STAT 350: Lecture 30

Power and Sample Size Calculations: Examples

SAMPLE SIZE NEEDED using t test: SAND and FIBRE example.

Now for the same assumed values of the parameters how many replicates of the basic design (using 9 combinations of sand and fibre contents) would I need to get a power of 0.95? The matrix $X^T X$ for m replicates of the design actually used is m times the same matrix for 1 replicate. This means that $a^T(X^T X)^{-1}a$ will be $1/m$ times the same quantity for 1 replicate. Thus the value of δ for m replicates will be $\sqrt{m}$ times the value for our design, which was 2. With m replicates the degrees of freedom for the t-test will be $18m - 4$. We now need to find a value of m so that in the row in Table B 5 across from $18m - 4$ degrees of freedom and the column corresponding to

$$\delta = 2\sqrt{m}$$

we find 0.95. To simplify we try just assuming that the solution m is quite large and use the last line of the table. We get δ between 3 and 4 – say about 3.75. Now set $2\sqrt{m} = 3.7$ and solve to find $m = 3.42$ which would have to be rounded to 4 meaning a total sample size of $4 \times 18 = 72$. For this value of m the non-centrality parameter is actually 4 (not the target of 3.75 because of rounding) and the power is 0.98. Notice that for this value of m the degrees of freedom for error is 68 which is so far down the table that the powers are not much different from the ∞ line.

Technically it would be pretty easy to imagine using 3.5 replicates – each combination of SAND and FIBRE would be tried 7 times giving 63-4=59 degrees of freedom for error. The achieved power would then be quite close to 0.95.

POWER of F test: SAND and FIBRE example.

Now consider the power of the test that all the higher order terms are 0 in the model

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 F_i + \beta_3 F_i^2 + \beta_4 S_i^2 + \beta_5 S_i F_i + \epsilon_i$$

that is the power of the F test of $\beta_3 = \beta_4 = \beta_5 = 0$.

You will need to specify the non-centrality parameter for this F test. In general the noncentrality parameter for a F test based on ν_1 numerator degrees of freedom is given by

$$E(\text{Extra SS})/\sigma^2 - \nu_1 \,.$$

This quantity needs to be worked out algebraically for each separate case, however, some general points can be made.

- Write the full model as

  $$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

  and the reduced model as

  $$Y = X_1\beta_1 + \epsilon$$

- The Extra SS is the difference between two Error sums of squares. One is for the full model and

  $$E(\text{ErrorSS}_{\text{FULL}}) = \text{ErrorDF}_{\text{FULL}}\sigma^2$$

  because we assume that the FULL model is correct.

- The Error SS for the reduced model is

  $$Y^T(I - H_1)Y$$

  where $H_1 = X_1(X_1^T X_1)^{-1}X_1^T$. Replace Y by its formula $X_1\beta_1 + X_2\beta_2 + \epsilon$ from the full model equation and take expected value. The answer is

  $$\sigma^2(n - p_1) + \beta_2^T X_2^T (I - H_1) X_2 \beta_2$$

  where $p_1 =$ is the rank of X_1. This makes the non-centrality parameter $\beta_2^T X_2^T (I - H_1) X_2 \beta_2 / \sigma^2$.

- The numerator can be calculated from the Error sum of squares when $X_2\beta_2$ is regressed on X_1.

Now consider the sand and fibre example and assume $\beta_3 = -0.004$, $\beta_4 = -0.005$ and $\beta_5 = 0.001$. The following SAS code computes the required numerator.

```
data plaster;
infile 'plaster.dat';
input sand fibre hardness strength;
newx = -0.004*fibre*fibre -0.005*sand*sand
  +0.001*sand*fibre;
proc reg  data=plaster;
 model newx = sand fibre ;
run;
```

The output shows that the error sum of squares regressing newx on sand, fibre and an intercept is 31.1875. Taking σ^2 to be 7 we get a noncentrality parameter of roughly 4.55. Now compute the quantity $\phi = \sqrt{4.55}/\sqrt{3+1} = 1.07$ needed for table B 11. For 3 numerator and 18-6=12 denominator degrees of freedom we get a power between 0.27 and 0.56 but close to 0.27. SAMPLE SIZE for F test: SAND and FIBRE example.

Now for the same basic problem and parameter values how many times would we need to replicate the design to get a power of 0.95? Again the non-centrality parameter for m replicates is m times that for 1 replicate; in terms of the parameter ϕ used in the tables the value is proportional to $\sqrt{m}$. With m replicates we now have $18m - 6$ denominator degrees of freedom. Again if $18m-6$ is reasonably large then we can use the ∞ line and see that ϕ_m must be around 2.2 making m roughly 4 ($\phi_m = \sqrt{m}\phi_1 = 1.07\sqrt{m}$).

Table B 12 can be used directly. Table 12 gives values of n/r where n is the total sample size, the degrees of freedom in the numerator of the F-test are $r-1$, the degrees of freedom for error are $n-r$ and the non-centrality parameter δ^2 is given by

$$\left(\frac{\Delta}{\sigma}\right)^2 \frac{n}{2}$$

If your basic design has n_1 data points and p parameters and your F test is based on ν_1 degrees of freedom then when you replicate the design m times you get mn_1 total data points, $mn_1 - p$ degrees of freedom for error and ν_1 degrees of freedom for the numerator of the F test.

To use the table take $r = \nu_1 + 1$. Then work out Δ/σ by taking the value of the noncentrality parameter δ_1^2 for one replicate of the basic design and computing

$$frac\Delta\sigma = \sqrt{2}\delta_1$$

. Look up n/r in the table and take that to be m. You will be making a small mistake unless $p = \nu_1 + 1$ (which is the case for the overall F test in the basic ANOVA table). The problem is that you will be pretending you have $(m-1)(\nu_1+1)$ degrees of freeedom for error instead of $mn_1 - p$. As long as these are both large all is well.

In our example for a power of 0.95 and m replicates of the 18 point design we have $\delta_1^2 = 4.55$ as above. We have $r = 3 + 1 = 4$. We get $\Delta/\sigma = \sqrt{2}\sqrt{4.55} = 3.02$. For a level 0.05 test we then look on page 1362 and get $m = 5$ for a total sample size of 90. The degrees of freedom for error will really be 84 but the table pretends that the degrees of freedom for error will be $(5-1)\times 4 = 16$. The latter is pretty small. The table supposes a small number of error df which would decrease the power of a test. This means that $m=5$ is probably an overestimate of the required sample size.

A better answer can be had by looking at replicates of the 9 point design. For 9 data points the nonecntrality parameter would have been $\delta_1^2 = 4.55/2 = 2.275$. This would give $\Delta/\sigma = 2.13$ and m of 9 or 10. For $m=10$ we would have the same design as before. For $m=9$ we would have only 72 data points. At this point you go back to Table B 11 to work out the power properly for 72 or 80 data points and see if 72 is enough.

Heteroscedastic Errors

If plots and/or tests show that the error variances $\sigma_i^2 = Var(\epsilon_i)$ depend on i there are several standard approaches to fixing the problem, depending on the nature of the dependence.

- If the variances are known except for a constant factor then we may write $\sigma_i^2 = \sigma^2/w_i$ and use a technique called **weighted least squares**. (See Chapter 10 in the text.)

  This usually arises realistically in the following situations:

  - Y_i is an average of n_i measurements where you know n_i. Then $w_i = n_i$.
  - Plots suggest that σ_i^2 might be proportional to some power of some covariate: $\sigma_i^2 = kx_i^\gamma$. Then $w_i = x_i^{-\gamma}$.

- If the variances are thought to depend on the mean of Y_i then two standard approaches are available:

  - The older approach is **transformation**. We compute $Y_i^* = g(Y_i)$ for some function g like the logarithm or square root. Then we regress Y_i^* on the covariates. This approach sometimes works for skewed response variables like income; after transformation we occasionally find the errors are more nearly normal, more homoscedastic and that the model is simpler. See page 130ff and check under transformations and Box-Cox in the index.
  - The newer approach is to use a **generalized linear model**; see STAT 402.

– Transformation uses the model

$$E(g(Y_i)) = x_i^T \beta$$

while generalized linear models use

$$g(E(Y_i)) = x_i^T \beta$$

Generally the latter approach offers more flexibility since it is then possible to model the variance as a general function of the mean while for transformation followed by ordinary least squares the transformed data must follow a *homoscedastic* linear model.

- It is also possible to use a hybrid approach in which the parameters are *estimated* by least squares but inference (estimation of standard errors, testing, confidence intervals) are based on a model in which the errors may be heteroscedastic.

Weighted Least Squares

If

$$E(Y_i) = x_i^T \beta$$

and

$$Var(Y_i) = \sigma^2 / w_i$$

and *the errors are independent with normal distributions* then the likelihood is

$$\prod_{i=1}^n \frac{\sqrt{w_i}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{w_i}{2\sigma^2}(Y_i - x_i^T \beta)^2\right]$$

To choose β to maximize this likelihood we minimize the quantity

$$\sum_{i=1}^n w_i (Y_i - x_i^T \beta)^2 .$$

The process is called weighted least squares.

Algebraically it is easy to see how to do the minimization. Rewrite the quantity to be minimized as

$$\sum_{i=1}^n \left[w_i^{1/2} Y_i - (w_i^{1/2} x_i)^T \beta\right]^2 .$$

This is just an ordinary least squares problem with the response variable being

$$Y_i^* = w_i^{1/2} Y_i$$

and the covariates being

$$x_i^* = w_i^{1/2} x_i \,.$$

The calculation can be written in matrix form. If $W^{1/2}$ is a diagonal matrix with $w_i^{1/2}$ in the ith diagonal position then put $Y^* = W^{1/2}Y$ and $X^* = W^{1/2}X$. Then

$$Y = X\beta + \epsilon$$

becomes

$$Y^* = X^*\beta = W^{1/2}\epsilon$$

If ϵ had mean 0, independent entries and $Var(\epsilon_i) = w_i/\sigma^2$ then $\epsilon^* = W^{1/2}\epsilon$ has mean 0, independent entries $\epsilon_i^* = w_i^{1/2}\epsilon$ and $Var(\epsilon_i^*) = \sigma^2$ so that ordinary multiple regression theory applies. The estimate of β is

$$\hat\beta_w = \left[(X^*)^T X^*\right]^{-1} (X^*)^T Y^* = (X^T W X)^{-1} X^T W Y$$

where now $W = W^{1/2}W^{1/2}$ is a diagonal matrix with w_i on the diagonal. This estimate is unbiased and has variance covariance matrix

$$\sigma^2 \left[(X^*)^T X^*\right]^{-1} = \sigma^2 (X^T W X)^{-1} \,.$$

Example

It is possible to do weighted least squares in SAS fairly easily. As an example we consider using the SENIC data set taking the variance of RISK to be proportional to 1/CENSUS. (Motivation: RISK is an estimated proportion; variance of a Binomial proportion is inversely proportional to the sample size. This makes the weight just CENSUS.

```
proc reg  data=scenic;
  model Risk = Culture Stay Nratio Chest Facil;
  weight Census;
run ;
```

EDITED OUTPUT

```
Dependent Variable: RISK
                      Analysis of Variance
                      Sum of         Mean
Source       DF      Squares       Square F Value  Prob>F
Model         5  12876.94280   2575.38856  17.819  0.0001
Error       107  15464.46721    144.52773
C Total     112  28341.41001
    Root MSE      12.02197     R-square       0.4544
    Dep Mean       4.76215     Adj R-sq       0.4289
    C.V.         252.44833
                          Parameter Estimates
              Parameter    Standard    T for H0:
Variable  DF  Estimate        Error   Parameter=0 Prob > |T|
INTERCEP   1  0.468108  0.62393433       0.750     0.4547
CULTURE    1  0.030005  0.00891714       3.365     0.0011
STAY       1  0.237420  0.04444810       5.342     0.0001
NRATIO     1  0.623850  0.34803271       1.793     0.0759
CHEST      1  0.003547  0.00444160       0.799     0.4263
FACIL      1  0.008854  0.00603368       1.467     0.1452
```

EDITED OUTPUT FOR UNWEIGHTED CASE

```
Dependent Variable: RISK
                       Analysis of Variance
                    Sum of       Mean
Source       DF    Squares      Square  F Value Prob>F
Model         5  108.32717    21.66543   24.913 0.0001
Error       107   93.05266     0.86965
C Total     112  201.37982
 Root MSE       0.93255     R-square       0.5379
 Dep Mean       4.35487     Adj R-sq       0.5163
 C.V.          21.41399
                        Parameter Estimates
             Parameter  Standard    T for H0:
Variable  DF  Estimate     Error Parameter=0 Prob > |T|
INTERCEP  1 -0.768043 0.61022741    -1.259    0.2109
CULTURE   1  0.043189 0.00984976     4.385    0.0001
STAY      1  0.233926 0.05741114     4.075    0.0001
NRATIO    1  0.672403 0.29931440     2.246    0.0267
```

```
CHEST      1  0.009179 0.00540681    1.698    0.0925
FACIL      1  0.018439 0.00629673    2.928    0.0042
```

- Notice many changes in significance levels.
- Weighted model would fail diagnostic tests – it would be clearly heteroscedastic.
- Can compute standardized residuals and so on from starred variables as usual.