed this chapter, you will be able to:

ONE

Describe the relationship between two or more independent variables and the dependent variable using a multiple regression equation.

TWO

Compute and interpret the multiple standard error of estimate and the coefficient of determination.

THREE

Interpret a correlation matrix.

FOUR

Setup and interpret an ANOVA table.

Chapter 14 *continued*

Multiple Regression and Correlation Analysis

GOALS

When you have completed this chapter, you will be able to:

FIVE

Conduct a test of hypothesis to determine if any of the set of regression coefficients differ from zero.

SIX

Conduct a test of hypothesis on each of the regression coefficients.

Multiple Regression Analysis

- For two independent variables, the general form of the multiple regression equation is:

$$Y' = a + b_1 X_1 + b_2 X_2$$

- X_1 and X_2 are the independent variables.
- a is the Y-intercept.
- b_1 is the net change in Y for each unit change in X_1
- holding X_2 constant. It is called a partial regression coefficient, a net regression coefficient, or just a regression coefficient.

Multiple Regression Analysis

- The general multiple regression with *k* independent variables is given by:

$$Y' = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

- The least squares criterion is used to develop this equation.
- Since estimating b_1, b_2, *etc*. is very tedious, there are many computer software packages that can be used to estimate these parameters.

Multiple Standard Error of Estimate

- The multiple standard error of estimate is a measure of the effectiveness of the regression equation.
- It is measured in the same units as the dependent variable.
- It is difficult to determine what is a large value and what is a small value of the standard error.

Multiple Standard Error of Estimate

- The formula is:

$$S_{Y\cdot 12\ldots k} = \sqrt{\frac{\Sigma(Y - Y')^2}{n - (k+1)}}$$

where n is the number of observations and

k is the number of independent variables.

Multiple Regression and Correlation (Assumptions)

- The independent variables and the dependent variable have a linear relationship.
- The dependent variable must be continuous and at least interval-scale.
- The variation in (Y-Y') or residual must be the same for all values of Y. When this is the case, we say the difference exhibits homoscedasticity.
- The residuals should be normally distributed with mean 0.
- Successive values of the dependent variable must be uncorrelated.

The ANOVA Table

- The ANOVA table gives the variation in the dependent variable (of both that which is and is not explained by the regression equation).

Correlation Matrix

- A correlation matrix is used to show all possible simple correlation coefficients between all variables.
  - The matrix is useful for locating correlated independent variables.
  - How strongly each independent variable is correlated to the dependent variable is shown in the matrix.

Global Test

- The global test is used to investigate whether any of the independent variables have significant coefficients. The hypotheses are:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

$$H_a: \textit{At least one of the regression coefficients is not zero}$$

Global Test *continued*

- The test statistic is the F distribution with *k*(number of independent variables) and *n-(k+1)* degrees of freedom, where *n* is the sample size.

Test for Individual Variables

- This test is used to determine which independent variables have nonzero regression coefficients.
- The variables that have zero regression coefficients are usually dropped from the analysis.
- The test statistic is the t distribution with n-(k+1)degrees of freedom.

EXAMPLE 1

- A market researcher for Super Dollar super markets is studying the yearly amount families of four or more spend on food. Three independent variables are thought to be related to food expenditures. Those variables are: total family income, size of family, and whether the family has children in college.

EXAMPLE 1 *continued*

Family	Food Expenditure	Income ($1000)	Family Size	College Student
1	3900	37.6	4	0
2	5300	51.5	5	1
3	4300	51.6	4	0
4	4900	46.8	5	0
5	6400	53.8	6	1
6	7300	62.6	7	1
7	4900	54.3	5	0
8	5300	43.7	4	0
9	6100	60.8	5	1
10	6400	51.3	6	1
11	7400	49.3	6	1
12	5800	56.3	5	0

EXAMPLE 1 *continued*

- Use a computer software package, such as MINITAB or Excel, to develop a correlation matrix.
- From the analysis provided by MINITAB, write out the regression equation:

$$Y' = 954 + 10.9\,X_1 + 748\,X_2 + 565\,X_3$$

- What food expenditure would you estimate for a family of 4, with no college students, and an income of $50,000?

EXAMPLE 1 *continued*

- Y'=954 + 10.9(50) + 748(4) + 565 (0) = 4491.
- Conduct a global test of hypothesis to determine if any of the regression coefficients are not zero.

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \quad H_1: \text{at least one } \beta \neq 0$$

- H_0 is rejected if F>4.07
- From the MINITAB output, the computed test statistic value is 10.94
- Decision: Since F=10.94>4.07, H_0 is rejected. Thus not all the regression coefficients are zero

EXAMPLE 1 *continued*

- Conduct an individual test to determine which coefficients are not zero.
- From the MINITAB output, the only significant variable is FSIZE (family size) using the p-values. The other variables can be omitted from the model.
- Thus, $H_0: \beta_2 = 0 \qquad H_1: \beta_2 \neq 0$
- Using the 5% level of significance, reject H_0 if the p-value<.05

EXAMPLE 1 *continued*

- Since p-value =.039<.05, reject H_0 and conclude that $\beta_2 \neq 0$. That is, the size of the family and the amount spent on food are significantly related.

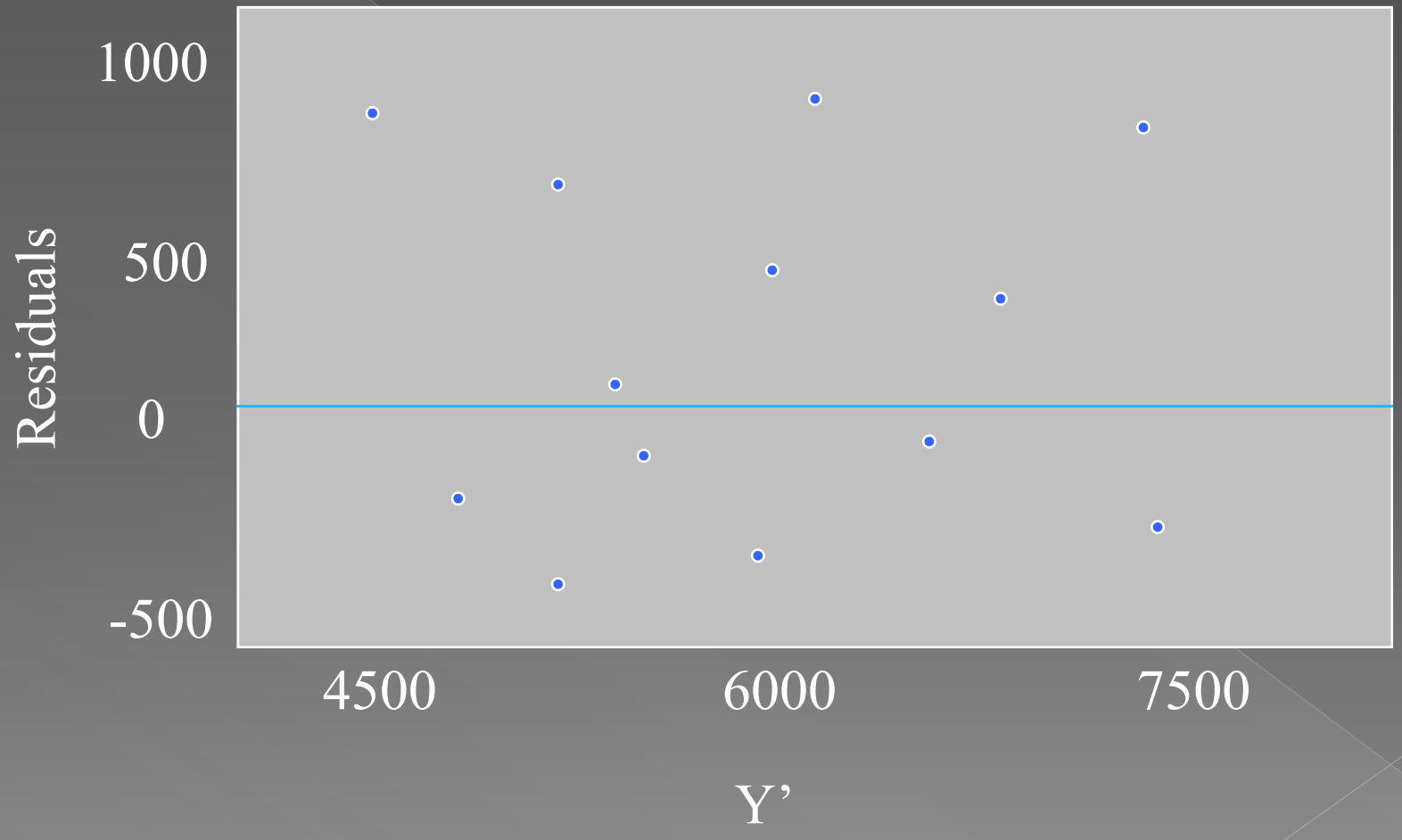
Qualitative Variables & Stepwise Regression

- Qualitative variables are nonnumeric and are also called *dummy variables*.
  - For a qualitative variable, there are only two conditions possible.
- Stepwise Regression leads to the most efficient regression equation.
  - Only independent variables with significant regression coefficients are entered into the analysis. Variables are entered in the order in which they increase R^2 the fastest.

Analysis of Residuals

- A residual is the difference between the actual value of Y and the predicted value Y'.
- Residuals should be approximately normally distributed. Histograms and stem-and-leaf charts are useful in checking this requirement.
- A plot of the residuals and their corresponding Y' values is used for showing that there are no trends or patterns in the residuals.

Residual Plot



Histograms of Residuals