Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Lecture 7:
Modeling Nonlinear Associations

January 29, 2014

1

Lecture Outline

- Modeling complex "dose response"
- Multiple comparisons
- Comparing statistical models
- Flexible methods

2

Modeling Complex "Dose-Response"

3

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

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

$$g\left[\theta \mid X_i, W_i\right] = \gamma_0 + \gamma_X \times X_i + \gamma_{X^2} \times X_i^2$$

4

Transformations of Predictors

- We transform predictors to answer scientific questions aimed at detecting nonlinear relationships
  - E.g., is the association between all cause mortality and LDL in elderly adults nonlinear?
  - E.g., is the association between all cause mortality and LDL in elderly adults U-shaped?
- We transform predictors to provide more flexible description of complex associations between the response and some scientific measure (especially confounders, but also precision and POI)
  - Threshold effects
  - Exponentially increasing effects
  - U-shaped functions
  - S-shaped functions
  - etc.

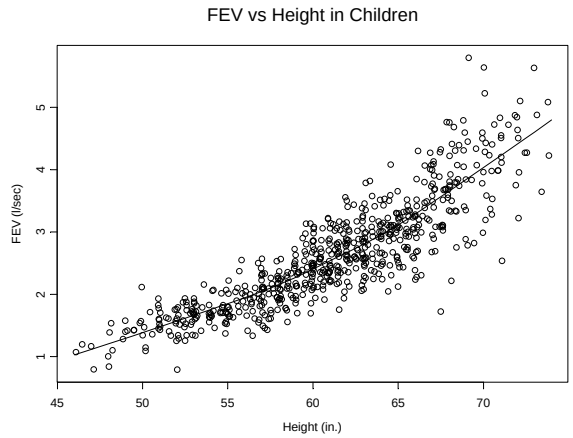
5

General Applicability

- The issues related to transformations of predictors are similar across all types of regression with linear predictors
  - Linear regression
  - Logistic regression
  - Poisson regression
  - Proportional hazards regression
  - Accelerated failure time regression
- However, it is easiest to use descriptive statistics to illustrate the issues in linear regression
- In other forms of regression we can display differences between fitted values, but display of the original data is more difficult
  - Binary data
  - Censored data
  - Models that use a log link

6

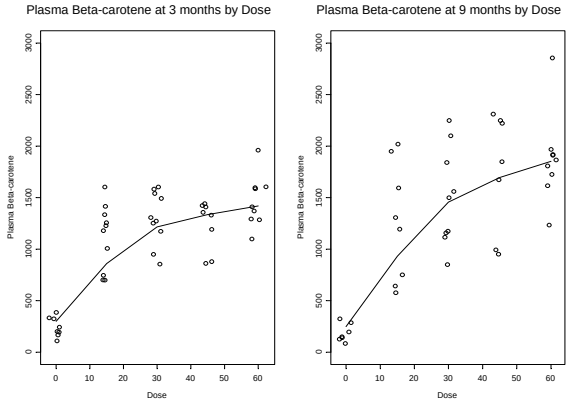
Ex: Cubic Relationship



7

Ex: Threshold Effect of Dose?

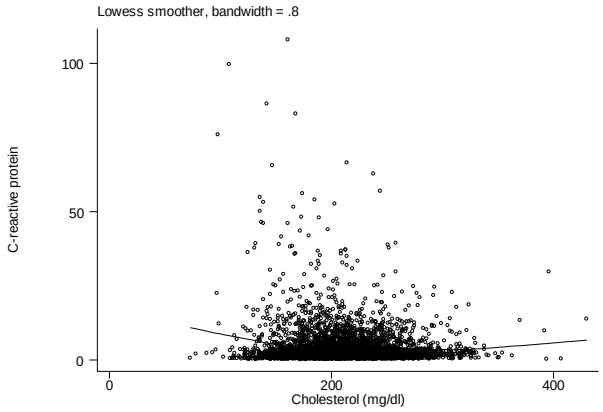
- RCT of beta carotene supplementation: 4 doses plus placebo



8

Ex: U-shaped Trend?

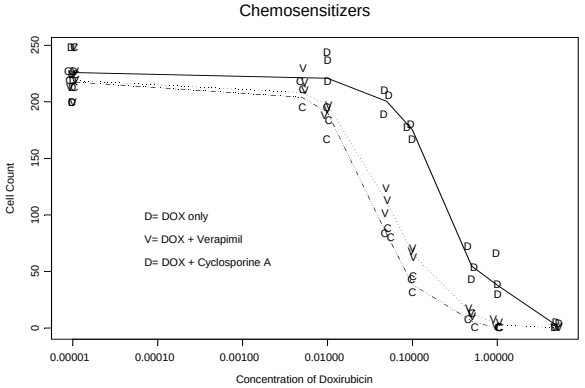
- Inflammatory marker vs cholesterol



9

Ex: S-shaped trend

- *In vitro* cytotoxic effect of Doxorubicin with chemosensitizers



10

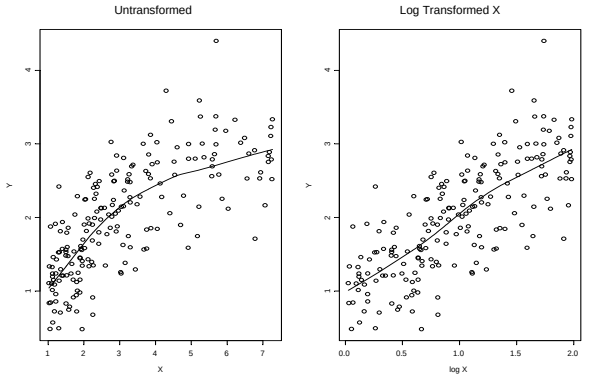
"1:1 Transformations"

- 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
- Ex: dichotomization of dose to detect efficacy in presence of strong "threshold effect" against placebo

11

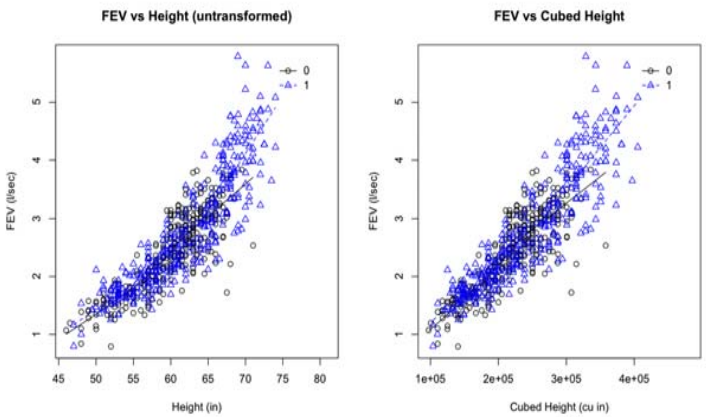
Log Transformations

- Simulated data where every doubling of *X* has same difference in mean of *Y*



12

Cubic Transformation: FEV vs Height



13

Linear Regression: Transformations of X

- Note that in each of the previous examples, the transformation of the predictor yielded a more linear smooth on the scatterplot
- However, transformation of the predictor did not change the within group variability
  - In the case of FEV versus height^3, there remains substantial heteroscedasticity
  - When analyzing geometric means, rather than means, there was less heteroscedasticity

14

Transforming Predictors: Interpretation

- When using a predictor that represents a transformed predictor, we try to use the same interpretation of slopes
  - Additive models:
    - Difference in $\theta_{Y|X}$ per 1 unit difference in modeled predictor
  - Multiplicative models:
    - Ratio of $\theta_{Y|X}$ per 1 unit difference in modeled predictor
- Such interpretations are generally easy for
  - Dichotomization of a measured variable
  - Logarithmic transformation of a measured variable
- Other univariate transformations are generally difficult to interpret
  - I tend not to use other transformations when interpretability of the estimate of effect is key (and I think it always is)

15

Diagnostics

- It is natural to wonder whether univariate transformations of some measured covariate are appropriate
- We can illustrate methods for investigating the appropriateness of a transformation using one of the more common "flexible methods" of modeling covariate associations
  - I consider polynomial regression to investigate whether some of the transformations we have talked about make statistical sense
  - I am not suggesting that we do "model building" by routinely investigating many different models
- I think questions about linearity vs nonlinearity of associations is an interesting scientific question in its own right and should be placed in a hierarchy of investigation
  - I revisit this later

16

"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"

17

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

18

Ex: FEV – Height Assoc Linear?

- We can try to assess whether any association between mean FEV and height follows a straight line association
  - I am presuming this was a prespecified scientific question
  - (We should not pre-test our statistical models)
- I fit a 3rd order (cubic) polynomial due to the known scientific relationship between volume and height

19

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   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .030603   .6350167     0.05   0.962    -1.217116    1.278322
       htsqr |  -.001502   .0107888    -0.14   0.888    -.0226777    .0196737
       htcub |   .000026   .0000599     0.43   0.671    -.0000916    .0001436
       _cons |   .457004   12.40116     0.04   0.971    -23.83966    24.75367
------------------------------------------------------------------------------
```

20

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
  - In all our regressions, we can use Wald tests
  - When using classical regressions (no robust SE) we can use likelihood ratio tests

21

Ex: FEV – Height Assoc Linear?

- Note that the P values for each term were not significant

```
. test htsqr htcub

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

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

22

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)

23

Ex: FEV – Height Associated?

- We have not addressed the question of whether FEV is associated with height
- This question could have been addressed in the cubic model by
  - 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
- Alternatively, fit a model with only the height
  - But generally bad to go fishing for models

24

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
------------------------------------------------------------------------------
```


Ex: FEV – Ht Associated?

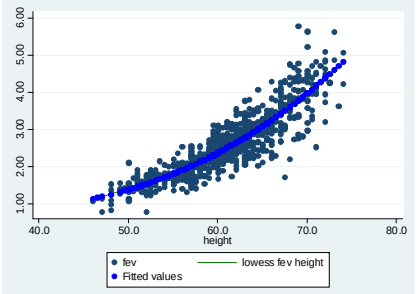
```
. test height htsqr htcub

 ( 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
```


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
  - We could superimpose a fitted line from the regression using the Stata `predict` command

Ex: FEV – Ht Associated? Interpretation

```
. predict cubicfit
(option xb assumed; fitted values)
. twoway (scatter fev height) (lowess fev height,
  color(green)) (scatter cubicfit height,
  color(blue))
```



Ex: log FEV – Height 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
- I will also use classical linear regression to illustrate use of the likelihood ratio test
  - We do expect something closer to homoscedasticity with the logarithmic transformation of the FEV
  - But, in real life I would still tend to use the robust SE, in which case I cannot use the likelihood ratio test

29

Ex: log FEV – Height 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   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .0707   .24835     0.28   0.776    -.417    .558
       htsqr |  -.0002   .00410    -0.04   0.964    -.0082   .008
       htcub |  3.2e-07  .00002     0.01   0.989    -.00004  .00004
       _cons |  -2.79    4.985     -0.56   0.576    -12.6    6.997
------------------------------------------------------------------------------
```

30

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

31

Ex: log FEV – Ht Assoc Linear?

```
. test htsqr htcub

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

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

32

Ex: log FEV – Ht Assoc Linear?

- We do not find clear evidence that the trend in log geometric mean FEV versus height is nonlinear
  - P = 0.7464
- 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)

33

Ex: log FEV – Height Associated?

- 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

34

Ex: log FEV – Height Associated?

```
. g logfev = log(fev)
. regress logfev height htsqr htcub, robust
Linear regression                 Number of obs =     654
                                  F(  3,   650) =  730.49
                                  Prob > F      =  0.0000
                                  R-squared     =  0.7958
                                  Root MSE      =  .15094
------------------------------------------------------------------------------
             |               Robust
      logfev |      Coef   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |      .0707   .24835     0.28   0.776    -.417    .558
       htsqr |    -.0002   .00410    -0.04   0.964    -.0082   .008
       htcub |  3.2e-07    .00002     0.01   0.989   -.00004   .00004
       _cons |    -2.79    4.985    -0.56   0.576    -12.6     6.997
```

35

Slide 88

Ex: log FEV – Height Association?

```
. testparm h*

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

       F(  3,   650) =  730.49
            Prob > F =    0.0000
```

36

Ex: log FEV – Height Association?

```
. 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
------------------------------------------------------------------------------
```

37

Multiple Comparisons

38

Comments

- When we tested for an association between log FEV and height using a cubic model, we had to test three parameters
  - "A three degree of freedom test" instead of just 1 df
- If the extra two parameters do not add substantial precision to the model, they detract greatly from the precision
  - The squared and cubic terms are highly correlated with the linear term
  - If they do not greatly reduce the RMSE, they lead to "variance inflation" through their correlation with the linear height term
- As a rule, separate your questions: Ask in following order
  - Is there a linear trend in geometric mean FEV with height?
    - Just fit the linear height term
  - Is any trend in geometric mean FEV by height nonlinear?
    - Fit quadratic or cubic polynomial and test for nonlinearity

39

Why Not Pre-Testing

- We are often tempted to remove parameters that are not statistically significant before proceeding to other tests
- Such data-driven analyses tend to suggest that failure to reject the null means equivalence
  - They do not
- Such a procedure will tend to underestimate the true standard error
  - Multiple testing problems

40

Multiple Comparisons in Biomedicine

- In this hierarchical testing, we are trying to avoid inflation of our type 1 error by multiple testing
- Observational studies
  - Observe many outcomes
  - Observe many exposures
  - Consequently: Many apparent associations
- Interventional experiments
  - Rigorous science: Well defined methods and outcomes
  - Exploratory science ("Drug discovery")
    - Modification of methods
    - Multiple endpoints
    - Restriction to subgroups

41

Why Emphasize Confirmatory Science?

"When you go looking for something specific, your chances of finding it are very bad, because of all the things in the world, you're only looking for one of them.

"When you go looking for anything at all, your chances of finding it are very good, because of all the things in the world, you're sure to find some of them."

- Darryl Zero in "The Zero Effect"

42

Why Emphasize Confirmatory Science?

"When you go looking for something specific, your chances of finding [a spurious association by chance] are very bad, because of all the things in the world, you're only looking for one of them.

"When you go looking for anything at all, your chances of finding [a spurious association by chance] are very good, because of all the things in the world, you're sure to find some of them."

43

Real-life Examples

- Effects of arrhythmias post MI on survival
  - Observational studies: high risk for death
  - CAST: anti-arrhythmics have higher mortality
- Effects of beta-carotene on lung CA and survival
  - Observational studies: high dietary beta carotene has lower cancer incidence and longer survival
  - CARET: beta carotene supplementation in smokers leads to higher lung CA incidence and lower survival
- Effects of hormone therapy on cardiac events
  - Observational studies: HT has lower cardiac morbidity and mortality
  - WHI: HT in post menopausal women leads to higher cardiac mortality

44

Mathematical Basis

- The multiple comparison problem is traced to a well known fact of probability

$$\Pr(A \text{ or } B) \geq \Pr(A)$$

$$\Pr(A \text{ or } B) \geq \Pr(B)$$

45

Statistics and Game Theory

- Multiple comparison issues
  - Type I error for each endpoint – subgroup combination
    - In absence of treatment effect, will still decide a benefit exists with probability, say, .025
- Multiple endpoints and subgroups increase the chance of deciding an ineffective treatment should be adopted
  - This problem exists with either frequentist or Bayesian criteria for evidence
  - The actual inflation of the type I error depends
    - the number of multiple comparisons, and
    - the correlation between the endpoints

46

Ex: Level 0.05 per Decision

- Experiment-wise Error Rate
  - Consider additional endpoints (typically correlated)
  - Consider effects in subgroups (at least some are independent)

Number Compared	Worst Case	Correlation 0.00	0.30	0.50	0.75	0.90
1	.050	.050	.050	.050	.050	.050
2	.100	.098	.095	.090	.081	.070
3	.150	.143	.137	.126	.104	.084
5	.250	.226	.208	.184	.138	.101
10	.500	.401	.353	.284	.193	.127
20	1.000	.642	.540	.420	.258	.154
50	1.000	.923	.806	.624	.353	.193

47

For Each Outcome Define “Tends To”

- In general, the space of all probability distributions is not totally ordered
- There are an infinite number of ways we can define a tendency toward a “larger” outcome
- This can be difficult to decide even when we have data on the entire population
- Ex: Is the highest paid occupation in the US the one with
  - the higher mean?
  - the higher median?
  - the higher maximum?
  - the higher proportion making $1M per year?

48

Statistical Issues

- Need to choose a primary summary measure or multiple comparison issues result
  - We cannot just perform many tests and choose smallest p value
- Example: Type I error with normal data
  - Any single test: 0.050
  - Mean, geometric mean 0.057
  - Mean, Wilcoxon 0.061
  - Mean, geom mean, Wilcoxon 0.066
  - Above plus median 0.085
  - Above plus Pr (Y > 1 sd) 0.127
  - Above plus Pr (Y > 1.645 sd) 0.169

49

Statistical Issues

- Need to choose a primary summary measure or multiple comparison issues result
  - We cannot just perform many tests and choose smallest p value
- Example: Type I error with lognormal data
  - Any single test: 0.050
  - Mean, geometric mean 0.074
  - Mean, Wilcoxon 0.077
  - Mean, geom mean, Wilcoxon 0.082
  - Above plus median 0.107
  - Above plus Pr (Y > 1) 0.152
  - Above plus Pr (Y > 1.645) 0.192

50

Ideal Results

- Goals of "scientific discovery" are similar to those of diagnostic testing in clinical medicine
- We want a "scientific discovery" process in which there is
  - A low probability of believing false hypotheses
    - High specificity (low type I error)
    - Type 1 error = probability of rejecting null when it is true
  - A high probability of believing true hypotheses
    - High sensitivity (low type II error; high power)
    - Power = probability of rejecting null when it is false
  - A high probability that adopted hypotheses are true
    - High positive predictive value
    - PPV = probability that null is truly false when it is rejected
    - Will depend on prevalence of "good ideas" among our ideas

51

Bayes Factor

- Bayes rule tells us that we can parameterize the positive predictive value by the type I error, power, and prevalence
- Maximize new information by maximizing Bayes factor
  - Relates prior odds of hypothesis being true to posterior odds of hypothesis being true
  - With simple hypotheses:

$$PPV = \frac{power \times prevalence}{power \times prevalence \;+\; type\ I\ err \times (1 - prevalence)}$$

$$\frac{PPV}{1-PPV} = \frac{power}{type\ I\ err} \times \frac{prevalence}{1-prevalence}$$

$$posterior\ \ odds \;=\; Bayes\ Factor \;\times\; prior\ \ odds$$

52

Bayes Factor: Most Important Point

- IMPORTANT POINT: A multiplicative effect
- If we inflate our type 1 error without increasing our power by a similar proportion, we decrease the credibility of our analyses
  - Suppose we aim for type 1 error 0.025, power of 0.95
    - Bayes Factor of 36 takes prevalence from 10% to 81%
  - Maybe multiple comparisons ➔ type 1 error 0.05, power 0.96
    - Bayes Factor of 19.4 takes prevalence from 10% to 67%
    - To have same PPV after multiple comparisons, we would need to increase power to the impossible value of 1.90

$$\frac{PPV}{1-PPV} = \frac{power}{type\ I\ err} \times \frac{prevalence}{1-prevalence}$$

$$posterior\ odds = Bayes\ Factor \times prior\ odds$$

53

Comparing Models

54

Hierarchical Models

- When testing for associations, we are implicitly comparing two models
- "Full" model
  - Usually corresponds to the alternative hypothesis
  - Contains all terms in the "restricted" model plus some terms being tested for inclusion
- "Restricted" model
  - Usually corresponds to the null hypothesis
  - Terms in the model are the subset of the terms in the full model that are not being tested

55

Scientific Interpretation

- The scientific interpretation of our statistical tests depends on the meaning of the restricted model compared to the full model
- What associations are possible with the full model that are not possible with the restricted model?

56

Example: Adjusted Effects

- Hierarchical models:
  - Full model: FEV vs smoking, age, height
  - Restricted model: FEV vs age, height
- If the full model provides no advantage over the restricted model, we conclude that there is insufficient evidence to suggest an association between FEV and smoking

57

Example: Tests of Linearity

- Hierarchical models:
  - Full model: Survival vs cholest, cholest^2
  - Restricted model: Survival vs cholesterol
- If the full model provides no advantage over the restricted model, we conclude that there is insufficient evidence to suggest a U shaped trend in survival with cholesterol

58

Likelihood Based Tests

- We have three distinct (but asymptotically equivalent) ways of making inference with maximum likelihood methods
  - Wald, score, likelihood ratio
- The tests differ in their exact formula, as well as how they handle the mean-variance relationship
  - I find that the handling of the mean-variance relationship tends to matter the most
- Wald statistic: estimate +/- critical value x std error
  - estimates variance using the estimated mean
- Score statistic: uses the efficient transformation of the data
  - estimates variance using the hypothesized null
- Likelihood ratio: uses ratio of probability under MLE and null
  - On the log scale and uses both variances

59

Example: One Sample Binomial

$$Y_i \sim B(1,p) \quad \Rightarrow \quad \hat{p} = \bar{Y} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

$$H_0 : p = p_0$$

$$\text{Wald}: \quad Z = \sqrt{n}\frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})}} \quad \sim N(0,1)$$

$$\text{Score}: \quad Z = \sqrt{n}\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)}} \quad \sim N(0,1)$$

$$\text{LR}: \quad \chi^2 = 2\log\left(\frac{\hat{p}^{n\bar{Y}}(1-\hat{p})^{n-n\bar{Y}}}{p_0^{n\bar{Y}}(1-p_0)^{n-n\bar{Y}}}\right) \quad \sim \chi^2_1$$

60

Regression Models

- With regression models, tformulas for statistics differ more
  - Wald statistic is based on regression parameter estimates
  - Score statistic is based on a transformation of the data
    - In special cases of GLM with a "canonical" link:
      - Like a contribution of each observation to a correlation
      - Includes linear, logistic, Poisson regression
  - Likelihood ratio involves the parametric density
- Still the statistics differ in their handling of the mean-variance
  - Wald based on MLE of means
  - Score uses null hypothesis
  - LR uses both

61

Statistical Methods: Wald Tests

- Can be used with all approximately normal regression parameter estimates (including when using robust SE estimates)
- We fit the full model, obtaining
  - Estimates and SE of all coefficients (typically printed)
  - Correlation matrix for coefficient estimates (typically not printed)
- We use matrix multiplication to simultaneously test that multiple coefficients are simultaneously zero
  - Quadratic form: Estimate x Inverse Covariance Matrix x Estimate
  - Asymptotic chi square distn (F distn if we use sample variance)
    - Degrees of freedom = number of parameters tested
- If only one coefficient, matrix multiplication reduces to division of estimate by standard errors
  - Square root of chi square distn w/ df=1 is normal
  - Square root of F stat w/ numerator df=1 is t distribution

62

Statistical Methods: LR Tests

- Likelihood ratio tests can be used with "regular" parametric and semi-parametric probability models
- We fit the full model, obtaining "full log likelihood"
- We fit a reduced model, obtaining "reduced log likelihood"
  - Models must be "hierarchical"
    - Every covariate in reduced model must also be in the full model
    - (But reparameterizations are okay)
  - Must be fit using the same data set
    - Need to make sure no cases with missing data are added when fitting the reduced model
- Compute LR statistic: 2 x (log L_{Full} – log L_{Red})
  - Asymptotically chi square distribution in "regular" problems
  - Degrees of Freedom = number of covariates removed from full model

63

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`*

64

Ex: log FEV – Ht Assoc Linear?

- I will also use classical linear regression to illustrate use of the likelihood ratio test
  - We do expect something closer to homoscedasticity with the logarithmic transformation of the FEV
  - But, in real life I would still tend to use the robust SE, in which case I cannot use the likelihood ratio test

65

Ex: log FEV – Height Assoc Linear?

```
. regress logfev height htcub htsqr

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  3,   650) =  844.50
       Model |  57.7177        3  19.239                Prob > F      =  0.0000
    Residual |  14.8082      650  .02278                R-squared     =  0.7958
-------------+------------------------------           Adj R-squared =  0.7949
       Total |  72.5259      653  .11107                Root MSE      =  .15094

------------------------------------------------------------------------------
      logfev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |    .070664   .232475     0.30   0.761    -.3858279    .5271558
       htcub |  3.24e-07    .000021     0.02   0.988    -.0000415    .0000422
       htsqr |  -.000183    .003866    -0.05   0.962    -.0077735    .0074079
       _cons |  -2.79183    4.63247    -0.60   0.547    -11.88823    6.304574
------------------------------------------------------------------------------


. est store cubic
```

66

Ex: log FEV – Height Assoc Linear?

```
. regress logfev height

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  1,   652) = 2537.94
       Model |  57.7021        1  57.7021               Prob > F      =  0.0000
    Residual |  14.8238      652   .02274               R-squared     =  0.7956
-------------+------------------------------           Adj R-squared =  0.7953
       Total |  72.5260      653  .111066               Root MSE      =  .15078

------------------------------------------------------------------------------
      logfev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |     .0521   .0010346    50.38   0.000     .0500876    .0541506
       _cons |   -2.271    .063531   -35.75   0.000    -2.396062   -2.146562
------------------------------------------------------------------------------


. lrtest cubic

Likelihood-ratio test                           LR chi2(2)  =      0.69
(Assumption: . nested in cubic)                 Prob > chi2 =    0.7100
```

67

Models with Interactions

- We also use this approach when modeling effect modification
- Best scientific approach:
  - Pre-specify the statistical model that will be used for analysis
- Sometimes we choose a relatively large model including interactions
  - Allows us to address more questions
    - E.g., effect modification
  - Sometimes allows greater precision
    - Tradeoffs between more parameters to estimate versus smaller within group variability

68

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?

69

Ex: Full Model w/ Interactions

- Full model:
  - Survival vs sex, smoking, sex-smoking interaction
- Question:
  - Is there effect modification?
- Restricted model:
  - Survival vs sex, smoking
- Test that parameter for sex-smoking interaction is zero

70

Ex: Full Model w/ Interactions

- Full model:
  - Survival vs sex, smoking, sex-smoking interaction
- Question:
  - Association between survival and sex?
- Restricted model:
  - Survival vs smoking
- Test that parameters for sex, sex-smoking interaction are zero

71

Ex: Full Model w/ Interactions

- Full model:
  - Survival vs sex, smoking, sex-smoking interaction
- Question:
  - Association between survival and smoking?
- Restricted model:
  - Survival vs sex
- Test that parameters for smoking, sex-smoking interaction are zero

72

Flexible Methods

73

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 predictor
  - Ex: “polynomial regression”
  - Ex: “dummy variables” (“factored variables”)
  - Ex: “piecewise linear”
  - Ex: “splines”

74

Flexible Methods

Polynomial Regression

75

Covered Earlier: 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

76

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

77

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

78

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

79

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

80

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

81

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

82

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

83

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)

84

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 = 730.49
  - regress height htcub htsqr, robust ➔ overall F = 730.55
  - regress htcub height htsqr, robust ➔ overall F = 730.48
  - regress htcub htsqr height, robust ➔ overall F = 730.48
  - regress htsqr height htcub, robust ➔ overall F = 730.48
  - regress htsqr htcub height, robust ➔ overall F = 730.49
- This showed up in the fifth significant digit of the overall F statistic
- R provided greater precision: about 3 extra significant digits

85

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

86

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
```

87

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 730.50
- The inference about the cubic term is unchanged from previous uncentered analysis (cf: Slide 35 from this lecture)

```
. regress logfev cheight chtsqr chtcub, robust

Linear regression                               Number of obs =     654
                                                F(  3,   650) =  730.50
                                                Prob > F      =  0.0000
                                                R-squared     =  0.7958
                                                Root MSE      =  .15094
------------------------------------------------------------------------------
             |               Robust
      logfev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cheight |   .0519429   .0020107    25.83   0.000     .0479947    .0558912
      chtsqr |  -.0001234   .0001633    -0.76   0.450    -.0004441    .0001974
      chtcub |   3.24e-07   .0000225     0.01   0.989    -.0000438    .0000445
       _cons |   .9194572   .0073612   124.91   0.000     .9050025    .9339118
```

88

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

89

Flexible Methods

Dummy Variables

90

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)"

91

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)
```

92

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   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        arts |  -1014.04   105.0102    -9.65   0.000    -1219.966   -808.1139
        prof |   1225.39   133.7599     9.16   0.000     962.9284    1487.852
       _cons |   6292.41   61.09911   103.03   0.000     6172.597    6412.223
------------------------------------------------------------------------------
```

93

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)

94

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

95

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

96

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
```

97

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*"

98

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
```

99

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

100

Stata: Dummy Variables

- Stata has historically had a facility to automatically create dummy variables
  - Prefix regression commands with "`xi: regcmnd …`"
  - Prefix 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)

101

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. Interval]
-------------+----------------------------------------------------------------
       field |
           2 |      1014   104.8302     9.67   0.000     807.9124    1219.188
           3 |      2239   146.3035    15.30   0.000     1951.624    2525.576
             |
       _cons |      5278   85.21139    61.94   0.000     5110.923    5445.154
------------------------------------------------------------------------------
```

102

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 |     -2239   146.3035   -15.30   0.000    -2525.576   -1951.624
           2 |     -1225   133.6911    -9.16   0.000    -1487.288   -962.7124
             |
       _cons |      7517   118.9346    63.20   0.000     7283.423    7750.577
------------------------------------------------------------------------------
```

103

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

104

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

105

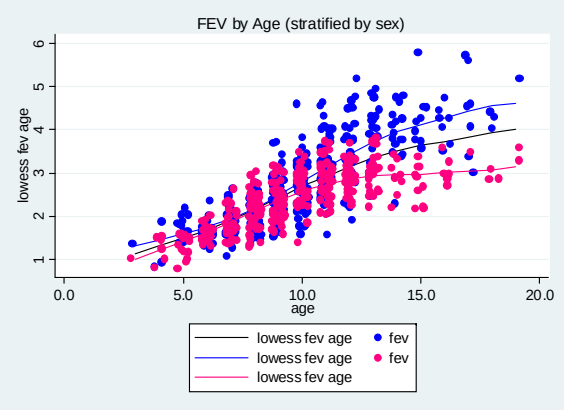
Relative Advantages

- 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

106

Example

- We can look at FEV regressed on age in children



107

Regression with Dummy Variables

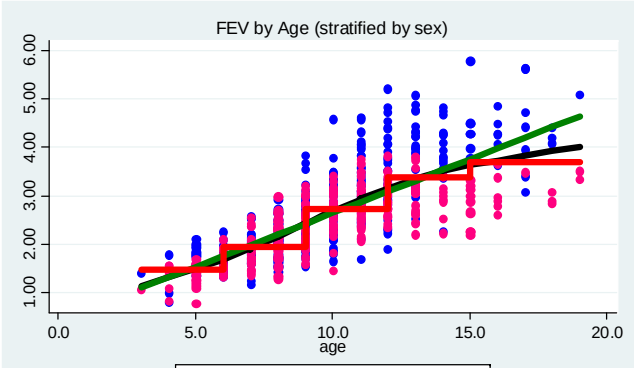
```
. egen ageCTG= cut(age), at(3 6 9 12 15 20)
. regress fev i.ageCTG, robust

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
```

108

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)
```


Fitted Values



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

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

Testing Linearity

- When using dummy variables with categorized continuous variables in a non-saturated model, a straight line is not a special case unless it is a flat line
  - To test linearity, we would have to add a linear term and then test the dummy variables together
- With a discretely sampled random variable, the dummy variable model is saturated, and a straight line is a special case
  - So we could use LR test in classical regression
  - Or we could add a linear term, though the software would discard one dummy variable

113

Testing Linearity: Example

- **regress fev age i.ageCTG**

```
      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  5,   648) =  188.68
       Model |   291.02238     5   58.204476           Prob > F      =  0.0000
    Residual |  199.897453   648  .308483724           R-squared     =  0.5928
-------------+------------------------------           Adj R-squared =  0.5897
       Total |  490.919833   653  .751791475           Root MSE      =  .55541

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2368507   .0260301     9.10   0.000     .1857372    .2879642
             |
      ageCTG |
           6 |  -.1459047   .1194333    -1.22   0.222    -.3804277    .0886182
           9 |  -.0147833   .1680416    -0.09   0.930     -.344755    .3151884
          12 |   -.000931   .2336329    -0.00   0.997       -.4597    .4578379
          15 |  -.4948871    .323106    -1.53   0.126    -1.129348     .139574
             |
       _cons |   .3670811   .1505513     2.44   0.015     .0714538    .6627085
------------------------------------------------------------------------------
```

114

Testing Linearity: Example

- Strong evidence for nonlinearity when using dummy variables to detect it

```
. testparm i.a*

 ( 1)  6.ageCTG = 0
 ( 2)  9.ageCTG = 0
 ( 3)  12.ageCTG = 0
 ( 4)  15.ageCTG = 0

       F(  4,   648) =    8.19
            Prob > F =    0.0000
```

115

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

116

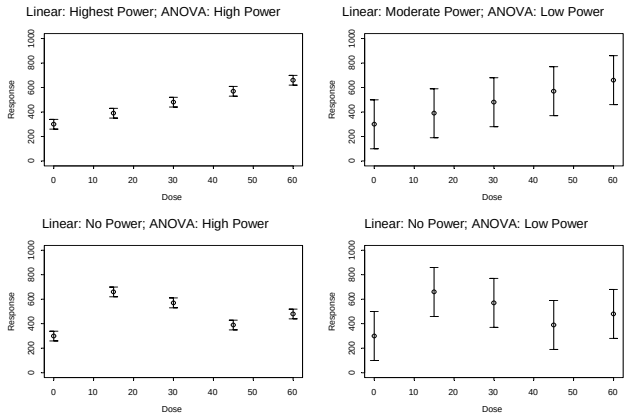
Linear Continuous Models

- 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

117

Hypothetical Settings

- Group means by dose with SE



118

Flexible Methods

Linear Splines

119

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”

120

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 $k_{J+1} \leq X$

121

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.

122

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

123

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)
```

124

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
```

125

Fitted Values with Linear Splines

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

Difference
0.13372
0.13372
0.13372
0.25943
0.25943
0.25943
0.29671
0.29671
0.29671
0.11080
0.11080
0.11080
0.09977
0.09977
0.09977

126

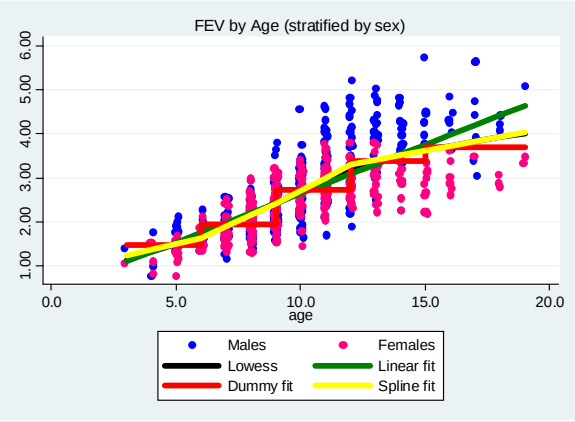
Linear Splines: Parameter Interpretation

- 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

127

Fitted Values

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



128

Testing Linearity

- A straight line is a special case of linear splines
- All the parameter coefficients would have to be equal
- Can use Stata's `test`

```
. test age3 = age6 = age9 = age12 = age15

 ( 1)  age3 - age6 = 0
 ( 2)  age3 - age9 = 0
 ( 3)  age3 - age12 = 0
 ( 4)  age3 - age15 = 0

       F(  4,   648) =    6.89
            Prob > F =    0.0000
```

129

Flexible Methods

Comments

130

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

131

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

132