Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Lecture 11:
Diagnosing Nonlinear Associations

February 27, 2015

1

Lecture Outline

- Modeling complex “dose-response”
- Diagnostics for shape of association
- Multiple comparisons
- Comparing models

2

Modeling Complex "Dose-Response"

Linear Predictors

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

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

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

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.

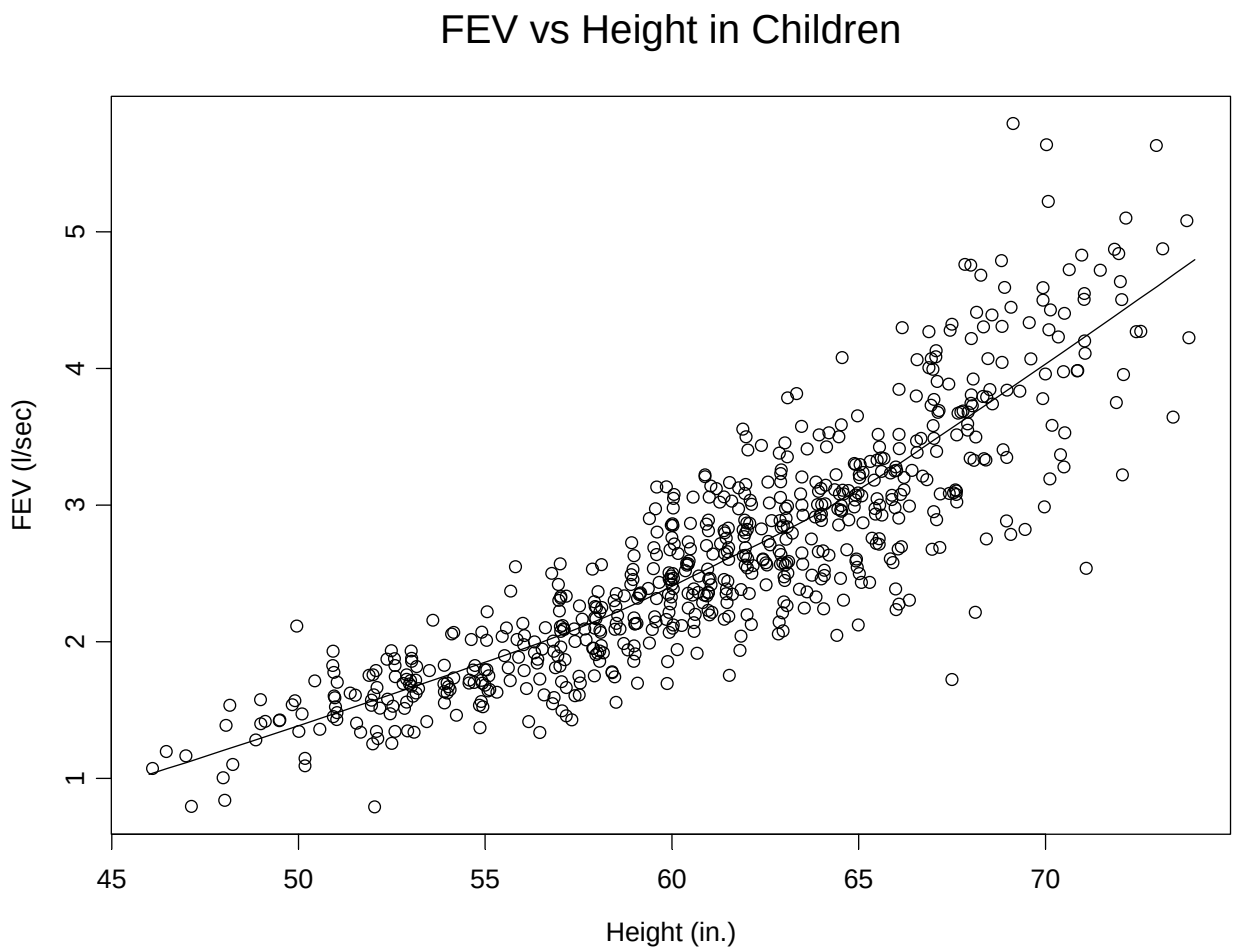
5

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

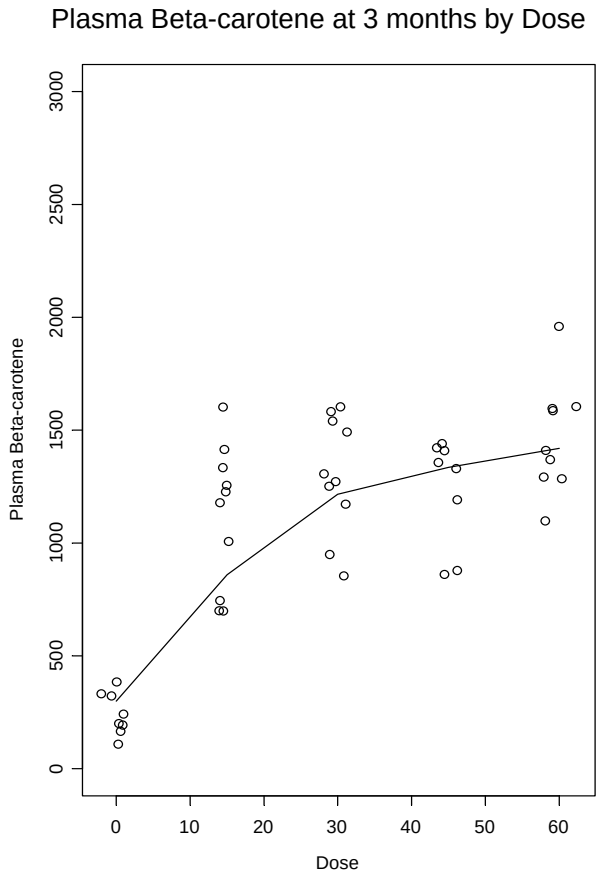
Ex: Cubic Relationship

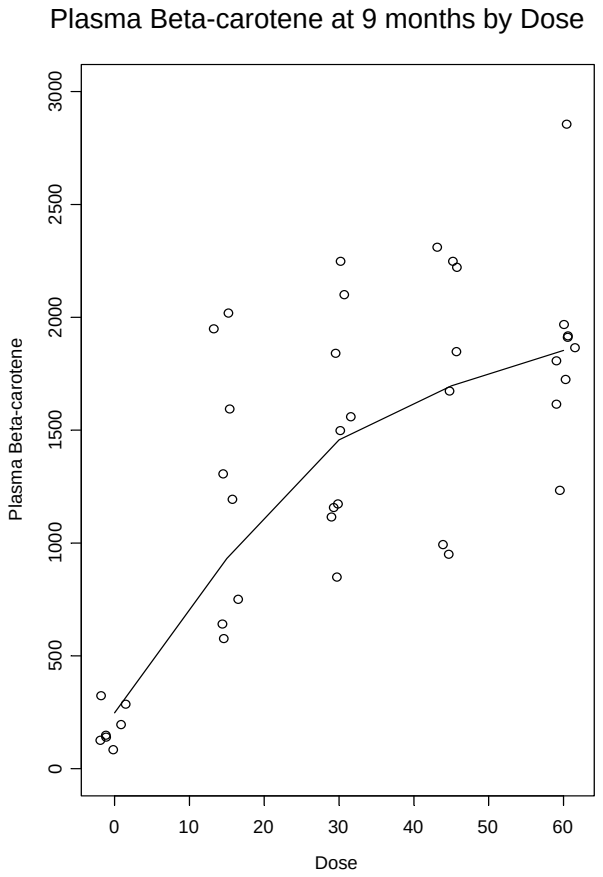


7

Ex: Threshold Effect of Dose?

- RCT of beta carotene supplementation: 4 doses plus placebo

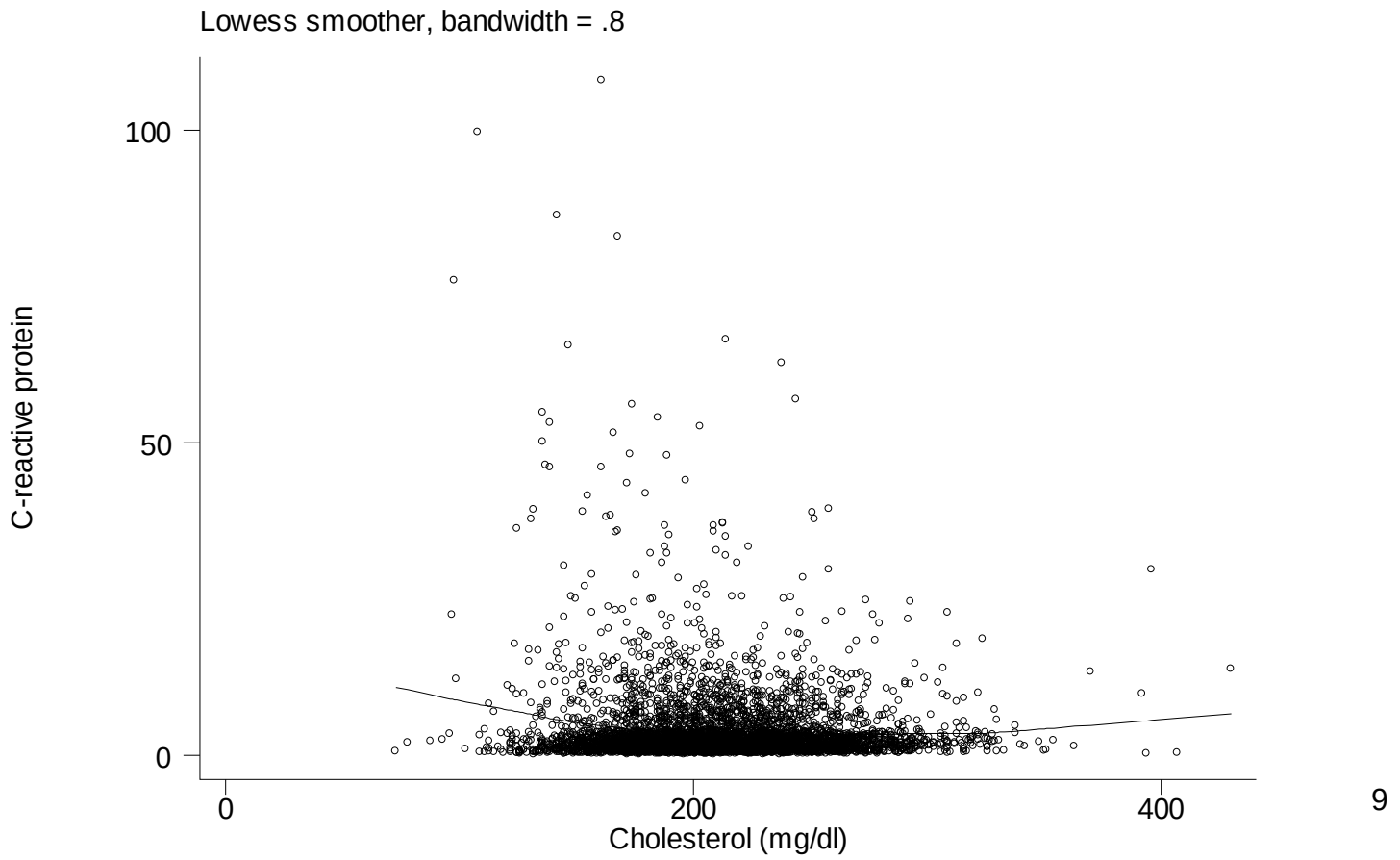




8

Ex: U-shaped Trend?

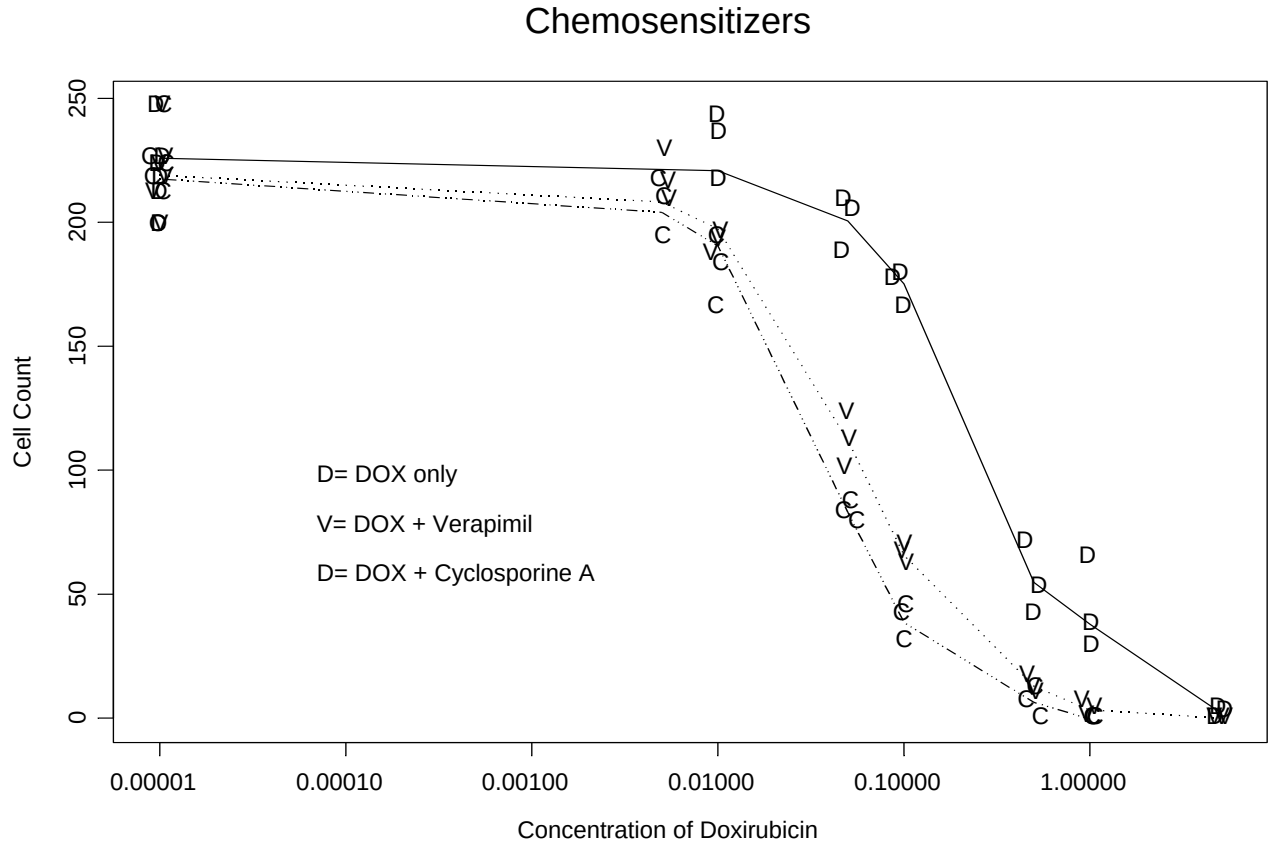
- Inflammatory marker vs cholesterol



9

Ex: S-shaped trend

- *In vitro* cytotoxic effect of Doxorubicin with chemosensitizers



10

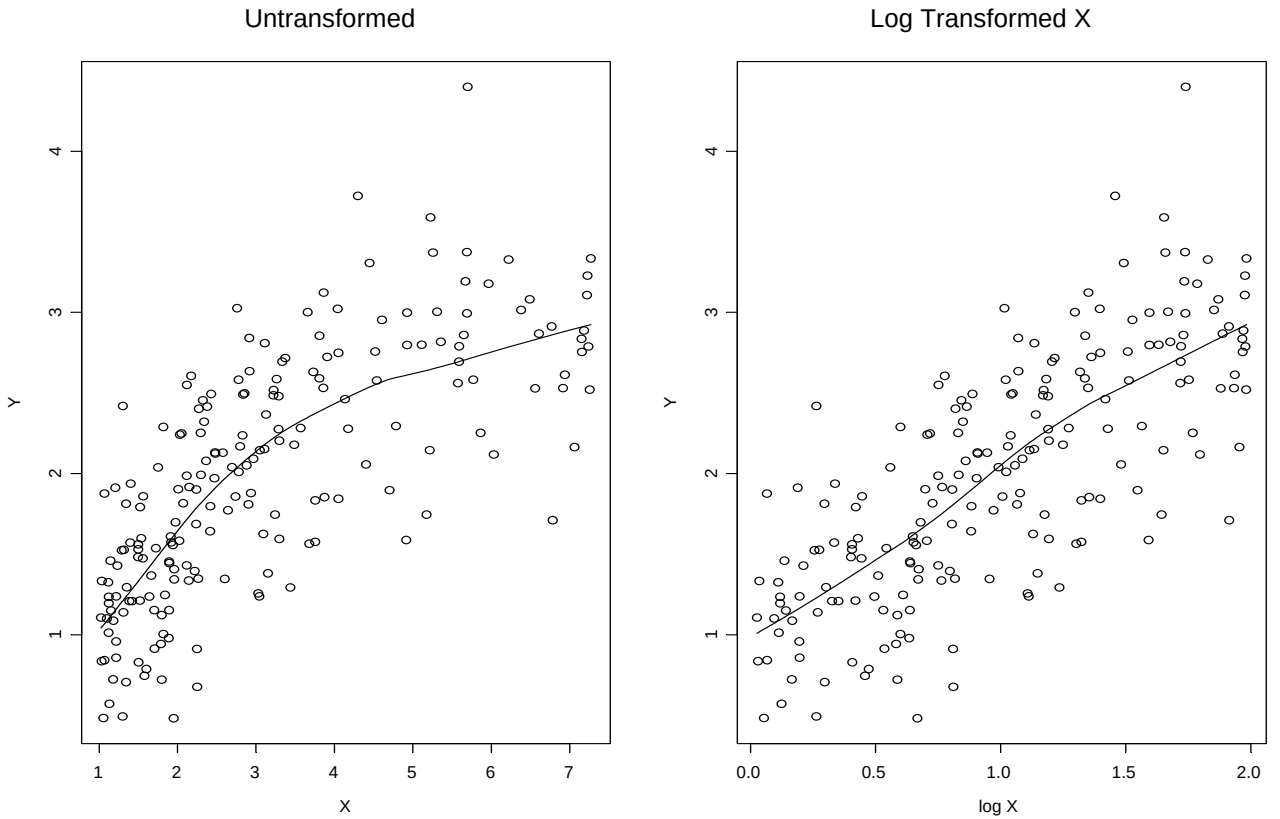
"1:1 Transformations"

- 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

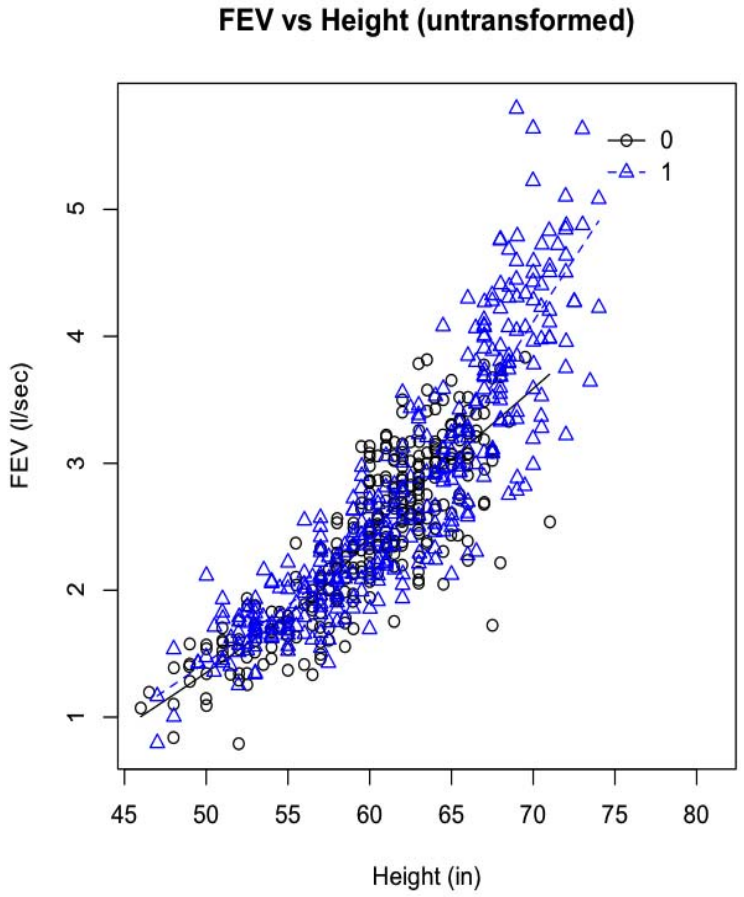
Log Transformations

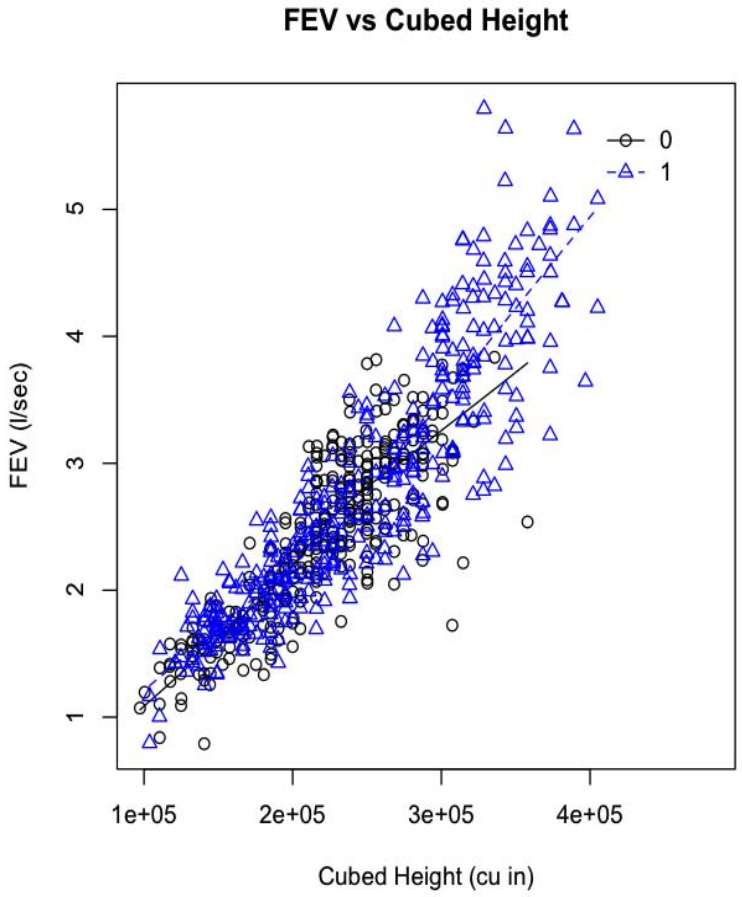
- Simulated data where every doubling of *X* has same difference in mean of *Y*



12

Cubic Transformation: FEV vs Height





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 for Shape of Association

16

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

17

Statistical Questions: Classification

1. Clustering of observations
   - Perhaps into groups that might be different diseases
2. Clustering of variables
   - Perhaps into groups representing biochemical pathways
3. Quantification of distributions
   - Perhaps reporting mean life expectancy after diagnosis
4. **Comparing distributions**
   - **Perhaps investigating associations between variables**
5. Prediction of individual observations
   - Perhaps diagnosing disease or estimating kidney function

18

4. Investigating Associations

- Our scientific questions can be at many different levels of detail
  1. Is there an association?
  2. What is the general (first order) trend in *Y* with higher *X?*
  3. Is their a nonlinear trend in the association?
  4. Is the general trend a particular shape?
     - Increasing exponentially?
     - Increasing to a threshold?
     - Constant then decreasing?
     - U-shaped?
     - S-shaped?
  5. What is the association at particular levels of *X?*
     - E.g., What is the difference in odds of mortality between subjects with LDL of 160 and 161 mg/dL?
- Any questions can be about associations independent of other mechanisms (i.e., adjusted for potential confounding)

19

"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"

20

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

21

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

22

Ex: FEV – Height Assoc Linear?

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

23

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

24

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
```

25

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)

26

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

27

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

28

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
```

29

Ex: FEV – Ht Associated? Interpretation

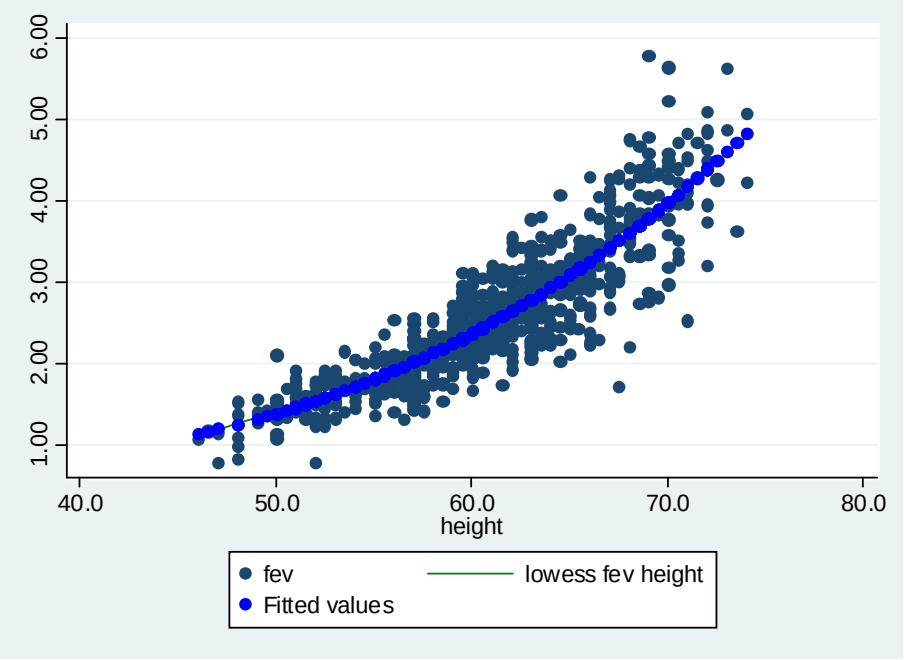
- We thus find strong evidence for a statistically significant association between FEV and height (P < 0.0001)
  - (Not a surprise)

- 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

30

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))
```



31

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

32

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    -.4170     .5584
       htsqr |    -.0002   .00410    -0.04   0.964    -.0082     .0078
       htcub |  3.20e-07   .00002     0.01   0.989    -.0000     .0000
       _cons |    -2.79    4.985    -0.56   0.576    -12.60     6.997
```

33

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

34

Ex: log FEV – Ht Assoc Linear?

```
. test htsqr htcub

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

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

35

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)

36

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

37

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    -.4170    .5584
       htsqr |    -.0002   .00410    -0.04   0.964    -.0082    .0078
       htcub |   3.2e-07   .00002     0.01   0.989    -.0000    .0000
       _cons |    -2.79    4.985    -0.56   0.576    -12.60    6.997
```

38

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
```

39

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

40

Testing Linearity with Dummy Variables

- 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

41

Ex: Testing Linearity w/ Dummy Variables

- **egen ageCTG= cut(age), at(3 6 9 12 15 20)**
- **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
------------------------------------------------------------------------------
```

42

Ex: Testing Linearity w/ Dummy Variables

- 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
```

43

Testing Linearity with Linear Splines

- 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
```

44

More Detailed Questions

- Testing particular questions about a shape is more difficult
- We have to distinguish between testing in a model that presumes a particular shape and testing in a model that includes a particular shape
  - E.g.: Quadratic model presumes U-shaped, but linear splines could include a U-shaped function as well as many others
- Have to isolate the properties of your question
  - E.g.: U-shaped functions will have different slopes at the extremes of the predictor values
- We will rarely have good power
  - And we need to avoid using too flexible a design, because then we will have very little information
  - We also have to worry about influential points

45

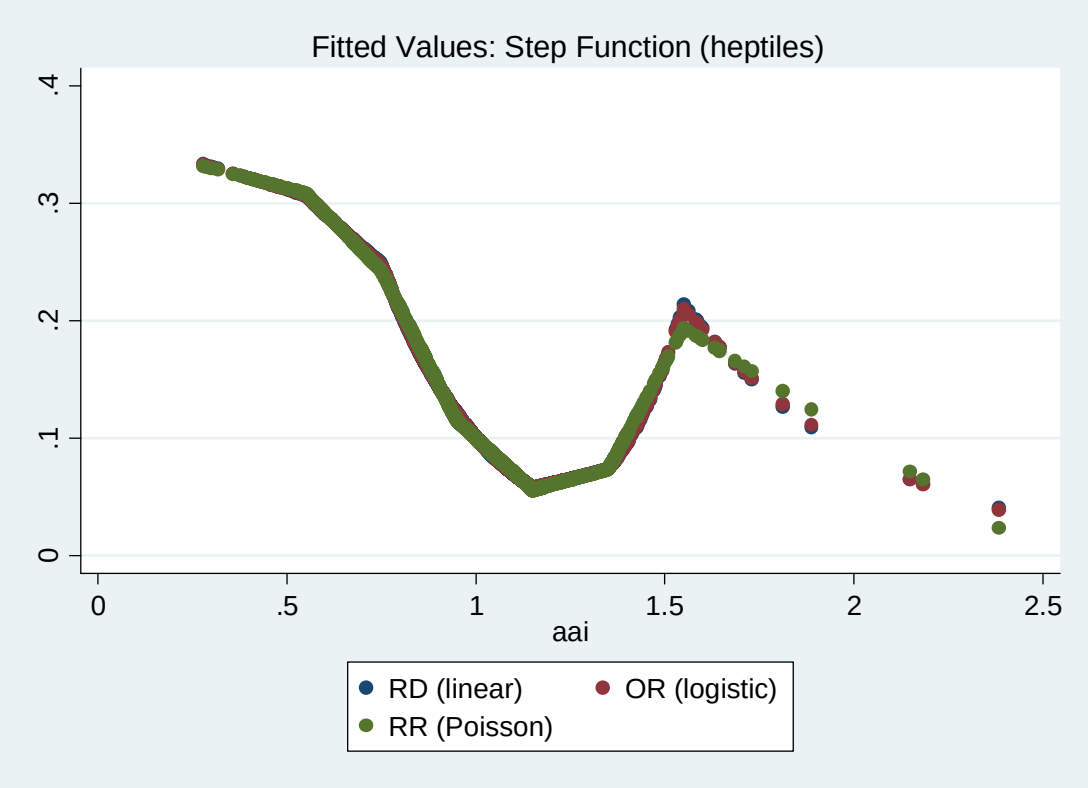
Example: Mortality and Ankle-Arm Index

- U-shaped functions using linear splines model
  - Linear splines in 7 intervals (heptiles)
  - We would need to test that simultaneously
    - saai025 slope is “negative” (respectively, “positive”)
    - saai155 slope is “positive” (respectively, ‘negative”)

```
. logistic deadin4 saai*
Logistic regression                     Number of obs   =       4879
                                        LR chi2(7)      =     152.99
                                        Prob > chi2     =     0.0000
Log likelihood = -1496.5831             Pseudo R2       =     0.0486
------------------------------------------------------------------------------
deadin4 |      OR   StdErr      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
saai025 |    .650    1.86    -0.15   0.880      .0024     177
saai055 |    .221    .370    -0.90   0.367       .008     5.88
saai075 |    .0125   .0148    -3.71  0.000     .00124     .127
saai095 |    .0180   .0172    -4.20  0.000      .0027     .117
saai115 |    3.67    5.67     0.84   0.399      .178      75.7
saai135 |    426     1381     1.87   0.062      .740     245402
saai155 |    .106    .368    -0.65   0.518      .00012    94.3
```

Example: Mortality vs AAI (Plots)

- Using linear splines, we find a suggestion of a nonlinear effect
  - I can use linear splines to mimic a smooth to the data
  - (Linear splines can handle the link functions in RR and OR)



47

Example: Mortality and Ankle-Arm Index

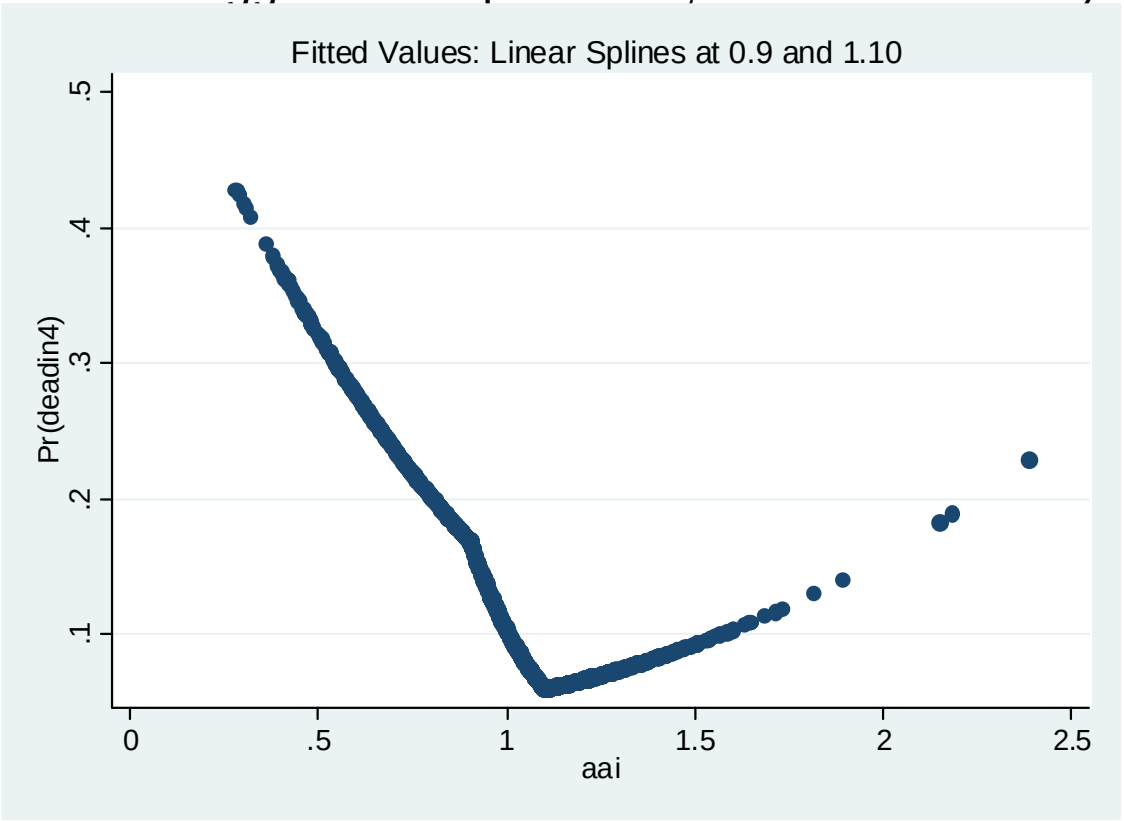
- U-shaped functions using linear splines model
  - Linear splines with three groups
  - We would need to test that simultaneously
    - s2aai025 slope is “negative” (respectively, “positive”)
    - s2aai110 slope is “positive” (respectively, ‘negative”)

```
. logistic deadin4 s2aai*
Logistic regression                     Number of obs   =       4879
                                        LR chi2(3)      =     150.39
                                        Prob > chi2     =     0.0000
Log likelihood = -1497.8815             Pseudo R2       =     0.0478
------------------------------------------------------------------------------
     deadin4 |      OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    s2aai025 |    .123   .0678    -3.80   0.000     .0418    .363
    s2aai090 |    .003   .0024    -7.27   0.000     .0006    .014
    s2aai110 |   3.330    2.27     1.76   0.078     .875     12.7
```

48

Example: Mortality vs AAI (Plots)

- We find suggestion of a nonlinear effect
  - Linear splines with 2 knots
  - Estimates suggest U-shaped trend, but not statistically significant



49

Multiple Comparisons

50

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

51

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

52

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

53

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"

54

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."

55

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

56

Mathematical Basis

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

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

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

57

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

58

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

59

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?

60

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

61

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

62

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

63

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:

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

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

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

64

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

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

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

65

Consider Alternative Models

- We can test each with a type I error of .05
  - Linear continuous
  - Quadratic
  - Dummy variables

- What if we take the lowest of p values from multiple models
  - Linear, quadratic ➔ 0.05 for each goes to 0.0749
  - Linear, dummy ➔ 0.05 for each goes to 0.0874
  - All three ➔ 0.05 for each goes to 0.1046

- Example simulations: With a true linear effect, power increases by less than .03
  - 0.228 ➔ 0.256 or 0.270
  - 0.664 ➔ 0.700 or 0694

66

Comparing Models

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

68

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?

69

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

70

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

71

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

72

Example: One Sample Binomial

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

$$H_0 : p = p_0$$

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

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

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

73

Regression Models

- With regression models, formulas 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

74

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

75

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: $2 \times (\log L_{Full} - \log L_{Red})$
  - Asymptotically chi square distribution in “regular” problems
  - Degrees of Freedom = number of covariates removed from full model

76

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`*

77

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

78

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
```

79

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
```

80

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

81

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?

82

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

83

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

84

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

85