Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Lecture 9:
Flexible Modeling of Associations

February 23, 2015

1

Lecture Outline

- Flexible methods
- Dummy variables
- Polynomial models
- Linear splines
- Use of flexible methods

2

Flexible Methods

Transformations

- Sometimes we transform 1 scientific measurement into 1 modeled predictor
  - Ex: dichotomization at some threshold
  - Ex: log transformation
  - Ex: square root transformation
  - Ex: square transformation
- Sometimes we transform 1 scientific measurement into several modeled predictors
  - Ex: “dummy variables” (“factored variables”)
  - Ex: “polynomial regression”
  - Ex: “piecewise linear”
  - Ex: “splines”

4

Testing for Associations

- When we model a **single** scientific factor with a **single** modeled covariate, testing for an association is usually based on the Wald test of the regression parameter
  - "Partial t test" ("Partial F test", "Partial Z test") that tests that the regression parameter is 0 while adjusting for all other covariates
  - If there are not other covariates, this will be exactly equivalent to the "Overall F test" ("Overall chi square test") that tests that **all** modeled covariates have a zero coefficient
- When we model a **single** scientific factor with **multiple** modeled covariates, out test for association must simultaneously consider all of the regression parameters that relate to the scientific factor
  - "Multiple partial F test" ("Multiple partial chi square test")
  - If all covariates in the model are derived from this single scientific factor, this will be exactly equivalent to the "Overall F test" ("Overall chi square test") that tests that **all** modeled covariates[5]

Testing in Stata

- Wald tests are performed using post regression commands
  - `test` (testing parameters or equality of parameters)
  - `testparm` (allows wildcards)
  - `lincom` (estimation and testing of linear combinations)
  - `testnl` (testing nonlinear combinations)
  - `nlcom` (estimation of nonlinear combinations)
- LR tests are performed using post regression commands
  - Fit a "full model"
    - Stata: save the results with a name `est store` *`modelname`*
  - Fit a "reduced model" by omitting 1 or more covariates
    - Must use same data: **watch about missing data**
  - Compare the two models
    - Stata: `lrtest` *`modelname`*

6

Which Parameters Do We Test?

- It can be difficult to decide the statistical test that corresponds to specific scientific questions
- Need to consider the restricted model that corresponds to your null hypothesis
- Which parameters need to be set to zero?

7

Flexible Methods

Dummy Variables

8

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
  - E.g., model each dose group separately in a 4 arm RCT
- Equivalent to “Analysis of Variance (ANOVA)”

9

Ex: Mean Salary by Field

- University salary data used to investigate sex discrimination
  - In my example, I consider mean salaries
- 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)
```

10

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
```

11

Ex: Interpretation of Intercept

- Try interpretation using “all other covariates are 0”
  - But will be 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
  - (not a surprise)

12

Ex: Interpretation of Slopes - arts

- Try interpretation using “one unit difference in covariate while holding all other covariates constant”
  - But will be based on coding used
  - There may be only one value at which I can hold other covariates constant
- 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 will 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

13

Ex: Interpretation of Slopes - other

- Try interpretation using “one unit difference in covariate while holding all other covariates constant”
  - But will be based on coding used
  - There may be only one value at which I can hold other covariates constant
- Slope for “prof” is difference in mean salary between “Professional” and “Other” fields
  - Professional faculty will have arts==0 and prof==1
  - “Other” fields will 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

14

Ex: Descriptive Statistics

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

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

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

15

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*"

16

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

17

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

18

Stata: Dummy Variables

- Stata has historically had a facility to automatically create dummy variables
  - Prefix regression commands with "`xi: regcmnd …`"
  - Prefix integer variables to be modeled as dummy variables with "`i.varname`"
  - (Stata will drop the lowest category)
- Modern versions allow you to automatically create dummy variables without using the prefix to the command
  - Prefix variables to be modeled as dummy variables with "`i.varname`"
  - (Stata will drop the lowest category by default)

19

Stata: Dummy Variables

- Stata will drop the lowest category by default

```
. regress salary i.field 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.   Std Err     t     P>|t|    [95% Conf. Intervl]
  field |
      2 |  1013.6   104.83   9.67   0.000     807.9    1219.2
      3 |  2238.6   146.30  15.30   0.000    1951.6    2525.6
        |
  _cons |  5278.1    85.21  61.94   0.000    5110.9    5445.2
```

20

Stata: Dummy Variables

- But you can specify an alternative baseline group using “ib#.”

```
. regress salary ib3.field 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.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       field |
          1  |  -2238.6   146.30   -15.30   0.000    -2525.6    -1951.6
          2  |  -1225.0   133.69    -9.16   0.000    -1487.3     -962.8
             |
       _cons |   7516.7   118.93    63.20   0.000     7283.4     7749.9
```

Ex: Correspondence

- This regression model is the exact same as the one in which I modeled “arts” and “prof”
  - Merely “reparameterized” (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

22

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

23

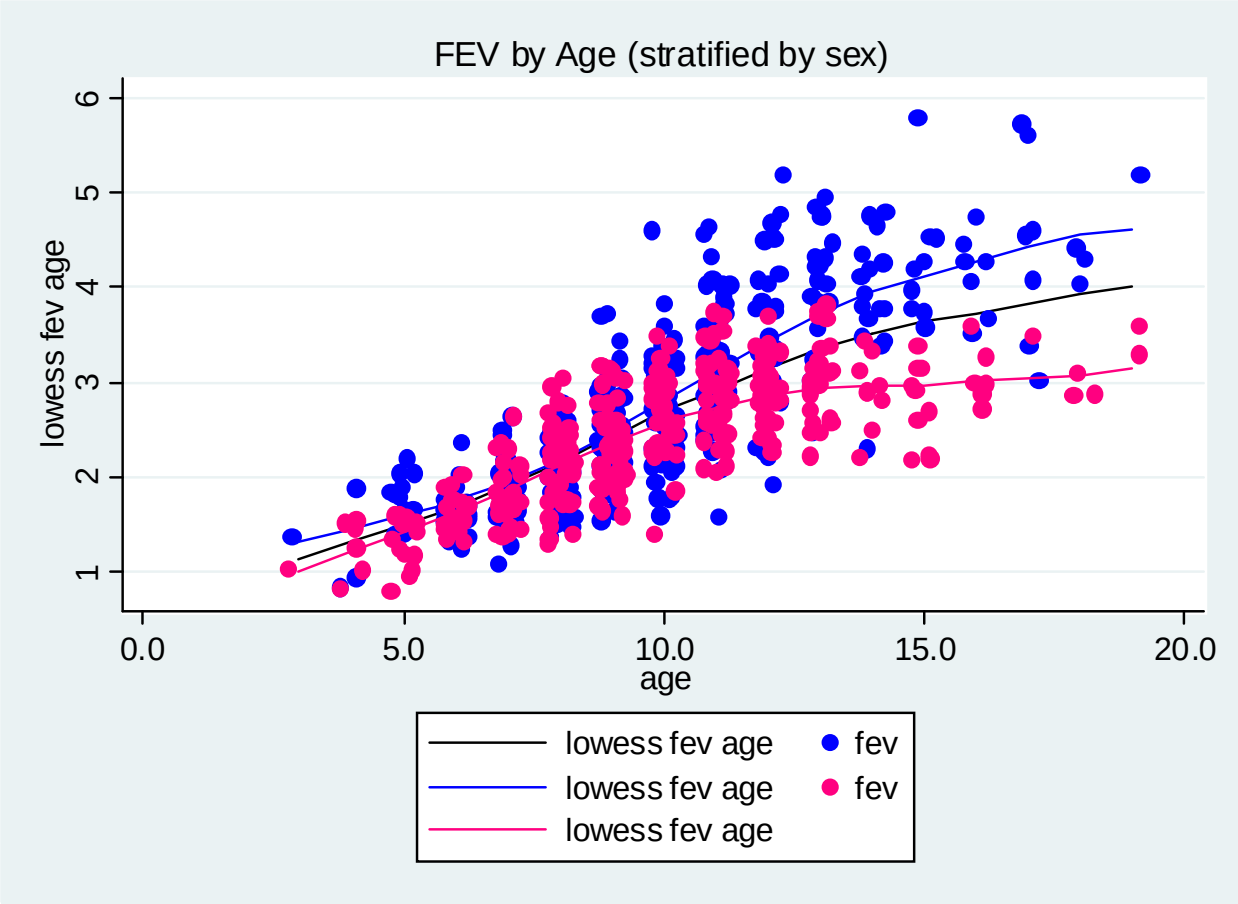
Relative Advantages

- 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

24

Example

- We can look at FEV regressed on age in children



25

Regression with Dummy Variables

```
. egen ageCTG= cut(age), at(3 6 9 12 15 20)
. regress fev i.ageCTG, robust

Linear regression                               Number of obs =     654
                                                F(  4,   649) =  231.05
                                                Prob > F      =  0.0000
                                                R-squared     =  0.5408
                                                Root MSE      =  .58937
------------------------------------------------------------------------------
             |               Robust
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ageCTG  |
          6  |    .47134   .060659     7.77   0.000     .35223    .59046
          9  |    1.2448   .064220    19.38   0.000     1.1188    1.3710
         12  |    1.9122   .084342    22.67   0.000     1.7466    2.0778
         15  |    2.2378   .135970    16.46   0.000     1.9708    2.5048
             |
       _cons |    1.4724   .053106    27.73   0.000      1.3681   1.5767
------------------------------------------------------------------------------

. predict dummyfit
```

26

Regression with Dummy Variables

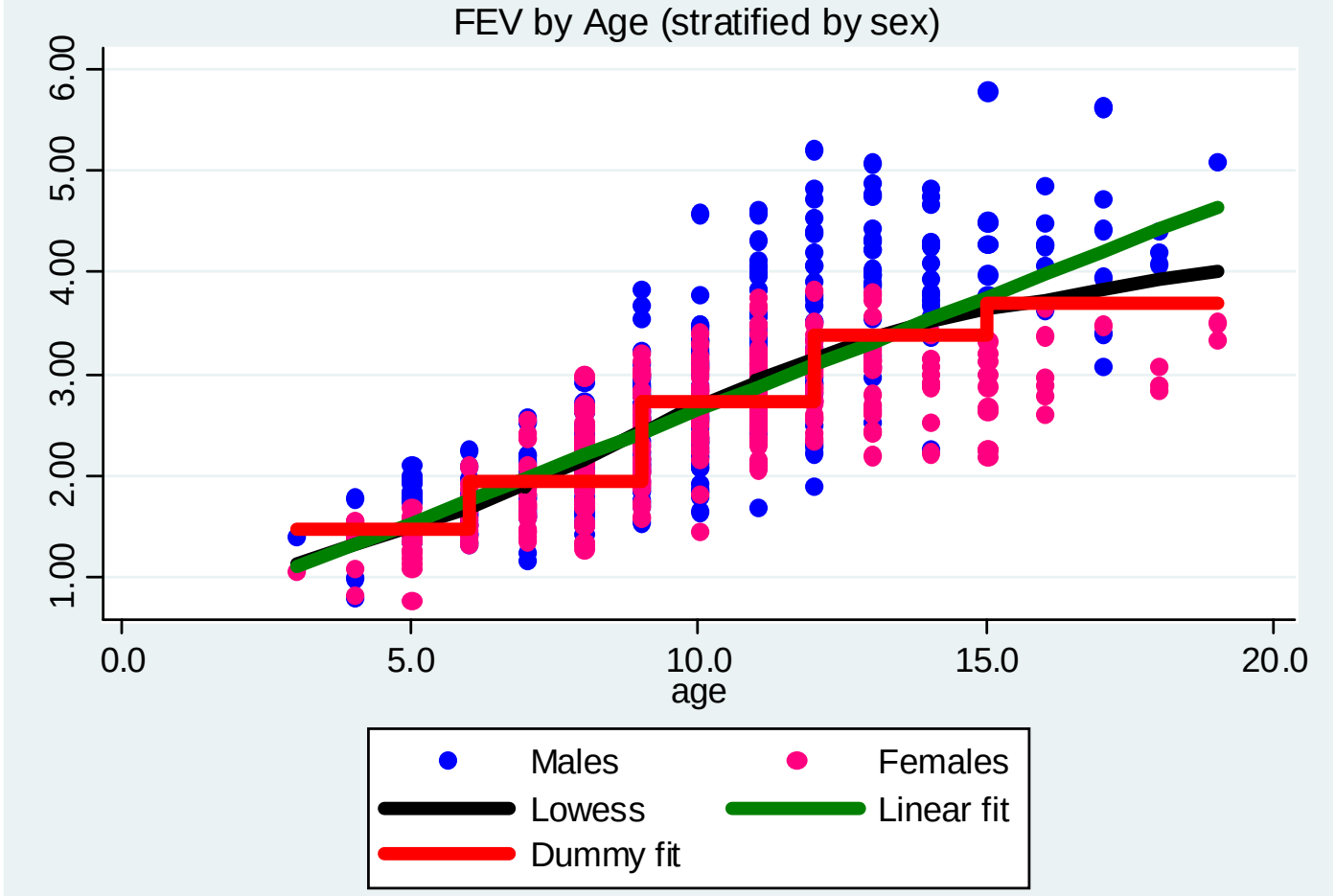
```
. regress fev age, robust

Linear regression                               Number of obs =     654
                                                F(  1,   652) =  608.29
                                                Prob > F      =  0.0000
                                                R-squared     =  0.5722
                                                Root MSE      =  .56753
------------------------------------------------------------------------------
             |               Robust
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |      .2220   .00900     24.66   0.000     .204363    .239719
       _cons |      .4316   .07992      5.40   0.000     .274707    .588589
------------------------------------------------------------------------------

. predict linear
(option xb assumed; fitted values)
```

27

Fitted Values



28

Comments

- Even though a relationship is nonlinear, the best fitting straight line may be a better approximation than dummy variables
- We can compare the RMSE
  - Measures the average standard deviation from the fitted model
  - Usually the RMSE will decrease with the addition of each variable
    - But these models are not hierarchical so can be worse with more variables
  - RMSE is lower in linear fit: 0.568 vs 0.589
- Similarly compare R^2 higher in linear fit: 0.572 vs 0.541
  - Measure of "explained variation"
  - What proportion of total variation is explained by fitted model's variation in the mean
- Adjustment for confounding better with linear fit in this case
- Detecting association will likely be more precise with linear fit

29

Comments

- Detecting association will likely be more precise with linear fit
  - Tendency to lower RMSE translates to more precision
  - Also uses ordering of groups

- This also holds true for discretely sampled data

30

ANOVA (dummy variables)

- Analysis of Variance (ANOVA) corresponds to fitting dummy variables to discretely sampled random variables
  - E.g., RCT with 4 dose groups and placebo
- Fits group means exactly
- Does not mix “random error” with “systematic error:
- Loses “degrees of freedom” to estimate nuisance parameters
- 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

31

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

32

Question

- When is it more powerful to model the POI multivariately rather than univariately?

33

Compare Linear vs Dummy Variables

- Suppose the truth is a straight line relationship:
- We can consider an example in logistic regression

Linear Continuous Power	Dummy Variables Power
.228	.142
.664	.452
.944	.802

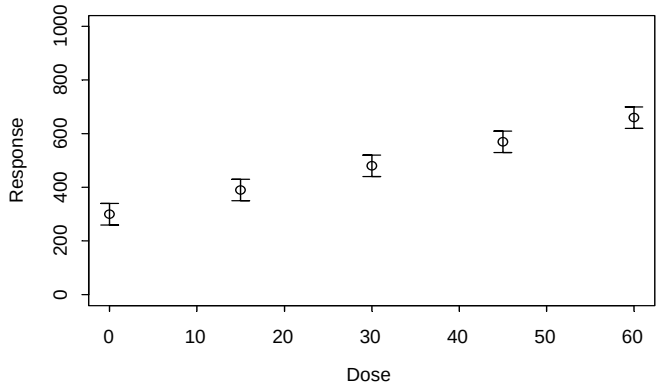
- The major loss of power is from the dummy variables ignoring the order of the groups
  - Had we used grouped linear, the power was .172, .576, .896

34

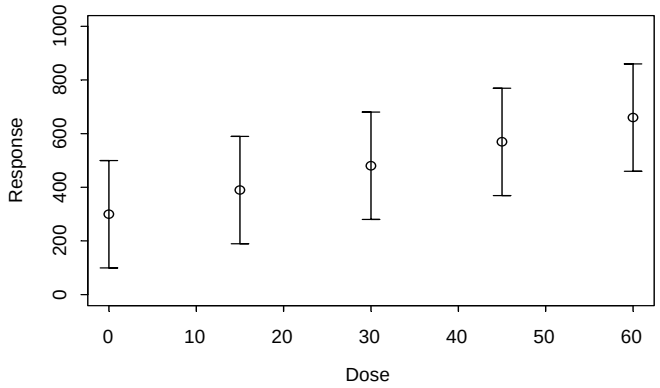
Hypothetical Settings

- Group means by dose with SE

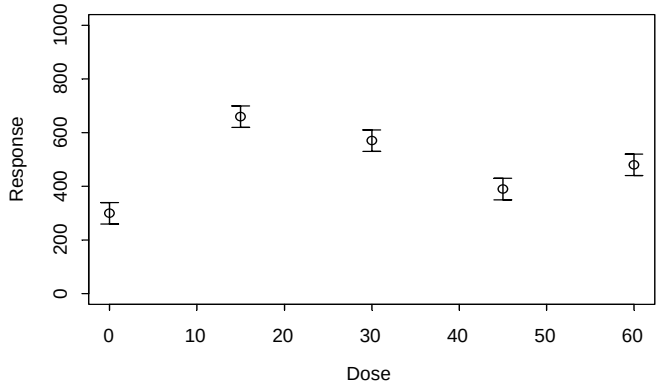
Linear: Highest Power; ANOVA: High Power



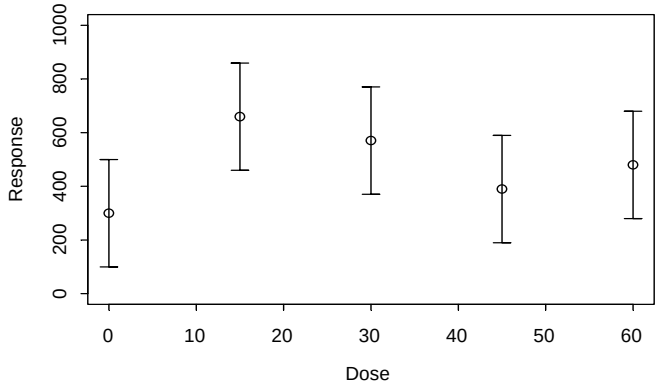
Linear: Moderate Power; ANOVA: Low Power



Linear: No Power; ANOVA: High Power



Linear: No Power; ANOVA: Low Power



35

Flexible Methods

Polynomial Regression

36

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
  - 2nd order (quadratic) model to look for U-shaped trend
  - Test for linearity by testing that all higher order terms have parameters equal to zero

37

Ex: FEV – Height Associated?

- Fit a model with only the height variable
  - Test that the coefficient for height is 0

- This question could have been addressed in a cubic model that includes a linear term, a squared term, and a cubed term
  - Testing all three height-derived variables simultaneously
    - Has to account for covariance among parameter estimates
  - OR (because only height-derived variables are included in the model) looking at the overall F test

- NOTE: it is generally bad to go fishing for models ➔ more later

38

Ex: FEV – Ht Associated?

```
. regress fev height htsqr htcub, robust

Linear regression                               Number of obs =     654
                                                F(  3,   650) =  773.63
                                                Prob > F      =  0.0000
                                                R-squared     =  0.7742
                                                Root MSE      =  .41299
------------------------------------------------------------------------------
             |               Robust
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |    .030594   .634607     0.05   0.962    -1.21553    1.27672
       htsqr |   -.001522   .010780    -0.14   0.888    -.022689    .019645
       htcub |    .000026   .000061     0.43   0.671    -.000093    .000145
       _cons |    .456930   12.3767     0.04   0.971     -23.846    24.7601
------------------------------------------------------------------------------
```

- Generally very difficult to interpret parameters
  - I usually graph function when I want an interpretation

39

Ex: FEV – Ht Associated?

```
. test height htsqr htcub

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

       F(  3,   650) =  773.63
            Prob > F =    0.0000

. testparm h*

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

       F(  3,   650) =  773.63
            Prob > F =    0.0000
```

40

Ex: FEV – Ht Associated? Interpretation

- We thus find strong evidence for a statistically significant association between FEV and height ($P < 0.0001$)
  - (Not a surprise)

- In fitting this larger model, however, we have lost our ability to easily interpret the magnitude of the association
  - It does not work to "consider 1 unit difference while holding all other covariates constant"
    - It is impossible to vary height without also varying height ^{2} and/or height ^{3}
  - We could superimpose a fitted line from the regression using the Stata `predict` command

41

Aside: Collinear Predictors

- When fitting high order polynomials, the various terms can end up being highly "collinear"
  - E.g., in FEV data, height, htsqr, and htcub are strongly correlated

```
. corr height htsqr htcub
(obs=654)
             |   height    htsqr    htcub
-------------+---------------------------
      height |   1.0000
       htsqr |   0.9984   1.0000
       htcub |   0.9937   0.9985   1.0000
```

- Collinear predictors cause two problems in regression
  - Scientifically:
    - We lack precision to distinguish "independent effects"
  - Computationally:
    - Roundoff error in computational algorithms provides fewer significant digits worth of accuracy

42

Regression with Collinear Predictors

- Some textbooks give the impression that high collinearity is to be avoided at all costs
  - I do not believe this to be the case

- Instead we need to consider why we are adjusting for the variable
  - Collinearity among variables modeling POI
    - E.g., polynomial regression
  - Collinearity with POI
    - Confounders vs variance inflation
  - Collinearities among confounders or precision variables

43

Collinear Predictors: Science Issues

- Scientifically **collinearity with the POI** does lead to confounding if the collinear variable is also associated with response
  - But avoiding adjustment for the collinear variable gives an inaccurate representation of the "independent effect" of a POI
- Solutions:
  - Either adjust for the confounder
  - Or abandon the analysis noting the lack of precision available to answer the question of interest, and try to design a future study without such problems
    - RCT vs selected sampling in an observational study vs larger sample size

44

Collinear Predictors: Science Non-Issues

- Scientifically **collinearity among the variables modeling the POI** does not typically lead to a problem
  - The association between the response and POI will be tested by considering all variables jointly
  - (When you are trying to separate, say, nonlinear effects from linear effects, I would consider only the terms modeling the nonlinear effects as the POI)
  - (Similarly, when you are trying to separate, say, current smoking intensity from smoking history, I would consider pack-years as a confounder and intensity as the POI)
- Scientifically **collinearity among the variables modeling confounding** does not typically lead to a problem
  - We generally do not need to test the effect of the confounders so we do not need to worry about precision

45

Collinear Predictors: Statistical Issues

- Statistically **collinearity with the POI** can lead to variance inflation if the collinear variable is not associated with response

- Adjustment for such a variable leads to less precision to detect an association between response and POI

- So do not adjust for variables that you know are not important
  - What is your scientific question?
  - Burden of proof might demand you adjust for a variable that is later proven to be unimportant, but you have to answer your critics

46

Collinear Predictors: Statistical Issues

- Statistically **collinearity among the variables modeling the POI** could lead to less precision
  - The association between the response and POI will be tested by considering all variables jointly
  - If you include terms that do not add substantial information over the other variables, you pay a penalty in precision
    - Terminology from the F tests used in linear regression
      - "Adding degrees of freedom " to the numerator
      - "Losing degrees of freedom" to estimate nuisance parameters
- Solution: "Parsimony" (use only those terms you really need)
- Quite often: Assessing linear trends is more precise than trying to model nonlinearities
    - But need to make this decision *a priori*, or inflate type 1 error

47

Collinear Predictors: Statistical Issues

- Scientifically **collinearity among the variables modeling confounding** could lead to less precision
- "Losing degrees of freedom" to estimate nuisance parameters
- If the confounders are highly collinear, you do not need all of them to adjust for the confounder
  - We are not scientifically interested in the confounders
  - Hence, it does not matter if we do not isolate the "independent effects" of the various confounders

48

Collinear Predictors: Computational Issues

- Computationally collinearity can lead to instability of algorithms
- The analyses can be less reproducible
- Only an issue with extreme collinearity when using double precision
- In the most extreme case, every statistical program will omit variables that are too collinear, because we often over-specify a model due to laziness (more with interactions)

49

Example: Computational Issues

- Stata apparently has less precision with robust SE
- It should not matter how we list variables, but…
  - regress height htsqr htcub, robust ➔ overall F = 773.63
  - regress height htcub htsqr, robust ➔ overall F = 773.67
  - regress htsqr height htcub, robust ➔ overall F = 773.51
  - regress htsqr htcub height, robust ➔ overall F = 773.63
  - regress htcub height htsqr, robust ➔ overall F = 773.60
  - regress htcub htsqr height, robust ➔ overall F = 773.65
- This showed up in the fifth significant digit of the overall F statistic
- R provided greater precision: about 3 extra significant digits

50

Minimizing Computational Issues

- We sometimes model a POI using multiple terms
  - Dummy variables, polynomials, more complex models
  - We test them jointly
- In polynomial regression, we often center variables before creating the higher order terms
  - This is just a reparameterization of the model
    - The fitted values will remain unchanged
  - This will not change the slope estimate for the highest order term,
    - But will change all other slope estimates due to the change in their interpretation
  - However, all but the highest order term are very hard to interpret anyway, so no great loss
  - (And the highest order term is not very easy to interpret either)
- If we center variables modeling polynomial effects at their mean, we can reduce (but not remove) the collinearities

51

Example: Using Centered Height

- In the old days the recommendation would be: center at the mean
  - `egen mht = mean(height)`
  - `g cheight = height – mht`
  - `g chtsqr = cheight^2`
  - `g chtcub = cheight^3`
- We now have less extreme correlation among the predictors modeling height

```
. corr cheight chtsqr chtcub
(obs=654)

             |  cheight   chtsqr   chtcub
-------------+---------------------------
     cheight |   1.0000
      chtsqr |  -0.1736   1.0000
      chtcub |   0.8487  -0.2963   1.0000
```

52

Example: Using Centered Height

- When we fit the regression, we have more reproducible results as we vary the order of the variables
  - Overall F statistic is always 773.65
- The inference about the cubic term is unchanged from previous uncentered analysis (cf: Slide 37 from this lecture)

```
. regress fev cheight chtsqr chtcub, robust

Linear regression                               Number of obs =     654
                                                F(  3,   650) =  773.65
                                                Prob > F      =  0.0000
                                                R-squared     =  0.7742
                                                Root MSE      =  .41299
------------------------------------------------------------------------------
             |               Robust
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cheight |   .1337729   .0051904    25.77   0.000     .1235809    .1439648
      chtsqr |   .0032097   .0005298     6.06   0.000     .0021694    .0042499
      chtcub |   .0000258   .0000607     0.43   0.671    -.0000933    .0001449
       _cons |   2.533546   .0195885   129.34   0.000     2.495081     2.57201
------------------------------------------------------------------------------
```

Collinearity: Final Comments

- Statistical software now uses double precision almost always
  - About 16 significant digits precision in a single operation
    - Depends on the hardware for the machine
  - But errors “propagate” through analyses
    - Final precision may be substantially less, as we have seen
- Older routines in newer programs may sometimes still have single precision
  - In Stata, numbers typed into the commands seems to be lower precision than data entered in files
- Just the same, I am usually happy with about 3 significant figures in my final output, so I usually do not bother with centering variables when constructing polynomials

54

Flexible Methods

Linear Splines

55

Flexible Modeling of Predictors

- We do have methods that can fit a wide variety of curve shapes
- Polynomials
  - If high degree: allows many patterns of curvature
  - Fractional polynomial: allows raising to a fractional power, often searching for best fit
    - (I will not be a party to the propagation of these methods)
- Dummy variables
  - A step function with tiny steps
    - Flat lines over each interval
- Piecewise linear or piecewise polynomial
  - Define intervals over which the curve is a line or polynomial
- Splines
  - Piecewise linear or piecewise polynomial but joined at "knots"

56

Linear Splines

- Draw straight lines between pre-specified "knots"
- Model intercept and *m+1* variables when using *m* knots
- Suppose knots are $k_1, \ldots, k_m$, for variable *X*
  - Define variables *Spline0* … *SplineM*
  - *Spline0 equals*
    - X for $X < k_1$
    - k_1 for $k_1 \leq X$
  - Then, for *J = 1.. m*, *SplineJ equals (*define $k_0=0$, $k_{m+1}=\infty$*)*
    - 0 for $X < k_J$
    - $X - k_J$ for $k_J \leq X \leq k_{J+1}$
    - $k_{J+1} - k_J$ for $k_{J+1} \leq X$

57

Stata: Linear Splines

- Stata will make variable that will fit piecewise linear curves

```
mkspline new0 #k1 new1 #k2 new2 … #kp newp= oldvar
```

- Regression on *newvar0 … newvarp*
  - Straight lines between min and k1; k1 and k2, etc.

58

Regression with Linear Splines: FEV, Age

```
. mkspline age3 6 age6 9 age9 12 age12 15 age15= age
. list age age3 age6 age9 age12 age15 in 1/15
```

age	age3	age6	age9	age12	age15
11.0	6	3	2	0	0
10.0	6	3	1	0	0
9.0	6	3	0	0	0
14.0	6	3	3	2	0
11.0	6	3	2	0	0
11.0	6	3	2	0	0
15.0	6	3	3	3	0
8.0	6	2	0	0	0
7.0	6	1	0	0	0
12.0	6	3	3	0	0
10.0	6	3	1	0	0
11.0	6	3	2	0	0
9.0	6	3	0	0	0
9.0	6	3	0	0	0
8.0	6	2	0	0	0

59

Regression with Linear Splines: FEV, Age

```
. mkspline age3 6 age6 9 age9 12 age12 15 age15= age
. regress fev age3 age6 age9 age12 age15, robust

Linear regression                               Number of obs =     654
                                                F(  5,   648) =  240.68
                                                Prob > F      =  0.0000
                                                R-squared     =  0.5945
                                                Root MSE      =  .55424
------------------------------------------------------------------------------
             |               Robust
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age3 |    .13372   .03942     3.39   0.001     .05632    .21113
        age6 |    .25943   .02001    12.97   0.000     .22015    .29872
        age9 |    .29671   .02764    10.74   0.000     .24245    .35098
       age12 |    .11080   .05309     2.09   0.037     .00654    .21505
       age15 |    .09977   .08604     1.16   0.247    -.06918    .26872
       _cons |    .82887   .21983     3.77   0.000     .39721    1.2605
------------------------------------------------------------------------------

. predict splinefit
(option xb assumed; fitted values)
```

60

Fitted Values with Linear Splines

```
. tabstat splinefit, by(age) stat(n mean sd min max)
```

age	N	mean	sd	min	max
3	2	1.2300	0	1.2300	1.2300
4	9	1.3638	0	1.3638	1.3638
5	28	1.4975	0	1.4975	1.4975
6	37	1.6312	0	1.6312	1.6312
7	54	1.8907	0	1.8907	1.8907
8	85	2.1501	0	2.1501	2.1501
9	94	2.4095	0	2.4095	2.4095
10	81	2.7062	0	2.7062	2.7062
11	90	3.0029	0	3.0029	3.0029
12	57	3.2997	0	3.2997	3.2997
13	43	3.4105	0	3.4105	3.4105
14	25	3.5213	0	3.5213	3.5213
15	19	3.6321	0	3.6321	3.6321
16	13	3.7318	0	3.7318	3.7318
17	8	3.8316	0	3.8316	3.8316
18	6	3.9314	0	3.9314	3.9314
19	3	4.0311	0	4.0311	4.0311

61

Fitted Values with Linear Splines

```
. tabstat splinefit, by(age)
```

age	N	mean	sd	min	max	Difference
3	2	1.2300	0	1.2300	1.2300	
4	9	1.3638	0	1.3638	1.3638	0.13372
5	28	1.4975	0	1.4975	1.4975	0.13372
6	37	1.6312	0	1.6312	1.6312	0.13372
7	54	1.8907	0	1.8907	1.8907	0.25943
8	85	2.1501	0	2.1501	2.1501	0.25943
9	94	2.4095	0	2.4095	2.4095	0.25943
10	81	2.7062	0	2.7062	2.7062	0.29671
11	90	3.0029	0	3.0029	3.0029	0.29671
12	57	3.2997	0	3.2997	3.2997	0.29671
13	43	3.4105	0	3.4105	3.4105	0.11080
14	25	3.5213	0	3.5213	3.5213	0.11080
15	19	3.6321	0	3.6321	3.6321	0.11080
16	13	3.7318	0	3.7318	3.7318	0.09977
17	8	3.8316	0	3.8316	3.8316	0.09977
18	6	3.9314	0	3.9314	3.9314	0.09977
19	3	4.0311	0	4.0311	4.0311	

62

Linear Splines: Parameter Interpretation

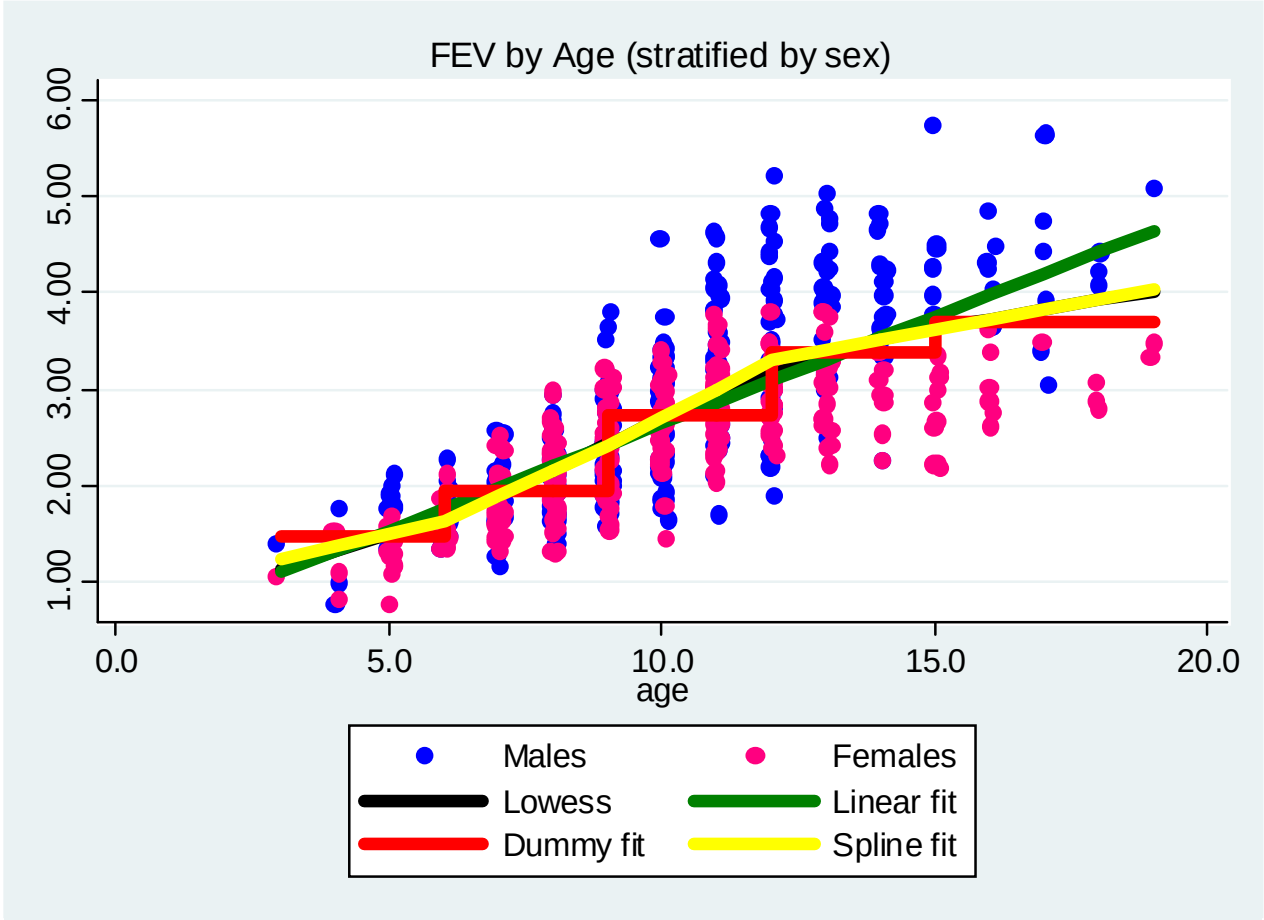
- With identity link
  - Intercept β_0:
    - $\theta_{Y|X}$ when $X = 0$
  - Slope parameters β_j:
    - Estimated difference in $\theta_{Y|X}$ between two groups both between the same knots but differing by 1 unit in X

- With log link
  - Exponentiated intercept $exp(\beta_0)$:
    - $\theta_{Y|X}$ when $X = 0$
  - Exponentiated slope parameters $exp(\beta_j)$:
    - Estimated ratio of $\theta_{Y|X}$ between two groups both between the same knots but differing by 1 unit in X

63

Fitted Values

- Lowess (largely hidden), linear, dummy variables, linear splines



64

Flexible Methods

Comments

65

Flexible Modeling of Predictors

- Commonly used “flexible models” include
  - Polynomials
  - Dummy variables
  - Linear splines

- Possibilities are limitless, but some you may encounter
  - Cubic splines
    - Makes curves smooth at knots
    - But for the ways I use splines, I cannot be bothered
  - Fractional polynomial: allows raising to a fractional power
    - Often searching for best fit over a grid of values
    - I will not be a party to the propagation of these methods

66

Uses of Flexible Modeling of Predictors

- For predictor of interest
  - When strong suspicion of a complex nonlinear fit
    - May provide greater precision due to better fit
    - Can test for linearity by including linear term, then testing all the other terms
  - When fit is fairly well approximated by a straight line of untransformed predictor or straight line with a univariate transformation of predictor, splines may result in loss of precision due to loss of "df"
  - "Keep an open mind, but not so open that your brains fall out"
    - Virginia Gildersleeve
- For confounders, ensures more accurately modeled effect of covariates
  - But, again, not wise to go overboard
- For precision variables, often not often worth the effort

67