Biost 518
Applied Biostatistics II

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 10:
Multiple Regression:
Choice of Model

January 29, 2007

Lecture Outline

- Choice of Model
  - Alternative Models
  - Effect of data driven selection of model

Choice of Model

Choice of Model for Analysis

- Compare power of linear continuous versus ANOVA as a function
  - of trend in means and
  - standard errors within groups

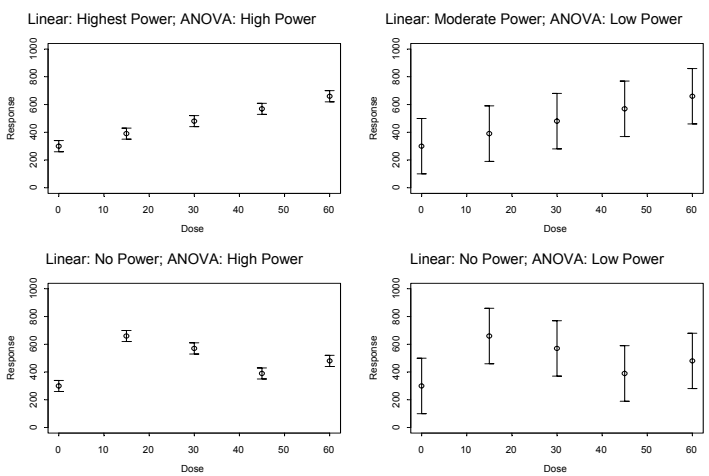
ANOVA (dummy variables)

- Fits group means exactly
- Does not mix “random error” with “systematic error:
- Ignores the ordering of the groups, so it gains no power from trends
  - The same level of significance is obtained no matter what permutation of dose groups is considered

Linear Continuous Models

- Borrows information across groups
  - Accurate, efficient if model is correct
- If model incorrect, mixes “random” and “systematic” error
- Can gain power from ordering of groups in order to detect a trend
  - But, no matter how low the standard error is, if there is no trend in the mean, there is no statistical significance

Hypothetical Settings



Other Options

- We can model continuous variables with other flexible models
  - Combinations of linear trends and indicator variables
  - Splines

Choice of Transformation

- The exact form used to model predictors should be based on the following criteria
  - Scientific issues
    - Scientific question to be addresses
    - Role of predictors in the scientific question
    - Ease of interpretation
  - Statistical issues
    - Accuracy of model
    - Precision of model
      - Parsimony

Scientific Issues

- The form used to model predictors must address the specific scientific question
  - Should be the next logical step in the process of investigating the overall goal: Role of binary search
    - Establishing some sort of an association
    - Detecting a first order trend
    - Detecting specific forms of nonlinearities
      - Effect above lowest dose?
      - U – shaped trend?
    - More complex models

Scientific Issues

- When the scientific question relates to prediction, it is imperative that the regression model accurately reflect the true relationship between predictors and the modeled summary measure of response
  - Failure to have the “correct” model will mean that some groups may not have the correct responsepre dicted

Scientific Issues

- When the scientific question relates to detection of associations, the importance of having the “true” model depends on the statistical role of the predictor

Scientific Issues

- With the predictor or interest, the most important issue is to protect the validity of statistical inference
  - Data driven decisions inflate the type I error
- With precision variables, it is not as crucial that the "true" relationship be modeled
  - Approximate models provide most of the precision
- With confounders, failure to accurately model the relationship between confounder and response may leave some residual confounding

Scientific Issues

- As the goal is to communicate your findings to the greater scientific community, it is also very important that your modeling of predictors be easy to understand
  - This is an issue that matters the most for the predictor of interest
    - We are not generally interested in making inference about precision variables or confounders

Statistical Issues

- The greatest statistical precision will be gained when the model reflects the true relationship between the predictor and response
  - Accurate modeling of the relationship will avoid including systematic error in the estimates of standard errors
  - Parsimony: Using the fewest parameters to model the relationship will allow greater precision

Statistical Issues

- We should select the form of modeling the predictor before looking at the data
  - Data driven selection of transformations will tend to lead to inaccurate (anti-conservative) statistical inference
    - "Overfitting" of the data leads to spuriously low estimates of within group variability
    - Thus standard error estimates are too low
    - Type I errors are inflated
    - Confidence intervals are too narrow
      - Inaccurate coverage probabilities

Ex: Beta Carotene Supplements

- Phase II cancer prevention study of beta carotene supplementation
  - Randomized clinical trial of placebo or four doses (15, 30, 45, 60 mg/day) of beta carotene
  - Plasma beta carotene measured at
    - baseline (`carot0`), and
    - after nine months of treatment (`carot3`)

Ex: Descriptive Statistics

```
. tabstat carot3, by(dose) stat(n mean sd min q max)
```

dose	N	mean	sd	min	p25	p50	p75	max
0	7	186	88	85	126	149	286	323
15	8	1254	570	577	695	1250	1771	2019
30	9	1505	479	849	1157	1499	1840	2249
45	7	1749	579	950	993	1848	2248	2310
60	9	1878	430	1233	1725	1865	1918	2855
Total	40	1350	734	85	800	1529	1915	2855

Response Variable

- In this randomized clinical trial, we can consider
  - Plasma level at the end of treatment,
  - Change in plasma level over the treatment period,
  - Either of the above adjusted for baseline.
    - "ANCOVA model"

Accounting for Baseline

- Notation in randomized clinical trial

Dose group i; subject j; time t

$$Y_{ijt} \sim \left(\mu_{it}, \sigma^2\right); \quad \operatorname{corr}\left(Y_{ij0}, Y_{ij9}\right) = \rho$$

$$\overline{Y}_{i\bullet 9} \sim \left(\mu_{i9}, \sigma^2 / n\right)$$

$$\overline{Y}_{i\bullet 9} - \overline{Y}_{i\bullet 0} \sim \left(\mu_{i9} - \mu_{i0}, 2\sigma^2\left(1-\rho\right)/ n\right)$$

$$\overline{Y}_{i\bullet 9} - \rho\overline{Y}_{i\bullet 0} \sim \left(\mu_{i9} - \rho\mu_{i0}, \sigma^2\left(1-\rho^2\right)/ n\right)$$

Contrast Across Dose Groups

- Equal means at baseline by randomization

By randomization : $\mu_{Tx,0} = \mu_{Plc,0}$

$$\bar{Y}_{Tx,\bullet 9} - \bar{Y}_{Plc,\bullet 9} \sim \left(\mu_{Tx,9} - \mu_{Plc,9}, 2\sigma^2 / n\right)$$
$$\left(\bar{Y}_{Tx,\bullet 9} - \bar{Y}_{Tx,\bullet 0}\right) - \left(\bar{Y}_{Plc,\bullet 9} - \bar{Y}_{Plc\bullet 0}\right) \sim \left(\mu_{Tx,9} - \mu_{Plc,9}, 4\sigma^2(1-\rho)/n\right)$$
$$\left(\bar{Y}_{Tx,\bullet 9} - \rho\bar{Y}_{Tx,\bullet 0}\right) - \left(\bar{Y}_{Plc,\bullet 9} - \rho\bar{Y}_{Plc\bullet 0}\right) \sim \left(\mu_{Tx,9} - \mu_{Plc,9}, 2\sigma^2\left(1-\rho^2\right)/n\right)$$

Simple Linear Regression

- Regress *Y* on *X*

$$Y_i \sim \left(\mu_Y, \sigma_Y^2\right); \quad X_i \sim \left(\mu_X, \sigma_X^2\right)$$
$$\operatorname{corr}\left(Y_i, X_i\right) = \rho$$

Regression model $E\left[Y_i \mid X_i\right] = \beta_0 + \beta_1 X_i$

$$\beta_0 = \mu_Y - \beta_1 \mu_X$$
$$\beta_1 = \rho \frac{\sigma_Y}{\sigma_X}$$

Analysis of Covariance

- Notation in randomized clinical trial

Dose group i; subject j; time t

$$Y_{ijt} \sim \left(\mu_{it}, \sigma^2\right); \quad \operatorname{corr}\left(Y_{ij0}, Y_{ij9}\right) = \rho$$

Regression model

$$E\left[Y_{ij9} \mid Y_{ij0}\right] = \beta_0 + \beta_1 Y_{ij0}$$
$$\beta_1 = \rho$$

Inferential Statistics

- In a randomized clinical trial, we will tend to have greatest precision if we adjust for baseline as a predictor in a linear regression model
  - An aside: If we constrain the maximum number of measurements
    - Better to have twice as many people with only a follow-up measurement if correlation less than 0.71

Modeling Dose Response

- A wide variety of models might be considered for examining the relationship between dose and plasma levels
  - Dummy variables (modeling each dose group independently without "borrowing information" across groups) (ANOVA)
  - Linear continuous predictors (untransformed or transformed)
  - Dichotomization (at any of several thresholds)
  - Polynomials, splines, etc.
  - Combinations of the above

Comparing Models

- I will compare possible models
  - Graphically
    - Show data and fitted values <u>without</u> adjustment for baseline
  - Numerically
    - Show regression estimates and tests <u>after</u> adjustment for baselin

Stata: "Predicted Values"

- After computing a regression model, Stata will provide "predicted values" for each case
  - Covariates times regression parameter estimates for each case
  - "`predict` *`varname`*"

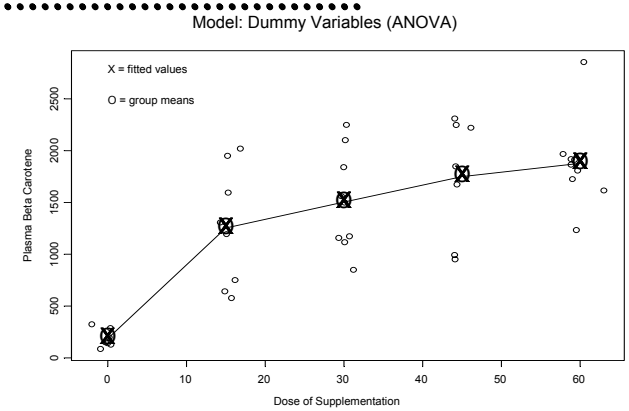
Dummy Variables (ANOVA)

- Fits each group independently
  - Does not use the ordering of the dose groups when looking for an effect
    - (But here we expect a continuous effect: Larger plasma levels with increasing dose.)
      - (We will have less power to detect an effect than if we looked for a trend)

Ex: ANOVA

```
. xi: regress carot3 i.dose carot0, robust
Linear regression                      Number of obs =      40
                                       F(  5,    34) =   47.68
                                       Prob > F      =  0.0000
                                       R-squared     =  0.7184
                                       Root MSE      =  417.46
                                       Robust
          |
    carot3 |  Coef   SE      t    P>|t|   [95% C I]
 _Idose_15 |  1224 214    5.73   0.000   790   1658
 _Idose_30 |  1440 156    9.24   0.000  1123   1756
 _Idose_45 |  1679 167   10.04   0.000  1339   2019
 _Idose_60 |  1791 153   11.71   0.000  1480   2102
    carot0 |  1.90 .54    3.54   0.001  .811   2.99
     _cons | -361. 168   -2.16   0.038  -702  -21.0
```


Ex: ANOVA



Testing for Dose Effect

- We must use special command, because model includes baseline
  - Not related to dose
- Stata allows "wildcards" when testing multiple parameters using **"testparm"**

```
. testparm _I*
 ( 1)  _Idose_15 = 0
 ( 2)  _Idose_30 = 0
 ( 3)  _Idose_45 = 0
 ( 4)  _Idose_60 = 0
       F(  4,    34) =   59.47
            Prob > F =    0.0000
```


ANOVA Comments

- We would have had the same fitted values (and thus inference) had we dropped a different dose group
  - Example: Making my own dummy variables indicating each dose

Ex: ANOVA

```
. regress carot3 dose0 dose15 dose30 dose45 carot0
Linear regression                      Number of obs =      40
                                       F(  5,    34) =   47.68
                                       Prob > F      =  0.0000
                                       R-squared     =  0.7184
                                       Root MSE      =  417.46
------------------------------------------------------------------------
       |      Robust
carot3 |  Coef   SE      t    P>|t|     [95% C I]
 dose0 | -1791  153  -11.71   0.000  -2102  -1480
dose15 |  -567  240   -2.36   0.024  -1055  -78.6
dose30 |  -351  189   -1.86   0.071   -734   32.1
dose45 |  -112  205   -0.55   0.587   -528    304
carot0 |  1.90  .54    3.54   0.001   .811   2.99
 _cons |  1430  177    8.08   0.000   1070   1789
```


Correspondence

- Note that the parameter estimates all will lead to the same fitted values
  - E.g., intercept 1430 = -361 + 1791
- Overall F statistic, R squared, Root MSE all the same
  - These all relate to the model as a whole
- Partial t tests tend to differ
  - Making comparisons to different reference groups

DANGER: Omitting Intercept

- We could have tried to fit all five dose groups
  - Stata would have dropped one (of its choice)
    - Due to fact that five groups fit perfectly by intercept and four additional parameters
- We can get Stata to include all five dose variables if we use the "`noconstant`" option
  - Changes the meaning of the overall F statistic and R squared measures

Regression with no Intercept

```
. regress carot3 dose0 dose15 dose30 dose45 dose60 carot0,
  robust noconstant
Linear regression                      Number of obs =      40
                                       F(  6,    34) =   89.14
                                       Prob > F      =  0.0000
                                       R-squared     =  0.9370
                                       Root MSE      =  417.46
------------------------------------------------------------------------
             |               Robust
      carot3 | Coef    S      t    P>|t|     [95% C I]
       dose0 | -361   168  -2.16   0.038   -702   -21.0
      dose15 |  863   242   3.57   0.001    371    1354
      dose30 | 1078   179   6.03   0.000    715    1442
      dose45 | 1318   223   5.90   0.000    864    1771
      dose60 | 1430   177   8.08   0.000   1070    1789
      carot0 | 1.90   .54   3.54   0.001   .811    2.99
```

Correspondence of Models

- This is the same model as before
  - No intercept means each dose group compared to a mean of 0
- Fitted values will be the same
- Test of dose effect will need to test equality of all 5 dose covariates
  - NOT that those parameters are 0

Testing Dose Effect

```
. test dose0=dose15=dose30=dose45=dose60

 ( 1)  dose0 - dose15 = 0
 ( 2)  dose0 - dose30 = 0
 ( 3)  dose0 - dose45 = 0
 ( 4)  dose0 - dose60 = 0

       F(  4,    34) =   59.47
            Prob > F =    0.0000
```


Difference in Interpretation

- Overall F statistic
  - Tests removing all covariates, leaving a mean of 0
    - cf: Leaving an intercept representing an overall arbitrary mean
- Multiple R squared
  - "Percent of explained variation"
    - With intercept compares to variance around overall mean
    - Without intercept compares to variance about 0

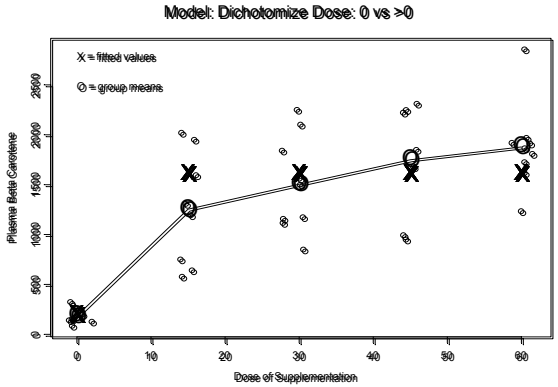
Binary: Placebo vs Active

- Dichotomizes into dose 0 versus dose > 0
  - Accurate if there is all (or virtually all) of the effect is attained at the lowest dose level
  - Often used when little is known about a treatment, or when dose is difficult to quantify
    - E.g., smoking
    - (In this case, we are relatively certain of an effect, and our major interest is probably related to dose response relationships above the lowest dose)

Binary Dose: Threshold above 0

```
. g tx= 0
. replace tx=1 if dose > 0
. regress carot3 tx carot0, robust
Linear regression          Number of obs =      40
                           F(  2,    37) =   84.00
                           Prob > F      =  0.0000
                           R-squared     =  0.6434
                           Root MSE      =   450.3
        |               Robust
carot3 | Coef    SE     t    P>|t|      [95% C I]
    tx | 1544    120 12.86   0.000   1301     1787
carot0 | 2.06   .709  2.90   0.00    .623     3.50
 _cons | -407    215 -1.89   0.067   -843     29.6
```


Ex: Beta Carotene Supplements



Testing Effect of Treatment

- Because only a single variable models dose, we can use the partial t test
  - We could also have used the "`test`" command to get the same answer

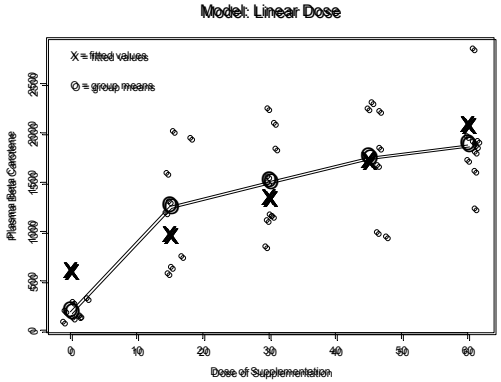
Linear Continuous Dose

- Estimates best fitting straight line to response
  - Accurate if dose-response is linear
  - Often used when little is known about a treatment and a general trend is expected
    - (In this case, we are relatively certain of an effect, and our major interest is probably related to dose response relationships above the lowest dose, but a linear relationship is not necessarily expected)
      - S – shaped trends are common in biology

Linear Continuous

```
. regress carot3 dose carot0, robust
Linear regression                      Number of obs =      40
                                       F(  2,    37) =   25.47
                                       Prob > F      =  0.0000
                                       R-squared     =  0.5622
                                       Root MSE      =  498.94
------------------------------------------------------------------------------
             |      Robust
       carot3 | Coef   SE     t    P>|t|     [95% C I]
-------------+----------------------------------------------------------------
        dose | 25.5  3.63  7.01   0.000   18.1   32.8
       carot0 | 1.33  .657  2.03   0.050   .003   2.67
        _cons |  245   223  1.10   0.279   -206.  696.
```


Ex: Beta Carotene Supplements



Testing Effect of Treatment

- Because only a single variable models dose, we can use the partial t test
  - We could also have used the "`test`" command to get the same answer

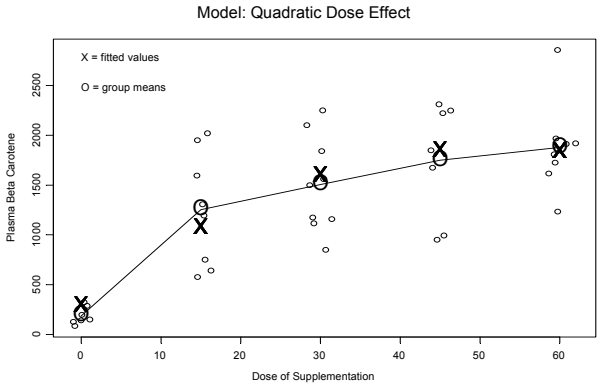
Polynomial Models of Dose

- Fit terms involving dose, dose squared
  - Higher order polynomials could be used
  - But with only k distinct levels of dose sampled, a polynomial of order k-1 is equivalent to dummy variables
  - Often used to detect U shaped trends
    - (But a quadratic is a pretty strong assumption: constant curvature- unlikely here)

Quadratic

```
. g dosesqr= dose ^2
. regress carot3 dose dosesqr carot0, robust
Linear regression          Number of obs =      40
                           F(  3,    36) =   59.30
                           Prob > F      =  0.0000
                           R-squared     =  0.6824
                           Root MSE      =  430.81
------------------------------------------------------------------------
        |               Robust
 carot3 |  Coef   SE     t     P>|t|      [95% C I]
   dose |  67.1  8.21   8.18   0.000    50.5  83.8
dosesqr | -.672  .145  -4.63   0.000   -.967 -.378
 carot0 |  1.73  .564   3.06   0.004    .584  2.87
  _cons |  -196  179.  -1.09   0.283   -559.  168.
```


Ex: Beta Carotene Supplements



Testing Effect of Treatment

- Because two variables model dose, we must use the “multiple-partial F test”

```
. test dose dosesqr

 ( 1)  dose = 0
 ( 2)  dosesqr = 0

       F(  2,    36) =   84.56
            Prob > F =    0.0000
```


Testing Linearity of Dose

- The partial t test for the dose squared term can be interpreted as a test for linear dose response
  - It is highly significantly different from 0

Ad hoc: Threshold at 0 & Linear

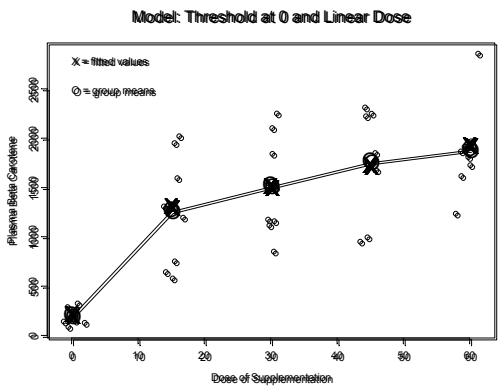
- Threshold at 0 and linear dose
  - Dummy variable for dose 0 plus dose
    - Fits group 0 by its group mean
    - Looks for a linear trend among all other dose groups
    - Allows us to address two questions:
      - Any effect of dose (test both slopes)
      - Benefit above lowest dose (test linear term's slope

Threshold and Linear Terms

```
. regress carot3 tx dose carot0, robust
Linear regression                      Number of obs =      40
                                       F(  3,    36) =   81.26
                                       Prob > F      =  0.0000
                                       R-squared     =  0.7170
                                       Root MSE      =  406.69
------------------------------------------------------------------------
       |      Robust
carot3 | Coef   SE     t    P>|t|     [95% C I]

    tx | 1051   223   4.70   0.000     598    1504
  dose | 12.8 4.79   2.67   0.011     3.09   22.5
carot0 | 1.90 .516   3.69   0.001     .857   2.95
 _cons | -362   161 -2.24   0.031    -689   -34.6
```


Ex: Beta Carotene Supplements



Testing Effect of Treatment

- Because two variables model dose, we must use the "multiple-partial F test"

```
. test tx dose

 ( 1)  tx = 0
 ( 2)  dose = 0

       F(  2,    36) =  121.88
            Prob > F =    0.0000
```

Testing Linearity of Dose

- The partial t test for the tx term can be interpreted as a test for linear dose response
  - It is highly significantly different from 0

Testing Trend Above 0

- The partial t test for the dose term can be interpreted as a test for any added effect above the lowest dose
  - It is significantly different from 0 (P=0.011)
    - (There is a multiple comparison issue here, but after testing for an overall dose effect, most people would be comfortable without adjusting for multiple comparisons)

Data Driven Selection of Model

- Suppose we look at a scatterplot before deciding which model we fit and choose a model that can fit the data well
  - If the data looks like a straight line, choose the model linear in dose
  - If the data looks like a U, choose the quadratic
  - If the data is a complicated pattern of differences among groups, we might choose dummy variables
  - Etc.

Data Driven Selection of Model

- This approach would tend to mimic the behavior of fitting several different models and choosing the lowest P values
  - When our eye sees some trend in the data, we would be most likely to pick the model giving the lowest P value

Simulated Repetitions

- I can use the 46 subjects in this study and consider 5,000 replications of the experiment in which I randomize them differently each time
  - This simulates 5,000 studies when the null hypothesis is true (no true effect of dose)
  - I then look at their baseline data using each of the five models (linear, quadratic, dichotomized at 0, dummy variables, and the combination of linear and dichotomized at 0)

Individual Model Results

- Empirical type I error (percent statistically significant) for each method of analysis individually
  - Linear dose: 0.0504
  - Quadratic dose: 0.0488
  - Dichotomized dose at 0: 0.0484
  - Dummy variables (ANOVA): 0.0498
  - Linear-dichotomized: 0.0480
  - (Precision of 95% CI for true type I error with 5,000 simulations: +/- 0.006)

Multiple Comparison Issue

- With five hypothesis tests at the nominal 0.05 level, experiment-wise type I error of 5 x 0.05 = 0.25 in worst case
  - But worst-case assumes that the errors would be mutually exclusive across analyses
    - E.g., if the linear dose model was significant, then no other analysis would be significant
  - This is unrealistic, because both the quadratic and the linear-dichotomized models also include the linear dose predictor

Multiple Comparison Results

- How many of the 5,000 simulated trials had at least one of the 5 P values less than 0.05
  - The empirical experiment-wise type I error was estimated to be 0.1204
  - 95% confidence interval for the true experiment-wise type I error (a CI for a binomial proportion when the sample size is 5,000)
    - 0.114, 0.1294

Statistical Issues

- The true type I error for such data-driven analyses will depend on several factors
  - The number of tests performed
  - The models considered
    - Similar models will tend to reject the null hypothesis on the same datasets
  - The distribution of the data
    - A tendency to heavy tailed distributions may affect the concordance between the tests

Multiple Comparison Problem

- In frequentist reasoning, we try to ensure that our error rate is at some specified level α
  - When only making one decision, this is relatively easy
  - When making multiple decisions, we must consider the "experiment-wise" error

Multiple Comparison Problem

- Worst case scenario: An error rate of α on each decision could lead to an experiment-wise error as high as $k\alpha$
  - This would be the situation if all of our errors were "mutually exclusive"
- If all our errors were independent of each other, our experiment-wise error is $1-(1-\alpha)k$

Ex: Level 0.05 per Decision

- Experiment-wise Error Rate

Number of Comparisons	Worst Case Scenario	Independent Errors
1	.0500	.0500
2	.1000	.0975
3	.1500	.1426
5	.2500	.2262
10	.5000	.4013
20	1.0000	.6415
50	1.0000	.9231

Multiple Comparison Problem

- Thus, when making multiple comparisons which all tend to address the exact same question, we tend to adjust our level of significance (or our P values) to protect our experiment-wise error

Bonferroni Correction

- Assume the worst case scenario
  - When making k comparisons
    - Test individual P values against α / k, OR
    - Multiply P values by k and compare to α
      - (But don't get absurd: P values can never be above 1)

Procedures After ANOVA

- Group comparisons after performing ANOVA
  - Tukey: All pairwise comparisons (but only pairwise comparisons)
    - May not agree with overall test in ANOVA
  - Scheffe: Arbitrary linear contrasts
    - Will agree with ANOVA
    - Allows fishing through all contrasts

General Comments

- The Bonferroni adjustment is easy and it can be applied in all settings
  - However, it is extremely conservative when the statistics from the various comparisons are positively correlated
    - Hence when making pairwise comparisons among individual levels of predictors fit as dummy variables (as in ANOVA) it is fairly typical to use other methods
      - (The positive correlation in this setting comes from making many pairwise comparisons among the same groups)

General Comments

- Post-ANOVA testing is generally done only if overall test was significant
  - Otherwise inflate type I error
- Post-ANOVA testing is subject to serious lack of power
  - We can rely on statistically significant differences
  - We cannot rely on equality of groups when not statistically significant

Choice of Model for Analysis

- So how to choose the model to use?
  - Scientific issues
    - Which question are you ready to answer?
  - Statistical issues
    - Power to detect an alternative that you care about