Lecture Notes

# Chapter 12: Regression and Correlation

## Learning Objectives

1. Describe linear relationships and prediction rules for bivariate and multiple regression models.
2. Construct and interpret straight-line graphs and best-fitting lines.
3. Calculate and interpret *a* and *b*.
4. Calculate and interpret the coefficient of determination (*r*2) and Pearson’s correlation coefficient (*r*).
5. Interpret multiple regression output.
6. Test the significance of *r*2 and *R*2 using ANOVA.

## Chapter Outline

1. Introduction
   1. **Correlation:** A measure of association used to determine the existence and strength of the relationship between variables and is similar to the proportional reduction of error (PRE).
   2. **Regression:** A linear prediction model, using one or more independent variables to predict the values of a dependent variable.
   3. **Bivariate regression:** Examines how changes in one independent variable affects the value of a dependent variable.
   4. **Multiple regression:** Estimates how several independent variables affect one dependent variable.
2. The Scatter Diagram
   1. **Scatter diagram** or **scatterplot:** Visual method used to display the relationship between two interval-ratio variables. It is used as a first exploratory step in regression analysis, a scatter diagram can suggest whether two variables are associated.
   2. Scatter diagrams may also reveal a negative association between two variables or no relationship at all.
3. Linear Relationships and Prediction Rules
   1. **Linear relationship:** Allows to approximate the observations displayed in a scatter diagram with a straight line.
   2. **Deterministic relationship:** In a perfectly linear relationship, all the observations (the dots) fall along a straight line (a perfect relationship is sometimes called a deterministic relationship, and the line itself provides a predicted value of *Y* (the vertical axis) for any value of *X* (the horizontal axis).
   3. Finding the Best-Fitting Line
      1. Most relationships we study in the social sciences are not deterministic, and we are not able to come up with a linear equation that allows us to predict *Y* from *X* with perfect accuracy.
      2. Formula for calculating relationship: 

where,

*Ŷ* = the predicted score on the dependent variable,

*X* = the score on the independent variable,

*a* = the ***Y*-intercept** or the point where the line crosses the *Y*-axis; therefore,

*a* is the value of *Y* when *X* is 0,

*b* = the **slope** of the regression line, or the change in *Y* with a unit change in *X.*

* + 1. Defining Error
       1. The best-fitting line is the one that generates the least amount of error, also referred to as the residual.
       2. There are two values for *Y*: (1) a predicted *Y*, which we symbolize as *Ŷ* and which is generated by the prediction equation, also called the linear regression equation , and (2) the observed *Y*, symbolized simply as *Y*.
       3. Residual is the difference between the observed *Y* and the predicted *Ŷ*. If we symbolize the residual as *e*, then: *e* = *Y* – *Ŷ*.
    2. The Residual Sum of Squares (Σ*e*2):

1. Statisticians prefer to work with the third method: Squaring and summing the residuals over all observations.
2. The result is the residual sum of squares or Σ*e*2. Symbolically, Σ*e*2 is expressed as: 
   * 1. The Least Squares Line
3. The best-fitting regression line is that line where the sum of the squared residuals, or Σ*e2*, is at a minimum.
4. Such a line is called the **least squares line,** and the technique that produces this line is called the **least squares method**.
5. The technique involves choosing *a* and *b* for the equation such that Σ*e*2 will have the smallest possible value.
   1. Computing *a* and *b*
      1. Formula: To figure out the values of *a* and *b* in a way that minimizes Σ*e*2, we need to apply the following formulas:
         1. 
         2. 

where,

 = the covariance of *X* and *Y,*

 = the variance of *X,*

 = the mean of *Y,*

 = the mean of *X,*

*a* = the *Y*-intercept,

*b* = the slope of the line.

* + 1. Variance of variable *X*: The denominator for *b* is the variance of the variable *X*. It is defined as follows: 
    2. Covariance of *X* and *Y*: The numerator, however, is a new term.

1. It is the covariance of *X* and *Y* and is defined as:
   * + - 1. 
     1. The covariance is a measure of the linear relationship between two variables, and its value reflects both the strength and the direction of the relationship.
2. The covariance will be close to zero when *X* and *Y* are unrelated.
3. It will be larger than zero when the relationship is positive and smaller than zero when the relationship is negative.
   1. Interpreting *a* and *b*
      1. The pattern suggested by the regression equation may not hold for every individual, it gives us a tool by which to make the best possible guess about how Internet usage is associated, on average, with educational attainment.
      2. The *Y-*intercept *a* is the predicted value of *Y*, when X = 0.
         1. Thus, it is the point at which the regression line and the *Y*-axis intersect.
      3. Plotting the regression equation: One can plot the regression equation with two points:
4. The mean of *X* and the mean of *Y.*
5. 0 and the value of *a*.
6. Methods for Assessing the Accuracy of Predictions
   1. Formula to calculate PRE: 

where,

= prediction errors made when the independent variable is ignored.

= prediction errors made when the prediction is based on the independent variable.

* 1. Error of prediction:
     1. The first prediction rule is to predict  in the absence of information on *X*. The error of prediction is defined as .
     2. The second rule of prediction uses *X* and the regression equation to predict . The error of prediction is defined as 
  2. Total sum of squares or SST: The sum of the squared deviations from the mean is called the *total sum of squares*, or *SST*.
     1. 
  3. **Residual sum of squares** or SSE: Sum of squared deviations from the regression line is denoted as residual sum of squares or SSE.
     1. 
  4. **Regression sum of squares** or SSR: By subtracting *SSE* from *SST*, we obtain the regression sum of squares or *SSR*, which reflects improvement in the prediction error resulting from our use of the linear prediction equation.
     1. *SSR* is defined as: 
  5. Calculating Prediction Errors
     1. Steps for calculating PRE:
        1. *SST* measures the prediction errors when the independent variable is ignored, we can define: *E*1 = SST.
        2. Similarly, because *SSE* measures the prediction errors resulting from using the independent variable, we can define: *E*2 = SSE.
        3. We are now ready to define the coefficient of determination, *r*2. It measures the *PRE* associated with using the linear regression equation as a rule for predicting .
     2. **Coefficient of determination (*r*2** Reflects the proportion of the total variation in the dependent variable, *Y*, explained by the independent variable, *X*.

1. The coefficient of determination ranges from 0.0 to 1.0.
   * 1. Calculating
2. Equation: 
3. This formula tells us to divide the square of the covariance of *X* and *Y* by the product of the variance of *X* and the variance of *Y*.
4. Testing the Significance of Using ANOVA
   1. ANOVA (analysis of variance) can easily be applied to determine the statistical significance of the regression model as expressed in .
   2. ANOVA and regression analysis can look very much the same.
      1. Both methods accounts for variation in the dependent variable in terms of the independent variable, except that in ANOVA the independent variable is a categorical variable and with regression, it is an interval-ratio variable.
   3. With ANOVA, the total variation in the dependent variable can be decomposed into portions explained (SSB) and unexplained (SSW) by the independent variable.
   4. The statistical test, *F*, is the ratio of the mean squares between to the mean squares within as shown in formula: .
   5. **Mean squares regression:** Ratio of regression sum of squares (SSR) to the degree of freedom (*df*r).
      1. 
   6. **Mean squares residual:** Ratio of total sum of squares (SST) to the degree of freedom (*df*e).
   7. The *F* statistic is the ratio of the mean squares regression to the mean squares residual.

..

* 1. Pearson’s Correlation Coefficient (*r*)
     1. **Pearson’s correlation coefficient:** The square root of , or *r* is most often used as a measure of association between two interval-ratio variables.
        1. 
     2. Pearson’s *r* is usually computed directly by using the following definitional formula: 
     3. *r* is defined as the ratio of the covariance of *X* and *Y* to the product of the standard deviations of *X* and *Y.*
     4. Characteristics of Pearson’s *r*

1. Pearson’s *r* is a measure of relationship or association for interval-ratio variables.
2. Like gamma, it ranges from 0.0 to ±1.0, with 0.0 indicating no association between the two variables.
3. An *r* of +1.0 means that the two variables have a perfect positive association; –1.0 indicates that it is a perfect negative association.
4. The absolute value of *r* indicates the strength of the linear association between two variables.
5. Unlike the *b* coefficient, *r* is a symmetrical measure. That is, the correlation between *X* and *Y* is identical to the correlation between *Y* and *X*.
6. Statistics in Practice: Multiple Regression and ANOVA
   1. **Multiple regression** is an extension of bivariate regression, allowing us to examine the effect of two or more independent variables on the dependent variable.
   2. The general form of the multiple regression equation involving two independent variables: .

where,

= the predicted score on the dependent variable,

= the score on independent variable ,

= the score on independent variable ,

*a* = the *Y*-intercept, or the value of *Y* when both and are equal to zero,

= the partial slope of *Y* and , the change in *Y* with a unit change in ,when the other independent variable is controlled,

= the partial slope of *Y* and , the change in *Y* with a unit change in ,when the other independent variable is controlled.

* 1. **Partial slopes:** Reflect the amount of change in *Y* for a unit change in a specific independent variable while controlling or holding constant the value of the other independent variables.
  2. **Standardized slope coefficient** or **beta**: Converts the values of each score into a *Z* score, standardizing the units of measurement so we can interpret their relative effects. Beta, also referred to as beta weights, range from 0 to ±1.0.
  3. **Multiple coefficient of determination:** Measure that reflects the proportion of the total variation in the dependent variable that is explained jointly by two or more independent variables), symbolized as (corresponding to in the bivariate case).
  4. **Pearson’s multiple correlation coefficient:** It measures the linear relationship between the dependent variable and the combined effect of two or more independent variables.
  5. The ANOVA summary table for multiple regression is nearly identical to the one for bivariate linear regression except that the degrees of freedom are adjusted to reflect the number of independent variables in the model.
  6. SPSS can produce correlation matrix, a table that presents the Pearson’s correlation efficient for all pairs of variables in the multiple regression model.
  7. Correlation matrix provides a baseline summary of the relationships between variables, identifying relationships or hypotheses that are usually the main research objective.
  8. Three pairs of extensive correlation:
     1. Internet hours with educational attainment,
     2. Internet hours with age, and
     3. Educational attainment with age.