Biost 518
Applied Biostatistics II

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

Lecture 6:
Multiple Regression:
Overview of Uses

January 30, 2006

Lecture Outline

- Adjustment for confounders / precision
- Effect modification
- Modeling complex "dose response"
- Testing for linearity

Adjustment for Confounders, Precision Variables

Adjustment for Covariates

- We "adjust" for other covariates
  - Define groups according to
    - Predictor of interest, and
    - Other covariates
  - Compare the distribution of response across groups which
    - differ with respect to the Predictor of Interest, but
    - are the same with respect to the other covariates
      - "holding other variables constant"

Unadjusted vs Adjusted Models

- Adjustment for covariates changes the scientific question
  - Unadjusted models
    - Slope compares parameters across groups differing by 1 unit in the modeled predictor
      - Groups may also differ with respect to other variables
  - Adjusted models
    - Slope compares parameters across groups differing by 1 unit in the modeled predictor but similar with respect to other modeled covariates

Interpretation of Slopes

- Difference in interpretation of slopes

$$\text{Unadjusted Model}: \quad g[\theta \mid X_i] = \beta_0 + \beta_1 \times X_i$$

- β_1 = Compares θ for groups differing by 1 unit in X
  - (The distribution of W might differ across groups being compared)

$$\text{Adjusted Model}: \quad g[\theta \mid X_i, W_i] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

- γ_1 = Compares θ for groups differing by 1 unit in X, but agreeing in their values of W

Comparing models

$$\text{Unadjusted} \quad g[\theta \mid X_i, W_i] = \beta_0 + \beta_1 \times X_i$$

$$\text{Adjusted} \quad g[\theta \mid X_i, W_i] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

When is $\gamma_1 = \beta_1$?

When is $\hat{\gamma}_1 = \hat{\beta}_1$?

When is $se(\hat{\gamma}_1) = se(\hat{\beta}_1)$?

When is $\hat{se}(\hat{\gamma}_1) = \hat{se}(\hat{\beta}_1)$?

Linear Regression

- Difference in interpretation of slopes

$$\text{Unadjusted Model}: \quad E[Y_i \mid X_i] = \beta_0 + \beta_1 \times X_i$$

- β_1 = Diff in mean Y for groups differing by 1 unit in X
  - (The distribution of W might differ across groups being compared)

$$\text{Adjusted Model}: \quad E[Y_i \mid X_i, W_i] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

- γ_1 = Diff in mean Y for groups differing by 1 unit in X, but agreeing in their values of W

Relationships: True Slopes

- The slope of the unadjusted model will tend to be

$$\beta_1 = \gamma_1 + \rho_{XW} \frac{\sigma_W}{\sigma_X} \gamma_2$$

- Hence, true adjusted and unadjusted slopes for X are estimating the same quantity only if
  - $\rho_{XW} = 0$ (X and W are truly uncorrelated), OR
  - $\gamma_2 = 0$ (no association between W and Y after adjusting for X)

Relationships: Estimated Slopes

- The estimated slope of the unadjusted model will be

$$\hat{\beta}_1 = \hat{\gamma}_1 \left(1 + \hat{\gamma}_2 r_{XW} \left[\frac{s_W}{s_X \left(r_{YX} - r_{YW} r_{XW} \right)} \right] \right)$$

- Hence, estimated adjusted and unadjusted slopes for X are equal only if
  - $r_{XW} = 0$ (X and W are uncorrelated in the sample, which can be arranged by experimental design), OR
  - $\hat{\gamma}_2 = 0$ (which cannot be predetermined, because Y is random)

Relationships: True SE

Unadjusted Model $\left[se\left(\hat{\beta}_1\right) \right]^2 = \frac{Var\left(Y \mid X\right)}{nVar\left(X\right)}$

Adjusted Model $\left[se\left(\hat{\gamma}_1\right) \right]^2 = \frac{Var\left(Y \mid X, W\right)}{nVar\left(X\right)\left(1 - r_{XW}^2\right)}$

$$Var\left(Y \mid X\right) = \gamma_2^2 Var\left(W \mid X\right) + Var\left(Y \mid X, W\right)$$

$$\sigma_{Y|X}^2 = \gamma_2^2 \sigma_{W|X}^2 + \sigma_{Y|X,W}^2$$

Relationships: True SE

Unadjusted Model $\left[se\left(\hat{\beta}_1\right) \right]^2 = \frac{Var\left(Y \mid X\right)}{nVar\left(X\right)}$

Adjusted Model $\left[se\left(\hat{\gamma}_1\right) \right]^2 = \frac{Var\left(Y \mid X, W\right)}{nVar\left(X\right)\left(1 - r_{XW}^2\right)}$

$$Var\left(Y \mid X\right) = \gamma_2^2 Var\left(W \mid X\right) + Var\left(Y \mid X, W\right)$$

Thus, $se\left(\hat{\beta}_1\right) = se\left(\hat{\gamma}_1\right)$ if

$r_{XW} = 0$

AND

$\gamma_2 = 0$ OR $Var\left(W \mid X\right) = 0$

Relationships: Estimated SE

Unadjusted Model $\left[\hat{se}(\hat{\beta}_1)\right]^2 = \dfrac{SSE(Y|X)/(n-2)}{(n-1)s_X^2}$

Adjusted Model $\left[\hat{se}(\hat{\gamma}_1)\right]^2 = \dfrac{SSE(Y|X,W)/(n-3)}{(n-1)s_X^2(1-r_{XW}^2)}$

$$SSE(Y \mid X) = \sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \times X_i\right)^2$$
$$SSE(Y \mid X,W) = \sum\left(Y_i - \hat{\gamma}_0 - \hat{\gamma}_1 \times X_i - \hat{\gamma}_2 \times W_i\right)^2$$

Relationships: Estimated SE

Unadjusted Model $\left[\hat{se}(\hat{\beta}_1)\right]^2 = \dfrac{SSE(Y|X)/(n-2)}{(n-1)s_X^2}$

Adjusted Model $\left[\hat{se}(\hat{\gamma}_1)\right]^2 = \dfrac{SSE(Y|X,W)/(n-3)}{(n-1)s_X^2(1-r_{XW}^2)}$

Thus, $\hat{se}(\hat{\beta}_1) = \hat{se}(\hat{\gamma}_1)$ if

$r_{XW} = 0$

AND

$SSE(Y|X)/(n-2) = SSE(Y|X,W)/(n-3)$

Residual Squared Error

$$SSE(Y \mid X) = \sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \times X_i\right)^2$$
$$SSE(Y \mid X,W) = \sum\left(Y_i - \hat{\gamma}_0 - \hat{\gamma}_1 \times X_i - \hat{\gamma}_2 \times W_i\right)^2$$

When calculated on the same data :

$SSE(Y \mid X) \geq SSE(Y \mid X,W)$

Relationships: Estimated SE

$$SSE(Y \mid X) = \sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \times X_i\right)^2$$
$$SSE(Y \mid X,W) = \sum\left(Y_i - \hat{\gamma}_0 - \hat{\gamma}_1 \times X_i - \hat{\gamma}_2 \times W_i\right)^2$$

Now $\hat{\beta}_1 = \hat{\gamma}_1$ if

$\hat{\gamma}_2 = 0$, in which case $SSE(Y \mid X) = SSE(Y \mid X,W)$

OR

$r_{XW} = 0$, and $SSE(Y \mid X) > SSE(Y \mid X,W)$ if $\hat{\gamma}_2 \neq 0$

Special Cases

- Behavior of unadjusted and adjusted models according to whether
  - *X* and *W* are uncorrelated
  - *W* is associated with *Y* after adjustment for *X*

	$r_{XW} = 0$	$r_{XW} \neq 0$
$\gamma_2 \neq 0$	Precision	Confounding
$\gamma_2 = 0$	Irrelevant	Var Inflation

Precision Variables

- E.g., independence in population, or completely randomized experiment

$\rho_{XW} = 0 \qquad \gamma_2 \neq 0$

	True Value	Estimates
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 \approx \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) > se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) > s\hat{e}(\hat{\gamma}_1)$

Stratified Randomization

- Stratified randomization in a designed experiment

$r_{XW} = 0 \qquad \gamma_2 \neq 0$

	True Value	Estimates
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 = \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) = se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) > s\hat{e}(\hat{\gamma}_1)$

Confounding

- Causally associated with response and associated with POI in sample

$r_{XW} \neq 0 \qquad \gamma_2 \neq 0$

	True Value	Estimates
Slopes	$\beta_1 = \gamma_1 + \rho_{XW} \frac{\sigma_X}{\sigma_W} \gamma_2$	$\hat{\beta}_1 = \hat{\gamma}_1 \left(1 + \hat{\gamma}_2 r_{XW} \left[\frac{s_W}{s_X (r_{YX} - r_{YW} r_{XW})} \right] \right)$
Std Errs	$se(\hat{\beta}_1) \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} s\hat{e}(\hat{\gamma}_1)$

Variance Inflation

- Associated with POI in sample, but not associated with response

$r_{XW} \neq 0 \qquad \gamma_2 = 0$

	True Value	Estimates
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 = \hat{\gamma}_1 \left(1 + \hat{\gamma}_2 r_{XW} \left[\frac{s_W}{s_X (r_{YX} - r_{YW} r_{XW})} \right] \right)$
Std Errs	$se(\hat{\beta}_1) < se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) < s\hat{e}(\hat{\gamma}_1)$

Irrelevant Variables

- Uncorrelated with POI in sample, and not associated with response

$r_{XW} = 0 \qquad \gamma_2 = 0$

	True Value	Estimates
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 = \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) = se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) < s\hat{e}(\hat{\gamma}_1)$

Stata: Multiple Regression

- In Stata, we use the same commands as were used for simple regression
  - We just list more variable names
  - Interpretation of CI, P values for coefficient estimates now relate to new scientific interpretation of intercept and slopes
  - Test of entire regression model also provided
    - A test that all slopes are equal to 0

Ex: FEV and Smoking

```
. regress logfev smoker if age>=9, robust

                                   Number of obs =      439
                                   F(  1,    437) =    10.45
                                   Prob > F      =   0.0013
                                   R-squared     =   0.0212
                                   Root MSE      =   .24765
-----------------------------------------------------------
       |          Robust
logfev |    Coef. St Err      t     P>|t|    [95% CI]
smoker |    .102  .0317    3.23  0.001    .040    .165
 _cons |   1.058  .0129   81.82  0.000   1.033   1.084
```

Unadjusted Interpretation

- Intercept
  - –Geometric mean of FEV in nonsmokers is 2.88 l/sec
    - The scientific relevance is questionable here, because we do not really know the population our sample represents
      - Comparing smokers to nonsmokers is more useful than looking at either group by itself
    - (Calculations: $e^{1.058}$= 2.881)
    - (The P value is of no importance whatsoever, it is testing that the log geometric mean is 0 or that the geometric mean is 1. Why would we care?)
  - –(Because *smoker* is a binary variable, the estimate corresponds to the sample geometric mean)

Unadjusted Interpretation

- Smoking effect
  - Geometric mean of FEV is 10.8% higher in smokers than in nonsmokers (95% CI: 4.1% to 17.9% higher)
    - These results are atypical of what we might expect with no true difference between groups: P = 0.001
    - (Calculations: $e^{0.102}$= 1.108; $e^{0.040}$= 1.041; $e^{0.165}$= 1.179)
      - (Note that exp (x) is approx 1+x for x close to 0)
  - (Because smoker is a binary (0-1) variable, this analysis is nearly identical to a two sample t test allowing for unequal variances)

Ex: Adjusted for Age

```
. regress logfev smoker age if age>=9, robust

                                Number of obs =      439
                                F(  2,   437) =    82.28
                                Prob > F      =   0.0000
                                R-squared     =   0.3012
                                Root MSE      =   .20949
----------------------------------------------------------
        |         Robust
 logfev |   Coef. St Err     t    P>|t|    [95% CI]
 smoker |  -.051  .0344  -1.49  0.136  -.119     .016
    age |   .064  .0051  12.37  0.000   .053     .074
  _cons |  0.352  .0575   6.12  0.000   .239     .465
```


Age Adjusted Interpretation

- Intercept
  - –Geometric mean of FEV in newborn nonsmokers is 1.42 l/sec
    - Intercept corresponds to the log geometric mean in a group having all predictors equal to 0
    - There is no scientific relevance is here, because we are extrapolating outside our data
    - (Calculations: $e^{0.352}$= 1.422)

Age Adjusted Interpretation

- Age effect
  - –Geometric mean of FEV is 6.6% higher for each year difference in age between two groups with similar smoking status(95% CI: 5.5% to 7.6% higher for each year difference in age)
    - These results are highly atypical of what we might expect with no true difference in the geometric mean FEV between age groups having similar smoking status: P < 0.0005

Age Adjusted Interpretation

- Smoking effect
  - Geometric mean of FEV is 5.0% lower in smokers than in nonsmokers of the same age (95% CI: 12.2% lower to 1.6% higher)
    - These results are not atypical of what we might expect with no true difference between groups of the same age: P = 0.136
      - Lack of statistical significance is also evident because the confidence interval contains 1 (as a ratio) or 0 (as a percent difference)
    - (Calculations: $e^{-0.051}$= 0.950; $e^{-0.119}$= 0.888; $e^{0.016}$= 1.016)
      - (Note that exp (x) is approx 1+x for x close to 0)

Age Adjusted Comments

- Comparing unadjusted and age adjusted analyses
  - Marked difference in effect of smoking suggests that there was indeed confounding
    - Age is a relatively strong predictor of FEV
    - Age is associated with smoking in the sample
      - Mean (SD) of age in analyzed smokers: 11.1 (2.04)
      - Mean (SD) of age in analyzed nonsmokers: 13.5 (2.34)
  - Effect of age adjustment on precision
    - Lower Root MSE (.209 vs .248) would tend to increase precision of estimate of smoking effect
    - Association between smoking and age tends to lower precision
    - Net effect: Less precision (adj SE 0.034 vs unadj SE 0.031)

Ex: Adjusted for Age, Height

```
. regress logfev smoker age loght if age>=9, robust

                              Number of obs =     439
                              F(  3,   437) =  284.22
                              Prob > F      =  0.0000
                              R-squared     =  0.6703
                              Root MSE      =  .14407
-------------------------------------------------------
       |          Robust
logfev |    Coef. St Err     t    P>|t|     [95% CI]
smoker |   -.054  .0241  -2.22  0.027    -.101   -.006
   age |    .022  .0035   6.18  0.000     .015    .028
 loght |   2.870  .1280  22.42  0.000    2.618   3.121
 _cons |-11.095  .5153 -21.53  0.000  -12.107 -10.082
```

Age, Ht Adjusted Interpretation

- Intercept
  - Geometric mean of FEV in newborn nonsmokers who are 1 inch high is 0.000015 l/sec
    - Intercept corresponds to the log geometric mean in a group having all predictors equal to 0
      - Nonsmokers
      - Age 0 (newborn)
      - Log height 0 (height 1 inch)
    - There is no scientific relevance is here, because there are no such people in our sample OR the population

Age, Ht Adjusted Interpretation

- Age effect
  - Geometric mean of FEV is 2.2% higher for each year difference in age between two groups with similar height and smoking status (95% CI: 1.5% to 2.9% higher for each year difference in age)
    - These results are highly atypical of what we might expect with no true difference in the geometric mean FEV between age groups having similar height and smoking status: $P < 0.0005$
  - Note that there is clear evidence that height confounded the age effect estimated in the analysis which modeled only smoking and age
    - But there is a clear independent effect of age on FEV

Age, Ht Adjusted Interpretation

- Height effect
  - Geometric mean of FEV is 31.5% higher for each 10% difference in height between two groups with similar ages and smoking status (95% CI: 28.3% to 34.6% higher for each 10% difference in height)
    - These results are highly atypical of what we might expect with no true difference in the geometric mean FEV between height groups having similar age and smoking status: $P < 0.0005$
    - (Calculations: $1.1^{2.867} = 1.315$)
  - Note that the regression coefficient of 2.870 (95% CI 2.618 to 3.121) is consistent with the scientifically derived value of 3.0

Age, Ht Adjusted Interpretation

- Smoking effect
  - Geometric mean of FEV is 5.2% lower in smokers than in nonsmokers of the same age and height (95% CI: 9.6% to 0.6% lower)
    - These results are atypical of what we might expect with no true difference between groups of the same age and height: $P = 0.027$
    - (Calculations: $e^{-0.054} = .948$; $e^{-0.101} = .904$; $e^{-0.006} = .994$)
  - Note the wording “same age and height” even though I adjusted using a log transformation of height.
    - Equal log heights lead to equal heights

Age, Ht Adjusted Comments

- Comparing age and age-height adjusted analyses
  - No difference in effect of smoking suggests there was no more confounding after age adjustment
  - Effect of height adjustment on precision
    - Lower Root MSE (.144 vs .209) would tend to increase precision of estimate of smoking effect
    - Little association between smoking and height after adjustment for age will not tend to lower precision
    - Net effect: Higher precision (adj SE 0.024 vs unadj SE 0.034)

Effect Modification

Effect Modifier

- The association between Response and POI differs in strata defined by effect modifier
    - Statistical term: "Interaction"
  - Depends on the measurement of effect
    - Summary measure
      - Mean, geometric mean, median, proportion, odds, hazard, etc.
    - Comparison across groups
      - Difference, ratio

Analysis of Effect Modification

- When the scientific question involves effect modification, analyses must be within each stratum separately
  - If we want to estimate degree of effect modification or test for its existence:
    - A regression model will typically include
      - Predictor of interest
      - Effect modifier
      - A covariate modeling the interaction (usually product)

Model for Effect Modification

- Typical model for effect modification
  - Include "main effects" (can be bad not to)
    - *X* (or predictors that involve only *X*)
    - *W* (or predictors that involve only *W*)
  - Include "interactions"
    - Predictor(s) derived from both *X* and *W*

$$g\left[\theta \mid X_i, W_i\right] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times (XW)_i$$
$$= \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

Interpretation of Parameters

$$g\left[\theta \mid X_i, W_i\right] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Usual approach a bit more difficult
  - We can try using the idea of "comparison of θ across groups differing by 1 unit in corresponding predictor but agreeing in other modeled predictors"
  - However, terms involving two scientific variables makes this approach difficult

Intercept

$$g\left[\theta \mid X_i, W_i\right] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Interpretation of intercept straightforward
  - β_0 corresponds to *X*= 0, *W*= 0
    - May not be scientifically meaningful

Slopes for Main Effects

$$g\left[\theta \mid X_i, W_i\right] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Interpretation of main effects
  - β_X corresponds to 1 unit difference in *X* holding *W* and *(X×W)* constant
    - So 1 unit difference in *X* when *W*= 0
    - May not be scientifically meaningful
  - β_W corresponds to 1 unit difference in *W* holding *X* and *(X×W)* constant
    - So 1 unit difference in *W* when *X*= 0
    - May not be scientifically meaningful

Slope for interaction

$$g[\theta \mid X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Interpretation of interaction difficult
  - β_{XW} corresponds to 1 unit difference in *(X×W)* holding *X* and *W* constant
    - Impossible, so we need another way to interpret this slope parameter

Consider Scientific Predictors

$$g[\theta \mid X_i, w] = \beta_0 + \beta_X \times X_i + \beta_W \times w + \beta_{XW} \times X_i \times w$$
$$= (\beta_0 + \beta_W \times w) + (\beta_X + \beta_{XW} \times w) \times X_i$$

In stratum with $W = w$

Intercept : $(\beta_0 + \beta_W \times w)$ corresponds to $X_i = 0$

Slope : $(\beta_X + \beta_{XW} \times w)$ compares groups differing by 1 unit in X

β_{XW} is difference in X slope per 1 unit difference in W

Consider Scientific Predictors

$$g[\theta \mid x, W_i] = \beta_0 + \beta_X \times x + \beta_W \times W_i + \beta_{XW} \times x \times W_i$$
$$= (\beta_0 + \beta_X \times x) + (\beta_W + \beta_{XW} \times x) \times W_i$$

In stratum with $X = x$

Intercept : $(\beta_0 + \beta_X \times x)$ corresponds to $W_i = 0$

Slope : $(\beta_W + \beta_{XW} \times x)$ compares groups differing by 1 unit in W

β_{XW} is difference in W slope per 1 unit difference in X

Symmetry of Effect Modification

- Note that if X modifies the association between Y and W, then W modifies the association between Y and X
  - Aside: Confounding need not be symmetric
    - W can confound the association between Y and X, but X not confound the association between Y and W
      - W and X associated in the sample
      - Y and X not associated after adjusting for W
      - Y and W associated after adjusting for X

Inference for Effect Modification

$$g[\theta \mid X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- No effect modification if $\beta_{XW} = 0$
  - Hence, inference about existence of effect modification tests that $\beta_{XW} = 0$
    - We can perform such inference using standard regression output for the corresponding slope parameter

Inference for Main Effect Slope

$$g[\theta \mid X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Interpretation of $\beta_X = 0$
  - Same intercept in all strata defined by *W*
  - Generally a very uninteresting question
  - We rarely make inference on main effect slopes by themselves

Inference About Effect of X

$$g[\theta \mid X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Response parameter not associated with *X* if $\beta_X = 0$ AND $\beta_{XW} = 0$
  - We will need to construct special tests that both parameters are simulataneously 0
    - The t tests given in regression output consider only one slope parameter at a time

Stata: Testing Multiple Slopes

- Stata has easy method for performing test that multiple parameters are simultaneously 0
  - Perform any regression command
  - Then use "`test` *`var1 var2`* …"
    - Provides P value of the hypothesis test based on most recently executed regression command ***of any type of regression***

Ex: Salary by Sex and Admin

- Does sex modify the association between mean salary and administrative duties
  - With two binary variables, modeling interaction by product is the obvious choice

$$\mathrm{E}\left[Sal \mid Fem, Adm\right] = \beta_0 + \beta_A \times Adm_i + \beta_F \times Fem_i + \beta_{AF} \times Adm_i \times Fem_i$$

Ex: Stata output

```
. g admfem= admin * female
. regress salary admin female admfem if year==95,
Linear regression              Number of obs =    1597
                               F(  3,  1593) =  125.26
                               Prob > F      =  0.0000
                               R-squared     =  0.1615
                               Root MSE      =  1866.9
-----------------------------------------------------
       |            Robust
salary |     Coef.  StdErr  t     P>|t|    [95% CI]
 admin |  1951.378    176  11.06 0.000  1605   2297
female | -1226.234     95 -12.86 0.000 -1413  -1039
admfem | -461.9072    342  -1.35 0.177 -1132    208
 _cons |  6506.607     62 105.25 0.000  6385   6627
```


Ex: Descriptive Statistics

- Note that with two binary variables, the regression parameters agree exactly with the corresponding group sample means

```
. table admin female if year==95, co(mean salary)

          |       female
    admin |     Male    Female
----------------------------
 Nonadmin | 6506.607  5280.373
    Admin | 8457.985  6769.844
```


Ex: Inference About Eff Mod

- Does sex modify association between mean salary and administrative duties?
  - Estimate that the "administrative supplement" averages $462 less for women than men
    - 95% CI: $1132 less to $208 more
    - Not statistically significant: P = 0.177

```
|                   Robust
salary |     Coef.  StdErr  t     P>|t|    [95% CI]
 admin |  1951.378    176  11.06 0.000  1605   2297
female | -1226.234     95 -12.86 0.000 -1413  -1039
admfem | -461.9072    342  -1.35 0.177 -1132    208
 _cons |  6506.607     62 105.25 0.000  6385   6627
```

Ex: Inference About Sex Assoc

- Is sex associated with mean salary?
  - Need to test that slope parameters for `female` and `admfem` are simultaneously 0

```
. test female admfem
 ( 1)   female = 0
 ( 2)   admfem = 0

        F(  2,  1593) =   95.90
             Prob > F =    0.0000
```


Ex: Inference for Admin Assoc

- Are administrative duties associated with mean salary?
  - Need to test that slope parameters for `admin` and `admfem` are simultaneously 0

```
. test admin admfem
 ( 1)   admin = 0
 ( 2)   admfem = 0

        F(  2,  1593) =   74.15
             Prob > F =    0.0000
```


Continuous Predictors

- Modeling interactions with continuous predictors is conceptually more complicated
  - Is a multiplicative interaction at all a reasonable model for the data?
  - Nonetheless, this is the most common way we detect interactions
    - I would caution against using the model as predictions without carefully examining the data
      - But this can be difficult, too

Example: SEP “Normal Ranges”

- We want to find normal ranges for somatosensory evoked potential (SEP)
  - As a first step, we want to consider important predictors of nerve conduction times
    - If any variables such as sex, age, height, race, etc. are important predictors of nerve conduction times, then it would make most sense to obtain normal ranges within such groups

Example: SEP “Normal Ranges”

- Scientifically, we might expect that height, age, and sex are related to the nerve conduction time
  - Nerve length should matter, and height is a surrogate for nerve length
  - Age might affect nerve conduction times: People slow down with age
  - Sex: Men are SO fragile

Example: SEP “Normal Ranges”

- Prior to looking at the data, we can also consider the possibility that interactions between these variables might be important
  - Height - age interaction?
    - Do we expect the difference in conduction times between 6 foot tall and 5 foot tall 20 year olds to be the same as the difference in conduction times between 6 foot tall and 5 foot tall 80 year olds?

Example: SEP “Normal Ranges”

- We might suspect such an interaction due to the fact that height may not be as good a surrogate for nerve length in older people
  - With age, some people tend to shrink due to osteoporosis and compression of intervertebral discs
    - It is not clear that nerve length would be altered in such a process

Example: SEP “Normal Ranges”

- Thus, in young people, differences in height probably are a better measure of nerve length than in old people
    - Tall old people probably have been tall always
    - Short old people will include some who were much taller when they were young

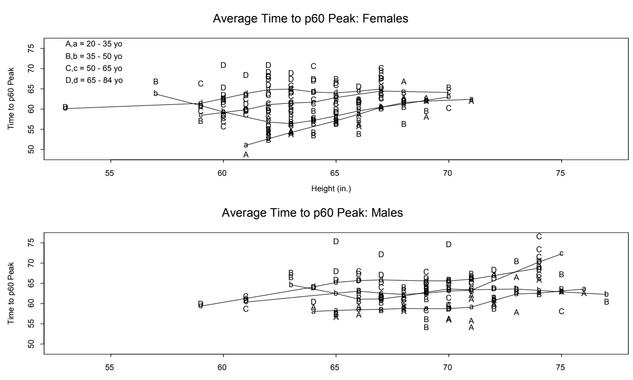
Example: SEP "Normal Ranges"

- We can also consider the possibility of three way interactions between height, age, and sex
  - Osteoporosis affects women far more than men
    - Hence, we might expect the height - age interaction to be greatest in women and not so important in men

Example: SEP "Normal Ranges"

- A two way interaction between height and age that is different between men and women defines a three way interaction between height, age, and sex

Stratified Scatterplots



Example: SEP "Normal Ranges"

- Defining a regression model with interactions
  - We must create variables to model the three way interaction term

Example: SEP "Normal Ranges"

- Furthermore, it is a VERY GOOD idea to include all "main effects" and "lower order interactions" in the model as well
    - "main effects": the individual variables which contribute to the interaction
    - "lower order terms": all interactions that involve some combination of the variables which contribute to the interaction

Example: SEP "Normal Ranges"

- Most often, we lack sufficient information to be able to guess what the true form of an interaction might be
    - The most popular approach is thus to consider multiplicative interactions
        - Create a new variable by merely multiplying the two (or more) interacting predictors

Example: SEP "Normal Ranges"

- Thus for this problem we could create variables
    - HA = Height * Age
    - HM = Height * Male
    - AM = Age * Male
    - HAM = Height * Age * Male

Example: SEP "Normal Ranges"

- Interpretation of the model parameters
    - In the presence of higher order terms (powers, interactions) interpretation of parameters is not easy
        - We can no longer use "the change associated with a 1 unit difference in predictor holding other variables constant"
            - It is generally impossible to hold other variables constant when changing a covariate involved in an interaction
            - If not impossible, it is often uninteresting

Example: SEP “Normal Ranges”

Interpretation of the model in terms of the SEP height relationship within age-sex strata

73

Example: SEP “Normal Ranges”

$$E(p60 \mid Ht, Age, Male) = \beta_0 + \beta_H Ht + \beta_A Age + \beta_M Male$$
$$+ \beta_{HA} HA + \beta_{HM} HM + \beta_{AM} AM + \beta_{HAM} HAM$$

p60 - Height relationship for Age = a :

Sex	Intercept	Slope
F	$(\beta_0 + \beta_A a)$	$(\beta_1 + \beta_{HA} a)$
M	$(\beta_0 + \beta_M + (\beta_A + \beta_{AM})a)$	$(\beta_H + \beta_{HM} + (\beta_{HA} + \beta_{HAM})a)$

74

Example: SEP “Normal Ranges”

- From the above, we see the importance of including the main effects and lower order terms
  - E.g., leaving out the height - sex interaction is tantamount to claiming that the p60 - height relationship among newborns is the same for the two sexes
    - (It might be, but the chance that our lines would predict the truth is very slight-- we are trying to approximate relationships in other age ranges)

75

Example: Regression Output

```
. regress p60 height age male HA HM AM HAM
```

p60	Coef	SE	t	P>|t|	[95%	CI]
height	1.38	.363	3.81	0.000	.666	2.09
age	1.13	.425	2.66	0.008	.292	1.97
male	75.0	32.3	2.32	0.021	11.3	138
HA	-.015	.007	-2.26	0.025	-.028	-.0019
HM	-1.12	.483	-2.34	0.020	-2.08	-.176
AM	-1.16	.582	-2.00	0.047	-2.31	-.0170
HAM	.0175	.009	2.00	0.047	.0002	.0347
_cons	-36.4	23.5	-1.55	0.122	-82.7	9.82

76

Aside: Subgroup Analysis

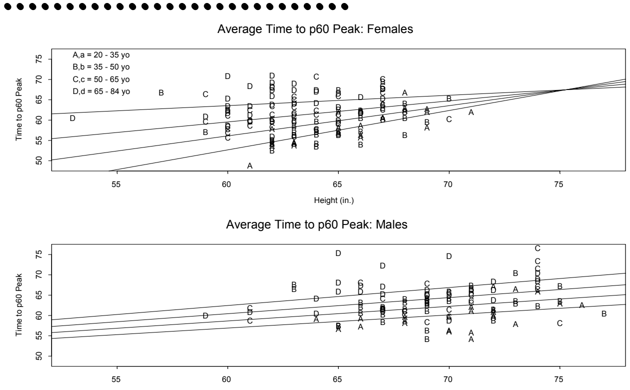
- If I restrict analysis to females, estimates are the same in this “saturated” model
  - (Restricting by age or height would differ due to “borrowing information across groups)
    - Inference can differ due to the estimate of the residual standard error

```
. regress p60 height age HA   if male==0
   p60 |  Coef   SE     t   P>|t|    [95% CI]
height |  1.38 .361   3.82 0.000  .665  2.10
   age |  1.13 .424   2.67 0.009  .292  1.97
    HA | -.015 .007 -2.27 0.025 -.028 -.002
 _cons | -36.4 23.4 -1.56 0.122 -82.7  9.86
```


Interpreting Estimates

- Figuring out what all these estimates mean is nearly impossible
  - I find it easiest to graph the predicted values

Lines Predicted By Model



Example: SEP “Normal Ranges”

- From the inference, we find a statistically significant three way interaction
  - P= .0471
- This would argue that I should make predictions based on a model including the 3-way interaction
  - But…

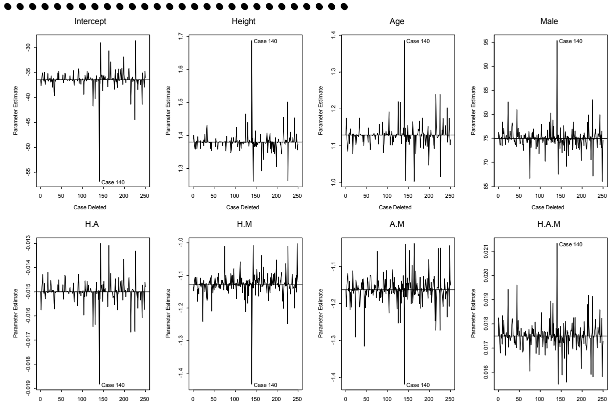
Influence of Individual Cases

- I always worry that interactions might be significant only because of a single “outlier”
  - If that were the case, I might choose not to include the interaction (but I always include the case)
  - Looking ahead: I can “diagnose” such a problem by assessing the influence of each case

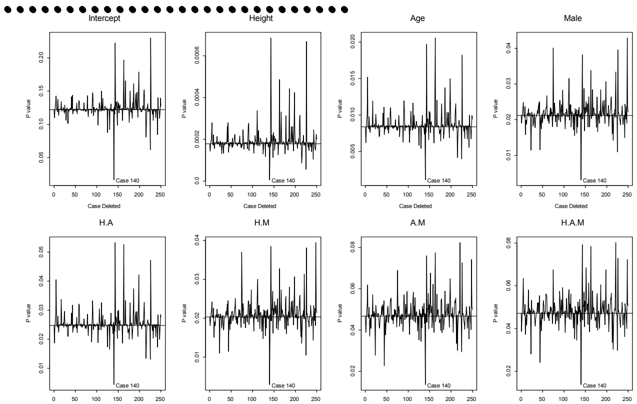
Example: SEP “Normal Ranges”

- I am now interested in ensuring that the evidence for an interaction is not based solely on a single person’s observation
  - Hence, I consider 250 different regressions in which I leave out each case in turn
  - I plot the slope estimates and P values for each variable as a function of which case I left out
    - Case 0 corresponds to using the full data set

Influence on Estimates



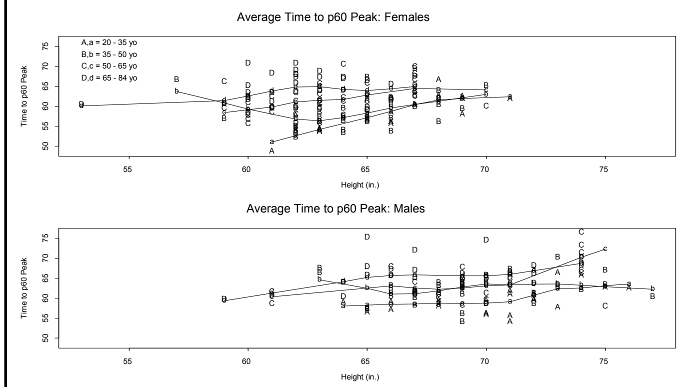
Influence on P values



Example: SEP “Normal Ranges”

- Contrary to what I was afraid of, the only influential case actually lessened the evidence of an interaction
  - When Case 140 is removed from the data, the evidence for an interaction is a larger estimate and a lower P value
  - We can examine the scatterplot to see why Case 140 might be so influential

Stratified Scatterplots



Example: SEP “Normal Ranges”

- So now what do I do with Case 140
  - From the influence diagnostics, I now feel comfortable with the fact that the data really do suggest a three way interaction

Example: SEP “Normal Ranges”

- Personally, I do NOT remove the case from the dataset when making my prediction intervals
  - I do not know why Case 140 is so unusual
  - It is possible that people like her are actually more prevalent in the population than my sample would suggest
    - My best guess is that she represents 0.4% of the population, so leave her in

Modeling Complex "Dose-Response"

Linear Predictors

- The most commonly used regression models use "linear predictors"
  - "Linear" refers to linear in the parameters
  - The modeled predictors can be transformations of the scientific measurements
    - Examples

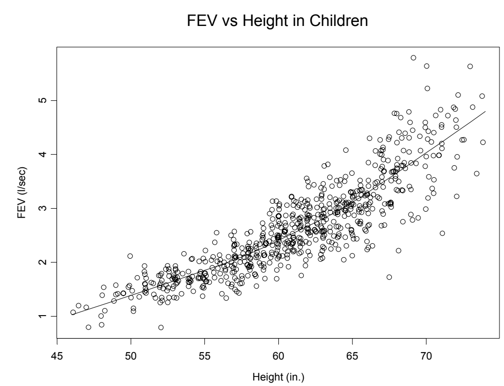
$$g\left[\theta \mid X_i, W_i\right] = \beta_0 + \beta_{\log X} \times \log\left(X_i\right)$$

$$g\left[\theta \mid X_i, W_i\right] = \beta_0 + \beta_X \times X_i + \beta_{X^2} \times X_i^2$$

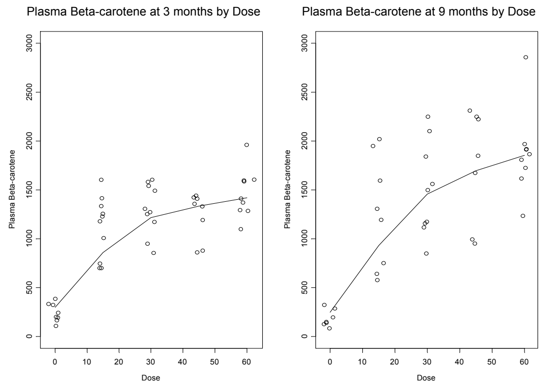
Transformations of Predictors

- We transform predictors to provide more flexible description of complex associations between the response and some scientific measure
    - Threshold effects
    - Exponentially increasing effects
    - U-shaped functions
    - S-shaped functions
    - etc.

Ex: Cubic Relationship



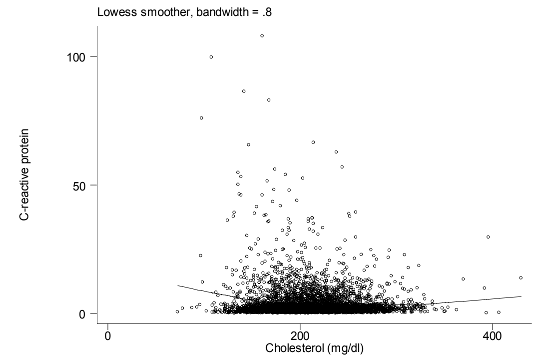
Ex: Threshold Effect of Dose?



93

Ex: U-shaped Trend?

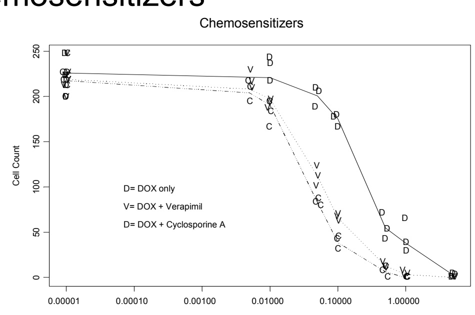
- Inflammatory marker vs cholesterol



94

Ex: S-shaped trend

- *In vitro* cytotoxic effect of Doxorubicin with chemosensitizers



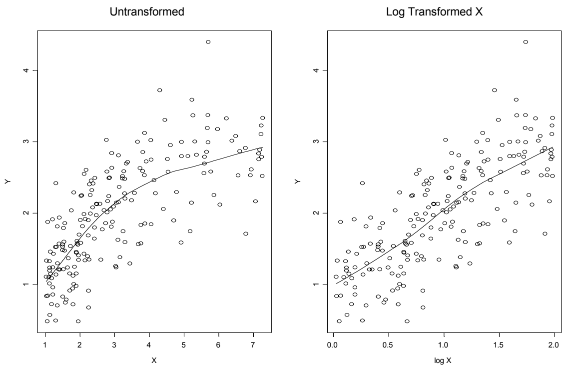
95

"1:1 Transformations"

- Sometimes we transform 1 scientific measurement into 1 modeled predictor
  - Ex: log transformation will sometimes address apparent "threshold effects"
  - Ex: cubing height produces more linear association with FEV

96

Log Transformations



"1:Many Transformations"

- Sometimes we transform 1 scientific measurement into several modeled predictor
  - Ex: "polynomial regression"
  - Ex: "dummy variables" ("factored variables")
  - Ex: "piecewise linear"
  - Ex: "splines"

Polynomial Regression

- Fit linear term plus higher order terms (squared, cubic, …)
  - Can fit arbitrarily complex functions
    - An n-th order polynomial can fit n+1 points exactly
  - Generally very difficult to interpret parameters
    - I usually graph function when I want an interpretation
  - Special uses
    - 2^{nd} order (quadratic) model to look for U-shaped trend
    - Test for linearity by testing that all higher order terms have parameters equal to zero

Ex: FEV – Height Assoc Linear?

- We can try to assess whether any association between mean FEV and height follows a straight line association
  - I fit a 3^{rd} order (cubic) polynomial due to the known scientific relationship between volume and height

Ex: FEV – Height Assoc Linear?

```
. g htsqr= height^2
. g htcub = height^3
. regress fev height htsqr htcub, robust
Linear regression                   Number of obs =      654
                                    Prob > F      =   0.0000
                                    R-squared     =   0.7742
                                    Root MSE      =   .41299
---------------------------------------------------------------
       |        Robust
   fev |  Coef    SE       t  P>|t|     [95% C I]
---------------------------------------------------------------
height |  .0306 .635    0.05 0.962 -1.22    1.28
 htsqr | -.0015 .0108 -0.14 0.888 -.0227   .0196
 htcub | .00003 .00006 0.43 0.671 -.00009 .0001
 _cons |  .457  12.4    0.04 0.971 -23.8   24.76
```


Ex: FEV – Height Assoc Linear?

- Note that the P values for each term were not significant
  - But these are addressing irrelevant questions:
    - After adjusting for 2nd and 3rd order relationships, is the linear term important?
    - After adjusting for linear and 3rd order relationships, is the squared term important?
    - After adjusting for linear and 2nd order relationships, is the cubed term important
  - We need to test 2nd and 3rd order terms simultaneously

Ex: FEV – Height Assoc Linear?

```
. test htsqr htcub

 ( 1)  htsqr = 0
 ( 2)  htcub = 0

       F(  2,    650) =    30.45
            Prob > F =     0.0000
```


Ex: FEV – Height Assoc Linear?

- We find clear evidence that the trend in mean FEV versus height is nonlinear
  - (Had we seen P > 0.05, we could not be sure it was linear– it could have been nonlinear in a way that a cubic polynomial could not detect)

Ex: log FEV – Ht Assoc Linear?

- We can try to assess whether any association between mean log FEV and height follows a straight line association
  - I again fit a 3rd order (cubic) polynomial, but don't really have a good reason to do this rather than some other polynomial

105

Ex: log FEV – Ht Assoc Linear?

```
. g logfev = log(fev)
. regress logfev height htsqr htcub, robust
Linear regression                  Number of obs =      654
                                   F(  3,    650) =  730.53
                                   Prob > F      =  0.0000
                                   R-squared     =  0.7958
                                   Root MSE      =  .15094

       |         Robust
logfev |   Coef    SE       t    P>|t|   [95% C I]
height |  .0707 .24835   0.28 0.776 -.417    .558
 htsqr | -.0002 .00410  -0.04 0.964 -.0082   .008
 htcub | 3e-07   .00002   0.01 0.989 -.00004 .00004
 _cons | -2.79   4.985  -0.56 0.576 -12.6    6.997
```

106

Ex: log FEV – Ht Assoc Linear?

- Note that again that the P values for each term were not significant
  - But these are addressing irrelevant questions:
  - We need to test 2nd and 3rd order terms simultaneously

107

Ex: log FEV – Ht Assoc Linear?

```
. test htsqr htcub

 ( 1)  htsqr = 0
 ( 2)  htcub = 0

       F(  2,   650) =    0.29
            Prob > F =    0.7464
```

108

Ex: log FEV – Ht Assoc Linear?

- We do not find clear evidence that the trend in mean FEV versus height is nonlinear
  - This does not prove linearity, because it could have been nonlinear in a way that a cubic polynomial could not detect
    - (But I would think that the cubic would have picked up most patterns of nonlinearity likely to occur in this setting)

Ex: log FEV – Ht Assoc Linear?

- We have not addressed the question of whether log FEV is associated with height
  - This question could have been addressed in the cubic model by
    - Testing all three height-derived variables simultaneously
    - OR (because only height-derived variables are included in the model) looking at the overall F test
  - Alternatively, fit a model with only the height
    - But generally bad to go fishing for models

Ex: log FEV – Ht Assoc?

```
. regress logfev height, robust
Linear regression             Number of obs =     654
                              F(  1,   652) = 2155.08
                              Prob > F      =  0.0000
                              R-squared     =  0.7956
                              Root MSE      =  .15078
__________________________________________________________
       |                Robust
logfev |       Coef.   Std. Err.      t    P>|t|
  [95% Conf. Interval]
height | .0521 .0011  46.42 0.000  .0499    .0543
 _cons | -2.27 .0686 -33.13 0.000 -2.406   -2.137
```


Dummy Variables

- Indicator variables for all but one group
  - This is the only appropriate way to model nominal (unordered) variables
    - E.g., for marital status
      - Indicator variables for
        - » married (married = 1, everything else = 0)
        - » widowed (widowed = 1, everything else = 0)
        - » divorced (divorced = 1, everything else = 0)
        - » (single would then be the intercept)
  - Often used for other settings as well
  - Equivalent to “Analysis of Variance (ANOVA)”

Ex: Mean Salary by Field

- Field is a nominal variable, so we must use dummy variables
  - I decide to use "Other" as a reference group, so generate new indicator variables for Fine Arts and Professional fields

```
. g arts= 0
. replace arts=1 if field==1
(2840 real changes made)
. g prof= 0
. replace prof=1 if field==3
(3809 real changes made)
```


Ex: Mean Salary by Field

```
. regress salary arts prof if year==95, robust
Linear regression                Number of obs =    1597
                                 F(  2,  1594) =  120.85
                                 Prob > F      =  0.0000
                                 R-squared     =  0.1021
                                 Root MSE      =  1931.2
-----------------------------------------------------------
       |        Robust
salary |  Coef   SE       t   P>|t|    [95% CI]
-----------------------------------------------------------
  arts | -1014  105    -9.67 0.000 -1219   -808
  prof |  1225  134     9.16 0.000   963   1487
 _cons |  6292  61.1 103.03 0.000  6172   6411
```


Ex: Interpretation of Intercept

- Based on coding used
  - Intercept corresponds to mean salary for faculty in "Other" fields
    - These faculty will have arts==0 and prof==0
  - Estimated mean salary is $6,292 / month
  - 95% CI: $6,172 to $6,411 / month
  - Highly statistically different from $0 / month

Ex: Interpretation of Slopes

- Based on coding used
  - Slope for "arts" is difference in mean salary between "Fine Arts" and "Other" fields
    - Fine arts faculty will have arts==1 and prof==0; "Other" fields wil have arts==0 and prof==0
  - Estimated difference in mean monthly salary is $1,014 lower for fine arts
  - 95% CI: $808 to $1,219 / month lower
  - Highly statistically different from $0

Ex: Interpretation of Slopes

- Based on coding used
  - Slope for "prof" is difference in mean salary between "Professional" and "Other" fields
    - Professional faculty will have arts==0 and prof==1; "Other" fields wil have arts==0 and prof==0
  - Estimated difference in mean monthly salary is $1,225 higher for professional
  - 95% CI: $963 to $1,487 / month higher
  - Highly statistically different from $0

Ex: Descriptive Statistics

- Because we modeled the three groups with two predictors plus intercept, the estimates agree exactly with sample means

```
. table field if year==95, co(mean salary)

    field | mean(salary)
    ------+-------------
     Arts |     5278.082
    Other |     6291.638
     Prof |     7516.67
```


Ex: Hypothesis Test

- To test for different mean salaries by field
  - We have modeled field with two variables
    - Both slopes would have to be zero for there to be no association between field and mean salary
  - Simultaneous test of the two slopes
    - We can use the Stata "test" command

```
. test arts prof
            F(  2,  1594) =  120.85
            Prob > F =    0.0000
```

    - OR because only field variables are in the model, we can use the overall F test

Stata: Dummy Variables

- Stata has a facility to automatically create dummy variables
  - Prefix regression commands with "`xi: …`"
  - Prefix variables to be modeled as dummy variables with "`i.`*`varname`*"
  - (Stata will drop the lowest category)

Stata: Dummy Variables

```
. xi: regress salary i.field if year==95, robust
i.field _Ifield_1-3(ntrlly coded; _Ifield_1 omitted)
Linear regression             Number of obs =    1597
                              F(  2,  1594) =  120.85
                              Prob > F      =  0.0000
                              R-squared     =  0.1021
                              Root MSE      =  1931.2
___________________________________________________
         |     Robust
   salary | Coef  SE    t   P>|t|    [95% C I]
_Ifield_2 | 1014 105   9.67 0.000  808     1219
_Ifield_3 | 2239 146  15.30 0.000 1952     2526
    _cons | 5278 85.2 61.94 0.000 5111     5445
```


Ex: Correspondence

- This regression model is the exact same as the one in which I modeled “arts” and “prof”
  - Merely “parameterized” (coded) differently
- Two models are equivalent if they lead to the exact same estimated parameters
  - Inference about corresponding parameters will be the same no matter how it is parameterized

Continuous Variables

- We can also use dummy variables to represent continuous variables
  - Continuous variables measured at discrete levels
    - E.g., dose in an interventional experiment
  - Continuous variables divided into categories

Relative Advantages

- Dummy variables fits groups exactly
  - If no other predictors in the model, parameter estimates correspond exactly with descriptive statistics
- With continuous variables, dummy variables assume a “step function” is true
- Modeling with dummy variables ignores order of predictor of interest

Choice of Model for Analysis

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

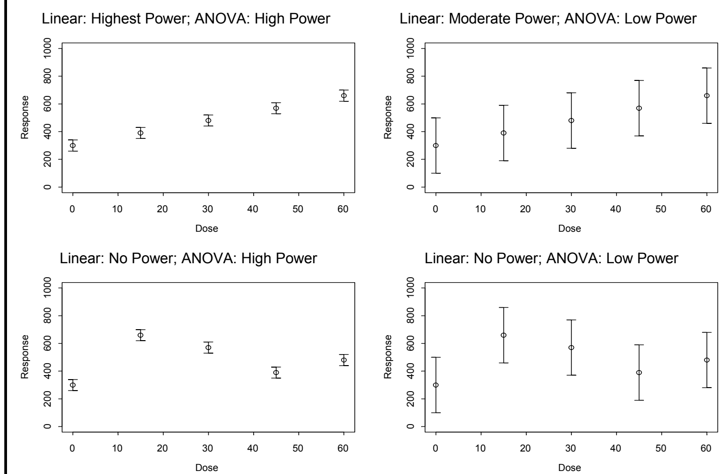
ANOVA (dummy variables)

- 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