

Multiple Regression

James H. Steiger

February 13, 2006

Goals for this Module

The Multiple ...

The Multiple ...

The Partial ...

The Semi-Partial ...

Statistical ...

Sample Formulas

Least Squares ...

Bias of the Sample R^2

Statistical Tests in ...

Regression Diagnostics

[Home Page](#)

[Print](#)

[Title Page](#)



Page 1 of 27

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Home Page

Print

Title Page



Page 2 of 27

Go Back

Full Screen

Close

Quit

1. Goals for this Module

In this module, we will discuss:

1. The general multiple linear regression model.
2. Statistical assumptions of multiple regression
3. The “best estimate” of the multiple regression equation
4. Statistical tests in multiple regression
5. Regression diagnostics

[Home Page](#)[Print](#)[Title Page](#)[◀](#) [▶](#)[◀](#) [▶](#)

Page 3 of 27

[Go Back](#)[Full Screen](#)[Close](#)[Quit](#)

2. The Multiple Regression Model

In *bivariate linear regression*, we learned to predict a single dependent variable y from a single independent variable x with the equations

$$\begin{aligned}y &= \hat{y} + \varepsilon \\ &= b_1x + b_0 + \varepsilon\end{aligned}$$

In multiple linear regression, we predict the dependent variable from several independent variables $x_1 \dots x_k$ using the equation

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_kx_k + b_0 + \varepsilon \quad (1)$$

Dealing with multiple predictors is considerably more challenging than dealing with only a single predictor. Some of the problems include

1. *Choosing the best model.* In multiple regression, often several different sets of variables perform equally well in predicting a criterion. Which set should you use?
2. *Interactions between variables.* In some cases, independent variables interact, and the regression equation will not be accurate unless this interaction is taken into account.
3. *Much greater difficulty visualizing the regression relationships.* With only one independent variable, the regression line can be plotted

neatly in two dimensions. With two predictors, there is a *regression surface* instead of a regression line, and with 3 predictors and one criterion, you run out of dimensions for plotting.

4. *Model interpretation becomes substantially more difficult.* The multiple regression equation changes as each new variable is added to the model. Since the regression weights for each variable are modified by the other variables, and hence depend on what is in the model, the substantive interpretation of the regression equation is problematic.

As an example consider the following data from the Kleinbaum, Kupper and Miller text on regression analysis. These data show weight, height, and age of a random sample of 12 nutritionally deficient children.

Suppose we wish to investigate how weight is related to height and age for these children. We may want to consider only the simple model

$$y = b_1x_1 + b_2x_2 + b_0 + \varepsilon$$

but we have several other alternatives. For example, we might want to examine both first and second order terms for x_1 , in which case our model would be

$$\begin{aligned} y &= b_1x_1 + b_2x_2 + b_3x_1^2 + b_0 + \varepsilon \\ &= \hat{y} + \varepsilon \end{aligned}$$

WGT(y)	HGT(x_1)	AGE(x_2)
64	57	8
71	59	10
53	49	6
67	62	11
55	51	8
58	50	7
77	55	10
57	48	9
56	42	10
51	42	6
76	61	12
68	57	9

Table 1: Data for 12 children

Goals for this Module

The Multiple ...

The Multiple ...

The Partial ...

The Semi-Partial ...

Statistical ...

Sample Formulas

Least Squares ...

Bias of the Sample R^2

Statistical Tests in ...

Regression Diagnostics

Home Page

Print

Title Page



Page 5 of 27

Go Back

Full Screen

Close

Quit

Goals for this Module

The Multiple . . .

The Multiple . . .

The Partial . . .

The Semi-Partial . . .

Statistical . . .

Sample Formulas

Least Squares . . .

Bias of the Sample R^2

Statistical Tests in . . .

Regression Diagnostics

Home Page

Print

Title Page

◀◀ ▶▶

◀ ▶

Page 6 of 27

Go Back

Full Screen

Close

Quit

Note, however, that this nonlinear model can also be written in the form

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_0 + \varepsilon$$

where $x_3 = x_1^2$, and so it can be viewed, in a sense, through the “lens” of the more basic linear model.

Home Page

Print

Title Page



Page 7 of 27

Go Back

Full Screen

Close

Quit

3. The Multiple Correlation Coefficient

The correlation between the predicted scores and the criterion scores is called the “multiple correlation coefficient,” and is almost universally denoted with the value R . Curiously, many writers use this notation whether a sample or a population value is referred to, which creates some problems for some readers. We can eliminate this ambiguity by using either ρ^2 or R_{pop}^2 to signify the population value. Since R is always positive, and R^2 is the “percentage of variance in y accounted for by the predictors” (in the colloquial sense), most discussions center on R^2 rather than R . When it is necessary for clarity, one can denote the squared multiple correlation as $R_{y|x_1x_2}^2$ to indicate that variates x_1 and x_2 have been included in the regression equation.

4. The Partial Correlation Coefficient

The *partial correlation coefficient* is a measure of the strength of the linear relationship between two variables after the contribution of other variables has been “partialled out” or “controlled for” using linear regression. We will use the notation $r_{yx|w_1, w_2, \dots, w_p}$ to stand for the partial correlation between y and x with the w 's partialled out. This correlation is simply the Pearson correlation between the regression residual $\varepsilon_{y|w_1, w_2, \dots, w_p}$ for y with the w 's as predictors and the regression residual $\varepsilon_{x|w_1, w_2, \dots, w_p}$ of x with the w 's as predictors.

Goals for this Module

The Multiple ...

The Multiple ...

The Partial ...

The Semi-Partial ...

Statistical ...

Sample Formulas

Least Squares ...

Bias of the Sample R^2

Statistical Tests in ...

Regression Diagnostics

Home Page

Print

Title Page

◀▶

◀▶

Page 8 of 27

Go Back

Full Screen

Close

Quit

5. The Semi-Partial (Part) Correlation

This is similar to the partial correlation, except that the variables “controlled for” are only partialled out of one of the two variables. We use the notation $r_{Y(X_1|X_2)}$ to stand for the correlation between y and the residual of x_1 after x_2 has been partialled from it.

Goals for this Module

The Multiple...

The Multiple...

The Partial...

The Semi-Partial...

Statistical...

Sample Formulas

Least Squares...

Bias of the Sample R^2

Statistical Tests in...

Regression Diagnostics

Home Page

Print

Title Page



Page 9 of 27

Go Back

Full Screen

Close

Quit

6. Statistical Assumptions of Multiple Regression

1. *Homoscedasticity.* The conditional variance of y given any specific combination of values of the $x_1 \dots x_k$ is the same, i.e., σ_ε^2
2. *Existence.* For each combination of values of the basic independent variables $x_1 \dots x_k$, y is a univariate random variable having a certain probability distribution with finite mean and variance.
3. *Independence.* The y observations are statistically independent
4. *Linearity.* The expected value of y conditional on all specific combinations of values of the $x_1 \dots x_k$ is a linear function of the x 's, and follows the linear regression rule. For example, if $k = 2$,

$$\mu_{y|x_1=a_1, x_2=a_2} = b_1 a_1 + b_2 a_2 + b_0$$

5. *Normality.* The conditional distribution of y for any combination of values of the $x_1 \dots x_k$ is normal, or Gaussian.

Note how these assumptions are quite similar to those for the bivariate case. Again, the conditional distribution of y given x is simply normal, with a mean that may be computed from the regression equation, and a variance that remains constant over all conditional values of x . A

mnemonic for the above suggested by Kleinbaum, Kupper, and Miller (1989) in their textbook on regression is HEIL GAUSS.

Goals for this Module

The Multiple . . .

The Multiple . . .

The Partial . . .

The Semi-Partial . . .

Statistical . . .

Sample Formulas

Least Squares . . .

Bias of the Sample R^2

Statistical Tests in . . .

Regression Diagnostics

Home Page

Print

Title Page



Page 11 of 27

Go Back

Full Screen

Close

Quit

7. Sample Formulas

The sample equivalent to the population formula in Equation 1 assumes that the N scores on y and $x_1 \dots x_k$ are in column variate form, whence

$$\mathbf{y} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \dots + b_k\mathbf{x}_k + b_0 + \mathbf{e} \quad (2)$$

$$= \hat{\mathbf{y}} + \mathbf{e} \quad (3)$$

Note that \mathbf{y} , the \mathbf{x} 's, and \mathbf{e} are all $N \times 1$ vectors.

7.1. Matrix Formulation

Equation 2 can be rewritten in matrix form. Place all the \mathbf{x} 's in a matrix \mathbf{X} , but also include a column of 1's in order to include the constant b_0 . Then the sample MR model may be written as

$${}_N\mathbf{Y}_1 = {}_N\mathbf{X}_{k+1}\mathbf{b}_1 + {}_N\mathbf{e}_1$$

where

$$\mathbf{X} = [\mathbf{x}_1\mathbf{x}_2 \dots \mathbf{x}_k\mathbf{1}]$$

and

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \\ b_0 \end{bmatrix}$$

- Goals for this Module
- The Multiple ...
- The Multiple ...
- The Partial ...
- The Semi-Partial ...
- Statistical ...
- Sample Formulas
- Least Squares ...**
- Bias of the Sample R^2
- Statistical Tests in ...
- Regression Diagnostics

- Home Page
- Print
- Title Page
- ◀◀ ▶▶
- ◀ ▶
- Page 13 of 27
- Go Back
- Full Screen
- Close
- Quit

8. Least Squares Estimates of the Multiple Regression Equation

The least squares estimates minimize the sum of squared errors. In matrix formulation, given \mathbf{y} and \mathbf{X} , we must choose the elements of \mathbf{b} so that $\mathbf{e}^T\mathbf{e}$ is a minimum. Just as in bivariate linear regression, the solution to this problem is computed easily from calculus as

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Notice that since $\mathbf{S}_{\mathbf{X}\mathbf{X}} = \mathbf{X}^T\mathbf{X}/(N - 1)$ and $\mathbf{s}_{\mathbf{X}\mathbf{y}} = \mathbf{X}^T\mathbf{y}/(N - 1)$, it also follows that

$$\mathbf{b} = \mathbf{S}_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{s}_{\mathbf{X}\mathbf{y}}$$

Notice how, in the above equations, I have adopted a notation that explicitly displays whether a covariance matrix is computed on a matrix or a vector of values, and whether the covariance matrix is itself a matrix or a vector. Many authors will not find it necessary to maintain all these distinctions in their notation. A major reason for this is that the above results for a single criterion variable generalize immediately to the situation where you have more than one criterion, and you are trying to minimize the “overall sum of squared errors.”

In this latter case, the regression model becomes

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

- Goals for this Module
- The Multiple ...
- The Multiple ...
- The Partial ...
- The Semi-Partial ...
- Statistical ...
- Sample Formulas
- Least Squares ...**
- Bias of the Sample R^2
- Statistical Tests in ...
- Regression Diagnostics

- Home Page
- Print
- Title Page
- ◀◀
- ▶▶
- ◀
- ▶
- Page 14 of 27
- Go Back
- Full Screen
- Close
- Quit

and our task is to minimize $\text{Tr}(\mathbf{E}^T \mathbf{E})$. The solution for \mathbf{B} is

$$\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

or

$$\mathbf{B} = \mathbf{S}_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{S}_{\mathbf{X}\mathbf{Y}}$$

Specializing the notation can become tiresome, especially when the audience is expert. In addition, quite a few books on multiple regression will completely dispense with a matrix treatment, or relegate a very brief matrix treatment to an appendix.

- Goals for this Module
- The Multiple ...
- The Multiple ...
- The Partial ...
- The Semi-Partial ...
- Statistical ...
- Sample Formulas
- Least Squares ...
- Bias of the Sample R^2**
- Statistical Tests in ...
- Regression Diagnostics
- Home Page
- Print
- Title Page
- ◀◀
- ▶▶
- ◀
- ▶
- Page 15 of 27
- Go Back
- Full Screen
- Close
- Quit

9. Bias of the Sample R^2

When a population correlation is zero, the sample correlation is hardly ever zero. As a consequence, the R^2 value obtained in an analysis of sample data is a biased estimate of the population value. An unbiased estimator is available (Olkin and Pratt, 1958), but requires very powerful software like Mathematica to compute, and consequently is not available in standard statistics packages. As a result, these packages compute an approximate “shrunk” (or “adjusted”) estimate and report it alongside the uncorrected value. The adjusted estimator is

$$\tilde{R}^2 = 1 - (1 - R^2) \frac{N - 1}{N - k - 1}$$

- Goals for this Module
- The Multiple ...
- The Multiple ...
- The Partial ...
- The Semi-Partial ...
- Statistical ...
- Sample Formulas
- Least Squares ...
- Bias of the Sample R^2
- Statistical Tests in ...**
- Regression Diagnostics
- Home Page
- Print
- Title Page
- ◀◀ ▶▶
- ◀ ▶
- Page 16 of 27
- Go Back
- Full Screen
- Close
- Quit

10. Statistical Tests in Multiple Regression

Frequently multiple regressions are at least partially exploratory in nature. You gather data on a large number of predictors, and try to build a model for explaining (or predicting) y from a number of x 's. A key aspect of this is choosing which x 's to retain. A key problem is that, especially when N is small and the number of x 's is large, there will be a number of *spuriously large correlations* between the criterion and the x 's. You can capitalize on chance, as it were, and build a regression equation using variables that have high correlations with the criterion, but this equation will not generalize to any new situation.

There are a number of statistical tests available in multiple regression, and they are printed routinely by statistical software such as SPSS, SAS, Statistica, SPLUS, and R. It is important to realize that these test do *not* in general correct for post hoc selection. So, for example, if you have 90 potential predictors that all actually correlate zero with the criterion, you can choose the predictor with the highest absolute correlation with the criterion in your current sample, and invariably obtain a “significant” result. Strangely, this fact is seldom brought to the forefront in regression texts, and so people actually believe that the F statistics and associated probability values somehow determine whether the regression equation is significant in the sense most relatively naive users would expect.

10.1. F Test for Significant Overall Regression

You have a random sample of N observations on y and a set of x 's. Is the the population R_{pop}^2 greater than zero? To test the null hypothesis that $R^2 = 0$, we can calculate an F statistic

$$\begin{aligned} F &= \frac{R^2/k}{(1 - R^2)/(N - k - 1)} \\ &= \frac{MS_{\hat{y}}}{MS_{\varepsilon}} = \frac{SS_{\hat{y}}/k}{SS_{\varepsilon}/(N - k - 1)} \end{aligned}$$

where

$$SS_{\hat{y}} = \sum_{i=1}^N (\hat{y}_i - \bar{y}_i)^2$$

and

$$SS_{\varepsilon} = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \sum_{i=1}^N e_i^2$$

10.2. The Partial F Test

Often, we are building up a model, and we want to assess whether a new predictor actually improves the quality of prediction, i.e., actually increases R_{pop}^2 beyond its current value. Suppose you have 3 potential predictors, x_1 , x_2 , and x_3 , and that you begin by predicting y from x_1 alone. The

question then becomes, will x_2 add any additional predictive ability to the regression equation? More generally, you have k predictors x_1, x_2, \dots, x_k already and you add a new predictor w .

This test can be computed as

$$\begin{aligned}
 F_{1, N-k-2} &= \frac{\left(R_{y|x_1, x_2, \dots, x_k, w}^2 - R_{y|x_1, x_2, \dots, x_k}^2 \right)}{\left(1 - R_{y|x_1, x_2, \dots, x_k, w}^2 \right) / (N - k - 2)} \\
 &= \frac{\left(SS_{\hat{y}|x_1, x_2, \dots, x_k, w} - SS_{\hat{y}|x_1, x_2, \dots, x_k} \right)}{SS_{\varepsilon|x_1, x_2, \dots, x_k, w} / (N - k - 2)}
 \end{aligned}$$

As each variable is added to the regression equation, it adds to $SS_{\hat{y}}$. These non-overlapping amounts of variance add up, so, for example, after you have a total of 3 variables in your prediction equation, you can look back and see that the total $SS_{\hat{y}}$ is equal to the sums of the unique amounts of prediction variance contributed by each variable. A numerical example may help make this clear.

10.3. An Example Analysis

Consider the simple data set in Table 1. Suppose we wish to examine the contribution of HGT, AGE, and $(AGE)^2$ to the prediction of WGT. We power up SPSS and quickly add a new variable AGE_2 and compute it as $(AGE)^2$ using the *transform->compute* facility.

SPSS has several options for selecting the variables to use in a regression analysis. Here are 3.

1. *Forward selection.*

- (a) You select a group of independent variables to be examined.
- (b) The variable with the highest squared correlation with the criterion is added to the regression equation
- (c) The partial F statistic for each possible remaining variable is computed.
- (d) If the variable with the highest F statistic passes a criterion, it is added to the regression equation, and R^2 is recomputed.
- (e) Keep going back to step c, recomputing the partial F statistics until no variable can be found that passes the criterion for significance.

2. *Backward elimination.*

- (a) You start with all the variables you have selected as possible predictors *included in the regression equation.*
- (b) You then compute partial F statistics for each of the variables remaining in the regression equation.

Goals for this Module

The Multiple . . .

The Multiple . . .

The Partial . . .

The Semi-Partial . . .

Statistical . . .

Sample Formulas

Least Squares . . .

Bias of the Sample R^2

Statistical Tests in . . .

Regression Diagnostics

Home Page

Print

Title Page



Page 19 of 27

Go Back

Full Screen

Close

Quit

- (c) Find the variable with the *lowest* F .
 - (d) If this F is low enough to be below a criterion you have selected, remove it from the model, and go back to step b.
 - (e) Continue until no partial F is found that is sufficiently low.
3. *Stepwise regression* This works like forward regression except that you examine, at each stage, the possibility that a variable entered at a previous stage has now become superfluous because of additional variables now in the model that were not in the model when this variable was selected. To check on this, at each step a partial F test for each variable in the model is made as if it were the variable entered last. We look at the lowest of these F 's and if the lowest one is sufficiently low, we remove the variable from the model, recompute all the partial F 's, and keep going until we can remove no more variables.

Suppose we try forward elimination. Since our sample size is so low, let's adopt a less stringent than normal criterion for entry into the formula, i.e. a p value less than .10 instead of the typical .05. We obtain the following results from SPSS

This table, by itself, really isn't very informative. It tells us that two models were found acceptable, one with only HGT as a predictor, one with HGT and AGE as predictors. Both were significant, in the sense that the

Home Page

Print

Title Page



Page 20 of 27

Go Back

Full Screen

Close

Quit

ANOVA ^c						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	588.923	1	588.923	19.675	.001 ^a
	Residual	299.327	10	29.933		
	Total	888.250	11			
2	Regression	692.823	2	346.411	15.953	.001 ^b
	Residual	195.427	9	21.714		
	Total	888.250	11			

a. Predictors: (Constant), HGT
b. Predictors: (Constant), HGT, AGE
c. Dependent Variable: WGT

Figure 1: ANOVA Results

Goals for this Module

The Multiple ...

The Multiple ...

The Partial ...

The Semi-Partial ...

Statistical ...

Sample Formulas

Least Squares ...

Bias of the Sample R^2

Statistical Tests in ...

Regression Diagnostics

Home Page

Print

Title Page

⏪ ⏩

◀ ▶

Page 21 of 27

Go Back

Full Screen

Close

Quit

Model Summary									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.814 ^a	.663	.629	5.4711	.663	19.675	1	10	.001
2	.883 ^b	.780	.731	4.6598	.117	4.785	1	9	.056

a. Predictors: (Constant), HGT
b. Predictors: (Constant), HGT, AGE

Figure 2: Summary of Partial F Tests

hypothesis of zero prediction could be rejected. To test whether model 2 is significantly better than model 1, you need to partition the sum of squares by subtraction. Notice that $SS_{\hat{y}} = 588.923$ for model 1, but $SS_{\hat{y}} = 692.823$ for model 2. The difference is 103.9. Divide this by MS_{ϵ} for model 2 to get the partial F statistic, i.e.

$$F = \frac{103.9}{21.714} = 4.785$$

This has a p value of .056, which is significant under our relaxed criterion. The partial F tests are summarized in this table of results

The final model selected predicts WGT as a linear function of HGT and AGE. The regression coefficients and their standard errors are shown below.

Coefficients ^a									
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	6.190	12.849		.482	.640			
	HGT	1.072	.242	.814	4.436	.001	.814	.814	.814
2	(Constant)	6.553	10.945		.599	.564			
	HGT	.722	.261	.548	2.768	.022	.814	.678	.433
	AGE	2.050	.937	.433	2.187	.056	.770	.589	.342

a. Dependent Variable: WGT

Figure 3: Regression Coefficients and their Standard Errors for Two Models

10.4. Forward Selection with Completely Random Data

In the previous session, I pointed out that the standard statistical tests in linear regression do not correct for *post hoc* selection. To demonstrate this, I used the random number generation system in SPSS to generate 100 observations on 91 variables. These data are completely independent normal random numbers, so that all the predictor-criterion correlations are actually zero. However, if we declare the first variable to be the criterion and the other 90 to be predictors, with the small N and large number of variables, we expect around 4 or 5 predictor-criterion correlations to be “significant” at the .05 level. As it turns out, in this case 7 of the predictors correlate significantly with the criterion. When we perform forward selection, we obtain the output in Figure 4. As you can see, we get a highly significant R^2 with $p < .001$.

Goals for this Module

The Multiple ...

The Multiple ...

The Partial ...

The Semi-Partial ...

Statistical ...

Sample Formulas

Least Squares ...

Bias of the Sample R^2

Statistical Tests in ...

Regression Diagnostics

Home Page

Print

Title Page

◀◀

▶▶

◀

▶

Page 24 of 27

Go Back

Full Screen

Close

Quit

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.329 ^a	.108	.099	.96759	.108	11.904	1	98	.001
2	.405 ^b	.164	.147	.94169	.056	6.467	1	97	.013
3	.465 ^c	.216	.192	.91655	.052	6.393	1	96	.013
4	.504 ^d	.254	.222	.89914	.037	4.754	1	95	.032
5	.540 ^e	.291	.254	.88066	.038	5.029	1	94	.027

a. Predictors: (Constant), VAR61

b. Predictors: (Constant), VAR61, VAR15

c. Predictors: (Constant), VAR61, VAR15, VAR54

d. Predictors: (Constant), VAR61, VAR15, VAR54, VAR6

e. Predictors: (Constant), VAR61, VAR15, VAR54, VAR6, VAR75

Figure 4: Regression Analysis of Completely Random Normal Data

11. Regression Diagnostics

11.1. Residual Analysis

Residuals tell us much about the suitability of a regression model. Large residuals, for example, are multivariate outliers, and may be indicative of a recording error, unusual observation, or violation of assumptions.

If a model fits well and the (HEIL GAUSS) statistical assumptions are met, then the residuals should be normally distributed, independent, have a zero mean, constant variance σ_ε^2 . Regression diagnostics examine the data in detail to see if they depart in a meaningful way from these specifications. A full discussion of these would take several lectures — I recommend Cohen, Cohen, West, and Aiken Chapter 10. Some types of residuals typically examined:

1. *Standardized Residual.*

$$z_i = \frac{e_i}{S}$$

where

$$S^2 = \frac{1}{N - k - 1} \sum_{i=1}^N e_i^2$$

is the unbiased estimate of σ_ε^2 .

2. Studentized Residual

$$r_i = \frac{z_i}{\sqrt{1 - h_i}}$$

These follow approximately a Student t distribution with $N - k - 1$ degrees of freedom if the data meet the HEIL GAUSS assumptions. h_i is the “leverage” of the i th observation. A typical guideline is to declare an observation to be an outlier if this value is greater than 2 for small samples and 3 for large samples.

3. *Leverage* is a measure of how far away an observation is from the means of the predictor variables. Typical values for rejecting an observation are $2(k + 1)/N$ for large N , and $3(k + 1)/N$ for small N .
4. *Cook’s D*. This measures the influence of an observation by aggregating the change in the \hat{y}_i when the i th observation is omitted from the data.

11.2. Collinearity Analysis

As predictors become more highly correlated, it becomes increasingly difficult to obtain accurate estimators of the regression coefficients.

The *tolerance* of the i th variable is defined as $1 - R_i^2$ where R_i^2 is the squared multiple correlation between x_i and the other x ’s. If tolerance is less than .1, it indicates that the variable is quite redundant with the other variables.