Power Analysis for Correlation & Multiple Regression

- Sample Size & multiple regression
- Subject-to-variable ratios
- Stability of correlation values
- Useful types of power analyses
  - Simple correlations
  - Full multiple regression
- Considering Stability & Power
- Sample size for a study

Sample Size & Multiple Regression

The general admonition that “larger samples are better” has considerable merit, but limited utility...

- $R^2$ will always be 1.00 if $k = N-1$ (it’s a math thing)
- $R^2$ will usually be “too large” if the sample size is “too small” (same principle but operating on a lesser scale)
- $R^2$ will always be larger on the modeling sample than on any replication using the same regression weights
- $R^2$ & b-values will replicate better or poorer, depending upon the stability of the correlation matrix values
- $R^2$ & b-values of all predictors may vary with poor stability of any portion of the correlation matrix (any subset of predictors)
- F- & t-test p-values will vary with the stability & power of the sample size – both modeling and replication samples

Subject-to-Variable Ratio

How many participants should we have for a given number of predictors? -- usually refers to the full model

The subject/variable ratio has been an attempt to ensure that the sample is “large enough” to minimize “parameter inflation” and improve “replicability”.

Here are some common admonitions..

- 20 participants per predictor
- a minimum of 100 participants, plus 10 per predictor
- 10 participants per predictor
- 200 participants for up to $k=10$ predictors and 300 if $k>10$
- 1000 participants per predictor
- a minimum of 2000 participants, + 1000 for each 5 predictors
As is often the case, different rules of thumb have grown out of different research traditions, for example:

- chemistry, which works with very reliable measures and stable populations, calls for very small s/v ratios
- biology, also working largely with “real measurements” (length, weight, behavioral counts) often calls for small s/v ratios
- economics, fairly stable measures and very large (cheap) databases often calls for huge s/v ratios
- education, often under considerable legal and political scrutiny, (data vary in quality) often calls for fairly large s/v ratios
- psychology, with self-report measures of limited quality, but costly data-collection procedures, often “shaves” the s/v ratio a bit

Problems with Subject-to-variable ratio

#1 neither n, N nor N/k is used to compute R² or b-values
- R² & b/-values are computed from the correlation matrix
- N is used to compute the significance test of the R² & each b-weight

#2 Statistical Power Analyses involves more than N & k
- We know from even rudimentary treatments of statistical power analysis that there are four attributes of a statistical test that are inextricably intertwined for the purposes of NHST...
  - acceptable Type I error rate (chance of a “false alarm”)
  - acceptable Type II error rate (chance of a “miss”)
  - size of the effect being tested for
  - sample size

We will “forsake” the subjects-to-variables ratio for more formal power analyses & also consider the stability of parameter estimates (especially when we expect large effect sizes).

NHST Power “vs.” Parameter estimate stability

NHST power → what’s the chances of rejecting a “false null” vs. making a Type II error?

Statistical power is based on...
- size of the effect involved (“larger effects are easier to find”)
- amount of power (probability of rejecting H0: if effect size is as expected or larger)

Stability → how much error is there in the sample-based estimate of a parameter (correlation, regression weight, etc.)?

Stability is based on ...
- “quality” of the sample (sampling process & attrition)
- sample size

\[ \text{Std of } r = \frac{1}{\sqrt{(N-3)}} \]

<table>
<thead>
<tr>
<th>N</th>
<th>r +/-</th>
<th>N</th>
<th>r +/-</th>
<th>N</th>
<th>r +/-</th>
<th>N</th>
<th>r +/-</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>.146</td>
<td>100</td>
<td>.101</td>
<td>200</td>
<td>.07</td>
<td>1000</td>
<td>.031</td>
</tr>
<tr>
<td>300</td>
<td>.058</td>
<td>500</td>
<td>.045</td>
<td>1000</td>
<td>.031</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The power table only tells us the sample size we need to reject H0: r=0!! It does not tell us the sample size we need to have a good estimate of the population r !!!!!

Partial Power Table (taken & extrapolated from Friedman, 1982)

<table>
<thead>
<tr>
<th>r</th>
<th>.15</th>
<th>.20</th>
<th>.25</th>
<th>.30</th>
<th>.35</th>
<th>.40</th>
<th>.45</th>
<th>.50</th>
<th>.55</th>
<th>.60</th>
<th>.65</th>
<th>.70</th>
</tr>
</thead>
<tbody>
<tr>
<td>power</td>
<td>.30</td>
<td>93</td>
<td>53</td>
<td>34</td>
<td>24</td>
<td>18</td>
<td>14</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td>132</td>
<td>74</td>
<td>47</td>
<td>33</td>
<td>24</td>
<td>19</td>
<td>15</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>.50</td>
<td>170</td>
<td>95</td>
<td>60</td>
<td>42</td>
<td>30</td>
<td>23</td>
<td>18</td>
<td>14</td>
<td>12</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td>257</td>
<td>143</td>
<td>90</td>
<td>62</td>
<td>45</td>
<td>34</td>
<td>24</td>
<td>20</td>
<td>16</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>.70</td>
<td>300</td>
<td>167</td>
<td>105</td>
<td>72</td>
<td>52</td>
<td>39</td>
<td>29</td>
<td>23</td>
<td>18</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>.80</td>
<td>343</td>
<td>191</td>
<td>120</td>
<td>82</td>
<td>59</td>
<td>44</td>
<td>33</td>
<td>26</td>
<td>20</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>.90</td>
<td>459</td>
<td>255</td>
<td>160</td>
<td>109</td>
<td>78</td>
<td>58</td>
<td>44</td>
<td>27</td>
<td>21</td>
<td>17</td>
<td>13</td>
</tr>
</tbody>
</table>

How can a sample have “sufficient power” but “poor stability”? Notice it happens for large effect sizes!! e.g., For a population with r = .30 & a sample of 100 …

• Poor stability of r estimate → +/- 1 std is .20-.40
• Large enough to reject H0: that r = 0 → power almost .90

Power Analysis for Simple Correlation

Post hoc
I found r (22) = .30, p > .05, what’s the chance I made a Type II error ??

N = 24    Power = .30    Chance Type II error .70

A priori
#1 I expect my correlation will be about .25, & want power = .90
sample size should be = 160

#2 Expect correlations of .30, .45, and .20 from my three predictors & want power = .80
sample size should be = 191, based on lowest r = .20

Power Analysis for Multiple regression

Power analysis for multiple regression is about the same as for simple regression, we decide on values for some parameters and then we consult a table …

Remember the F-test of H0: R² = 0 ??

\[
F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{N-k-1}}
\]

Which corresponds to:

significance test = effect size * sample size

So, our power analysis will be based not on R² per se, but on the power of the F-test of the H0: R² = 0
Using the power tables (post hoc) for multiple regression (single model) requires that we have four values:

- **a** = the p-value we want to use (usually .05)
- **u** = df associated with the model (we've used “k”)
- **v** = df associated with F-test error term (N - u - 1)
- **f²** = (effect size estimate) = \( \frac{R^2}{1 - R^2} \)

\[ \lambda = f^2 \times (u + v + 1) \]

This is the basis for determining power.

E.g., N = 96, and 5 predictors, \( R^2 = .10 \) was found

- a = .05
- u = 5
- v = 96 - 5 - 1 = 90

\[ f^2 = \frac{.1}{1 - .1} = .1111 \]

\[ \lambda = .1111 \times (5 + 89 + 1) = 10.6 \]

Go to table -- a = .05, & u = 5

<table>
<thead>
<tr>
<th>v</th>
<th>60</th>
<th>63</th>
<th>72</th>
</tr>
</thead>
<tbody>
<tr>
<td>power</td>
<td>.68</td>
<td>120</td>
<td>65</td>
</tr>
</tbody>
</table>

Another N = 48, and 6 predictors, \( R^2 = .20 \) (p < .05)

- a = .05
- u = 6

\[ f^2 = \frac{.2}{1 - .2} = .25 \]

\[ \lambda = .25 \times (6 + 41 + 1) = 12 \]

Go to table -- a = .05 & u = 6

<table>
<thead>
<tr>
<th>v</th>
<th>20</th>
<th>59</th>
</tr>
</thead>
<tbody>
<tr>
<td>power</td>
<td>.59</td>
<td>68</td>
</tr>
</tbody>
</table>

This sort of post hoc power analysis is, as before, especially helpful when the H0: has been retained -- to determine whether the result is likely to have been a Type II error.

Remember that one has to decide how small of an effect is “meaningful”, and adjust the sample size to that decision.

**a priori** power analyses for multiple regression are complicated by...

- Use of \( \lambda \) (combo of effect & sample size) rather than \( R^2 \) (just the effect size) in the table.
- This means that sample size enters into the process TWICE
  - when computing \( \lambda = f^2 \times (u + v + 1) \)
  - when picking the “v” row to use \( v = N - u - 1 \)
- So, so the \( \lambda \) of an analysis reflects the combination of the effect size and sample size, which then has differential power depending (again) upon sample size (v).

E.g.#1, \( R^2 = .20 \)

\[ f^2 = \frac{.2}{1 - .2} = .25 \]

\[ \lambda = .25 \times (50) = 12.5 \]

with u = 10, and v = N - 10 - 1 = 39 -- power is about .50

E.g.#2, \( R^2 = .40 \)

\[ f^2 = \frac{.4}{1 - .4} = .67 \]

\[ \lambda = .67 \times (19) = 12.5 \]

with u = 10, and v = 19 - 19 - 1 = 8 -- power is about .22

So, for a priori analyses, we need the sample size estimate to compute the \( \lambda \) to use to look up the sample size estimate we need for a given level of statistical power ????
Perhaps the easiest way to do *a priori* sample size estimation is to play the “what if game” . . .

I expect that my 4-predictor model will account for about 12% of the variance in the criterion -- what sample size should I use ???

\[
\alpha = .05 \quad u = 4 \quad f^2 = R^2 / (1 - R^2) = .12 / (1 - .12) = .136
\]

“what if..”

\[
\begin{array}{lll}
N = 25 & N = 65 & N = 125 \\
\nu = (N - u - 1) = & 20 & 60 & 120 \\
\lambda = f^2 * (\nu + 1) = & 3.4 & 8.8 & 17.0 \\
\end{array}
\]

Using the table…

- power = about .21
- power = about .62
- power = about .915

If we were looking for power of .80, we might then try N = 95

so \( \nu = 90 \), \( \lambda = 12.2 \), power = about .77 (I’d go with N = 100-110)

Putting Stability & Power together to determine the sample size

1. Start with stability – remember …

\[
\text{Std of } r = 1 / \sqrt{(N-3)}
\]

\[
\begin{array}{llll}
N = 50 & r +/- .146 & N = 100 & r +/- .101 & N = 200 & r +/- .07 \\
N = 300 & r +/- .058 & N = 500 & r +/- .045 & N = 1000 & r +/- .031 \\
\end{array}
\]

... suggesting that 200-300 is a good guess for most analyses (but more is better).

2. Then for the specific analysis, do the power analysis …

For the expected \( r/R^2 \) & desired power, what is the required sample size?

3. Use the larger of the stability & power estimates !

An example ….

We expect a correlation of .60, and want only a 10% risk of a Type II error if that is the population correlation

Looking at the power table for \( r = .60 \) and power = .90..

… the suggested sample size is 21

\[
N = 21, \text{ means the std of the correlation estimates (if we took multiple samples from the target population is}
\]

\[
1 / \sqrt{(21-3)} = .35
\]

With \( N = 21 \) we’ve a 90% chance of getting a correlation large enough to reject the Null 😊

→ on average, our estimate of the population correlation will be wrong my .35. We’d certainly interpret a .25 and a .95 differently 😔

In this case we’d go with the 200-300 estimate, in order to have sufficient stability – we’ll have lots of power!
Another example …. We expect a correlation of .10, and want only a 20% risk of a Type II error if that is the population correlation

Considering stability – let’s say we decide to go with 300

Looking at the power table for $r = .10$ and power = .80.. … the suggested sample size is 781

With $N = 300$, we’d only have power of about .40 … 60% chance of a Type II error.

In this case we’d go with the 781 estimate (if we can afford it), in order to have sufficient power – we’ll have great stability of +/- .036!

Considering the sample size for the **Study**

Really a simple process, but sometimes the answer is daunting!

First: For each analysis ($r$ or $R^2$)
   -> perform the power analysis
   -> consider the “200-300” suggestion & resulting stability
   -> pick the larger value as the $N$ estimate for that analysis

Then: Looking at the set of $N$ estimates for all the analyses …
   -> The largest estimate is the best bet for the study

This means we will base our **study** sample size on the sample size required for the least powerful significance test!

Usually this is the smallest simple correlation or a small $R^2$ with a large number of predictors.