Chapter 858

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_M 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: *conditional* and *unconditional*. Both of these approaches are available in this procedure.

Multiple Regression Model

Multiple regression uses a *linear model* to approximate the relationship between a *dependent variable* Y and one or more *independent variables* X_1, …, X_M. The theoretical model for the *i^th^* observation is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_M X_{Mi} + e_i$$

where $\beta_0, \beta_1, \dots, \beta_M$ are the unknown regression coefficients to be estimated with ordinary least squares and e_i is the error term or *residual*.

For convenience in the presentation, we will use a slightly different notation. We divide the X's into two, non-overlapping subsets which are T (tested) and C (covariates). We reorder the X's so that all of the subset T variables occur before the subset C variables. The theoretical model then becomes

$$Y_i = \beta_0 + \beta_1 T_{1i} + \cdots + \beta_K T_{Ki} + \lambda_1 C_{1i} + \cdots + \lambda_L C_{Li} + e_i$$

where $T_1 = X_1, \dots, T_K = X_K, C_1 = X_{K+1}, \dots, C_L = X_M$.

In what follows, we will present results for tests of the regression coefficients associated with the T variables $T_1, \dots, T_K$. Specifically, we will present a power analysis of tests that $\beta_1 = \beta_2 = \cdots = \beta_K = 0$. Since the subset C variables are not included in the test, but are fit in the regression analysis, the test is said to control for (adjust for or hold constant) the covariates $C_1, \dots, C_L$.

Conditional and Unconditional Power Calculation

In this section, we will define the two types of power calculation methods. Note that both methods anticipate the same test statistics.

These two methods are *conditional* and *unconditional.*

Conditional

The conditional expectation assumption is usually used to justify significance tests and so it is the approach that is we recommend.

In this method, conditional expectation of Y is estimated given the realized values of the T's and C's. Since these values are almost always impossible to know, this method suggests using a range of R^2 values.

The focus in the significance test is how much R^2 increases when test variables are added to a regression model that already contains the covariates.

Define $R^2_{T|C} = R^2_{T,C} - R^2_C$ to be the amount that R^2 increases 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 $R^2_{T,C}$ is the R^2 when *Y* is regressed on the variables in both sets.

Test Statistic in the Conditional Formula

F-tests can easily be constructed that will test whether the regression coefficients corresponding to certain subsets of *X's* are simultaneously zero while controlling for other variables. For example, Rencher (2000) shows that 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

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

where *K* is the number of variables in *T* and *L* is the number of variables in *C*.

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 $F_{K,N-K-L-1,\alpha}$ 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** uses an algorithm that gives exact answers.

Unconditional (Multivariate Normality)

In the unconditional 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* and their corresponding regression coefficients are all zero. Note that we will use ρ^2 to represent ρ_{YX}^2.

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.

Test Statistic in the Unconditional Model

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

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

Suppose the independent variables are divided into two sets: C containing *L* variables and T containing the remaining *K* = *M* – *L* 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

$$F_{K,N-M-1} = \frac{\left(\rho_{YX_T|X_C}^2\right)/K}{\left(1-\rho_{YX_T|X_C}^2\right)/(N-M-1)}$$

Cohen (1988) shows that $\rho_{YX_T|X_C}^2$ 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

$$\rho_{YX_T|X_C}^2 = \frac{R_{YX}^2 - R_{YX_C}^2}{1 - R_{YX_C}^2}$$

Calculating the Power in the Unconditional Method

In the unconditional method, the statistical hypotheses that is usually of most interest is the set

$$\mathrm{H}_0\colon \rho^2 \le \rho_0^2 \quad \text{versus} \quad \mathrm{H}_1\colon \rho^2 > \rho_0^2$$

because you want to establish a lower bound for the value, not just established that it is greater than zero. However, the hypotheses

$$\mathrm{H}_0\colon \rho^2 \ge \rho_0^2 \quad \text{versus} \quad \mathrm{H}_1\colon \rho^2 < \rho_0^2$$

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 $\mathrm{P}(R^2 \le r_\alpha | N, K, \rho_0^2) = 1 - \alpha$. Note that we use the value of ρ^2 specified in the null hypothesis.
2. Compute the power using $\text{Power} = 1 - \mathrm{P}(R^2 \le r_\alpha | N, K, \rho_1^2)$.

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

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

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 $N - L$ in the above CDF.

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.

Example 1 – Finding Sample Size in the Conditional Model

Suppose researchers are planning a multiple regression study to look at the significance of a specific independent variable. They want to assume that a conditional model will be used to analyze the data.

The data will come from a survey that includes four other continuous demographic variables.

They want a sample size large enough to detect an R^2 of between 0.1 and 0.4 in one of the four variables. They assume that the other three variables will account for an R^2 of 0.3.

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

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For	**Sample Size**
Power Calculation Method	**Conditional (Recommended) - Uses R^2**
Power	**0.8 0.9**
Alpha	**0.05**
K (Number Tested)	**1**
$R^2(T\|C) = R^2(T,C) - R^2(C)$	**0.1 0.2 0.3 0.4**
L (Number of Covariates)	**4**
$R^2(C)$	**0.3**

Output

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

Numeric Reports

```
Numeric Results
────────────────────────────────────────────────────────────────────────────────────────────
Solve For:       Sample Size
Power Method:    Conditional (Recommended)
────────────────────────────────────────────────────────────────────────────────────────────
                              Independent                Covariates (Independent
                           Variables Tested             Variables Controlled For)
                         ──────────────────────        ─────────────────────────
           Sample        Number                        Number of
             Size        Tested       R-Squared        Covariates      R-Squared
Power           N             K         R²(T|C)                 L          R²(C)      Alpha
────────────────────────────────────────────────────────────────────────────────────
0.8060         50             1             0.1                 4            0.3       0.05
0.8155         23             1             0.2                 4            0.3       0.05
0.8094         14             1             0.3                 4            0.3       0.05
0.8605         11             1             0.4                 4            0.3       0.05
0.9037         66             1             0.1                 4            0.3       0.05
0.9033         29             1             0.2                 4            0.3       0.05
0.9007         17             1             0.3                 4            0.3       0.05
0.9118         12             1             0.4                 4            0.3       0.05
────────────────────────────────────────────────────────────────────────────────────────────
Power      The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N          The number of observations on which the multiple regression is computed.
K          The number of independent variables tested for zero coefficients.
R²(T|C)    The amount that R² is increased when the test variables are added to a model that contains the control variables.
             R²(T|C) = R²(T,C) - R²(C).
L          The number of covariates (independent variables not tested for zero regression coefficients).
R²(C)      The R² achieved when only the control variables are included in the model.
Alpha      The probability of rejecting a true null hypothesis.


Summary Statements
────────────────────────────────────────────────────────────────────────────────────────────
A multiple regression (Y versus X's) design, with 1 independent variable tested and 4 control covariates, will be
used to test whether the R² of the test variable (above the R² of the control variables) is greater than 0 (H0: R²(T|C)
= 0 versus H1: R²(T|C) > 0). This corresponds to a test of whether the regression coefficient of the test variable
(given the control covariates) is different from 0. The comparison will be made using a multiple regression
full-versus-reduced-model F-test with a Type I error rate (α) of 0.05. The sample X values are assumed to be fixed
and known (the test is conditional upon known X values). The 4 control covariates are assumed to have a
combined R² of 0.3 when modeled alone. To detect an R²(T|C) of 0.1 with 80% power, the number of needed
subjects will be 50.
────────────────────────────────────────────────────────────────────────────────────────────
```

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	50	63	13
20%	23	29	6
20%	14	18	4
20%	11	14	3
20%	66	83	17
20%	29	37	8
20%	17	22	5
20%	12	15	3

Dropout Rate The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.

N The evaluable sample size at which power is computed. If N subjects are evaluated out of the N' subjects that are enrolled in the study, the design will achieve the stated power.

N' The total number of subjects that should be enrolled in the study in order to obtain N evaluable subjects, based on the assumed dropout rate. After solving for N, N' is calculated by inflating N using the formula N' = N / (1 - DR), with N' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)

D The expected number of dropouts. D = N' - N.

Dropout Summary Statements

Anticipating a 20% dropout rate, 63 subjects should be enrolled to obtain a final sample size of 50 subjects.

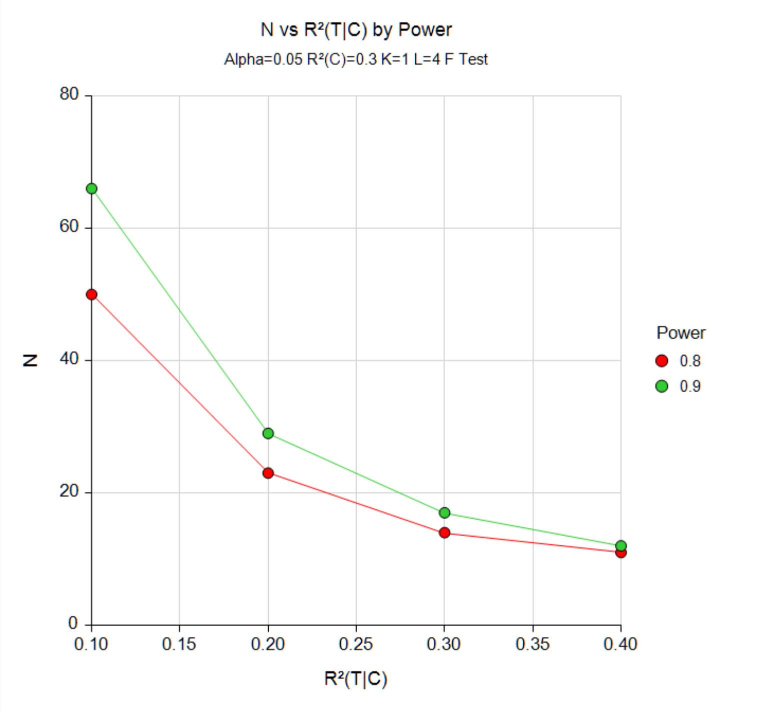
References

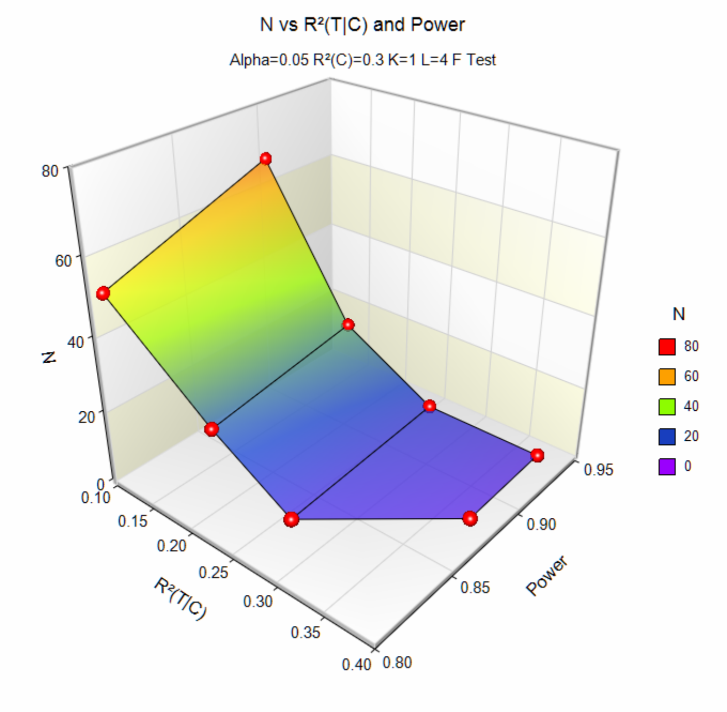
Cohen, Jacob. 1988. Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, Hillsdale, New Jersey.
Gatsonis, C. and Sampson, A.R. 1989. 'Multiple Correlation: Exact Power and Sample Size Calculations.' Psychological Bulletin, Vol. 106, No. 3, Pages 516-524.

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

Plots Section

Plots





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

Example 2 – Validation using the Unconditional Power Calculation in Shieh and Kung (2007)

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, L is 0, K is 5, $\rho0^2$ is 0.2, and $\rho1^2$ is 0.05. They calculate a sample size of 153.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

```
Design Tab
Solve For ......................................................Sample Size
Power Calculation Method .............................Unconditional (Assumes Multivariate Normality) - Uses ρ²
Power.............................................................0.90
Alpha.............................................................0.05
K (Number Tested) .......................................5
ρ0² (Null).......................................................0.2
ρ1² (Actual) ...................................................0.05
L (Number of Covariates)..............................0
```

Output

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

```
Numeric Results

Solve For:       Sample Size
Power Method:    Unconditional (Assumes Multivariate Normality)
Hypotheses:      H0: ρ² ≤ ρ0²   vs.   H1: ρ² > ρ0² if ρ0² < ρ1²    or    H0: ρ² ≥ ρ0²   vs.   H1: ρ² < ρ0² if ρ0² > ρ1²

                            Independent Variables Tested
                                        Squared
                                        Multiple
                                       Correlation
                                       Coefficient        Number of
                                                          Covariates
                                                         (Independent
          Sample       Number                             Variables
            Size       Tested       Null     Actual    Controlled For)
Power          N            K       ρ0²        ρ1²                   L     Alpha
0.9011       153            5       0.2       0.05                   0      0.05
```

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 Power Calculation

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 who's combined R^2 value is 0.5. Sample sizes from 10 to 150 will be investigated. The significance level is 0.05.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For	**Power**
Power Calculation Method	**Conditional (Recommended) - Uses R²**
Alpha	**0.05**
N (Sample Size)	**10 to 150 by 20**
K (Number Tested)	**1**
R²(T|C) = R²(T,C) - R²(C)	**0.05**
L (Number of Covariates)	**4**
R²(C)	**0.5**

Output

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

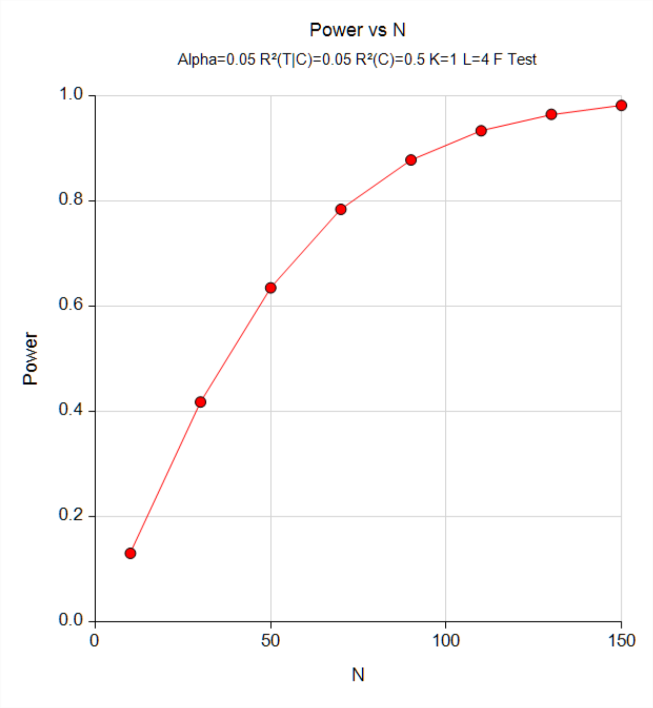
Numeric Reports

Numeric Results

Solve For: Power
Power Method: Conditional (Recommended)

		Independent Variables Tested		Covariates (Independent Variables Controlled For)		
Power	Sample Size N	Number Tested K	R-Squared $R^2(T\mid C)$	Number of Covariates L	R-Squared $R^2(C)$	Alpha
0.1304	10	1	0.05	4	0.5	0.05
0.4180	30	1	0.05	4	0.5	0.05
0.6351	50	1	0.05	4	0.5	0.05
0.7843	70	1	0.05	4	0.5	0.05
0.8782	90	1	0.05	4	0.5	0.05
0.9337	110	1	0.05	4	0.5	0.05
0.9649	130	1	0.05	4	0.5	0.05
0.9819	150	1	0.05	4	0.5	0.05

Plots



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.

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 4** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

```
Design Tab
Solve For ......................................................Effect Size (ρ1² or R²(T|C))
Power Calculation Method ............................Conditional (Recommended) - Uses R²
Power............................................................0.8 0.9
Alpha............................................................0.05
N (Sample Size)............................................30
K (Number Tested) .......................................1
L (Number of Covariates)..............................4
R²(C) ............................................................0.5
```

Output

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

```
Numeric Results

Solve For:       Effect Size (ρ1² or R²(T|C))
Power Method:    Conditional (Recommended)

                           Independent            Covariates (Independent
                        Variables Tested          Variables Controlled For)

           Sample    Number                       Number of
             Size    Tested      R-Squared       Covariates       R-Squared
Power           N         K        R²(T|C)                L           R²(C)     Alpha
0.8            30         1          0.111                4             0.5      0.05
0.9            30         1          0.138                4             0.5      0.05
```

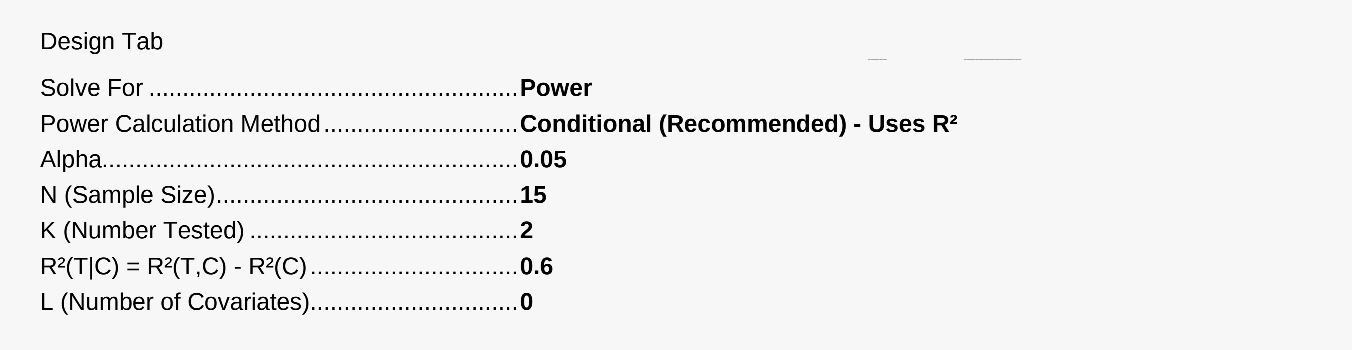
This report shows that at 90% power, a sample size of 30 cannot detect an R^2(T|C) less than 0.138.

Example 5 – Validation using a Conditional Power Calculation

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

Setup

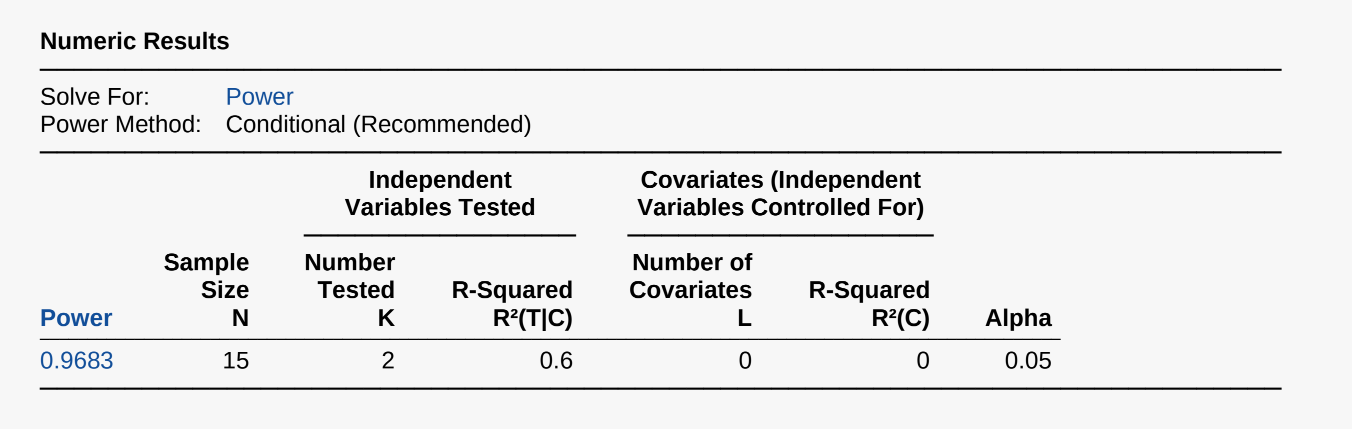
If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 5** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.



```
Design Tab
Solve For ......................................................Power
Power Calculation Method............................Conditional (Recommended) - Uses R²
Alpha...........................................................0.05
N (Sample Size)............................................15
K (Number Tested) .......................................2
R²(T|C) = R²(T,C) - R²(C)..............................0.6
L (Number of Covariates)..............................0
```

Output

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



```
Numeric Results

Solve For:        Power
Power Method:     Conditional (Recommended)

                            Independent              Covariates (Independent
                         Variables Tested           Variables Controlled For)

            Sample      Number                       Number of
              Size      Tested      R-Squared       Covariates      R-Squared
Power            N           K        R²(T|C)                L          R²(C)      Alpha
0.9683          15           2            0.6                0              0       0.05
```

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