METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

The A Priori Procedure (APP) for Estimating Regression Coefficients in Linear Models

Tingting Tong [1], David Trafimow [2], Tonghui Wang [1], Cong Wang [3], Liqun Hu [1], Xiangfei Chen [1]

[1] *Department of Mathematical Sciences, New Mexico State University, Las Cruces, NM, USA.* **[2]** *Department of Psychology, New Mexico State University, Las Cruces, NM, USA.* **[3]** *Department of Mathematics, University of Nebraska, Omaha, NE, USA.*

Methodology, 2022, Vol. 18(3), 203–220, https://doi.org/10.5964/meth.8245

Received: 2022-01-27 • **Accepted:** 2022-09-02 • **Published (VoR):** 2022-09-30

Handling Editor: Francisco J. Abad, Universidad Autónoma de Madrid, Madrid, Spain

Corresponding Author: David Trafimow, SH 337, Department of Psychology, MSC 3452, New Mexico State University, Las Cruces, 88003-8001, NM, USA. E-mail: dtrafimo@nmsu.edu

Abstract

Regression coefficients are crucial in the sciences, as researchers use them to determine which independent variables best explain the dependent variable. However, researchers obtain regression coefficients from data samples and wish to generalize to populations; without reason to believe that sample regression coefficients are good estimates of corresponding population regression coefficients, their usefulness would be curtailed. In turn, larger sample sizes provide better estimates than do smaller ones. There is much recent literature on the a priori procedure (APP) that was designed for the general purpose of determining the sample sizes needed to obtain sample statistics that are good estimates of corresponding population parameters. We provide an extension of the APP to regression coefficients, which works for standardized or unstandardized regression coefficients. A simulation study and real data example support the mathematical derivations. Also, we include a free and user-friendly computer program to aid researchers in making the calculations.

Keywords

regression coefficients, a priori procedure, sample size

For over a century, researchers have performed multiple regression analyses to obtain regression coefficients that, in turn, provided valuable information about the ability of independent variables to explain dependent variables. However, until the present

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, CC BY 4.0, which permits unrestricted use, distribution, and reproduction, provided the original work is properly cited.

work, sample size determination has been left to power analyses (e.g., Cohen, 2013). From the point of view of determining sample sizes needed to render good chances of obtaining statistical significance, power analyses make sense. But from the point of view of determining sample sizes needed so that sample statistics can be trusted to estimate corresponding population parameters, power analysis is insufficient (Trafimow et al., 2020; Trafimow & Myüz, 2019). This is because power analysis is a function not only of the sample size, but of the expected effect size too. For example, consider the simple case of a single mean under the typical prescription in psychology that researchers should attempt to detect a 'medium' effect size (e.g., Cohen's $d = 0.50$) with a 0.80 probability of rejecting the null hypothesis at alpha equals 0.05. In that case only 31 participants are required. However, that sample size implies that the researcher has a probability of 0.95 of obtaining a sample mean that is within 0.35 standard deviations of the population mean it is intended to estimate (Trafimow et al., 2020), which many would consider insufficiently precise. Trafimow (2018) recommended having a probability of 0.95 of having sample statistics be within 0.20 or 0.10 of corresponding population parameters for 'good' or 'excellent' precision, respectively, though also indicating that such designations could change depending on study contexts. Clearly, an alternative to power analysis is desirable, and the a priori procedure (APP), to be explained presently, provides it. The present goal is to expand the APP so that researchers can determine the sample sizes necessary to meet their requirements for obtaining sample regression coefficients that provide good estimates of corresponding population regression coefficients.

To set up the present work, it is useful to briefly consider the issue of regression coefficient size in the context of research that is exploratory or that is beyond exploratory, keeping in mind that the definitions of 'small' or 'large' regression coefficients depend on substantive areas and researcher goals. In exploratory research, variables with larger regression coefficients are typically considered better candidates for future investigation than are independent variables with smaller regression coefficients. In research that is beyond the initial exploration phase, independent variables with larger regression coefficients are typically considered better candidates for intervention, policy recommendations, or theoretical explanation than are independent variables with smaller regression coefficients. This is because policy makers must have reason not only to believe in the empirical relationship, but also that the relationship is sufficiently large to justify the costs of implementing a policy (Trafimow & Osman, 2022). Even for basic research, large regression coefficients are less susceptible to trivial alternative explanations than are smaller ones, all else being equal. Whether the research is at the exploratory level, or beyond that, researchers must have some reason to believe that the regression coefficients they obtain from their sample generalize to the population of interest; it is necessary to assume that sample regression coefficients are good estimates of corresponding population regression coefficients. Most researchers are aware that, in general, the larger the sample size, the better the estimation. However, at present,

PsychOpen GOLD

there is no way to know the minimum sample size needed to meet criteria for quality of estimation. Our goal is to derive a procedure to accomplish this. The procedure to be described works equally well for those researchers who prefer unstandardized regression coefficients to standardized ones.

Recently, a general methodology—the APP—has been developed for determining required sample sizes so that sample statistics provide good estimates of corresponding population parameters. There are two main criteria that are bullet-listed below:

- **Precision**: Researchers must specify the distance within which they wish their sample statistics to be of corresponding population parameters.
- **Confidence**: Researchers must specify the probability they wish to have of meeting the precision specification.

For example, in the case of a single mean, under normality, Trafimow (2017) showed that it is necessary to obtain a sample size of 385 to be 95% confident that the sample mean will be within one-tenth of a standard deviation of the population mean. Note the contrast between a sample size of 385 participants versus 31 participants sufficient to satisfy a typical power analysis. Even dropping the criterion to a precision level of 0.20 implies a sample size of 97, which still exceeds 31, thereby exemplifying that the APP is very different from power analysis.

Recent APP work has gone well beyond single means under normality. For example, Trafimow et al. (2019) extended the APP to work for skew normal distributions; Wang et al. (2019a) extended the APP to work for comparisons between independent groups under skew normal distributions; and Wang et al. (2019c) provided an APP extension to two dependent groups (matched data). Moreover, Wang et al. (2020) extended the APP to one-way analysis of variance paradigms. Then, too, there are APP extensions pertaining to standard deviations or scales (e.g., Wang et al., 2022), distribution shapes (Wang et al., 2019b), and correlation coefficients (Wang et al., 2021). There is even a Bayesian APP extension for estimating the normal mean (Wei et al., 2020) and proportion based on skew normal approximations and the Beta-Bernoulli process (Cao et al., 2021). The foregoing citations indicate that a sizable APP literature already exists and that it is growing quickly.

The present goal is to add to the APP literature by extending that literature to address regression coefficients assuming a multivariate normal distribution. Our aim is to derive a procedure by which researchers can specify the desired degree of precision and confidence, as well as the number of independent variables, to determine the minimum sample sizes needed to meet the specifications. In other words, the work to be presented provides a way for researchers to determine the sample size necessary so that they can trust that their sample regression coefficients are good estimates of corresponding population regression coefficients.

PsychOpen GOLD

Review of Multiple Linear Regression Model

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent (response) variable and one or more independent (explanatory) variables. The most common form of regression analysis is linear regression that most closely fits the data according to a specific mathematical criterion.

The multiple linear regression model for n observations can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

which is equivalent to the matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathscr{E} \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$ is the vector of dependent variable,

$$\mathbf{X} \equiv (\boldsymbol{j}_n, \mathbf{X}_1) = (\boldsymbol{j}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p)$$

is $n \times (p+1)$ design matrix with $\boldsymbol{j}_n = (1, \ldots, 1)'$ in $\Re^n$ and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})'$ is the vector of observations for jth variable for $j = 1, \ldots, p$. The random vector $\mathscr{E} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is the error term in the model and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients. For estimating $\boldsymbol{\beta}$, we need general assumptions that the mean and the covariance matrix of $\mathscr{E}$ are, respectively, given by $E(\mathscr{E}) = \mathbf{0}$ and $Cov(\mathscr{E}) = \sigma^2\mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of order n and $\sigma > 0$. Note that the model (1) can be written in centered form:

$$\mathbf{y} = (\boldsymbol{j}_n, \mathbf{X}_c)\begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \mathscr{E} \tag{2}$$

where $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_p)'$, $\alpha = \beta_0 + \beta_1\bar{x}_1 + \beta_2\bar{x}_2 + \cdots + \beta_k\bar{x}_p$, and

$$\mathbf{X}_c = (\mathbf{I}_n - \bar{\mathbf{J}}_n)\mathbf{X}_1, \qquad \bar{\mathbf{J}}_n = \frac{1}{n}\boldsymbol{j}_n\boldsymbol{j}_n'$$

The following lemma will be used in constructing our APP method for estimating $\boldsymbol{\beta}$ and its proof is given in Rencher and Schaalje (2008).

Lemma 1

Consider the regression model given in (1) and (2) and assume that $\mathbf{y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 I_n)$, *the n-dimensional multivariate normal distribution with mean* $\boldsymbol{X\beta}$ *and covariance matrix* $\sigma^2\boldsymbol{I}_n$. *Then the maximum likelihood estimators of* α, $\boldsymbol{\beta}_1$, *and* σ^2 *are, respectively, given by*

$$\hat{\alpha} = \bar{y}, \quad \hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y}, \quad \hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{3}$$

PsychOpen GOLD

where $\bar{y} = \mathbf{j}_n'\mathbf{y}/n$. *Furthermore, these estimators have the following properties*:

a. $\hat{\boldsymbol{\beta}}_1 \sim N_p\left(\boldsymbol{\beta}_1, \sigma^2(\mathbf{X_c}'\mathbf{X_c})^{-1}\right)$.
b. $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p-1}$, *the chi-square distribution with* $n-p-1$ *degrees of freedom.*
c. $\hat{\boldsymbol{\beta}}_1$ *and* $\hat{\sigma}^2$ *are independent.*

Let $\mathbf{S}_{xx} = (s_{ij})$ be the sample covariance matrix of $(\mathbf{x_1}, ..., \mathbf{x_p})'$ and $\mathbf{s}_{yx} = (s_{y1}, ..., s_{yp})'$ be the vector of sample covariances between $\mathbf{y}$ and the $\mathbf{x}_j$'s, where s_{yj}'s are sample covariance between y and x_j for $j = 1, ..., p$. It can be shown that

$$\mathbf{X_c}'\mathbf{X_c} = (n-1)\mathbf{S}_{xx}, \quad \mathbf{X_c}'\mathbf{y} = (n-1)\mathbf{s}_{yx}, \quad \text{and} \quad \hat{\boldsymbol{\beta}}_1 = \mathbf{S}_{xx}^{-1}\mathbf{s}_{yx} \tag{4}$$

We can also express the vector of regression coefficients $\hat{\boldsymbol{\beta}}_1$ in terms of sample correlations. Let $\mathbf{R}$ be the sample correlation matrix between y and $x_1, ..., x_p$'s and $\mathbf{S}$ be its corresponding sample covariance matrix as

$$\mathbf{R} = \begin{pmatrix} 1 & \boldsymbol{r}_{yx}' \\ \boldsymbol{r}_{yx} & \mathbf{R}_{xx} \end{pmatrix} \quad \text{and} \quad \mathbf{S} = \begin{pmatrix} s_y^2 & \boldsymbol{s}_{yx}' \\ \boldsymbol{s}_{yx} & \mathbf{S}_{xx} \end{pmatrix}$$

where $\boldsymbol{r}_{yx}$ is the vector of correlations between y and x_j's and $\mathbf{R}_{xx}$ is the correlation matrix for the $x_1, ..., x_p$, and s_y^2 is the sample variance of $y_1, y_2, ..., y_n$. Let

$$\mathbf{D} = [\text{diag}(\mathbf{S})]^{1/2} \equiv \text{diag}\left(s_y, s_1, ..., s_p\right) \quad \text{and} \quad \mathbf{D}_x = \text{diag}\left(s_1, ..., s_p\right)$$

Then it is easy to verify from (4) that

$$\hat{\boldsymbol{\beta}}_1 = s_y\mathbf{D}_x^{-1}\mathbf{R}_{xx}^{-1}\boldsymbol{r}_{yx} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_1^* = (\hat{\beta}_1^*, ..., \hat{\beta}_p^*)' = \frac{1}{s_y}\mathbf{D}_x\hat{\boldsymbol{\beta}}_1 \tag{5}$$

where $\hat{\boldsymbol{\beta}}_1^*$ is the estimator of the vector of standardized coefficient $\boldsymbol{\beta}_1^* = (\frac{s_1}{s_y}\beta_1, ..., \frac{s_p}{s_y}\beta_p)'$. For references of standardized regression coefficients, please see Cohen (1968), Darlington and Hayes (2016), Green (1991), and Viswesvaran (1998).

Example 1

Consider the regression model with two independent variables x_1 *and* x_2. *We have*

$$\mathbf{D}_x = \begin{pmatrix} s_1 & 0 \\ 0 & s_2 \end{pmatrix}, \quad \mathbf{S}_{xx} = \begin{pmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{pmatrix} \quad \text{and} \quad \mathbf{s}_{yx} = \begin{pmatrix} s_{y1} \\ s_{y2} \end{pmatrix}$$

By (4) *and* (5), *we obtain, after simplification, that the estimators of* $\boldsymbol{\beta}_1$ *and* $\boldsymbol{\beta}_1^*$ *are*

PsychOpen GOLD

$$\hat{\boldsymbol{\beta}}_1 = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \frac{1}{s_1^2 s_2^2 - s_{12}^2} \begin{pmatrix} s_2^2 s_{y1} - s_{12} s_{y2} \\ s_1^2 s_{y2} - s_{21} s_{y1} \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_1^* = \begin{pmatrix} \hat{\beta}_1^* \\ \hat{\beta}_2^* \end{pmatrix} = \frac{1}{1 - r_{12}^2} \begin{pmatrix} r_{y1} - r_{12} r_{y2} \\ r_{y2} - r_{12} r_{y1} \end{pmatrix}$$

respectively. □

Remark 1

Note that the $\hat{\beta}_j^*$'s can be compared to each other, whereas the $\hat{\beta}_j$'s cannot be so compared. The division by s_y is customary but not necessary. That is, the relative values of $s_1\hat{\beta}_1$ and $s_2\hat{\beta}_2$ are the same as those of $s_1\hat{\beta}_1/s_y$ and $s_2\hat{\beta}_2/s_y$. Therefore, in stead of finding the necessary sample size needed to trust standardized regression weights $\hat{\boldsymbol{\beta}}_1^*$, people look for the sample size needed for estimating $\boldsymbol{\gamma} = s_y\boldsymbol{\beta}_1^* = \mathbf{D}_x\boldsymbol{\beta}_1$, which is equivalent to estimating $\boldsymbol{\beta}_1$ as $\mathbf{D}_x$ are calculated by the observations of independent variables $X_1, ..., X_p$. Therefore we will focus on setting our APP method on the estimation of $\boldsymbol{\beta}_1$.

The Necessary Sample Size Needed for Estimating the Vector of Regression Coefficients

In this section, we will establish the APP for estimating $\boldsymbol{\beta}_1$, vector of the regression coefficients given in (2) under normal assumption.

Theorem 1

Assume that $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, *where* $\mathbf{X}$ *is* $n \times (p+1)$ *of rank* $p+1$ *and* $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)' = (\beta_0, \boldsymbol{\beta}_1')'$. *Then for specified precision* f *and confidence level* c, *the necessary sample size needed for estimating regression coefficients* $\boldsymbol{\beta}_1$ *(or regression weights) can be obtained by solving the following equation*

$$\int_0^{(n-1)f^2} f_U(u)du = c \tag{6}$$

where

$$U \equiv \frac{\left\| \mathbf{S}_{xx}^{1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \right\|^2}{ps^2/(n-1)} \sim \mathrm{F}_{p, n-p-1} \tag{7}$$

the F-distribution with numerator degrees of freedom p *and denominator degrees of freedom* $n-p-1$ *and* $f_U(u)$ *is the density of* U.

The proof of Theorem 1, together with the density of U is given in the Appendix.

PsychOpen GOLD

Remark 2

Note that Theorem 1 still hold for finding the necessary sample size needed to trust $\hat{\boldsymbol{\gamma}}$ since

$$\mathbf{R}^{1/2}\left(\hat{\boldsymbol{\gamma}} - E(\hat{\boldsymbol{\gamma}})\right) = \mathbf{R}^{1/2}(\mathbf{D}_x\hat{\boldsymbol{\beta}}_1 - \mathbf{D}_x\boldsymbol{\beta}_1) = \mathbf{R}^{1/2}\mathbf{D}_x(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) = \mathbf{S}_{xx}^{1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)$$

Therefore, the necessary sample size, n, for estimating $\hat{\boldsymbol{\beta}}_1^*$, the vector of standardized regression weights will be same as that for estimating $\boldsymbol{\beta}_1$.

Remark 3

If the previous data sets are available, we can use them to obtain $\boldsymbol{\beta}_{10}$ so that the random variable U given in (10) has a noncentral F distribution with non-centrality parameter

$$\lambda_0 = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})'\mathbf{S}_{xx}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})$$

and degrees of freedoms p and $n - p - 1$, denoted by $U \sim F_{p,n-p-1}(\lambda_0)$. The density function of $W \equiv U \sim F_{k_1,k_2}(\lambda)$ is given in the Appendix. Therefore the required sample size n can be obtained by specifying an extra λ_0 with f_U given in (6) replaced by f_W given above. □

The density curves of U with different values of p and n are listed in Figures 1 and 2. In Figure 1, the density curves of U are given for $p = 2$ and different values of $n - p - 1 = 27, 47, 97, 297$, which do not change much as n increases. In Figure 2, the density curves of U are graphed for $n = 50$ and different values of $p = 2, 3, 5, 10$, which change substantially when p changes.

PsychOpen GOLD

Figure 1

The Density Curves of U for $p = 2$ and $n = 30, 50, 100, 300$

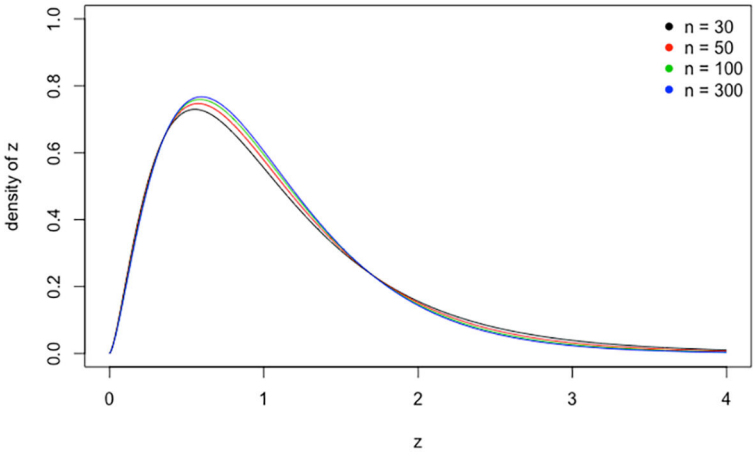
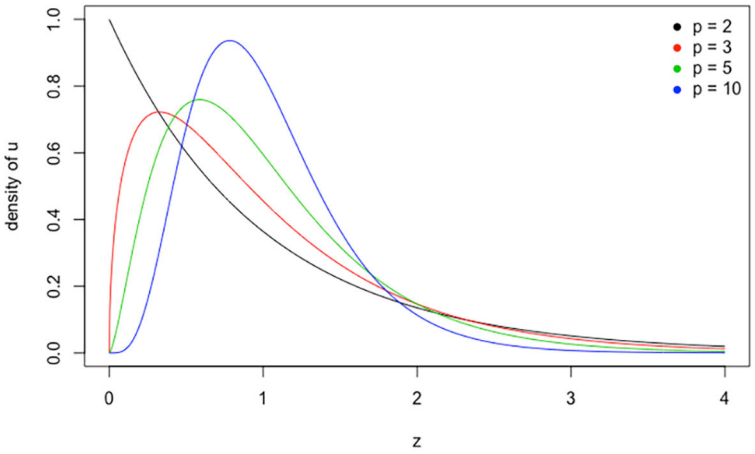


Figure 2

The Density Curves of U for $n = 50$ and $p = 2, 3, 5, 10$

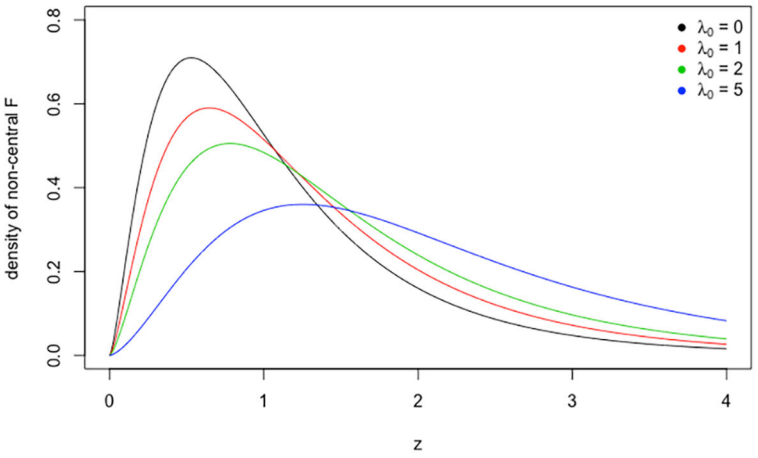


If the non-central parameter λ_0 is not 0, the density curves for $p = 5$, $n = 50$, and different values of $\lambda_0 = 0, 1, 2, 5$ are listed in Figure 3. We can see that the non-central parameter λ_0 do effect density curves.

PsychOpen GOLD

Figure 3

The Density Curves of Non-Central F-Distribution for $p = 5, n - p - 1 = 15$ *and Non-Central Parameter* $\lambda_0 = 0, 1, 2, 5$



Remark 4

From Theorem 1, we can construct a confidence region for $\boldsymbol{\beta}_1$ or $\boldsymbol{\gamma}$ with give confidence level $c100\%$ and precision f, the confidence region for $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}$ are given by

$$\mathscr{C}_{\boldsymbol{\beta}_1}(c, f) = \left\{\boldsymbol{\beta}_1 : \left\| \mathrm{S}_{xx}^{1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \right\| \leq \sqrt{p} f \sigma \right\} \tag{8}$$

and

$$\mathscr{C}_{\boldsymbol{\gamma}}(c, f) = \left\{\boldsymbol{\gamma} : \left\| \mathrm{R}_{xx}^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \right\| \leq \sqrt{p} f \sigma \right\} \tag{9}$$

respectively.

To illustrate the above results for case where $p = 2, c = 0.95$, the confidence regions of $\boldsymbol{\beta}_1 = (-4, 2.5)'$ with $n = 308, 138$ and $f = 0.1, 0.15$ are given in Figures 4 and 5, respectively. Here data observations of x_1 and x_2 are generated from the uniform distribution with mean 0 and variance 1, and with true values $(\beta_0, \beta_1, \beta_2) = (2, -4, 2.5)$.

PsychOpen GOLD

Figure 4

Confidence Regions of $\boldsymbol{\beta}_1 = (\beta_1, \beta_2)' = (-4, 2.5)'$ *(Green Point) Enclosed by Blue Dashed Line for* $C = 0.95$ *and* $f = 0.1, 0.15$ *From Left to Right and the Corresponding Point Estimators (Red Points)*

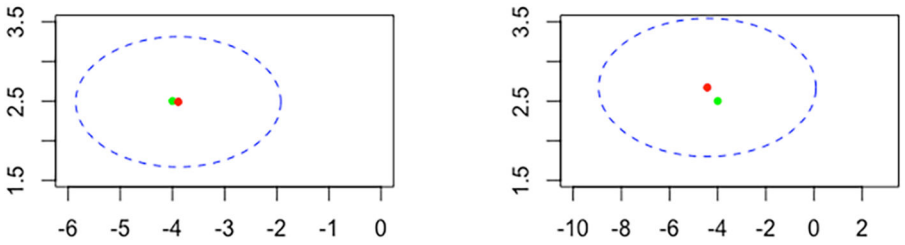
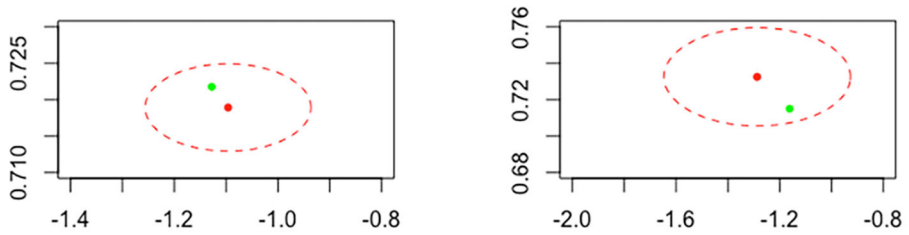


Figure 5

Confidence Regions of $\boldsymbol{\gamma} = (-1.128, 0.721)'$ *(Left Green Point) and* $\boldsymbol{\gamma} = (-1.162, 0.715)'$ *(Right Green Point) Enclosed by Red Dashed Line for* $C = 0.95$ *and* $f = 0.1, 0.15$ *From Left to Right and the Corresponding Point Estimates (Red Points)*



Simulation Study and Real Data Example

In this section, we conduct a simulation study and present a real data analysis to evaluate the performance of the APP proposed above. The necessary sample sizes (n) obtained by using Theorem 1 are provided in Tables 1, 2, 3, 4 for different values of p when $f = 0.1, 0.15, 0.2, 0.25$ and $c = 0.9, 0.95$.

Table 1

The Necessary Sample Sizes and Coverage Rates for p = 2

f	0.1		0.15		0.2		0.25	
c	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9
n	308	244	138	114	84	62	58	49
cr	0.95027	0.89956	0.95006	0.89995	0.95005	0.89968	0.94963	0.90019

PsychOpen GOLD

Table 2

The Necessary Sample Sizes and Coverage Rates for p = 3

f	0.1		0.15		0.2		0.25	
c	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9
n	275	222	129	104	74	60	49	38
cr	0.94976	0.90035	0.94999	0.90042	0.94974	0.90039	0.94991	0.90003

Table 3

The Necessary Sample Sizes and Coverage Rates for p = 5

f	0.1		0.15		0.2		0.25	
c	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9
n	229	195	113	96	71	51	43	38
cr	0.95009	0.90000	0.94962	0.90004	0.95003	0.89946	0.94992	0.90030

Table 4

The Necessary Sample Sizes and Coverage rates for p = 10

f	0.1		0.15		0.2		0.25	
c	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9
n	192	172	98	78	58	46	38	35
cr	0.95009	0.89961	0.94995	0.89959	0.94991	0.89957	0.94971	0.89984

The tables indicate the following. First, the required sample size n decreases as values of precision f increase for all $p = 2, 3, 5$, and 10. Second, with $M = 100,000$ runs (samples) for required sample size n, the coverage rates (the percentage of the constructed confidence intervals that include the true parameters) are very close to the corresponding confidence levels $c = 0.9, 0.95$. Third, as the number p of independent variables increases, the required sample size n decreases for fixed f and c. It is reasonable since the multiple correlation coefficient R^2 is increased as p increases so that the sample size decreases as p increases. Fourth, the effect of increasing the number p of independent variables is smaller for low precision setting (e.g., $f = 0.25$: $p = 2$, $c = 0.9$, n = 49; $f = 0.25$, $p = 10$,

PsychOpen GOLD

$c = 0.9$, $n = 35$) than for high-precision setting (e.g., $f = 0.10$: $p = 2$, $c = 0.9$, $n = 224$; $f = 0.10$, $p = 10$, $c = 0.9$, $n = 172$).

For calculating the necessary sample sizes needed to estimate the regression coefficients $\boldsymbol{\beta}_1$, a freely available online calculator can be found at the Supplementary Materials.

Introduction to the Link

To use the program for finding the sample size needed to estimate the regression coefficients, it is necessary to make three entries. In the first box, type in the number (p) of independent variables included in the model. In the second box, type in the desired degree of precision (f). In the third box, type in the desired confidence level (c). The last input is the noncentrality parameter (λ_0), which can be determined by using previous data. The default value of λ_0 is 0. Then click "update" to obtain the sample size needed to meet your specifications for precision and confidence.

Example 2

The data set was obtained from the R Package named *datarium* (Kassambara, 2019). The data sets list the impact of three advertising media (Youtube, Facebook and newspaper) on sales. Data are the advertising budget in thousands of dollars along with the sales. The advertising experiment has been repeated 200 times. Now, we construct a regression model to predict sales (y) on the basis of advertising budget spent in Youtube media (x_1) and newspaper (x_2). By the online calculator provided in the above, we obtain the necessary sample size needed for estimating the standardized regression weights is 138 with precision $f = 0.15$, confidence level $c = 0.95$. So we randomly choose a sample of size $n = 138$ from the row data. After calculation, the least-squares estimate of $\boldsymbol{\beta}_1$ in equation (2) and $\boldsymbol{\gamma}$ are $\hat{\boldsymbol{\beta}}_1 = (0.04680, 0.04472)'$ and $\hat{\boldsymbol{\gamma}} = (4.74557, 1.16442)'$, respectively. Also the estimate of the standardized regression weights is $\hat{\boldsymbol{\beta}}_1^*$=$(0.76823, 0.18850)'$. If we use the whole data set as a sample ($n = 200$), the estimates of $\boldsymbol{\beta}_1$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_1^*$ are $\hat{\boldsymbol{\beta}}_{10} = (0.04690, 0.04422)'$, $\hat{\boldsymbol{\gamma}}_0 = (4.83200, 1.15565)'$ and $\hat{\boldsymbol{\beta}}_{10}^* = (0.77177, 0.18458)'$, respectively. For comparison, the difference between $\hat{\boldsymbol{\beta}}_{10}$ and $\hat{\boldsymbol{\beta}}_1$ is $(0.0001, -0.0005)'$, which indicates that our proposed method for required $n = 138$ is consistent.

Remark 5

The verification of the assumptions of normality, homoscedasticity and influential values is provided in the C section of the Appendix.

PsychOpen GOLD

Discussion

In the introduction, we explained why the size of regression coefficients, not just whether they are statistically significant, is important especially for applied research. Even if a regression coefficient is statistically significant, it might not be sufficiently large to justify expenditures necessary for a policy change (Trafimow & Osman, 2022). However, once the importance of regression coefficient size is acknowledged, there remains the crucial issue of the accuracy with which sample regression coefficients estimate population regression coefficients. Even a large sample regression coefficient may not justify a policy change if it cannot be trusted to be a good estimator of the corresponding population regression coefficient. Consequently, it is useful to have a procedure to enable valid judgments of the degree of trust consumers of research can place in sample regression coefficients as estimators of corresponding population regression coefficients. The present APP expansion provides that procedure.

In turn, there are two ways the present work, with the free and user-friendly program, can be used. One use concerns the original purpose of the APP, which is to plan sample sizes necessary for achieving researcher goals pertaining to precision and confidence. Secondly, however, the present APP expansion can be used post data collection, such as evaluating an already published regression coefficient. If a researcher reports a seemingly impressive regression coefficient, the trust that regression coefficient deserves can be assessed using the present program. If the reported sample size is less than what is necessary to meet assessors', reviewers', or policy makers' criteria for precision and confidence, the applicability of the sample regression coefficient can be discounted accordingly. Alternatively, if the reported sample size exceeds that which is necessary to meet criteria for precision and confidence, trust in the sample regression coefficient can be augmented accordingly.

Also, we wish to be upfront about an important limitation, which is the assumption of multivariate normality. Future work, that we intend to perform, could include commencing from more general assumptions. For example, instead of assuming a multivariate normal distribution, it would be a further advance to extend the APP to regression coefficients under a multivariate skew normal distribution. In the meantime, the present work is nevertheless useful even if the assumption of multivariate normality is violated. To see why, consider that skewness decreases sample sizes necessary to meet specifications for precision and confidence (e.g., Trafimow et al., 2019; Wang et al., 2019a; Wang et al., 2019c). Thus, when multivariate normality is violated, the present computer program will overestimate necessary sample sizes needed to meet specifications for precision and confidence. Hence, the results the program produces can be considered conservative sample size estimates; if a researcher collects the sample size indicated by the computer program, he or she can be assured that precision and confidence are at the specified level, or better. Of course, in those cases where multivariate normality

PsychOpen GOLD

does apply, the results produced by the computer program should be very accurate and neither conservative nor liberal.

Finally, applied researchers should consider potential applications of their research, and explicitly consider how accurate the estimation needs to be to base an intervention or policy change on the regression coefficients they obtain. They can render specifications for precision and confidence accordingly. Also, the total cost of collecting a sample with required sample size n should be considered in selecting f and c. In our real data example, $n = 138$ for $f = 0.15$ and 0.95. If we use $f = 0.10$ and 0.15 with $p = 2$ instead, the required sample size are 308 and 1014, respectively. It is impossible to have such sample sizes because the whole data size is 200.

In conclusion, we hope and expect that the present contribution provides an alternative to power analysis for researchers who use correlational designs that feature regression coefficients. If the goal is to attain statistical significance, power analysis makes sense; but if the goal is to obtain sample regression coefficients that are trustworthy estimators of corresponding population regression coefficients, the present APP extension is best.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Data Availability: Data is freely available at Supplementary Materials.

Supplementary Materials

For this article, a Shiny App can be found online to calculate the necessary sample sizes needed to estimate the regression coefficients $\boldsymbol{\beta}_1$ (for access see Index of Supplementary Materials below).

- App: Necessary Sample Sizes to Estimate Regression Coefficients

Index of Supplementary Materials

Tong, T., Trafimow, D., Wang, T., Wang, C., Hu, L., & Chen, X. (2022). *Supplementary materials to "The A Priori Procedure (APP) for estimating regression coefficients in linear models"* [Shiny App]. https://probab.shinyapps.io/linearregressionweights/

References

Cao, L., Wang, C., Wang, T., & Trafimow, D. (2021). The APP for estimating population proportion based on skew normal approximations and the Beta-Bernoulli process. *Communications in*

PsychOpen GOLD

Statistics-Simulation and Computation. Advance online publication. https://doi.org/10.1080/03610918.2021.2012192

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*(6, Pt.1), 426–443. https://doi.org/10.1037/h0026714

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences.* Academic Press.

Darlington, R. B., & Hayes, A. F. (2016). *Regression analysis and linear models: Concepts, applications, and implementation.* Guilford Publications.

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research, 26*(3), 499–510. https://doi.org/10.1207/s15327906mbr2603_7

Kassambara, A. (2019). *datarium: Data bank for statistical analysis and visualization* (Version 3.1.0) [R package]. https://github.com/kassambara/datarium

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics.* John Wiley & Sons.

Trafimow, D. (2017). Using the coefficient of confidence to make the philosophical switch from a posteriori to a priori inferential statistics. *Educational and Psychological Measurement, 77*(5), 831–854. https://doi.org/10.1177/0013164416667977

Trafimow, D. (2018). An a priori solution to the replication crisis. *Philosophical Psychology, 31*(8), 1188–1214. https://doi.org/10.1080/09515089.2018.1490707

Trafimow, D., Hyman, M. R., & Kostyk, A. (2020). The (im)precision of scholarly consumer behavior research. *Journal of Business Research, 114*, 93–101. https://doi.org/10.1016/j.jbusres.2020.04.008

Trafimow, D., & Myüz, H. A. (2019). The sampling precision of research in five major areas of psychology. *Behavior Research Methods, 51*(5), 2039–2058. https://doi.org/10.3758/s13428-018-1173-x

Trafimow, D., & Osman, M. (2022). Barriers to converting applied social psychology to bettering the human condition. *Basic and Applied Social Psychology, 44*(1), 1–11. https://doi.org/10.1080/01973533.2022.2051327

Trafimow, D., Wang, T., & Wang, C. (2019). From a sampling precision perspective, skewness is a friend and not an enemy! *Educational and Psychological Measurement, 79*(1), 129–150. https://doi.org/10.1177/0013164418764801

Viswesvaran, C. (1998). Multiple regression in behavioral research: Explanation and prediction. *Personnel Psychology, 51*(1), 223–226.

Wang, C., Wang, T., Trafimow, D., & Chen, J. (2019a). Extending A Priori Procedure to two independent samples under skew normal settings. *Asian Journal of Economics and Banking, 3*(2), 29–40.

Wang, C., Wang, T., Trafimow, D., Li, H., Hu, L., & Rodriguez, A. (2021). *Extending the A Priori Procedure (APP) to address correlation coefficients.* In N. N. Thach, V. Kreinovich, & N. D. Trung (Eds.), *Data science for financial econometrics* (pp. 141–149). Springer. https://doi.org/10.1007/978-3-030-48853-6_10

Wang, C., Wang, T., Trafimow, D., & Myüz, H. A. (2019b). *Desired sample size for estimating the skewness under skew normal settings.* In V. Kreinovich & S. Sriboonchitta (Eds.), *Structural*

PsychOpen GOLD

changes and their econometric modeling. TES 2019. Studies in computational intelligence (pp. 152–162). Springer. https://doi.org/10.1007/978-3-030-04263-9_11

Wang, C., Wang, T., Trafimow, D., & Myüz, H. A. (2019c). Necessary sample sizes for specified closeness and confidence of matched data under the skew normal setting. *Communications in Statistics–Simulation and Computation, 51*(5), 1–12. https://doi.org/10.1080/03610918.2019.1661473

Wang, C., Wang, T., Trafimow, D., & Talordphop, K. (2020). Extending the A Priori Procedure to one-way analysis of variance model with skew normal random effects. *Asian Journal of Economics and Banking, 4*(2), 77–90.

Wang, C., Wang, T., Trafimow, D., & Xu, Z. (2022). *A Priori Procedure (APP) for estimating the scale parameter in gamma populations for known shape parameter.* In S. Sriboonchitta, V. Kreinovich, & W. Yamaka (Eds.), *Credible asset allocation, optimal transport methods, and related topics* (pp. 285–298). Springer. https://doi.org/10.1007/978-3-030-04263-9_11

Wei, Z., Wang, T., Trafimow, D., & Talordphop, K. (2020). Extending the A Priori Procedure to normal Bayes models. *International Journal of Intelligent Technologies and Applied Statistics, 13*(2), 169–183. https://doi.org/10.6148/IJITAS.202006_13(2).0004

Appendix

A. Proof of Theorem 1

By Lemma 1 we know that $\hat{\boldsymbol{\beta}}_1 \sim N_p(\boldsymbol{\beta}_1, \sigma^2(\mathbf{X_c}'\mathbf{X_c})^{-1})$ so that

$$\mathbf{D}_x(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \sim N_p(\mathbf{0}, \sigma^2\mathbf{D}_x(\mathbf{X_c}'\mathbf{X_c})^{-1}\mathbf{D}_x)$$

From (4), we obtain $\mathbf{D}_x(\mathbf{X_c}'\mathbf{X_c})^{-1}\mathbf{D}_x = \mathbf{R}_{xx}^{-1}/(n-1)$, and hence

$$\mathbf{D}_x(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \sim N_p\left(\mathbf{0}, \frac{\sigma^2}{n-1}\mathbf{R}_{xx}^{-1}\right)$$

Let

$$\mathbf{Z} = \frac{\mathbf{R}_{xx}^{1/2}\mathbf{D}_x(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)}{\sigma/\sqrt{n-1}} \qquad \text{and} \qquad q = \mathbf{Z}'\mathbf{Z}$$

Note that it is easy to verify that $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ so that

$$\mathbf{Z}'\mathbf{Z} = \frac{(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)'\mathbf{D}_x\mathbf{R}_{xx}\mathbf{D}_x(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)}{\sigma^2/(n-1)} = \frac{(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)'\mathbf{S}_{xx}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)}{\sigma^2/(n-1)} \sim \chi_p^2$$

which is equivalent to, with the norm notation

PsychOpen GOLD

$$\frac{\left\|S_{xx}^{1/2}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)\right\|^2}{\sigma^2/(n-1)}\sim\chi_p^2$$

Let s^2 be the unbiased estimator of σ^2 so that $(n-p-1)s^2=n\hat{\sigma}^2$ by Lemma 1. Note that $(n-p-1)s^2/\sigma^2\sim\chi^2(n-p-1)$ and $\hat{\boldsymbol{\beta}}_1$, s^2 are independent. From the definition of F-distribution, we obtain

$$U\equiv\frac{\left\|S_{xx}^{1/2}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)\right\|^2}{ps^2/(n-1)}\sim F_{p,n-p-1} \tag{10}$$

the F-distribution with numerate degrees of freedom p and denominator degrees of freedom $n-p-1$ and the the density of U is given by

$$f_U(u)=\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{n-p-1}{2})}\left(\frac{p}{n-p-1}\right)^{p/2}u^{p/2-1}\left(1+\frac{pu}{n-p-1}\right)^{-(n-1)/2},\quad u>0$$

where $\Gamma(\cdot)$ is the gamma function.

Now, we set up the APP for estimating $\boldsymbol{\beta}_1$, the vector of regression coefficients using its unbiased estimator $\hat{\boldsymbol{\beta}}_1$. For a given precision f and confidence level c,

$$\begin{aligned}c=&\mathrm{P}\left(\left\|S_{xx}^{1/2}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)\right\|\le\sqrt{p}f\sigma\right)=\mathrm{P}\left(\left\|S_{xx}^{1/2}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)\right\|^2\le pf^2\sigma^2\right)\\=&\mathrm{P}\left(\frac{\left\|S_{xx}^{1/2}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1)\right\|^2}{p\sigma^2/(n-1)}\le(n-1)f^2\right)\end{aligned}$$

By strong law of large numbers, we know that s^2 is a consistent estimator of σ^2 so that for large n, the above expression can be reduced to

$$\mathrm{P}\left(U\le(n-1)f^2\right)\approx c$$

Therefore the required n can be obtained from (6).

B. The Density of W Given in Remark 3

The density of W, the non-central F-distribution with degrees of freedoms k_1 and k_2 and non-centralized parameter λ is given by

$$\begin{aligned}f_W(w;k_1,k_2,\lambda)=&\frac{k_1k_2^{k_2/2}}{\Gamma(\frac{1}{2})\Gamma(\frac{k_1-1}{2})\Gamma(\frac{k_2}{2})2^{(k_1+k_2)/2}}\int_0^\infty\exp\left\{-\frac{1}{2}(\lambda+(k_1w+k_2)u)\right\}u^{k_2/2}\\&\times\int_{-\sqrt{k_1wu}}^{\sqrt{k_1wu}}\exp(\lambda^{1/2}s)(k_1wu-s^2)^{(k_1-3)/2}dsdu\end{aligned} \tag{11}$$

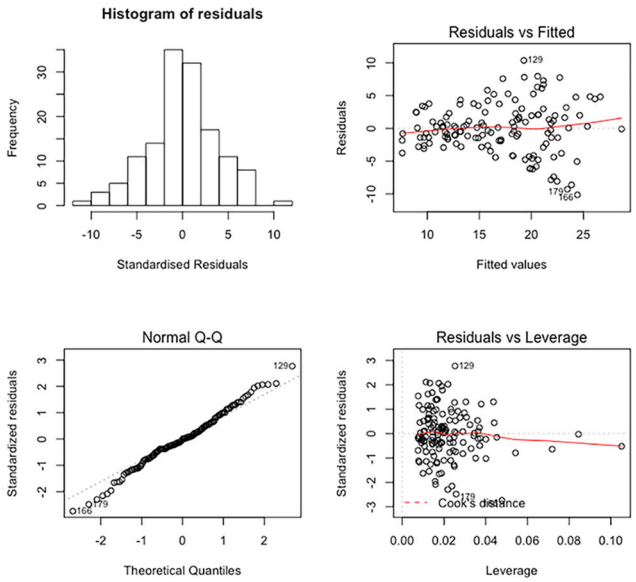
PsychOpen GOLD

C. The Verification of Normality Assumptions

Now we check the assumptions of normality, homoscedasticity and influential values by Figure 6. From the histogram and QQ plot of residuals, we can see that in our example, all the points fall approximately along this reference line, so we can assume normality. There is no pattern in the scatter of the fitted values and residuals and the width of the scatter as predicted values increase is roughly the same so the assumption of homoscedasticity has been met. The residual and leverage plot highlights the top 3 most extreme points (129, 166, and 179 are marked in Fitted values of Figure 6), with a standardized residuals above 2 or below −2. However, there is no outliers that exceed 3 standard deviations. Additionally, the data do not present much influential points. That is, most of data points have a leverage statistic below $2(p+1)/n = 6/138 = 0.04$.

Figure 6

Histogram of Standardized Residuals, Plot of Residuals vs Fitted Values, QQ Plot of Standardized Residuals, and Plot of Residuals vs Leverage



EUROPEAN ASSOCIATION OF METHODOLOGY

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.

PsychOpen GOLD