Choosing Sample size

Sample variation method

Consider a 2 x 3 analysis of covariance with 4 covariates. Call the factors A (2 values) and B (3 values). This is equivalent to a multiple regression with an intercept and 9 independent variables (so p = 10).

- 4 covariates
- 1 dummy variable for the main effect of A
- 2 dummy variables for the main effects of B
- 2 product terms representing the A by B interaction

Suppose we want the interaction to be significant provided that it explains 6% or more of the variation that remains after allowing for the covariates and the main effects.

```
/*************************  sampvar1.sas *************************/
/*      Finds n needed for significance, for a given proportion of */
/*      remaining variation                                        */
/*****************************************************************/

options linesize=79 noovp formdlim='_' pagesize = 200;
data explvar;     /* Can replace alpha, s, p, and a below.   */
   alpha = 0.05;  /* Significance level.                     */
   s = 2;         /* Numerator df = # IVs being tested.      */
   p = 10;        /* There are p beta parameters.            */
   a = .06  ;     /* Proportion of remaining variation after */
                  /* controlling for all other variables.    */

   /* Initializing ... */  pval = 1; n = p;
   do until (pval <= alpha);
      n = n+1 ;
      F = (n-p)/s * a/(1-a);
      df2 = n-p;
      pval = 1-probf(F,s,df2);
   end;
   F = (n-p)/s * a/(1-a); df2 = n-p;
   pval = 1-probf(F,s,df2);

   put ' *******************************************************';
   put ' ';
   put '  For a multiple regression model with ' p 'betas, ';
   put '  testing ' s ' variables controlling for the others,';
   put '  a sample size of ' n 'is needed for significance at the';
   put '  alpha = ' alpha 'level, when the effect explains a = ' a ;
   put '  of the remaining variation after allowing for all other ';
   put '  variables in the model. ';
   put '  F = ' F ', df = (' s ', ' df2 '), p = ' pval;
   put ' ';
   put ' *******************************************************';

tuzo > sas sampvar1 ; cat sampvar1.log
```

```
*****************************************************

  For a multiple regression model with 10 betas,
  testing 2  variables controlling for the others,
  a sample size of 107 is needed for significance at the
  alpha = 0.05 level, when the effect explains a = 0.06
  of the remaining variation after allowing for all other
  variables in the model.
  F = 3.0957446809 , df = (2 , 97 ), p = 0.0497394409

 *****************************************************
```

Since 107/6 = 17.83333, try 18 cases per cell, for a total n = 6*18 = 108. What value of a do we need for significance now?

```
/*************************  sampvar2.sas ***************************/
/*  Finds proportion of remaining variation needed for significance, */
/*  given sample size n                                              */
/*******************************************************************/

options linesize=79 noovp formdlim='_' pagesize = 200;
                    /************************************/
data explvar;       /* Replace alpha, s, p, and a below.  */
     alpha = 0.05;  /* Significance level.                */
     s = 2;         /* Numerator df = # IVs being tested. */
     p = 10;        /* There are p beta parameters.       */
     n = 108  ;     /* Sample size                        */
                    /************************************/
     oneminus = 1 - alpha; df2 = n-p;
     Fcrit = finv(oneminus,s,df2);
     a = s*Fcrit / (s*Fcrit + df2);

     put ' ***************************************************';
     put ' ';
     put '  For a multiple regression model with ' p 'betas, ';
     put '  testing ' s ' variables at significance level ';
     put '  alpha = ' alpha ' controlling for the other variables,';
     put '  and a sample size of ' n', the variables need to explain';
     put '  a = ' a ' of the remaining variation to be significant.';
     put '  Using critical value of F = ' Fcrit ', df = (' s ',' df2 ')';
     put '   ';
     put ' ***************************************************';
```

```
 *****************************************************

  For a multiple regression model with 10 betas,
  testing 2  variables at significance level
  alpha = 0.05  controlling for the other variables,
  and a sample size of 108 , the variables need to explain
  a = 0.0593060142  of the remaining variation to be significant.
  Using critical value of F = 3.089203013 , df = (2 ,98 )

 *****************************************************
```

Power (population variation method)

Now suppose we want the sample size to be large enough to detect the interaction as significant if the *population* proportion of remaining variation explained is 6% or more. For any given sample size, we might get unlucky and the test will not be significant (This is a Type II error). So let's say we want to find the sample size so that if the proportion of remaining variation explained is 6% or more, the test will be significant with probability 0.90.

```
/**********************  popvar1.sas ***************************/
options linesize=79 noovp formdlim='_' pagesize = 200;
                    /*************************************************/
data fpower;        /* Replace alpha, s, p, and wantpow below    */
     alpha = 0.05;  /* Significance level                        */
     s = 2;         /* Numerator df = # IVs being tested         */
     p = 10;        /* There are p beta parameters               */
     A = .06  ;     /* POPULATION effect size                    */
     wantpow = .90; /* Find n to yield this power.               */
                    /*************************************************/
     power = 0; n = p; oneminus = 1-alpha; /* Initializing ... */
     do until (power >= wantpow);
        n = n+1 ;
        ncp = (n-p)*A/(1-A);
        df2 = n-p;
        power = 1-probf(finv(oneminus,s,df2),s,df2,ncp);
     end;
     put ' ******************************************************';
     put '   ';
     put '   For a multiple regression model with ' p 'betas, ';
     put '   testing ' s 'independent variables using alpha = ' alpha ',';
     put '   a sample size of ' n 'is needed';
     put '   in order to have probability ' wantpow 'of rejecting H0';
     put '   for a POPULATION  effect of size A = ' A ;
     put '   ';
     put ' ******************************************************';
```

```
******************************************************

   For a multiple regression model with 10 betas,
   testing 2 independent variables using alpha = 0.05 ,
   a sample size of 212 is needed
   in order to have probability 0.9 of rejecting H0
   for a POPULATION  effect of size A = 0.06

 ******************************************************
```

For a power of 0.80, you only need n=164 .

We can turn this around and ask, for a particular sample size, what population effect size is required to have a specified power.

```
/*********************  popvar2.sas ***************************/
/* Given sample size, what effect size (population A) is required */
/* to have a specified power?                                     */
/****************************************************************/

options linesize=79 noovp formdlim='_' pagesize = 200;
                    /*****************************************/
data fpower;        /* Replace alpha, s, n, p, and wantpow below */
     alpha = 0.05;  /* Significance level                     */
     s = 2;         /* Numerator df = # IVs being tested      */
     n = 216;       /* Sample size                            */
     p = 10;        /* There are p beta parameters            */
     wantpow = .90; /* Find effect size A to yield this power.   */
                    /*****************************************/
     df2 = n-p; oneminus = 1 - alpha;
     critval = finv(oneminus,s,df2);
   /* Initializing ... */  A = 0;
   do until (power ge wantpow);
      A = A + .001 ;
      ncp = (n-p)*A/(1-A);
      power = 1-probf(critval,s,df2,ncp);
     end;
   put ' ***************************************************';
   put ' ';
   put '  For a multiple regression model with ' p 'betas, ';
   put '  testing ' s ' variables at significance level ';
   put '  alpha = ' alpha ' controlling for the other variables,';
   put '  and a sample size of ' n', the variables need to explain';
   put '  A = ' A ' of the remaining POPULATION variation to have a';
   put '  probability of '  wantpow  'of being significant';
   put '   ';
   put ' ***************************************************';
```

```
***************************************************

  For a multiple regression model with 10 betas,
  testing 2  variables at significance level
  alpha = 0.05  controlling for the other variables,
  and a sample size of 216 , the variables need to explain
  A = 0.059  of the remaining POPULATION variation to have a
  probability of 0.9 of being significant

 ***************************************************
NOTE: The data set WORK.FPOWER has 1 observations and 11 variables.
NOTE: DATA statement used:
      real time           0.18 seconds
      cpu time            0.06 seconds
```

Finally, you might just want to know the power for a particular sample size and effect size.

```
/**********************  whatispow.sas *********************************/
/* Given sample size n and effect size (population A), what is the power? */
/*************************************************************************/

options linesize=79 noovp formdlim='_' pagesize = 200;
                    /*****************************************/
data fpower;        /* Replace alpha, s, n, p, and A below       */
     alpha = 0.05;  /* Significance level                        */
     s = 2;         /* Numerator df = # IVs being tested         */
     n = 216;       /* Sample size                               */
     p = 10;        /* There are p beta parameters               */
     A = .06;       /* Population effect size                    */
                    /*****************************************/
     df2 = n-p; oneminus = 1 - alpha;
     critval = finv(oneminus,s,df2);
     ncp = (n-p)*A/(1-A);
     power = 1-probf(critval,s,df2,ncp);

   put ' ****************************************************';
   put ' ';
   put '  For a multiple regression model with ' p 'betas, ';
   put '  testing ' s ' variables at significance level ';
   put '  alpha = ' alpha ' controlling for the other variables,';
   put '  a sample size of ' n' and an effect size of A = 'A',';
   put '   ';
   put '                    Power = ' power;
   put '   ';
   put ' ****************************************************';
```

```
****************************************************

  For a multiple regression model with 10 betas,
  testing 2  variables at significance level
  alpha = 0.05  controlling for the other variables,
  a sample size of 216  and an effect size of A = 0.06 ,

                  Power = 0.9070776516

 ****************************************************
NOTE: The data set WORK.FPOWER has 1 observations and 10 variables.
NOTE: DATA statement used:
      real time           0.13 seconds
      cpu time            0.06 seconds


NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           0.48 seconds
      cpu time            0.24 seconds
```