Biost 518
Applied Biostatistics II

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

Lecture 9:
Multiple Regression:
Modeling Dose - Response

January 26, 2007

Lecture Outline

- Modeling complex "dose response"
- Flexible methods

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

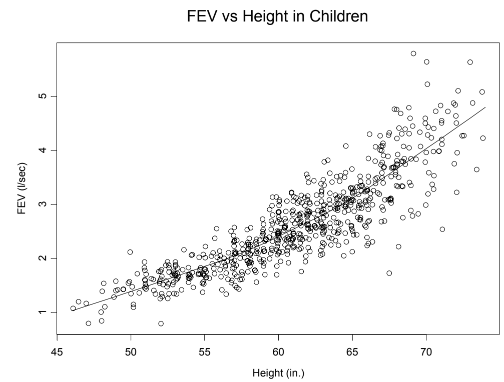
$$g\left[\theta \,|\, X_i, W_i\right] = \beta_0 + \beta_{\log X} \times \log(X_i)$$

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

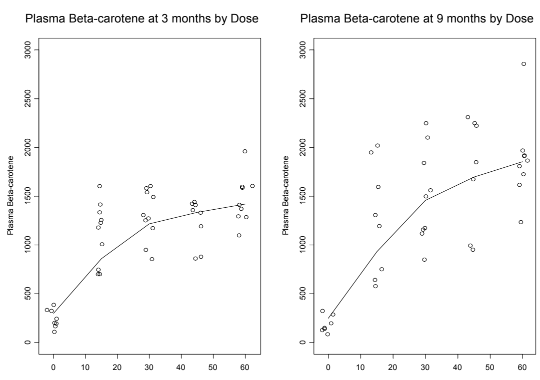
Transformations of Predictors

- 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

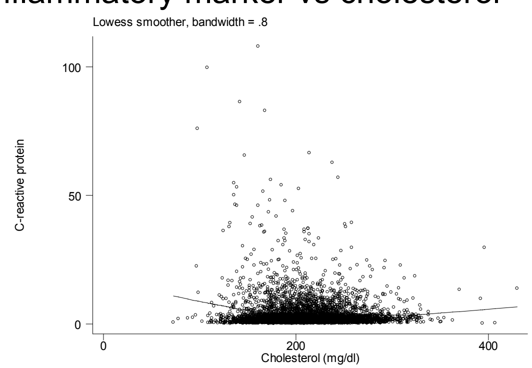


Ex: Threshold Effect of Dose?



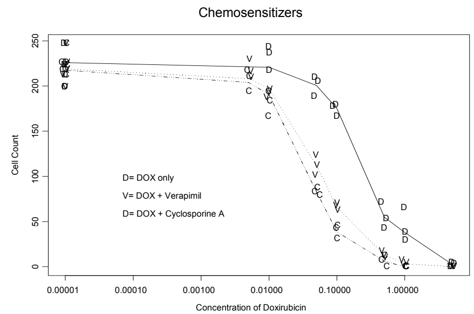
Ex: U-shaped Trend?

- Inflammatory marker vs cholesterol



Ex: S-shaped trend

- *In vitro* cytotoxic effect of Doxorubicin with chemosensitizers



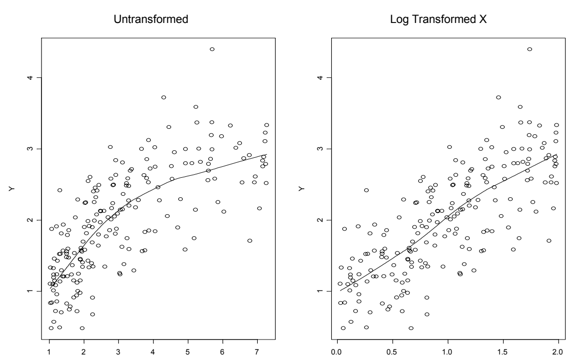
9

"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

10

Log Transformations



11

"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"

12

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 2^{nd} and 3^{rd} order relationships, is the linear term important?
    - After adjusting for linear and 3^{rd} order relationships, is the squared term important?
    - After adjusting for linear and 2^{nd} order relationships, is the cubed term important
  - We need to test 2^{nd} and 3^{rd} 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
```

17

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)

18

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

19

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
```

20

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

Ex: log FEV – Ht Assoc Linear?

```
. test htsqr htcub

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

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


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 StdErr   t     P>|t|      [95% CI]
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
```

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***”

Ex: Salary by Field

```
. predict fit
(option xb assumed; fitted values)
. bysort field: summ fit
-> field = Arts
Vrbl |  Obs      Mean     SD        Min         Max
 fit |  220  5278.082    0   5278.082    5278.082
-> field = Other
Vrbl |  Obs      Mean     SD       Min          Max
 fit | 1067  6291.638    0   6291.638    6291.638
-> field = Prof
Vrbl |  Obs      Mean     SD        Min         Max
 fit |  310  7516.67     0   7516.67     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

Flexible Methods

41

Flexible Modeling of Predictors

- We do have methods that can fit a wide variety of curve shapes
  - Dummy variables
    - A step function with tiny steps
  - Polynomials
    - If high degree: allows many patterns of curvature
  - Splines
    - Piecewise linear or piecewise polynomial
  - Fractional polynomial

42

Stata: Linear Splines

- Stata will make variable that will fit piecewise linear curves
  - Joined at "knots"
  - Lines in between
- **mkspline *newvar0 #k1 newvar1 #k2 newvar2 … #kp varp= oldvar***
  - Regression on *newvar0 … newvarp*
    - Straight lines between min and k1; k1 and k2, etc.

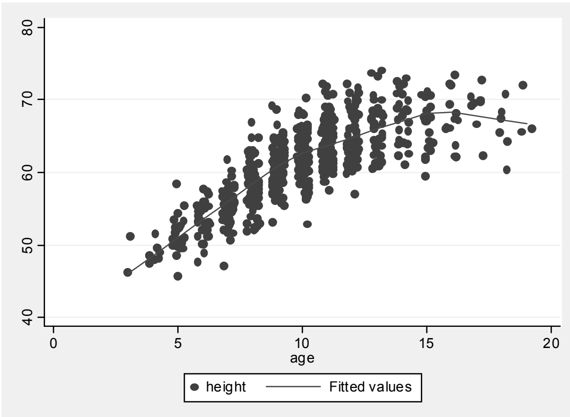
43

Ex: Height vs Age in Children

```
. mkspline  age6A 9.5 age12A  15.5 age17A=age

. regress height age6A age12A age17A
height |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
 age6A |   2.488554   .0999399    24.90   0.000     2.292311    2.684798
age12A |   1.085214    .088609    12.25   0.000     .9112196    1.259209
age17A |  -.5530841   .3713509    -1.49   0.137    -1.282276     .176108
 _cons |   38.49107   .8131791    47.33   0.000      36.8943    40.08785

. predict fitA

. sort age
. twoway (scatter height age, jitter(3)) (line fitA
  age)
```

44

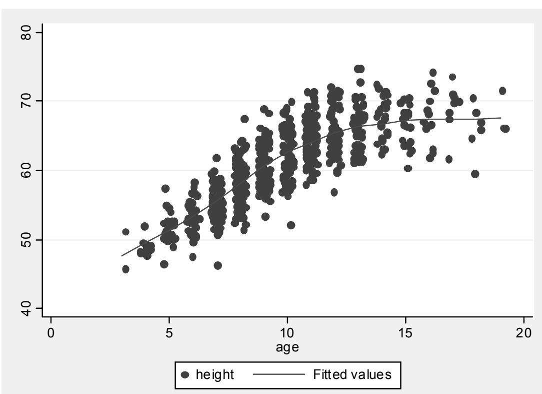
Ex: Height vs Age: 2 Knots



Ex: Height vs Age: 4 Knots

```
. mkspline age4 6.5 age8 9.5 age11  12.5 age14 15.5
  age17=age

. regress height age4 age8 age11 age14 age17


. predict fit

. sort age

. twoway (scatter height age, jitter(3)) (line fit
  age)
```


Ex: Height vs Age: 4 Knots



Stata: fracpoly

- Stata will make variables modeling "fractional polynomials"
  - Can fit many different shapes depending on degree of the fractional polynomials
  - Can ask Stata to find "best" degree of the fractional polynomials: "`fracpoly`"
  - Can ask Stata to make new variables to model fractional polynomial of desired degree: "`fracgen`"

fracpoly

- Command

fracpoly *regressioncommand yvar xvar*, degree(*#*)

Example

fracpoly regression logslry yrdeg, degree(*3*)

Ex: 3rd Degree fracpoly

```
-> gen double Iyrde__1 = X^3-374.9107994 if
   e(sample)
-> gen double Iyrde__2 = X^3*ln(X)-740.6597949 if
   e(sample)
-> gen double Iyrde__3 = X^3*ln(X)^2-1463.219871 if
   e(sample)
    (where: X = yrdeg/10)

(Regression output omitted)

Deviance: 21329.66.
Best powers of yrdeg among 164 models fit: 3 3 3.
```


Adjusting for Confounding

```
. regress logslry female Iyrde__1 Iyrde__2 Iyrde__3,
  robust
```

logslry	Coef	Robust SE	t	P>|t|	[95% C I]	
female	-.119	.007	-16.87	0.000	-.133	-.106
Iyrde__1	-.111	.014	-7.68	0.000	-.139	-.082
Iyrde__2	.088	.012	7.30	0.000	.065	.112
Iyrde__3	-.018	.003	-6.91	0.000	-.023	-.013
_cons	8.36	.004	1983.98	0.000	8.35	8.36