Chapter 7

7.1

The most simple linear regression is a straight line, Y= bX + a. in this formula is X the predictor variable. This one Is used to predict the criterion variable Y. The slope of the line Is denoted by b, and this indicated the number of Y units the line changes for a one-unit change in X. The Y-intercept is denoted by a and is the point at which the line intersects or crosses the Y-axis. We just use the term intercept. The slope can be calculated as follows:

$$b = \frac{\Delta Y}{\Delta X} = \frac{Y_2 - Y_1}{X_2 - X_1} \qquad (51)$$

If you have two points of the line, you can calculate the slope with the previous formula. You know the b now, the next point you can fill in in the formula Y = bX + a, and you can find a.

We will now combine this theory with correlation. When the slope of a line is positive, as X increases, Y also increases, than the correlation will be positive. When the slope of the line is 0, then when X increases, Y will remain constant. Then the correlation will be 0. When the slope is negative, when X increases Y decreases. Then the correlation is negative. So this shows that the sign of the slope corresponds to the sign of the correlation.

7.2

We will now apply these concepts to the simple linear regression. We define the linear regression model as the equation for a straight line. So the population regression model for Y (dependent variable) being predicted by X (independent variable) is:

$$Y_i = \beta_{yx} X_i + \alpha_{yx} + \varepsilon_i \qquad (52)$$

In this formula:
Y is the criterion variable
X is the predictor variable
β_{yx} is the population slope for Y predicted by X
α_{yx} is the population intercept for Y predicted by X
ε_i are the population residuals or errors of prediction (the part of Y_i not predicted form X_i)
i represents an index for a particular case

The index i can take on values from 1 to N, where N is the size of the population, so i = 1,..,N.

The population prediction model is:

$$Y'_i = \beta_{yx} X_i + \alpha_{yx} \quad (53)$$

Where Y'_i is the predicted value of Y for a specific value of X. That is, Y_i is the actual or observed score, while Y'_i is the predicted score. Thus, the population predictor error is:

$$\varepsilon_i = Y_i - Y'_i \quad (54)$$

The difference between the regression and prediction models is that the regression model explicitly includes prediction error as ε_i, whereas the prediction model includes prediction error implicitly as part of the predicted score, Y'_i.

A simple method for determining the population slope (β_{yx}) and intercept (α_{yx}) is computed as:

$$\beta_{yx} = \rho_{xy} \frac{\sigma_Y}{\sigma_x}$$

$$\alpha_{yx} = \mu_y - \beta_{yx}\mu_x \quad (54)$$

Where:
σ_Y and σ_x are the population standard deviations for Y and X respectively
ρ_{xy} is the population correlation between X and Y (simply the Pearson correlation coefficient, rho)
μ_y and μ_x are the population means for Y and X respectively

However this method is not appropriate for determining the slope and intercept of a straight line in a regression analysis with real data!

7.3

We will now return to the real world of sample statistics and we consider the sample simple linear regression model. We use as always, Greek letters for the population parameters, and English letters for the sample statistics. The sample regression model is as follows:

$$Y_i = b_{yx} X_i + a_{yx} + e_i \quad (55)$$

Where
Y and X are as before
b_{yx} is the sample slope for Y predicted by X
a_{yx} is the sample intercept for Y predicted by X
e_i are sample residuals or errors of prediction
i represents an index for a case (individual or object). Values can range from $i = 1 \ldots n$.

The sample prediction model is computed as follows:

$$Y'_i = b_{yx} X_i + a_{yx} \quad (56)$$

Again the residual (or error) is computed as follows:

$$e_i = Y_i - Y'_i \quad (57)$$

The same as in the population prediction model, and population regression model, the only difference is that we now deal with samples instead of populations.

The sample slope,beta , is also referred to as (a) the expected or predicted change in Y for a one-unit change in X and (b) the unstandardized or raw regression coefficient.

The sample intercept, alpha, is also referred to as (a) the point at which the regression line intersects (or crosses) the Y-axis and (b) the value of Y when X is 0.

So the sample slope (b_{yx}) and intercept (a_{yx}) can be determined by:

$$b_{yx} = r_{xy} \frac{s_y}{s_x}$$

$$a_{yx} = \bar{Y} - b_{yx}\bar{X} \qquad (58)$$

where
s_y and s_x are the sample standard deviations for Y and X respectively
r_{yx} is the sample correlation between X and Y
$\bar{Y}$ and $\bar{X}$ are the sample means for Y and X respectively

Until now we have looked at computations in the simple linear regression that involved the use of raw scores. So we call this the unstandardized regression model. The slope estimate is an unstandardized or raw regression slope because it is the predicted change in Y raw score units for a one raw score unit change in X. We can also express regression in standard z-score units for both X and Y as:

$$z(X_i) = \frac{X_i - \bar{X}}{s_x}$$

$$z(Y_i) = \frac{Y_i - \bar{Y}}{s_Y} \qquad (59)$$

The sample standardized linear prediction model becomes the following, where $z(Y'_i)$ is the standardized prediction value of Y:

$$z(Y'_i) = b^*_{yx}\, z(X_i) = r_{xy}\, z(X_i) \qquad (60)$$

The *standardized regression slope*, b^*_{yx}, sometimes referred to as a *beta weight*, is equal to r_{xy}. And the *standardized intercept* is equal to 0.

A perfect prediction of Y from X is extremely unlikely, only whit a perfect correlation between X and Y is occurs (so correlation coefficient = 1.0). The residuals ei, are also known as errors of estimate, or prediction errors, and are that portion of Yi that is not predictable from Xi. The residual terms are random values that are unique to each individual or object.

In figure 7.2 on page 330, is a scatterplot of a regression example shown. You see the straight diagonal line. Individuals that fall above the regression line have positive residuals, in other words, the difference between the observed score is greater in value than the predicted value, which is represented by the regression line). Individuals that fall below the regression line have negative residuals, in other words, the difference between the observed score is less in value than the predicted value, which is represented by the regression line.

There are statistical criteria that help us decide which method to use in determining the slope and intercept. The criterion usually used in linear regression analysis is the least squares criterion. According to this criterion, the sum of the squared prediction errors or residuals is smallest. So we want to find a regression line, with a particular slope and intercept that results in the smallest sum of the squares residuals. We often refer to this method as least squares estimation, because b and a represent sample estimates of the population parameters obtained using the least squares criterion.

We can determine the utility of a predictor variable by the partitioning the total sum of squares in Y, which is denotes as SStotal. This process is much like partitioning the sum of squares in ANOVA.

In simple linear regression, we can partition SStotal into:

SStotal = SSreg + SSres

$$\sum_{i=1}^{n}(Y-\bar{Y})^2 = \sum_{i=1}^{n}(Y'-\bar{Y})^2 + \sum_{i=1}^{n}(Y-Y')^2 \qquad (61)$$

Where:

- SStotal is the total sum of squares in Y
- SSreg is the sum of squares of the regression of Y predicted by X
- SSres is the sum of squares of the residuals

In easy words, SStotal represents the total variation in the observed Y scores, SSreg the variation in Y predicted by X, and SSres the variation in Y not predicted by X.

SSreg examines how much better the line of best fit is as compared to the mean of Y. SSres provides an indication of how "off" or inaccurate the model is. When it is close to 0, the better the model fit.

$r_{xy}^2 = SS_{reg}/SS_{total}$, we can write SS_{total}, SS_{reg}, and SS_{res} as follows:

$$SS_{total} = n\sum_{i=1}^{n}Y^2 - (\sum_{i=1}^{n}Y^2)/n$$
$$SS_{reg} = r_{xy}^2\ SS_{total}$$
$$SS_{res} = (1 - r_{xy}^2)\ SS_{total} \qquad (62)$$

Where , is the squared sample correlation between X and Y, commonly referred to as the coefficient of determination. It also tells us the proportion of the total variation of the dependent variable, that has been explained by the regression model.

The coefficient of determination can be used both as a measure of effect size and as a test of significance. According to the subjective standards of Cohen, a small effect size is defined as r = 0.10, or r2 = 0.01, a medium effect size as r = 0.30, or r2 = 0.09, and a large effect size as r = 0.50 or r2 =0.25.

We will now discuss four procedures used in the simple linear regression context. The first two are tests of statistical significance that generally involve testing whether or not X is a significant predictor of Y. Then we consider two confidence interval (CI) techniques.

Test of significance of .

It is important that be different from 0 in order to have reasonable prediction. The null and alternative hypotheses are as follows:

$H_0: \rho_{xy}^2 = 0$
$H_1: \rho_{xy}^2 \neq 0$

The test is based on the following test statistic:

$$F = \frac{r^2/m}{(1-r^2)/(n-m-1)} \qquad (63)$$

Where:

- F indicates that this is an F statistic
- r2 is the coefficient of determination
- r2 is the proportion of variation in Y that is not predicted by X
- m is the number of predictors (in case of simple linear regression it is always 1)
- n is sample size

This F-statistic is compared to the F critical value, always a one-tailed test (given that a squared value cannot be negative) and at the designated level of significance, alpha, with degrees of freedom equal to m and (n-m-1), as taken from the F table in Table A.4. That is, the tabled critical value is Fm, (n – m – 1).

Test of significance of

This is the test of the slope or regression coefficient. In other words, is the unstandardized regression coefficient statistically significantly different from 0? This is actually the same as the test of b*, the standardized regression coefficient. The null and alternative hypotheses are:

$H_0: \beta_{yx} = 0$
$H_1: \beta_{yx} \neq 0$

To test whether the regression coefficient is equal to 0, we need a standard error for the slope b. First we need to develop some new concepts. The first new concept is the variance error of estimate (variance of the residuals), defined as:

$$s_{res}^2 = \sum e_i^2 / df_{res} = SS_{res}/df_{res} = MS_{res} \qquad (64)$$

Where dfres = (n – m- 1). The variance error of estimate indicates the amount of variation among the residuals. A relatively large variance of error, shows that there are some extremely large residuals, indicating a poor prediction. A relatively small variance of error, indicates a good prediction overall.

The next new concept is the standard error of estimate (root mean square error). This is simply the positive square root of the variance error of estimate, and thus is the standard deviation of the residuals or error of estimate. We denote the standard error of estimate as sres.

Final new concept is the standard error of b. We denote the standard error of b as sb and define it as:

$$s_b = s_{res} / \sqrt{[n \sum X^2 - (\sum X)^2]/n} = SS_{res} / \sqrt{SS_x} \qquad (65)$$

We want sb to be small to reject Ho, so we need sres to be small and SSx to be large. In other words, we want there to be a large spread of scores in X.

We can put these concepts together into a test statistics to test the significance of the slope b. It is as follows:

$$t = \frac{b}{s_b} \qquad (66)$$

We compare this to the critical value of table A.2. A two-tailed test for a non-directional H1, at the designated level of significance, , and with degrees of freedom of (n – m – 1).

We can also form a CI around the slope b. It follows the from of the sample estimate plus or minus the tabled critical value multiplied by the standard error:

$$CI\ (b) = b \pm (\alpha/2)\ t_{(n-m-1)sb} \qquad (67)$$

Confidence interval for the predicted mean value of Y

Third procedure is to develop a CI for the predicted mean value of Y, denoted by $\bar{Y}'_0$.
The standard error of $\bar{Y}'_0$ is:

$$s(\bar{Y}'_0) = s_{res} \sqrt{(1/n) + [(X_0 - \bar{X})^2 / SS_X} \qquad (68)$$

From this definition we can see that we expect to make our best predictions at the centre of the distribution of X scores and to make our poorest predictions for extreme values of X. A CI around is formed as follows:

values of X. A CI around $\bar{Y}'_0$ is formed as follows:

$$CI(\bar{Y}'_0) = \bar{Y}'_0 \pm (\alpha/2)\ t_{(n-2)} s(\bar{Y}'_0) \qquad (69)$$

Prediction interval for individual values of Y

Final procedure is to develop a prediction interval (PI) for an individual predicted value of . That is, the predictor score for a particular individual is known, but the criterion score for that individual has not yet been observed. This is in contrast to the CI just discusses, where the individual Y scores have already been observed. Thus, the CI deals with the mean of the predicted values, while the PI deals with an individual predicted value not yet observed.

The standard error of is:

$$s(Y'_0) = s_{res}\sqrt{1 + \left(\frac{1}{n}\right) + [(X_0 - \bar{X})^2/SS_X]} \qquad (70)$$

The PI around Y'_0 is formed as follows:

$$PI\ (Y'_0) = Y'_0 \pm (\alpha/2)\ t_{(n-2)}s(Y'_0) \qquad (71)$$

We will now have a look at the assumptions involved in simple linear regression: (a) independence, (b) homogeneity, (c) normality, (d) linearity, and (e) fixed X.

Independence

We already now this assumption from the ANOVA model. Another way of thinking of this assumption in the regression analysis is that the errors in prediction or the residuals are assumed to be random and independent. So there is no systematic pattern of the errors. We need to note that there are different types of residuals. The ei is known as raw residuals, for the same reason that Xi and Yi are called raw scores, so all being in their original scale. Some researchers dislike raw residuals as their scale depends on the scale of Y, and therefore, they must temper their interpretation of the residual values. That's why there are different types of standardized residuals developed. These values are measured along the z score scale with a mean of 0 and a variance of 1, and approximately 95% of the values are within 2 units of 0. Later in the SSPS explanation we will use studentized residuals. Studentized residuals are a type of standardized residual that are more sensitive to detecting outliers.

The easiest way for assessing this assumption is to examine a scatterplot (Y vs. X) or a residual plot. If the independence assumption is satisfied, there should be a random display of points. If the assumption is violated, the plot will display some type of pattern. As we know from ANOVA, violation of this assumptions generally occurs in the following three situations: (a) when observations are collected over time, (b) when observations are made within blocks, such that the observations within a particular block are more similar then observations in different blocks; or (c) when observation involves replication. Lack of independence affect the estimated standard errors, being under- or overestimated. For serious violations, you could consider using generalized or weighted least squares as the method of estimation.

Homogeneity

Second assumption is homogeneity of variance. This assumption must be reframed a bit in the regression context by examining the concept of a conditional distribution. In regression analysis, a conditional distribution is defined as the distribution of Y for a particular value of X. So the homogeneity assumption is that the conditional distributions have a constant variance for all values of X. In a plot of the Y scores or the residuals versus X, the consistency of the variance of the conditional distributions can be examined. A common violation of this assumption occurs when the conditional residuals variance increases as X increases. Then the residual plot is cone- or fan-shaped.

If the homogeneity assumption is violated, estimates of the standard error are larger, and although the regression coefficients remain unbiased, the validity of the significance tests is affected. Also with larger standard errors it is harder to reject H0, therefore this results in a larger number of Type II errors.

If this assumption is seriously violated, the simplest solution is to use some sort of transformation, known as variance stabilizing transformations. Commonly used transformations are the log or square root of Y. These can also improve the non-normality of the conditional distributions. A second solution is to use generalized or weighted least squares. A third solution is to use a form of robust estimation.

Normality

In the regression the normality assumption is that the conditional distributions of either Y or the prediction errors (residuals) are normal in shape. That is, for all values of X, the scores on Y or the prediction errors are normally distributed. Outliers often cause non-normality. The regression estimates are quite sensitive to outlying observations such that the precision of the estimates is affected, particularly the slope. Also the coefficient of determination can be affected. In general, the regression line will be pulled towards the outlier, because the least squares principle always attempts to find the line that best fits all of the points. Outlier observation may be a result of (a) a simple recording or data entry error, (b) an error in observation, (c) an improperly functioning instrument, (d) inappropriate use of administration instructions, or (e) a true outlier.

A simple procedure to use for single case outliers is to perform two regression analyses, both with and without the outlier being included. You can compare these results.

There are two commonly used procedures to detect the violation of the normality assumption. The simplest involves checking for symmetry in a histogram, frequency distribution, boxplot, or skewness and kurtosis statistics. Although, nonzero kurtosis (i.e. a distribution that is either flat, platykurtic, or has a sharp peak, leptokurtic) will have a minimal effect on the regression estimates. Nonzero skewness (i.e. a distribution that is not symmetrical with either a positive or a negative skew) will have much more impact on these estimates. One rule of thumb is to be concerned if the skewness value is larger than 1.5 or 2.0 in magnitude.

Another useful graphical technique is the normal probability plot (or quantile-quantile plot). With normality distributed data or residuals, the points on the normal probability plot will all along a straight diagonal line, whereas non-normal data will not. There are also several statistical procedures available for the detection of non-normality. Also transformations can be used. The most commonly used transformations to correct for non-normality in regression analysis are to transform the dependent variable using the log (to correct for positive skew) or the square root (to correct for positive or negative skew).

Linearity

This assumption indicates that there is a linear relationship between X and Y, which is also assumed for most types of correlations. If the relationship between X and Y is linear, then the sample slope and intercept will be unbiased estimators of the population slope and intercept. Looking at the scatterplot of Y versus X can often do detecting violation of the linearity assumption. If the linearity assumption is met, we expect to see no systematic pattern of points. If the assumption is violated, we expect to see a systematic pattern between e and X. Therefore; we recommend you examine both the scatterplot and the residual plot.There are two options how to deal with the violation of linearity assumption. The first option is to transform either one or both of the variables to achieve linearity. Then the method of least squares can be used to perform a linear regression analysis on the transformed variables. However, when dealing with transformed variables measured along a different scale, results need to be described in terms of the transformed rather then the original variables. A second option is to use a nonlinear model to examine the relationship between the variable in their original scale.

Fixed X

This means X is a fixed variable rather than a random variable. This result in the regression model being valid only for those particular values of X that was actually observed and uses in the analysis. Two obvious situations that come to mind are extrapolation and interpolations of values of X.

In general, we may not want to make predictions about individuals having X scores that are outside of the range of values used in developing the prediction model; this is defined as extrapolating beyond the sample predictor data. On the other hand, we are not quite as concerned in making predictions about individuals having X scores within the range of values used in developing the prediction model; this is defined as interpolating within the range of the sample predictor data. In the interpolation situation, we expect the prediction errors to be somewhat smaller as compared to the extrapolation situation because there are at least some similar supportive prediction data for the former.

This is a table that give a summary of the assumptions and the effects of violation.

Assumption	Effect of Assumption Violation
Independence	Influences standard errors of the model
Homogeneity	Bias in variances of errors May inflate the standard errors and thus increase likelihood of a Type II error May result in non-normal conditional distributions
Normality	Less precise slope, intercept, and R2
Linearity	Bias in slope and intercept Expected change in Y is not a constant and depends on value of X Reduced magnitude of coefficient of determination
Values of X fixed	Extrapolating beyond the range of X: prediction errors are larger, may also bias slope and intercept Interpolating within the range of X: smaller effect than when extrapolating; if other assumptions met, negligible effect.

7.4

To conduct a simple linear regression you need to have data that consists of two variables, a dependent and independent variable. Now we will discuss the steps to conduct a simple linear regression in SPSS:

1. Go to “Analyze” and select “regression” and select “Linear”.
2. Click the dependent variable and move it into the “dependent” box. And click the independent variable into the “Independent(s)” box.
3. From the “Linear regression” dialog box, clicking on “statistics” will provide the option to select various regression coefficients and residuals. From the “Statistics” dialog box place a checkmark in the box nest to the following: (1) estimates, (2) confidence intervals, (3) model fit, (4) descriptive, (5) Durbin-Watson, (6) case wise diagnostics. Click on “continue”.
4. Clicking on “plots” will provide the option to select various residual plots. From this dialog box, check the following: (1) histogram and (2) normal probability plot. Click on “continue”.

5. Clicking on “save” will provide the option to save everything. From the “save” dialog box under the heading of predicted values, place a checkmark in the box nest to the following: unstandardized. Under the heading of residuals, place a checkmark in the box next to the following (1) unstandardized and (2) studentized. Under the heading of distances, place a checkmark in the box next to the following: (1) mahalanobis and (2) Cook’s. Under the heading Influence Statistics, place a checkmark in the box next to the following: (1) DFBETA(s), and (2) Standardized DFBETA(s). Click on “continue”, then click “Ok” to generate output.

On the pages 348 en 349 you can see the output that is generated and how the interpret the results.

We will now review the values that we requested to be saved in our data file (see page 351):

- PRE_1, are the unstandardized predicted values
- RES_1 are the unstandardized residuals, simply the difference between the observed and predicted values.
- SRE_1 are the studentized residuals, a type of standardized residual that is more sensitive to outliers as compared to standardized residuals. Studentized residuals are computed as the unstandardized residual divided by an estimate of the standard deviation with that case removed. As a rule of thumb, studentized residuals with an absolute value greater than 3 are considered outliers.
- MAH_1 are Mahalanobis distance values that can be helpful in detecting outliers. These values can be reviewed to determine cases that are exerting leverage. Squared Mahalanobis distances divided by the number of variables which are greater than 2.5 (small samples) or 3-4 (large samples) are suggestive of outliers.
- COO_1 are Cook’s distance values and provide an indication of influence of individual cases. As a rule of thumb, Cook’s values greater than 1.0 suggest that case is potentially problematic.
- DFB0_1 and DFB1_1 are unstandardized DFBETA values for the intercept and slope. These values provide estimates of intercept and slope when that case is removed.
- SDB0_1 and SDB1_1 are standardized DFBETA values for the intercept and slope and are easier to interpret as compared to their unstandardized counterparts. Standardized DFBETA values greater than an absolute value of 2 suggest that the case may be exerting undue influence on the parameters of the model.

We can plot the studentized residuals against the values of X to examine the extent to which independence was met. If the assumption of independence is met, the points should fall randomly within a band of -2.0 and +2.0.

We can use the same plot to examine the extent to which homogeneity was met. Evidence of meeting the assumption of homogeneity is a plot where the spread of residuals appear fairly constant over the range of X values. If the spread of the residuals increases or decreases across the plot from left to right, this may indicate that the assumption has been violated.

When we have only one independent variable, a simple bivariate scatterplot of the dependent variable (on Y axis) and the independent variable (on X axis) will provide a visual indication of the extent to which linearity is reasonable. Additionally the plot of studentized residuals against X values can be used to examine the extent to which linearity was met. Here a random display of points within an absolute value of 2 or 3 suggests evidence.

We examine residuals for normality. Also we can use the skewness and kurtosis. When both the values are within the range of an absolute value of 2,0 suggest evidence of normality. We can also use the Shapiro-Wilk (S-W) statistic. Also Q-Q plots can be used.

7.5

Again a priori and post hoc power could be determined using G*power. In G*power you need to select the correct test family. Here we conduct simple linear regression. To find regression, select “tests”, then “correlation and regression”, and then “ Linear bivariate regression: one group, size of slope”. Once that selection is made the “test family” automatically changes to “t tests”. Now the input parameters for the Post hoc test are as follows:

(1) Number of tails, (2) effect size, slope H1, (3) alpha level, (4) total sample size, (5) slope H0, (6) standard deviation of X, (7) standard deviation of Y.