Chapter 13: Regression

I. Overview

a. Regression is a very powerful statistical technique that allows researchers to examine how two or more continuous (i.e., intervally scaled) variables are associated with each other.

b. Regression is based on the strength and direction of the correlations among these variables.

c. In a regression analysis, there is one dependent variable and one or more independent variables, often called predictor variables.

d. Regression can be used to predict a value for the dependent variable for any value of the independent (i.e., predictor) variable(s). It can also be used to tell you how of the variance in the dependent variable is explained by the predictor variable(s).
    i. Regression, particularly multiple regression, is being used more and more in "analytics" research that affects many aspects of our daily lives, from how much we pay for automobile insurance to which adds appear in our social media.

e. There are many kinds of regression techniques. In this chapter, we focus on simple linear regression with two variables and a brief introduction to multiple linear regression.

II. Simple Linear Regression

a. Purposes:
    i. To examine the strength of the association between two intervally scaled variables.
    ii. To determine how much of the variance in the dependent variable is explained by the independent, predictor variable.
    iii. To predict values of the dependent variable based on values of the independent variable.
        1. E.g., If we know a person's height, what would we predict his weight to be?

b. How it works

    i. Regression uses the correlation coefficient between two variables, the means of those variables, and the standard deviations of those variables to produce an intercept and a slope for a regression line.
        1. The intercept is the point where the regression line intersects the Y axis. Here's the formula for calculating the intercept:

$$a = \overline{Y} - b\,\overline{X}$$

            a. The Y axis contains the scale for the dependent variable
            b. The intercept represents the predicted value of Y (the dependent variable) when the value of X (the independent variable) is 0.

2. The slope of the regression line is determined by calculating how much the value of the dependent variable (Y) is expected to change for each unit of change in the independent variable (X). Here's the formula for calculating the regression coefficient:

$$b = r \times \frac{s_y}{s_x}$$

   a. The value of Y would be expected to increase with each increase in X if the two variables are positively correlated and to decrease if the two variables are negatively correlated.
3. The intercept has the symbol *a* and the slope, also known as the *regression coefficient*, has the symbol *b*.

ii. Using the slope and the intercept, you can find a predicted value for Y, the dependent variable, for any value of X, the independent variable. Here is the formula for calculating the predicted value of Y for a given value of X:

$$\hat{Y} = bX + a$$

iii. The percentage of variance in Y that is explained by X depends on the strength of the correlation coefficient.
   1. For a simple linear regression with two variables, this value (R^2) is simply the correlation coefficient squared (i.e., the coefficient of determination).

iv. The line of Ordinary Least Squares (OLS)
   1. Although regression allows you to predict values of Y for any given value of X, these predicted values of Y will not always be accurate.
      a. The accuracy of the predicted values depends on how strongly the two variables are correlated and how large the standard deviations of the two variables are.
   2. The difference between a predicted value of Y for a given value of X, and the observed value of Y for a given value of X, is considered to be error, and in regression this error is called a *residual*.
   3. If you were to take all of the residuals in an analysis, square them, and then add up the squared residuals, the regression line is the straight line that produces the smallest sum of the squared residuals.
      a. That is why this kind of regression analysis is called Ordinary Least Squares (OLS) regression.

III. Multiple Regression

a. Another powerful regression technique is called multiple regression.

b. In a multiple regression, one dependent variable is predicted by multiple independent (i.e., predictor) variables. Here is what the regression equation looks like with two independent variables:

$$\hat{Y} = a + bX_1 + bX_2$$

c. Using multiple predictor variables allows researchers to examine several interesting properties of the associations among the variables, including

i. How much of the overall variance in the dependent variable is explained by the *combined* predictor variables (i.e., multiple R^2).
ii. How strongly is each predictor variable related to the dependent variable *when controlling for the other predictor variables***.**
    1. In other words, what is the *unique* proportion of variance in the dependent variable that is explained by each of the independent variables?
    2. This is very powerful because it allows researchers to separate the effects of predictor variables that are correlated with each other.
iii. Where there is an interaction between predictor variables on the dependent variable.
iv. The predicted value of the dependent variable given any combination of values on the predictor variables.

IV. Summary

a. Regression allows researchers to test the associations among several intervally scaled variables at once.

b. Simple linear regression is like a fancy correlation analysis that allows researchers to find predicted values of the dependent variable.

c. Multiple regression is a very powerful technique and is often used to analyze complex associations among multiple variables.
    i. Especially powerful because it allows researchers to isolate the *unique* predictive power of each independent variable while controlling for all of the other independent variables in the model.

d. Simple linear regression is just one of many kinds of regression techniques.