Another way to select sample size

The sample R-squared method

ANOVA Summary Table

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	$p-1$	SSR	$MSR = SSR/(p-1)$	$F = \frac{MSR}{MSE}$	p-value
Error	$n-p$	SSE	$MSE = SSE/(n-p)$		
Total	$n-1$	$SSTO$			

Proportion of variation in the dependent variable that is explained by the independent variables

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}}$$

Full vs. Reduced Model

- You have 2 sets of variables, A and B
- Want to test B controlling for A
- Fit a model with both A and B: Call it the **Full Model**
- Fit a model with just A: Call it the **Reduced Model**

$$R_F^2 \geq R_R^2$$

When you add independent variables, R^2 can only go up

- By how much? Basis of F test.
- Same as testing H_0: All betas in set B (there are *s* of them) equal zero

$$F = \frac{(SSR_F - SSR_R)/s}{MSE_F}$$

F test is based not just on change in R^2, but upon

$$a = \frac{R_F^2 - R_R^2}{1 - R_R^2}$$

Increase in explained variation expressed as a fraction of the variation that the reduced model does *not* explain.

$$F = \left(\frac{n-r}{q}\right)\left(\frac{a}{1-a}\right)$$

- For any given sample size, the bigger *a* is, the bigger *F* becomes.
- For any a ≠0, *F* increases as a function of *n.*
- So you can get a large *F* from strong results and a small sample, or from weak results and a large sample.

$$F = \left(\frac{n-r}{q}\right)\left(\frac{a}{1-a}\right)$$

The sample variation method is to choose a value of a that is just large enough to be interesting, and increase n, calculating F and its p-value each time until $p < 0.05$; then stop. The final value of n is the smallest sample size for which an effect explaining that much of the remaining variation will be significant. With that sample size, the effect will be significant if and only if it explains a or more of the remaining variation.

That's all there is to it. You tell me a proportion of remaining variation that you want to be statistically significant, and I'll tell you a sample size.

Example

Suppose we are planning a 2x3x4 analysis of covariance, with two covariates, and factors named A, B and C. We are setting it up as a regression model, with one dummy variable for A, 2 dummy variables for B, and 3 for C. Interactions are represented by product terms, and there are 2 products for the AxB interaction, 3 for AxC, 6 for BxC, and 1*2*3 = 6 for AxBxC. The regression coefficients for these plus two for the covariates and one for the intercept give us r = 26. The null hypothesis is that of no BxC interaction, so q = 6. The "other effects in the model" for which we are "controlling" are represented by 2 covariates and 17 dummy variables and products of dummy variables.

```
samprsq1 <- function(r,q,a,alpha=0.05)
# Find n so remaining proportion of SS explained will be significant
#   r   Number of IVs in full model
#   q   Numerator df = number of linear constraints being tested
#   a   Sample proportion of remaining variation explained.
#   alpha       Significance level (default = 0.05)
    {
    pval <- 1 ; n <- r+1
    while(pval > alpha)
        {
        n <- n+1
        F <- (n-r)/q * a/(1-a)
        df2 <- n-r
        pval = 1-pf(F,q,df2)
        }#End while
     samprsq1 <- n
     samprsq1
     } # End of function samprsq1
```

```
> samprsq1(r=26,q=6,a=0.10) # Using default value of alpha=0.05
[1] 144
```

Cohen's Population R^2 Method

$$\begin{aligned}
\phi &= \frac{(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})}{\sigma^2} \\
F^* &= \frac{(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{h})}{q\,MSE} \\
&= \left(\frac{n-r}{q}\right)\left(\frac{a}{1-a}\right)
\end{aligned}$$

$$\phi/n = \frac{(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X}/n)^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})}{\sigma^2}$$

Let $n \to \infty$ and call the result "population effect size." Write it as $\frac{a}{1-a}$. Call a the proportion of remaining variation in the *population*.

Population R^2 Method

- It's a way to choose an effect size without having to guess true beta (mu) or sigma values
- To get a non-centrality parameter for power analysis, Cohen multiplies by n-r instead of n.
- That's because he thinks of phi as q times a population F statistic.

```
poprsq <- function(r,q,a,wantpow=0.80,alpha=0.05)
# Cohen's Popularion R-squared Method
#   r   Number of IVs in full model
#   q   Numerator df = number of linear constraints being tested
#   a   Population proportion of remaining variation explained.
#       This is Cohen's "effect size."
#   wantpow     Desired power (default = 0.80)
#   alpha       Significance level (default = 0.05)
    {
    pow <- 0 ; nn <- r+1 ; oneminus <- 1 - alpha
    while(pow < wantpow)
        {
        nn <- nn+1
        phi <- (nn-r) * a/(1-a)
        ddf <- nn-r
        pow <- 1 - pf(qf(oneminus,q,ddf),q,ddf,phi)
        }#End while
     poprsq <- nn
     poprsq
     } # End of function poprsq
```

```
> samprsq1(r=26,q=6,a=0.10)
[1] 144
> poprsq(r=26,q=6,a=0.10)
[1] 155
>
> samprsq1(r=26,q=6,a=0.05)
[1] 270
> poprsq(r=26,q=6,a=0.05)
[1] 292
```