STAT 350: Lecture 28

INCLUDING CATEGORICAL COVARIATES

```
options pagesize=60 linesize=80;
data scenic;
 infile 'scenic.dat' firstobs=2;
 input Stay  Age Risk Culture Chest Beds School
      Region Census Nurses Facil;
 Nratio = Nurses / Census  ;
 R1 = -(Region-4)*(Region-3)*(Region-2)/6;
 R2 = (Region-4)*(Region-3)*(Region-1)/2;
 R3 = -(Region-4)*(Region-2)*(Region-1)/2;
 S1 = School-1;
proc reg  data=scenic;
  model Risk = S1 Culture Stay Nurses Nratio { R1 R2 R3 }
  Chest Beds Census Facil / selection=stepwise
  groupnames = 'School' 'Culture' 'Stay' 'Nurses' 'Nratio'
  'Region' 'Chest' 'Beds' 'Census' 'Facil';
run ;
```

EDITED SAS OUTPUT

```
      Stepwise Procedure for Dependent Variable RISK
Step 1 Group Culture Entered R-square=0.312659 C(p)=58.3641
            Parameter   Standard    Type II
Variable     Estimate      Error Sum of Squares   F  Prob>F
INTERCEP   3.19789965 0.19376813 339.64905575 272.37 0.0001
--- Group Culture  --             62.96314170  50.49 0.0001
CULTURE    0.07325862 0.01030975  62.96314170  50.49 0.0001
-----------------------------------------------------------
Step 2 Group Stay Entered R-square=0.45040256 C(p)=26.82419
            Parameter   Standard   Type II
Variable     Estimate      Error Sum of Squares F  Prob>F
INTERCEP   0.80549102 0.48775579  2.74400250  2.73 0.1015
--- Group Culture  ---           33.39687778 33.19 0.0001
CULTURE    0.05645147 0.00979843 33.39687778 33.19 0.0001
--- Group Stay     ---           27.73884588 27.57 0.0001
STAY       0.27547211 0.05246473 27.73884588 27.57 0.0001
```

```
-----------------------------------------------------------
Step 3 Group Facil Entered R-square = 0.4934 C(p)=18.3545
            Parameter   Standard   Type II
Variable     Estimate      Error Sum of Squares F  Prob>F
INTERCEP   0.49133226 0.48163614  0.97401801  1.04 0.3099
--- Group Culture  ---           30.59827862 32.69 0.0001
CULTURE    0.05419997 0.00947933 30.59827862 32.69 0.0001
--- Group Stay     ---           16.47664606 17.60 0.0001
STAY       0.22390748 0.05336561 16.47664606 17.60 0.0001
--- Group Facil    ---            8.65883687  9.25 0.0029
FACIL      0.01963027 0.00645392  8.65883687  9.25 0.0029
---------------------------------------------------------
Step 4 Group Nratio Entered R-square=0.52548 C(p)=12.5433
            Parameter   Standard   Type II
Variable     Estimate      Error Sum of Squares F  Prob>F
INTERCEP  -0.49505513 0.59376426  0.61507231  0.70 0.4063
--- Group Culture  ---           22.84513509 25.82 0.0001
CULTURE    0.04818092 0.00948204 22.84513509 25.82 0.0001
--- Group Stay     ---           21.44995791 24.24 0.0001
STAY       0.26758404 0.05434637 21.44995791 24.24 0.0001
--- Group Nratio   ---            6.46014750  7.30 0.0080
NRATIO     0.79262357 0.29333869  6.46014750  7.30 0.0080
--- Group Facil    ---            6.75349077  7.63 0.0067
FACIL      0.01747585 0.00632554  6.75349077  7.63 0.0067
---------------------------------------------------------
Step 5 Group Chest Entered R-square=0.53792 C(p)=11.513
            Parameter  Standard    Type II
Variable   Estimate     Error   Sum of Squares  F  Prob>F
INTERCEP  -0.76804342 0.61022741  1.37763165  1.58 0.2109
--- Group Culture  ---           16.71979631 19.23 0.0001
CULTURE    0.04318856 0.00984976 16.71979631 19.23 0.0001
--- Group Stay     ---           14.43814950 16.60 0.0001
STAY       0.23392650 0.05741114 14.43814950 16.60 0.0001
--- Group Nratio   ---            4.38883521  5.05 0.0267
NRATIO     0.67240318 0.29931440  4.38883521  5.05 0.0267
--- Group Chest    ---            2.50619510  2.88 0.0925
CHEST      0.00917860 0.00540681  2.50619510  2.88 0.0925
--- Group Facil    ---            7.45710068  8.57 0.0042
```

```
FACIL      0.01843860 0.00629673  7.45710068  8.57 0.0042
---------------------------------------------------------
Step 6 Group Region Entered R-square=0.56826 C(p)=10.1269
            Parameter  Standard   Type II
Variable     Estimate     Error  Sum of Squares F Prob>F
INTERCEP  -0.66156855 0.68931767  0.77004723  0.92 0.3394
--- Group Culture  ---          19.41848300 23.23 0.0001
CULTURE    0.04717749 0.00978882 19.41848300 23.23 0.0001
--- Group Stay     ---          18.64724032 22.31 0.0001
STAY       0.28408192 0.06015054 18.64724032 22.31 0.0001
--- Group Nratio   ---           1.86769604  2.23 0.1380
NRATIO     0.47735146 0.31936579  1.86769604  2.23 0.1380
--- Group Region   ---           6.10861501  2.44 0.0689
R1        -0.91152625 0.33831556  6.06877293  7.26 0.0082
R2        -0.61170886 0.30630883  3.33408744  3.99 0.0484
R3        -0.54005754 0.30531855  2.61565335  3.13 0.0799
--- Group Chest    ---           3.10587423  3.72 0.0566
CHEST      0.01029102 0.00533912  3.10587423  3.72 0.0566
--- Group Facil    ---           7.66252029  9.17 0.0031
FACIL      0.01883340 0.00622080  7.66252029  9.17 0.0031
--------------------------------------------------------
Step 7 Group School Entered R-square=0.5783 C(p)= 9.68028
            Parameter  Standard    Type II
Variable     Estimate     Error Sum of Squares F Prob>F
INTERCEP  -1.29313397 0.79443852  2.18445103  2.65 0.1066
--- Group School   ---           2.02343484  2.45 0.1203
S1         0.45874175 0.29282732  2.02343484  2.45 0.1203
--- Group Culture  ---          21.14238169 25.64 0.0001
CULTURE    0.05016596 0.00990650 21.14238169 25.64 0.0001
--- Group Stay     ---          19.90843811 24.15 0.0001
STAY       0.29583936 0.06020399 19.90843811 24.15 0.0001
--- Group Nratio   ---           1.42881407  1.73 0.1909
NRATIO     0.42026288 0.31924279  1.42881407  1.73 0.1909
--- Group Region   ---           7.09035688  2.87 0.0402
R1        -0.99737538 0.34041455  7.07745167  8.58 0.0042
R2        -0.64425716 0.30489819  3.68115979  4.46 0.0370
R3        -0.59950685 0.30557155  3.17349874  3.85 0.0525
--- Group Chest    ---           2.85453005  3.46 0.0656
```

```
CHEST      0.00987802 0.00530873  2.85453005  3.46 0.0656
--- Group Facil    ---            9.68526975 11.75 0.0009
FACIL      0.02391008 0.00697611  9.68526975 11.75 0.0009
----------------------------------------------------------
Step 8 Group Nratio Removed R-square=0.57121 C(p)=9.40791
            Parameter   Standard    Type II
Variable     Estimate      Error Sum of Squares F Prob>F
INTERCEP  -0.83240584 0.71570292  1.12313185  1.35 0.2475
--- Group School   ---            2.46231681  2.97 0.0880
S1         0.50274483 0.29193670  2.46231681  2.97 0.0880
--- Group Culture  ---           23.66688888 28.50 0.0001
CULTURE    0.05233635 0.00980270 23.66688888 28.50 0.0001
--- Group Stay     ---           18.47964968 22.26 0.0001
STAY       0.27469386 0.05822575 18.47964968 22.26 0.0001
--- Group Region   ---            9.68716458  3.89 0.0111
R1        -1.10696516 0.33123989  9.27275385 11.17 0.0012
R2        -0.76673818 0.29137725  5.74922078  6.92 0.0098
R3        -0.75936643 0.28139304  6.04647398  7.28 0.0081
--- Group Chest    ---            3.92124933  4.72 0.0320
CHEST      0.01132621 0.00521177  3.92124933  4.72 0.0320
--- Group Facil    ---           11.30278424 13.61 0.0004
FACIL      0.02545939 0.00690031 11.30278424 13.61 0.0004
----------------------------------------------------------
All groups of variables left in the model are
significant at the 0.1500 level.  No other group
of variables met the 0.1500 significance level for
entry into the model.
  Summary of Stepwise Procedure for Dependent Variable RISK
      Group          Number Partial Model
Step Entered Removed  In    R**2    R**2    C(p)       F  Prob>F
 1   Culture          1    0.3127  0.3127 58.3641 50.4918 0.0000
 2   Stay             2    0.1377  0.4504 26.8242 27.5690 0.0000
 3   Facil            3    0.0430  0.4934 18.3545  9.2513 0.0029
 4   Nratio           4    0.0321  0.5255 12.5433  7.3012 0.0080
 5   Chest            5    0.0124  0.5379 11.5130  2.8818 0.0925
 6   Region           8    0.0303  0.5683 10.1269  2.4357 0.0689
 7   School           9    0.0100  0.5783  9.6803  2.4542 0.1203
 8           Nratio   8    0.0071  0.5712  9.4079  1.7330 0.1909
```

COMMENTS ON OUTPUT

- Final model selected has variables SCHOOL, CULTURE, STAY, REGION, CHEST and FACIL.
- Variable NRATIO included at step 4 was eliminated at step 8.
- `groupnames` assigns names to groups of variables.

Theory underlying C_p, R^2_{adj}

- C_p: Based on a trade off of bias and variance.
  - Start with full set of covariates $X_1, \ldots, X_{P-1}$. Choose subset of size $p-1$ of the possible $P-1$. Define

$$C_p = \frac{SSE_p}{MSE(X_1, \ldots, X_{P-1})} - (n-2p)$$

  - Motivation: assume full model is "correct" – there are coefficients $\beta_0, \ldots, \beta_{P-1}$ such that the errors in

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_{P-1} X_{P-1} + \epsilon_i$$

    are independent, mean 0 and homoscedastic. Consider fitted value $\hat{\mu}_i$ based on subset of regressors. Can work out total mean squared prediction error

$$\sum \left[(E(\hat{\mu}_i) - \mu_i)^2 + Var(\hat{\mu}_i) \right]$$

    and discover that C_p is a reasonable estimator of this quantity. Idea is: for model with too few parameters the fitted values are biased so first term large while for model with too many parameters subtracted term is smaller so C_p is bigger.
  - Note: if all β values but for set of $p-1$ are 0 then SSE_p should be about $\sigma^2(n-p)$ while $MSE(X_1, \ldots, X_{P-1})$ should be around σ^2 so that C_p is close to $(n-p)-(n-2p) = p$.

- R^2_{adj} is based on the idea of using the model which leads to the smallest estimate $ESS/(n-p)$ of σ^2. In general

$$R^2 = 1 - \frac{\text{Error SS}}{\text{Total SS}} = 1 - \frac{n-p}{n-1}\frac{\text{MSE}}{S_Y^2}$$

  The adjustment is to cancel the factor $(n-p)/(n-1)$ so that

$$R^2_{\text{adj}} = 1 - \frac{\text{MSE}}{S_Y^2}.$$

Power and Sample Size Calculations

Up to now our theory has been used to compute P-values or fix critical points to get desired α levels. We have assumed that all our null hypotheses are True. I now discuss power or Type II error rates of our tests. Read Chapter 26, section 4, 5 and 6.

Consider a t-test of $\beta_k = 0$. The test statistic is

$$\frac{\hat\beta_k}{\sqrt{MSE(X^TX)^{-1}_{kk}}}$$

which can be rewritten as the ratio

$$\frac{\hat\beta_k / \left[\sigma\sqrt{(X^TX)^{-1}_{kk}}\right]}{\sqrt{[SSE/\sigma^2]/(n-p)}}$$

When the null hypothesis that $\beta_k = 0$ is true the numerator is standard normal, the denominator is the square root of a chi-square divided by its degrees of freedom and the numerator and denominator are independent. When, in fact β_k is not 0 the numerator is still normal and still has variance 1 but its mean is

$$\delta = \frac{\beta_k}{\sigma\sqrt{(X^TX)^{-1}_{kk}}}.$$

This leads us to define the **non-central** t distribution as the distribution of

$$\frac{N(\delta,1)}{\sqrt{\chi^2_\nu/\nu}}$$

where the numerator and denominator are **independent**. The quantity δ is the **noncentrality parameter**.

Table B.5 on page 1346 gives the probability that the absolute value of a non-central t exceeds a given level. If we take the level to be the critical point for a t test at some level α then the probability we look up is the corresponding **power**, that is, the probability of rejection. Notice that the power depends on two unknown quantities, β_k and σ and on 1 quantity which is sometimes under the experimenter's control (in a designed experiment) and sometimes not (as in an observational study.)

Same idea applies to any linear statistic of the form $a^T\hat\beta$ – you get a non-central t distribution on the alternative. So, for example, if testing $a^T\beta = a_0$ but in fact $a^T\beta = a_1$ the non-centrality parameter is

$$\delta = \frac{a_1 - a_0}{\sigma\sqrt{a^T(X^TX)^{-1}a}}.$$

Sample Size determination

Before an experiment is run it is sensible, if the experiment is costly, to try to work out whether or not it is worth doing. You will nly do an experiment if the probability of Type I and II errors are both reasonably low. The simplest case arises when you prespecify a level, say $\alpha = 0.05$ and an acceptable probability of Type II error, β say 0.10. Then you need to specify

- The ratio β/σ; this value comes from a physically motivated understanding of what value of β would be important to detect and from some understanding of the roughly what values might be reasonable for σ.

- How the design matrix would depend on the sample size. The easiest thing is to fix some small set of say j values $x_1, \ldots, x_j$ and then use each member of that set say m times so that the aggregate sample size is mk. This gives a non-centrality parameter of the form

  $$\frac{\beta}{\sigma} \times \frac{\sqrt{m}}{\sqrt{(X^TX)^{-1}_{kk}}}$$

  The value $n = mk$ influences both the row in table B.5 which should be used and the value of δ. If the solution is large, however, then all the rows in B.5 at the bottom of the table are very similar so that effectively only δ depends on n; we can then solve for n.

F tests

The simplest example of the power of an F test arises in regression through the origin (that is, a model with no intercept term.) Consider the model

$$Y_i = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$$

To test $\beta_1 = \cdots = \beta_p = 0$ we use the F statistic

$$F = \frac{MSR}{MSE} = \frac{\hat{Y}^T \hat{Y}/p}{\hat{\epsilon}^T \hat{\epsilon}} = \frac{Y^T HY/p}{Y^T(I-H)Y/(n-p)}.$$

Suppose now that the null hypothesis is false. Substitute $Y = X\beta + \epsilon$ in the formula for the F statistic. Use the fact that $HX = X$ (and so $(I-H)X = 0$) to see that the denominator is

$$\frac{\epsilon^T(I-H)\epsilon}{n-p}$$

This shows that even when the null hypothesis is false the denominator divided by σ^2 has the distribution of a χ^2 on $n-p$ degrees of freedom divided by its degrees of freedom. It is also true that the numerator and denominator are independent of each other even when the null hypothesis is false.

The numerator, however, is

$$\frac{(\epsilon + X\beta)^T H(\epsilon + X\beta)}{p}$$

Dividing by σ^2 we can rewrite this as

$$W^T HW/p$$

where $W = (\epsilon + X\beta)/\sigma$ has a multivariate normal distribution with mean $X\beta/\sigma = \mu/\sigma$ and variance the identity matrix.
FACT:

If W is a $MVN(\tau, I)$ random vector and Q is idempotent with rank p then $W^T QW$ has a **non-central** χ^2 distribution with non-centrality parameter

$$\delta^2 = E(W) - p = \tau^T Q\tau$$

and p degrees of freedom. This is the same distribution as that of

$$(Z_1 + \delta)^2 + Z_2^2 + \cdots + Z_p^2$$

where the Z_i are iid standard normals. An ordinary χ^2 variable is called **central** and has $\delta = 0$.

FACT

If U and V are independent χ^2 variables with degrees of freedom ν_1 and ν_2, V is central and U is non-central with non-centrality parameter δ^2 then

$$\frac{U/\nu_1}{V/\nu_2}$$

is said to have a **non-central** F distribution with non-centrality parameter δ^2 and degrees of freedom ν_1 and ν_2.

POWER CALCULATIONS

Table B 11 gives powers of F tests for various small numerator degrees of freedom and a range of denominator degrees of freedom for $\alpha = 0.05$ or $\alpha = 0.01$. In the table ϕ is simply our $\delta/\sqrt{p+1}$ (that is, the square root of what I called the non-centrality parameter divided by the square root of 1 more than the numerator degrees of freedom.)

SAMPLE SIZE CALCULATIONS

Sometimes done with charts and sometimes with tables; see table B 12. This table depends on a quantity

$$\frac{\Delta}{\sigma} = \sqrt{\frac{(p+1)\delta^2}{n}}$$

To use the table you specify an α (one of 0.2, 0.1, 0.05 or 0.01) and a power ($= 1 - \beta$ in the notation of the table) which must be one of 0.7, 0.8, 0.9 or 0.95 and a value of non-centrality per data point, that is of δ^2/n. Then you look up n. Realistic specification of δ^2/n is **difficult**. in practice.