Chapter 869

Multiple Regression

Introduction

This procedure computes power and sample size for a multiple regression analysis in which the relationship between a dependent variable Y and a set independent variables X_1, X_2, …, X_k is to be studied. In multiple regression, interest usually focuses on the regression coefficients. However, since the X's are usually not available during the planning phase, little is known about these coefficients until after the analysis is run. Hence, this procedure uses the squared multiple correlation coefficient, R^2, as the measure of effect size upon which the power analysis and sample size is based. Gatsonis and Sampson (1989) present power analysis results for two approaches: *unconditional* and *conditional*. Both of these approaches are available in this procedure.

Unconditional (Random X's) Model

In the unconditional or random X's model, the X's and Y have a joint multivariate normal distribution with a specified mean vector and covariance matrix given by

$$\begin{bmatrix} \sigma_Y^2 & \Sigma_{YX}' \\ \Sigma_{YX} & \Sigma_X \end{bmatrix}$$

The study-specific values of X are unknown at the design phase, so the sample size determination is based on a single, effect-size parameter which represents the expected variations in the X's, their interrelationships, and their relationship with Y. This effect-size parameter is the *squared multiple correlation coefficient* which is defined in terms of the covariance matrix as

$$\rho_{YX}^2 = \frac{\Sigma_{YX}'\Sigma_X^{-1}\Sigma_{YX}}{\sigma_Y^2}$$

If this coefficient is zero, the variables *X* provide no information about the linear prediction of *Y*. Note that we will use ρ^2 to represent ρ_{YX}^2.

The sample statistic corresponding to this parameter is R^2, the *coefficient of determination*. Often, the primary hypothesis involves testing the significance of a subset of X's that have been statistically adjusted for a second set of X's. The population parameter is then called the *squared multiple **partial** correlation coefficient*, which is interpreted similarly.

This approach is more common because usually the independent variables are random variables that are observed during the study. If the study were conducted twice, the two set of X's would be different.

Test Statistic in the Unconditional Model

An *F*-test with *k* and *N-k-1* degrees of freedom can be constructed that will test whether all the regression coefficients simultaneously zero as follows

$$F_{k,N-k-1} = \frac{R^2/k}{(1-R^2)/(N-k-1)}$$

Suppose the independent variables are divided into two sets: C containing k_C variables and T containing the remaining $k_T = k - k_C$ variables. That is, we partition $X = X_T|X_C$. It can be shown that an F-test that tests the significance of the T variables adjusted for the C variables is

$$F_{k_T,N-k-1} = \frac{\left(R^2_{YX_T|X_C}\right)/k_T}{\left(1 - R^2_{YX_T|X_C}\right)/(N-k-1)}$$

The quantity $R^2_{YX_T|X_C}$ is the sample estimate of the population squared multiple partial correlation coefficient $\rho^2_{YX_T|X_C}$.

Cohen (1988) shows that $R^2_{YX_T|X_C}$ can be calculated from the R^2 of fitting all the variables and the R^2 of fitting just the set C variables as follows

$$R^2_{YX_T|X_C} = \frac{R^2_{YX} - R^2_{YX_C}}{1 - R^2_{YX_C}}$$

Calculating the Power in the Unconditional Model

In the unconditional model approach, the statistical hypotheses that is usually of most interest is the set $H_0: \rho^2 \leq \rho_0^2$ versus $H_1: \rho^2 > \rho_0^2$ because you want to establish a lower bound for the value, not just established that it is greater than zero.

However, the hypothesis $H_0: \rho^2 \geq \rho_0^2$ versus $H_1: \rho^2 < \rho_0^2$ is also valid. In the program, when $\rho_1^2 > \rho_0^2$ the former hypothesis set is assumed. Otherwise, the later set is assumed.

The calculation of the power of a particular test proceeds as follows:

1. Determine the critical value r_α from the CDF such that $\mathrm{P}(R^2 \leq r_\alpha|N,k,\rho_0^2) = 1-\alpha$. Note that we use the value of ρ^2 specified in the null hypothesis.
2. Compute the power using $\text{Power} = 1 - \mathrm{P}(R^2 \leq r_\alpha|N,k,\rho_1^2)$.

Krishnamoorthy and Xia (2003) give the CDF of R^2 as

$$\mathrm{P}(R^2 \leq x|N,k,\rho^2) = \sum_{i=0}^{\infty} \mathrm{P}(Y=i)\, I_x\left(\frac{k-1}{2}+i, \frac{N-k}{2}\right)$$

where

$$I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x t^{a-1}(1-t)^{b-1}dt$$

$$\mathrm{P}(Y=i) = \frac{\Gamma\left(\frac{N+1}{2}+i\right)}{\Gamma(i+1)\Gamma\left(\frac{N+1}{2}\right)}(\rho^2)^i(1-\rho^2)^{\frac{N+1}{2}}$$

This formulation does not admit $\rho^2 = 0$, so when this occurs, the program inserts $\rho^2 = 0.000000000001$.

Finally, when computing the squared multiple *partial* correlation coefficient, Gatsonis and Sampson (1989) indicate you simply need to replace N with $N - k_C$ in the above CDF.

Conditional (Fixed X's) Model

In this approach, the values of the X's are preset by the researchers and are assumed to be known at the planning stage. Since they are known constants, they are not treated as random variables with a probability distribution. Any hypotheses that are tested are conditional on the specific set of X values. The focus in this analysis is how much R^2 increases when a certain set of independent variables is added to the regression model.

We will adopt the following notation: suppose *C* (controlled) and *T* (tested) are two, non-overlapping subsets of X's. Define $R^2_{T|C} = R^2_{TC} - R^2_C$ to be the R^2 added when *Y* is regressed on the variables in set *T* after adjusting for the variables in set *C*. Here, R^2_C is the R^2 when *Y* is regressed on only those variables in set *C* and R^2_{TC} is the R^2 when *Y* is regressed on the variables in both sets.

Test Statistic in the Conditional Model

You can construct *F*-tests that will test whether the regression coefficients corresponding to certain sets of *X's* are simultaneously zero while controlling for other variables. For example, to test the significance of the *X's* in set *T* while removing the influence of the *X's* in set *C* from experimental error, you would use

$$F_{k_T,N-k_C-k_T-1} = \frac{\left(R^2_{T|C}\right)/k_T}{\left(1-R^2_C-R^2_{T|C}\right)/k_C}$$

where k_T is the number of variables in *T* and k_C is the number of variables in *C*. Most significance tests in regression analysis, correlation analysis, analysis of variance, and analysis of covariance may be constructed using these F-ratios.

Calculating the Power in the Conditional Model

In this case, power calculations are based on the noncentral-F distribution. The calculation of the power of a particular test proceeds as follows:

1. Determine the critical value $F_{T,N-k_T-k_C-1,\alpha}$ where α is the probability of a type-I error.
2. Calculate the noncentrality parameter λ using the formula:

$$\lambda = N\left(\frac{R^2_{T|C}}{1-R^2_C-R^2_{T|C}}\right)$$

3. Compute the power as the probability of being greater than $F_{u,v,\alpha}$ in a noncentral-*F* distribution with noncentrality parameter λ.

Note that the formula for λ is different from that used in ***PASS 6.0***. The algorithm used in ***PASS 6.0*** was based on formula (9.3.1) in Cohen (1988) which gives approximate answers. This version of ***PASS*** using an algorithm that gives exact answers.

Cohen's Effect Size

Cohen's (1988) measure of the effect size in multiple regression, f^2 is

$$f^2 = \frac{R^2}{1-R^2}$$

so that

$$R^2 = \frac{f^2}{1+f^2}$$

When the independent variables are divided into the two sets as outlined above, f^2 is

$$f^2 = \left(\frac{R^2_{YX_T|X_C}}{1-R^2_{YX_T|X_C}}\right)$$

Cohen (1988) defined values near 0.02 as small, near 0.15 as medium, and above 0.35 as large. In terms of R^2, these are about 0.02, 0.13, and 0.26.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design tab. For more information about the options of other tabs, go to the Procedure Window chapter.

Design Tab

The Design tab contains most of the parameters and options that you will be concerned with.

Solve For

Solve For

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *Power*, *Sample Size*, *Alpha*, and *Effect Size*. Under most situations, you will select either *Power* or *Sample Size*.

Select *Sample Size* when you want to calculate the sample size needed to achieve a given power and alpha level.

Select *Power* when you want to calculate the power of an experiment.

Power and Alpha

Power

This option specifies one or more values for power. Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha

This option specifies one or more values for the probability of a type-I error (alpha). A type-I error occurs when you reject the null hypothesis when in fact it is true.

Values of alpha must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Sample Size)

This option specifies the value(s) for N, the sample size. Note that $kc + k_T < N - 1$.

Effect Size

Regression Model Type

Choose between two approaches to the modelling and analysis of multiple regression data. Both approaches result in the same F-test for significance testing. However, they differ in the calculation of power.

- **Unconditional (Random X's)**

  This approach is by far the most realistic. It assumes that Y and the X's follow a multivariate normal distribution. Thus, the values of the X's are not known until they are observed during the study. They certainly are not known during the study planning phase.

  Power and sample size calculation is based on the distribution of the squared multivariate (partial) correlation coefficient. The extra variation that occurs because the X's are random variables is accounted for in the calculation.

- **Conditional (Fixed X's)**

  This approach assumes that the values of the X's are known at the planning phase, usually because they are set by the researchers. This seldom happens in practice, so this approach is usually unjustifiable.

  Power and sample size calculation is based on the noncentral-F distribution. It assumes that the X's values are preset by the experimenter.

C: Independent Variables (X's) Controlled

These options refer to the independent variables that are controlled for.

kc (Number of X's Controlled)

This option specifies the number of variables in set *C*, variables that are controlled for (or partialled out). This number must be greater than or equal to zero. Note that $kc + k_T < N - 1$.

R^2(C) *(Conditional Model Only)*

This option is only shown for the conditional model. It specifies the R^2 achieved by the variables in set *C* when they are fit alone in the regression equation. Note that this amount must be between zero and one and that the total of the two R^2 values must be less than one.

T: Independent Variables (X's) Tested

These options refer to the independent variables that are being tested for statistical significance.

k_T (Number of Independent Variables Tested)

This option specifies the number of X's in the set *T*, variables that are tested. This number must be greater than or equal to one. Note that $kc + k_T < N - 1$.

$\rho 0^2$ (Null) *(Unconditional Model Only)*

Enter one or more values of $\rho 0^2$, the *squared multiple correlation coefficient* used in the null hypothesis. This is the proportion of the variation in *Y* explained by the variation in the *X's* tested. Note that ρ^2 is the population value of R^2.

If there are control *X's* specified, this value becomes the squared multiple **partial** correlation coefficient.

If $\rho 0^2 < \rho 1^2$, the hypotheses tested is H0: $\rho^2 \leq \rho 0^2$ vs H1: $\rho^2 > \rho 0^2$.

If $\rho 0^2 > \rho 1^2$, the hypotheses tested is H0: $\rho^2 \geq \rho 0^2$ vs H1: $\rho^2 < \rho 0^2$.

This parameter represents a lower bound for R^2 in that values of R^2 below this are deemed unimportant.

The range is given by $0 \leq \rho 0^2 < 1$. Cohen interpreted these values as 0.02 = Small, 0.13 = Medium, 0.26 = Large.

ρ1² (Alternative) *(Unconditional Model Only)*

Enter one or more values of ρ1², the squared multiple correlation coefficient at which the power is calculated. This is the proportion of the variation in Y explained by the variation in the X's tested. ρ² is the population value of R².

If there are control X's, this is the squared multiple **partial** correlation coefficient.

If ρ0² < ρ1², the hypotheses tested is H0: ρ²≤ρ0² vs H1: ρ²>ρ0².

If ρ0² > ρ1², the hypotheses tested is H0: ρ²≥ρ0² vs H1: ρ²<ρ0².

The range is given by 0 < ρ1² < 1. Cohen interpreted these values as 0.02 = Small, 0.13 = Medium, 0.26 = Large.

$R^2(T|C)$ *(Conditional Model Only)*

This box specifies the increase in R^2 due to the variables in set *T* after including the variables in set *C* in the regression equation. Note that this amount must be between zero and one and that the total of the two R^2 values must be less than one.

Example 1 – Finding Sample Size in the Unconditional Model

Suppose researchers are planning a multiple regression study to look at the significance of a specific independent variable. The data come from a survey that includes four other continuous demographic variables.

They want a sample size large enough to detect an R^2 of at least 0.20 if it exists. Thus, the null and alternative hypotheses that they want to consider is H0: $\rho^2 \leq 0.2$ vs H1: $\rho^2 > 0.2$. They want to compute the power when the actual value of ρ^2 is between 0.25 and 0.40.

They want to consider power values of either 0.8 or 0.9 and a significance level is 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Regression** procedure. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Power	**0.8 0.9**
Alpha	**0.05**
Regression Model Type	**Unconditional (Random X's)**
kc	**4**
k_T	**1**
$\rho 0^2$	**0.2**
$\rho 1^2$	**0.25 0.3 0.35 0.4**

Annotated Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results
Model: Unconditional (Random X's)

Power	N	Number Controlled X's kc	Number Tested X's k_T	Squared Multiple Partial Correlation Coefficient under H0 $\rho 0^2$	Squared Multiple Partial Correlation Coefficient under H1 $\rho 1^2$	Alpha	Beta
0.8002	1340	4	1	0.200	0.250	0.050	0.1998
0.8003	351	4	1	0.200	0.300	0.050	0.1997
0.8017	163	4	1	0.200	0.350	0.050	0.1983
0.8019	95	4	1	0.200	0.400	0.050	0.1981
0.9000	1853	4	1	0.200	0.250	0.050	0.1000
0.9002	484	4	1	0.200	0.300	0.050	0.0998
0.9007	223	4	1	0.200	0.350	0.050	0.0993
0.9008	129	4	1	0.200	0.400	0.050	0.0992

References

Gatsonis, C. and Sampson, A.R. 1989. 'Multiple Correlation: Exact Power and Sample Size Calculations.' Psychological Bulletin, Vol. 106, No. 3, Pages 516-524.

Benton, D. and Krishnamoorthy, K. 2003. 'Computing discrete mistures of continuous distributions: noncentral chisquare, noncentral t and the distribution of the square of the sample multiple correlation coefficient.' Computational Statistics & Data Analysis, Vol. 43, Pages 249-267.

Multiple Regression

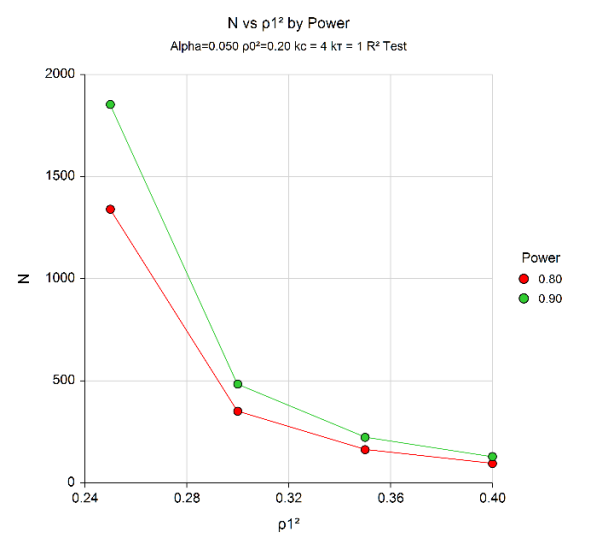
Krishnamoorthy, K. and Xia, Y. 2008. 'Sample Size Calculation for Estimating or Testing a Nonzero Squared Multiple Correlation Coefficient.' Multivariate Behavioral Research, Vol. 43, Pages 382-410.
Cohen, Jacob. 1988. Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

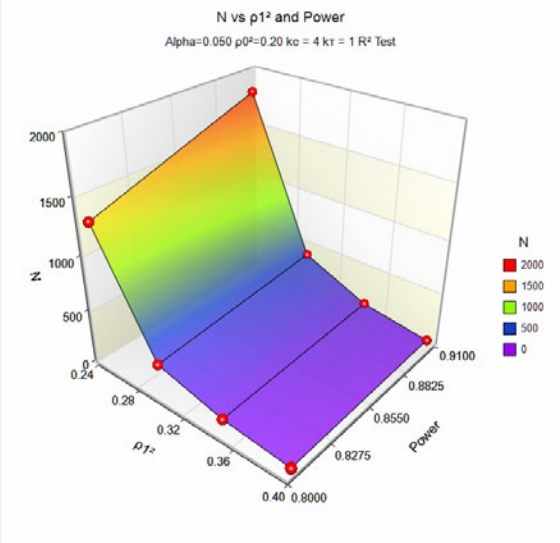
Report Definitions
Hypotheses: H0: $\rho^2 \le \rho 0^2$ versus H1: $\rho^2 > \rho 0^2$ if $\rho 0^2 < \rho 1^2$ or H0: $\rho^2 \ge \rho 0^2$ versus H1: $\rho^2 < \rho 0^2$ if $\rho 0^2 > \rho 1^2$.
Power is the probability of rejecting a false null hypothesis.
N is the number of observations on which the multiple regression is computed.
kc is the number of independent variables controlled.
kτ is the number of independent variables tested.
$\rho 0^2$ is the squared multiple correlation coefficient assumed by the null hypothesis.
$\rho 1^2$ is the squared multiple correlation coefficient at which the power is computed.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.

Summary Statements
A sample size of 1340 achieves 80% power to detect a ρ^2 of at least 0.200 attributed to 1 independent variable(s) when the significance level (alpha) is 0.050 and the actual value of ρ^2 is 0.250. The influence of an additional 4 independent variable(s) was removed.

This report shows the necessary sample sizes. The definitions of each of the columns is given in the Report Definitions section.

Plots Section





These plots show the relationship between sample size, effect size, and power.

Example 2 – Validation using an Unconditional Model

We will validate this procedure using an analysis published in Shieh and Kung (2007). In this example, the desired power is 0.90, alpha is 0.05, k_C is 0, k_T is 5, $\rho 0^2$ is 0.2, and $\rho 1^2$ is 0.05. They calculate a sample size of 153.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Regression** procedure. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Power	**0.9**
Alpha	**0.05**
kc	**0**
k_T	**5**
$\rho 0^2$	**0.2**
$\rho 1^2$	**0.25 0.3 0.35 0.4**

Annotated Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results
Model: Unconditional (Random X's)

Power	N	Number Controlled X's kc	Number Tested X's k_T	Squared Multiple Correlation Coefficient under H0 $\rho 0^2$	Squared Multiple Correlation Coefficient under H1 $\rho 1^2$	Alpha	Beta
0.9011	153	0	5	0.200	0.050	0.050	0.0989

PASS has also calculated the required sample size to be 153.

Example 3 – Testing the Addition or Deletion of a Single Variable in a Conditional Model

This example calculates the power of an F test constructed to test a fifth variable which adds 0.05 to R^2 after considering four other variables whose combined R^2 value is 0.50. Sample sizes from 10 to 150 will be investigated. The significance level is 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Regression** procedure window. You may then make the appropriate entries as listed below, or open **Example 3** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Alpha	**0.05**
N (Sample Size)	**10 to 150 by 20**
Regression Model Tested	**Conditional (Fixed X's)**
kc	**4**
$R^2(C)$	**0.50**
k_T	**1**
$R^2(T\|C)$	**0.05**

Annotated Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results
Model: Conditional (Fixed X's)

		Independent Variable(s) Controlled		Independent Variable(s) Tested			
Power	**N**	**kc**	**$R^2(C)$**	**k_T**	**$R^2(T\|C)$**	**Alpha**	**Beta**
0.1304	10	4	0.500	1	0.050	0.050	0.8696
0.4180	30	4	0.500	1	0.050	0.050	0.5820
0.6351	50	4	0.500	1	0.050	0.050	0.3649
0.7843	70	4	0.500	1	0.050	0.050	0.2157
0.8782	90	4	0.500	1	0.050	0.050	0.1218
0.9337	110	4	0.500	1	0.050	0.050	0.0663
0.9649	130	4	0.500	1	0.050	0.050	0.0351
0.9819	150	4	0.500	1	0.050	0.050	0.0181

References
Cohen, Jacob. 1988. Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, Hillsdale, New Jersey.
Gatsonis, C. and Sampson, A.R. 1989. 'Multiple Correlation: Exact Power and Sample Size Calculations.' Psychological Bulletin, Vol. 106, No. 3, Pages 516-524.

Report Definitions
Power is the probability of rejecting a false null hypothesis.
N is the number of observations on which the multiple regression is computed.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.
kc is the number of independent variables controlled.
k_T is the number of independent variables tested.

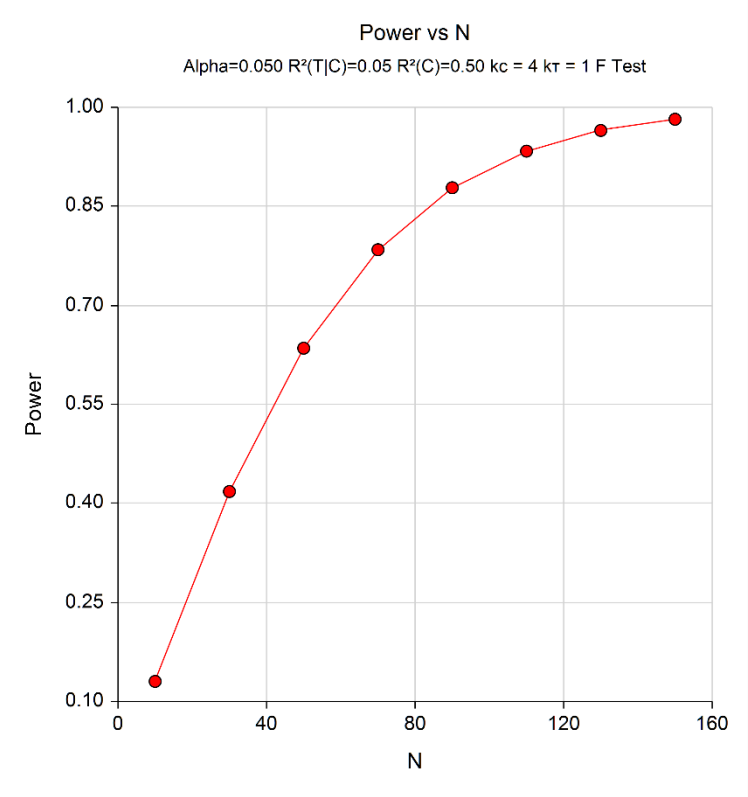
$R^2(C)$ is the R^2 achieved when only the control variables are included in the model.
$R^2(T|C)$ is the amount that R^2 is increased when the test variables are added to a model that contains the control variables.

Summary Statements
A sample size of 10 achieves 23% power to detect an R^2 of 0.100 attributed to 1 independent variable(s) using an F-Test with a significance level (alpha) of 0.050. The variables tested are adjusted for an additional 4 independent variable(s) with an R^2 of 0.500.

This report shows the values of each of the parameters, one scenario per row. The definitions of each of the columns is given in the Report Definitions section.

Note that in this particular example, a power of 0.90 is not reached until the sample size is about 110.

Plots Section



This plot shows the relationship between sample size and power.

Example 4 – Minimum Detectable R^2

Suppose the researchers in Example 3 can only afford a sample size of 30. They want to know the minimum detectable R^2 that can be detected if the power is 80% and 90%.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Regression** procedure window. You may then make the appropriate entries as listed below, or open **Example 4** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Effect Size (ρ1² or R²(T|C))**
Power	**0.8 0.9**
Alpha	**0.05**
N (Sample Size)	**30**
Regression Model Tested	**Conditional (Fixed X's)**
kc	**4**
R^2(C)	**0.50**
k_T	**1**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results
Model: Conditional (Fixed X's)

		Independent Variable(s) Controlled		Independent Variable(s) Tested			
Power	**N**	**kc**	**R²(C)**	**kT**	**R²(T|C)**	**Alpha**	**Beta**
0.8000	30	4	0.500	1	0.111	0.050	0.2000
0.9000	30	4	0.500	1	0.138	0.050	0.1000

This report shows that at 90% power, a sample size of 30 cannot detect an R^2(T|C) less than 0.138.

Example 5 – Validation using a Conditional Model

Ralph O'Brien, in a private communication to Jerry Hintze, gave the result that when Alpha = 0.05, $N = 15$, $k_T = 2$, and $R^2(T|C) = 0.6$, the power is 0.9683.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Regression** procedure window. You may then make the appropriate entries as listed below, or open **Example 5** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Alpha	**0.05**
N (Sample Size)	**15**
Regression Model Tested	**Conditional (Fixed X's)**
kc	**0**
k_T	**2**
$R^2(T|C)$	**0.6**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results
Model: Conditional (Fixed X's)

		Independent Variable(s) Controlled		Independent Variable(s) Tested			
Power	N	kc	$R^2(C)$	k_T	$R^2(T|C)$	Alpha	Beta
0.9683	15	0	0.000	2	0.600	0.050	0.0317

The power of 0.9683 matches O'Brien's result.