Functional Form:

Modeling Continuous Predictors

Greenland, 1995
Altman and Royston, 2006

Dave Glidden
20 September 2016

Lee Example

- Cohort of community-dwelling elders
- Predictors: age, co-morbidities, functional status: > 48 predictors
- Follow-up: mortality over 4 years
- Logistic regression model
- Age is a continuous variable

Lee, SJ et al. JAMA. 2006;295:801-808.

Summary Statistics

```
 summm age, detail

                 age at time of 1998 interview
-------------------------------------------------------------
      Percentiles      Smallest
 1%           51             50
 5%           53             50
10%           55             50       Obs               11701
25%           59             50       Sum of Wgt.       11701

50%           66                      Mean           67.22169
                        Largest       Std. Dev.      10.10853
75%           75            101
90%           82            101       Variance       102.1825
95%           85            102       Skewness       .4472938
99%           92            103       Kurtosis       2.458463
```


Distribution of Ages

- Looking for outliers
- Look to see where most values are
  90% of ages 53 and 85
- *DONT CARE IF THEIR NORMALLY DIST doesn't matter*
- *It doesn't.*

Modeling Challenge

Levels of Complexity

- age less than <65
- age 50-60, 60-70,70-80, 80-90,90+
- age as a "linear term"
- age + age^2
- other transformation (e.g. log)
- smoothing

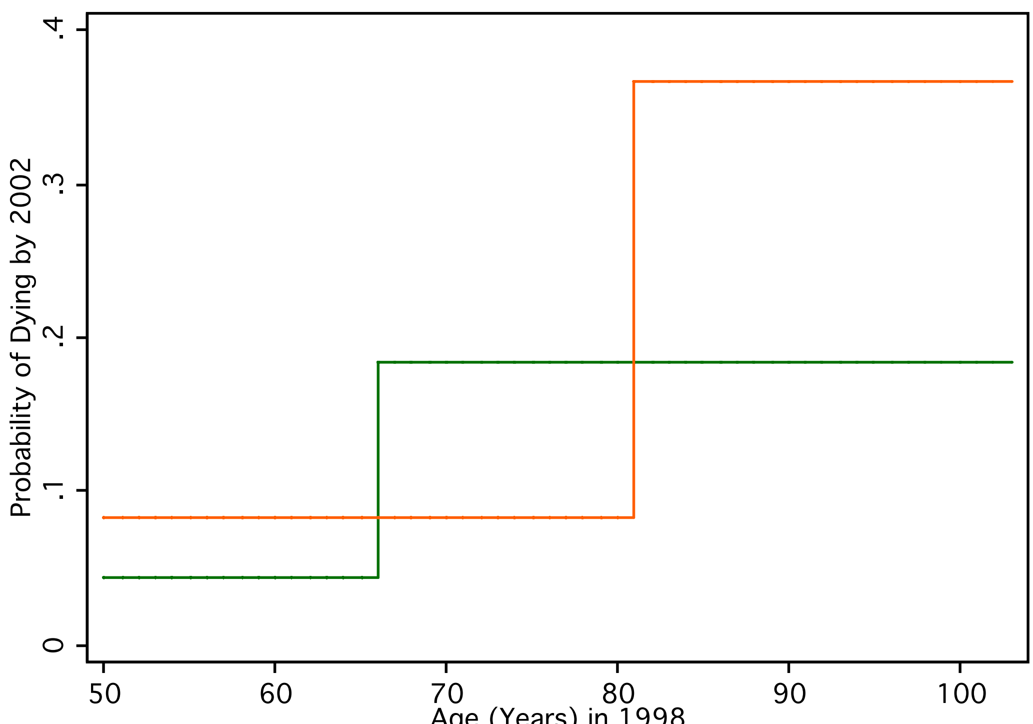
Age Cut at 65

```
Logistic regression                               Number of obs   =      11701
                                                  LR chi2(1)      =     592.71
                                                  Prob > chi2     =     0.0000
Log likelihood =  -3910.362                       Pseudo R2       =     0.0704

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      over65 |   4.829268   .3501685    21.72   0.000     4.189489    5.566749
------------------------------------------------------------------------------
```

Odds Ratios

Age 1	Age 2	Odds Ratio
60	64	1
64	66	4.8
71	79	1
79	81	1

Two Fits



How to Choose Cutpoint

- Strongly discourage making binary variable
- Lose a lot of information
- Cutpoint can be important choice
  - data: no unambiguous choice
  - avoid choosing "most significant" cutpoint

More Levels

- Cut age at quartiles: 25, 50, 75
  ages 59, 66, 75
- Use as a 4-level ordinal categorical variable
- Why use trend test?

Age by Quartile

```
. xi: logistic died i.age4
i.age4            _Iage4_1-4          (naturally coded; _Iage4_1 omitted)

Logistic regression                                 Number of obs   =      11701
                                                    LR chi2(3)      =     949.71
                                                    Prob > chi2     =     0.0000
Log likelihood = -3731.8609                         Pseudo R2       =     0.1129

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iage4_2 |   2.299158   .2939617     6.51   0.000     1.789523     2.95393
    _Iage4_3 |   3.842798   .4605231    11.23   0.000     3.038359    4.860222
    _Iage4_4 |   12.90059   1.454704    22.68   0.000      10.3425    16.09137
------------------------------------------------------------------------------
```

Learn anything about shape?

Interpretation

- OR is 12.9 comparing 4th quartile to first often need that 1 number
- Common summary for continuous *especially in NEJM*
- Quartile can be very specific to your data

Odds Ratios

Age 1	Age 2	Odds Ratio
60	64	1
64	66	1.68
71	79	3.34
79	81	1

Issues

Binary and Categorical

- Con:
  - potentially biased
  - subtle choices
  - residual confounding
- Pro
  - very simple to understand
  - easy to "act" on cutpoints

Continuous Predictor

```
. logistic died age

Logistic regression                               Number of obs   =      11701
                                                  LR chi2(1)      =    1147.48
                                                  Prob > chi2     =     0.0000
Log likelihood = -3632.9744                       Pseudo R2       =     0.1364

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.103388   .0034619    31.36   0.000     1.096623    1.110194
------------------------------------------------------------------------------
```

Interpretation?
Scale is huge issue here

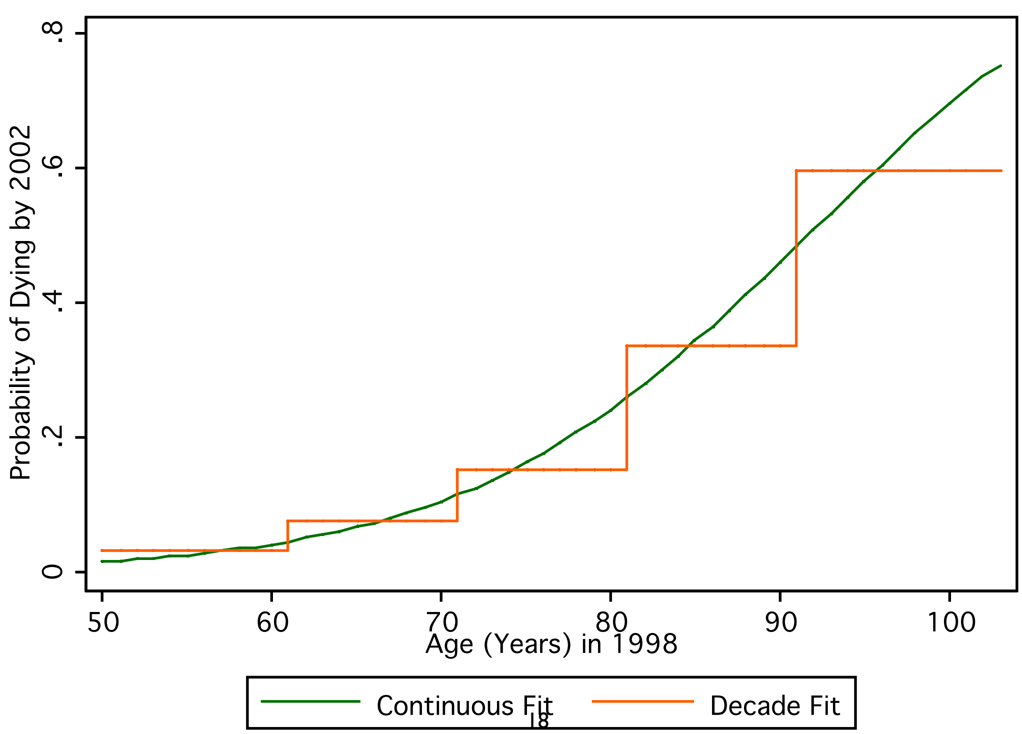
Changing Scale

```
lincom 5*age
 ( 1)  5 age = 0

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.635463   .0256563    31.36   0.000     1.585942    1.686529
------------------------------------------------------------------------------


lincom 10*age
( 1)  10 age = 0

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.674738     .08392    31.36   0.000     2.515213     2.84438
------------------------------------------------------------------------------
```

Induced Odds Ratios

Age 1	Age 2	Odds Ratio
60	64	1.48
64	66	1.22
71	79	2.2
79	81	1.22

Linear v. Decade Fit



Linear Fit

- Can give good easy predictions
- Requires only single df: parsimonious
- Avoids overfitting!!
  not big concern in this data
- Bias/variance tradeoff
  may give better predictor for very old
- Single parameter output
  every 10 year increases odds 2.67

Does the Model Fit

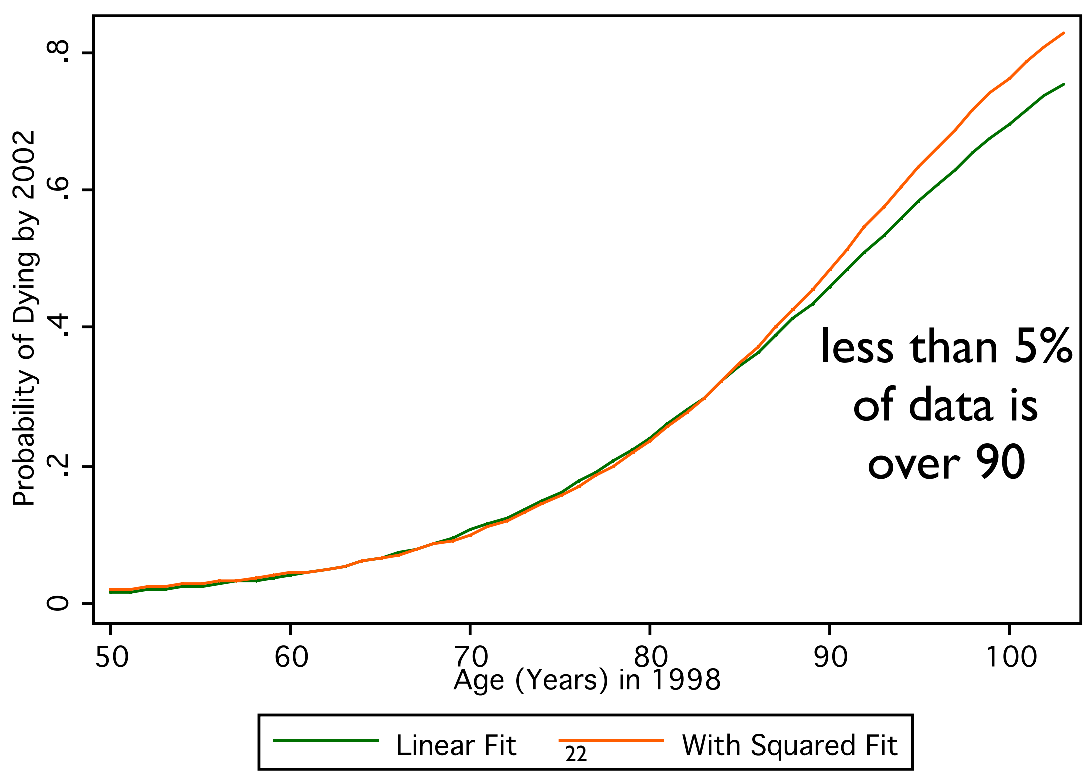
- Maybe the very simple model is missing a pattern
- Multiple categories don't have to give monotone risk
- One way to check: squared term
  squared terms have problems

Quadratic Fit

```
. logistic died age age2

Logistic regression                               Number of obs   =      11701
                                                  LR chi2(2)      =    1151.82
                                                  Prob > chi2     =     0.0000
Log likelihood = -3630.8075                       Pseudo R2       =     0.1369

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.016908   .0397754     0.43   0.668     .9418631    1.097932
        age2 |   1.000557   .0002666     2.09   0.037     1.000034    1.001079
------------------------------------------------------------------------------
```

Squared term “adds” to the model $p < 0.05$

Difference in Fits



Linear v. Quadratic Fit

Age 1	Age 2	OR Linear	OR Quad
60	64	1.48	1.41
64	66	1.22	1.2
71	79	2.2	2.23
79	81	1.22	1.24

Squared Terms

- Very prone to outliers
  outliers become magnified
  very non-robust models
  can affect fit far from outlying values
- Statistical v. practical significance in model improvement

Log transformation

- Does not add degrees of freedom per se
- But can handle "non-linear" patterns
- Produces a useful interpretation
- Interpretation in terms of % increase in predictor
- Particularly helpful when measurement is unfamiliar

Results of Log Fit

```
. logistic died lage

Logistic regression                               Number of obs   =      11701
                                                  LR chi2(1)      =    1132.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -3640.7019                       Pseudo R2       =     0.1346

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lage |   1166.966    270.697    30.44   0.000     740.6419     1838.69
------------------------------------------------------------------------------
```

Interpretation problematic

Results of Log Fit

```
. lincom 0.095*lage

 ( 1)  .095 lage = 0

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.956008   .0431042    30.44   0.000     1.873324    2.042341
------------------------------------------------------------------------------
```

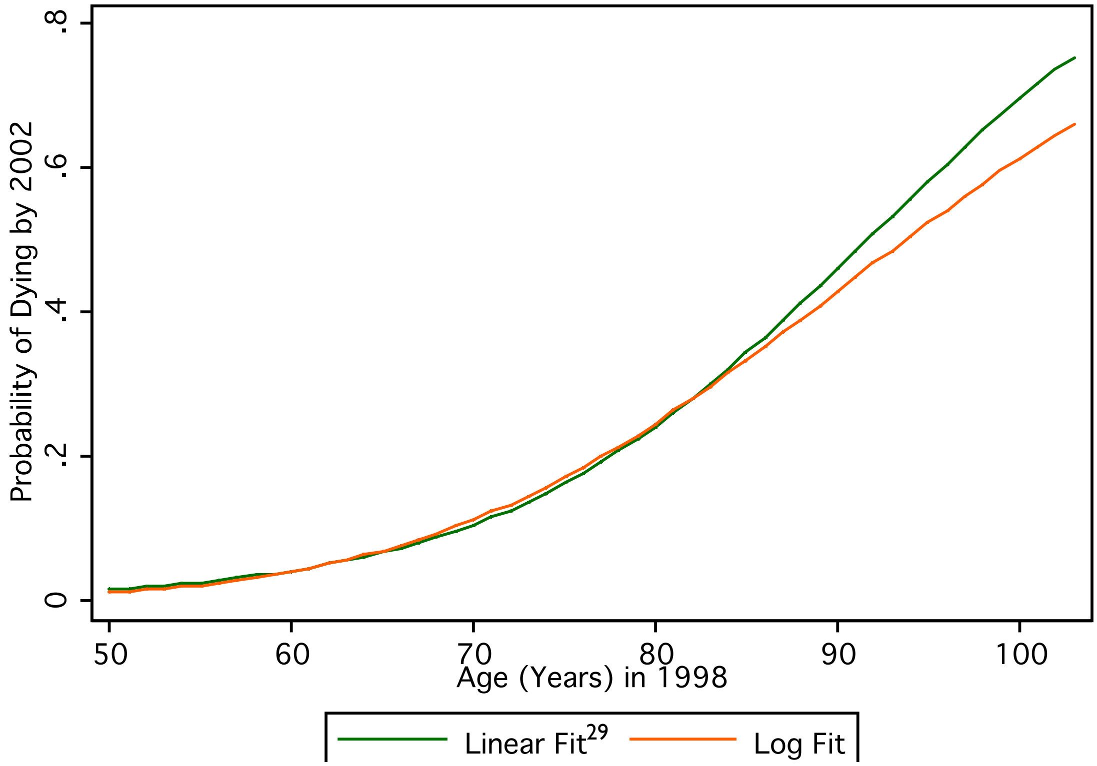
10% increase in age
same as multiplying age by 1.1
same as log(1.1) = 0.095 increase in log age

10% increase in age => HR 1.96

Log v. Linear Fit

Age 1	Age 2	OR Log	OR Linear
60	64	1.58	1.48
64	66	1.24	1.22
71	79	2.12	2.2
79	81	1.19	1.22

Linear v. Log Fit



Consider log transform

- Log likelihood reveals no improvement in fit
- In addition, interpretation not helpful
- Sometimes hear "twice your age"
- Generally talk about being 1,5, etc years older
- Additive scale more natural

Splines

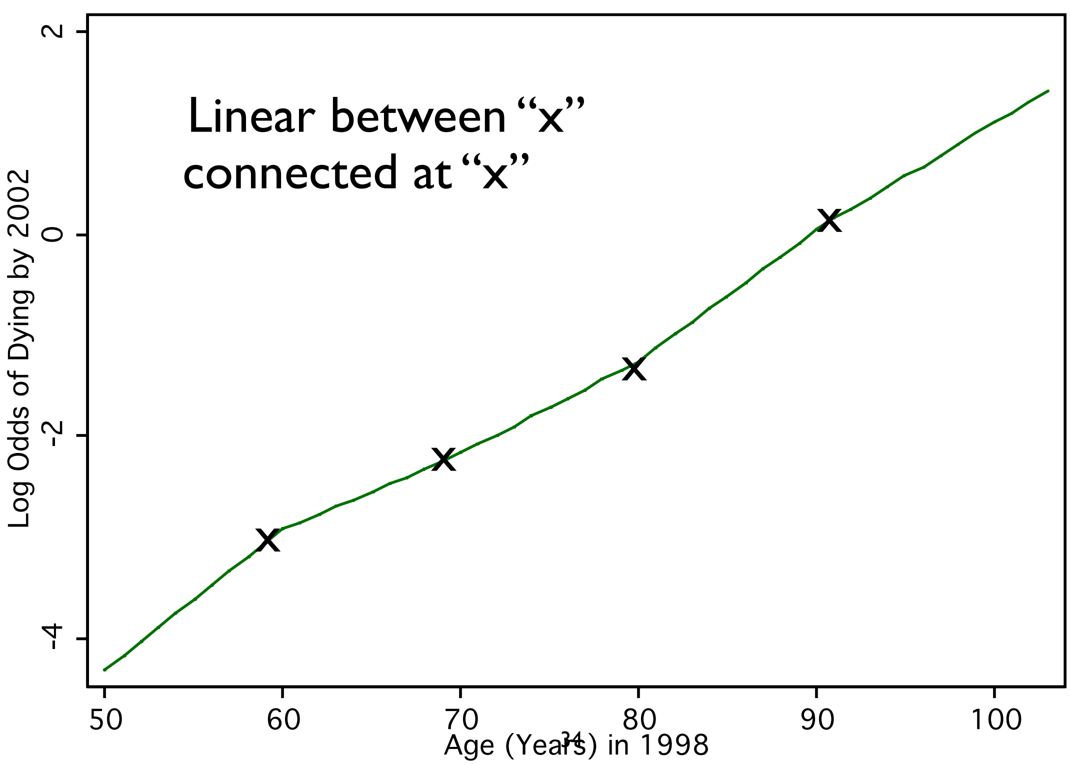
Splines

- A tool for producing smooth functions
- Very amenable to regression models
- Can help in describing complex shape *while adjusting for other variables*
- Similar across regression models

Heuristic

- Approximate complex function with
- Multiple simple functions
- Simplest example: linear spline
- Suppose use simple lines to approximate effect of age

Linear Spline



Knots

- “x” are called knots
  control how much shape can take
- Where to put knots?
- How many to put down?
- First, not a big effect
- Later can have big effect

of Knots

- Controls how eccentric curve can be
- DF = # Knots -1
- Too many knots: overfitting
  noisy curve
- Too few: bias
- Bias/variance trade off
- 3-5 knots is plenty

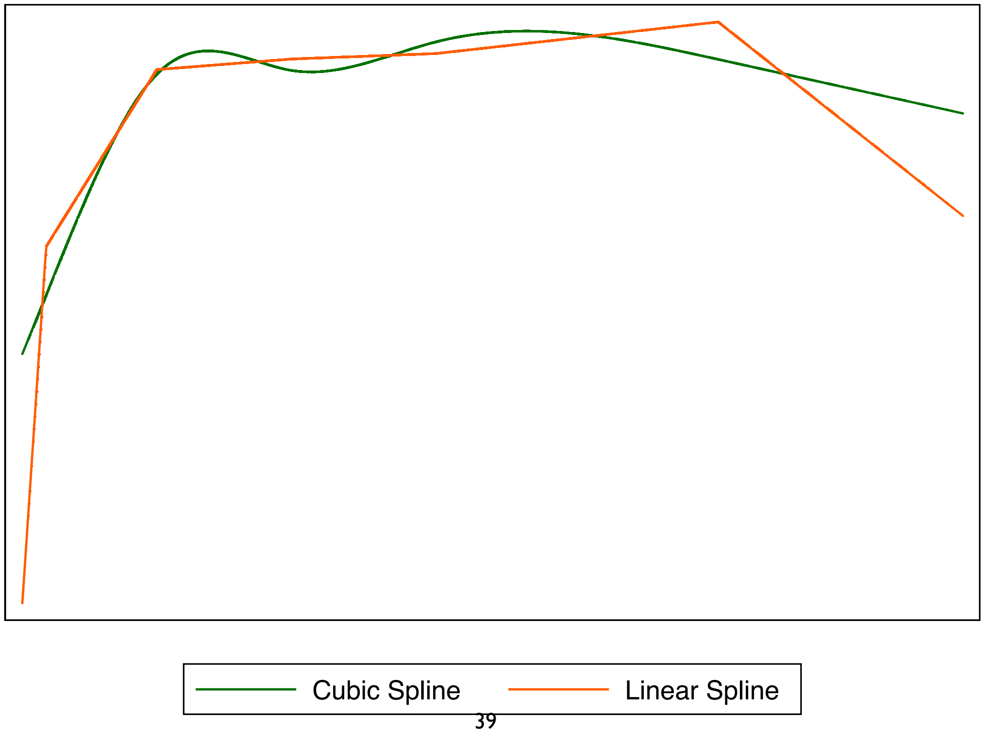
Advantage Over Cutpoint

- Get much more information for df
- Placement of points not as critical
- Extremely flexible methods

Cubic Splines

- Type of spline functions
- Fits cubic polynomial between knots
  more flexible than a line
- Local function: not susceptible to outliers
- Often need to constrain tails
- Produces nice, smooth looking functions

Cubic v. Linear Spline



Stata

```
. mkspline cube_age = age, cubic nknots(3) displayknots

             |     knot1      knot2      knot3
-------------+---------------------------------
         age |        55         66         82
```

```
. summ cube_age*

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   cube_age1 |     11701    67.22169    10.10853         50        103
   cube_age2 |     11701    5.493749    7.357163          0   43.18518
```

Model Fit

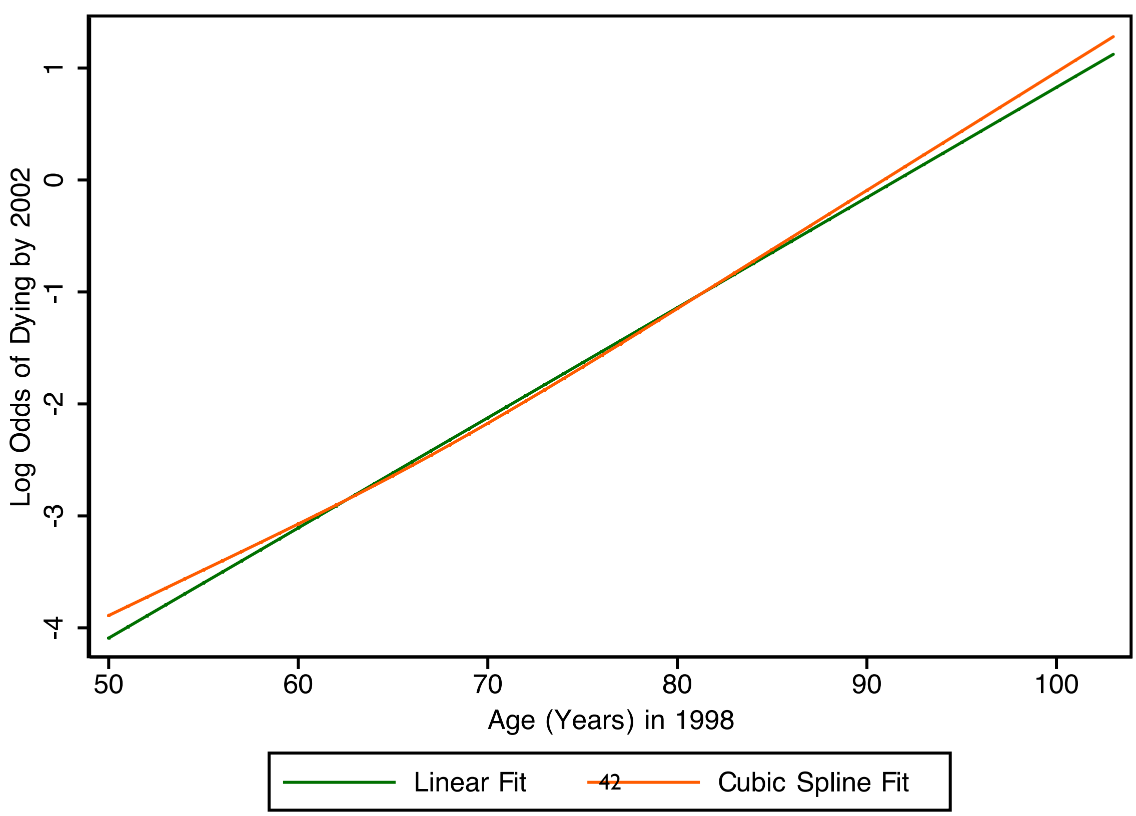
```
. logistic died cube_age*

Logistic regression                               Number of obs   =      11701
                                                  LR chi2(2)      =    1149.94
                                                  Prob > chi2     =     0.0000
Log likelihood = -3631.7479                       Pseudo R2       =     0.1367

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   cube_age1 |   1.084836   .0120655     7.32   0.000     1.061444    1.108743
   cube_age2 |   1.019988   .0128014     1.58   0.115     .9952037    1.045389
------------------------------------------------------------------------------
```

coefficient not interpretable

extract prediction

```
predict out_cubic
```


Spline Fit



How to test splines

flexible test for non-linear effects

- Fit spline model
  `logistic died cube_age1 cube_age2`
- Store results
  `est store splineresults`
- Fit model with linear effect
  `logistic died age`
- Apply likelihood ratio test
  `lrtest splineresults`
- LR test works because spline model always nested

How to test splines

simpler alternative for cubic splines

- Fit spline model
  `logistic died cube_age1 cube_age2`
- Note, "first" term is the linear effect
  first term: `cube_age1`
- Subsequent terms allow for non-linearity
  here cubic there is only `cube_age2`
- Test that subsequent terms 0
  `test cube_age2 = 0`

Two Tests

Likelihood Ratio Test

```
. lrtest splineresults

Likelihood-ratio test                                 LR chi2(1)  =      2.45
(Assumption: . nested in splineresults)               Prob > chi2 =    0.1173
```

Wald Test

```
. test cube_age2=0

 ( 1)  [died]cube_age2 = 0

           chi2(  1) =    2.49
         Prob > chi2 =    0.1148
```


Adjusted Splines

- Predicted outcome (prob, mean) by cts predictor
- All other variables "held constant"
- Race adjusted mortality: requires we specify racial composition of ref. population
- Requires that we margin race out of predictions
- According to mix in reference population

Race in Sample

race/ethnicity	Freq.	Percent	Cum.
white, non-hisp	9,467	80.99	80.99
black, non-hisp	1,225	10.48	91.47
other, non-hisp	225	1.92	93.40
hisp	772	6.60	100.00
Total	11,689	100.00	

I *set* reference population to resemble this: 81% white, 10% black, 2% other, 6% latino

Choice is *arbitrary* -- could have chose race mix for those under 65. Or the pop of California

Adjusted Model

```
. logistic died cube_age* i.raceeth

Logistic regression                               Number of obs   =      11689
                                                  LR chi2(5)      =    1165.06
                                                  Prob > chi2     =     0.0000
Log likelihood = -3618.6447                       Pseudo R2       =     0.1387

------------------------------------------------------------------------------
        died | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   cube_age1 |   1.086678   .0121674     7.42   0.000      1.06309    1.110789
   cube_age2 |   1.018119   .0128472     1.42   0.155     .9932481    1.043613
             |
     raceeth |
          2  |   1.416758   .1341638     3.68   0.000     1.176762      1.7057
          3  |   .9241177     .22599    -0.32   0.747     .5722275    1.492402
          4  |   .9216928   .1313255    -0.57   0.567     .6971151    1.218619
             |
       _cons |   .0003029   .0002135   -11.49   0.000     .0000761    .0012061
------------------------------------------------------------------------------
```

Key Code

```
gen adjlogodds = _b[_cons] + _b[cube_age1]*cube_age1 +
_b[cube_age2]*cube_age2 + _b[1.raceeth]*0.80 _b[2.raceeth]*0.10 +
_b[3.raceeth]*0.02 +  _b[3.raceeth]*0.07
```

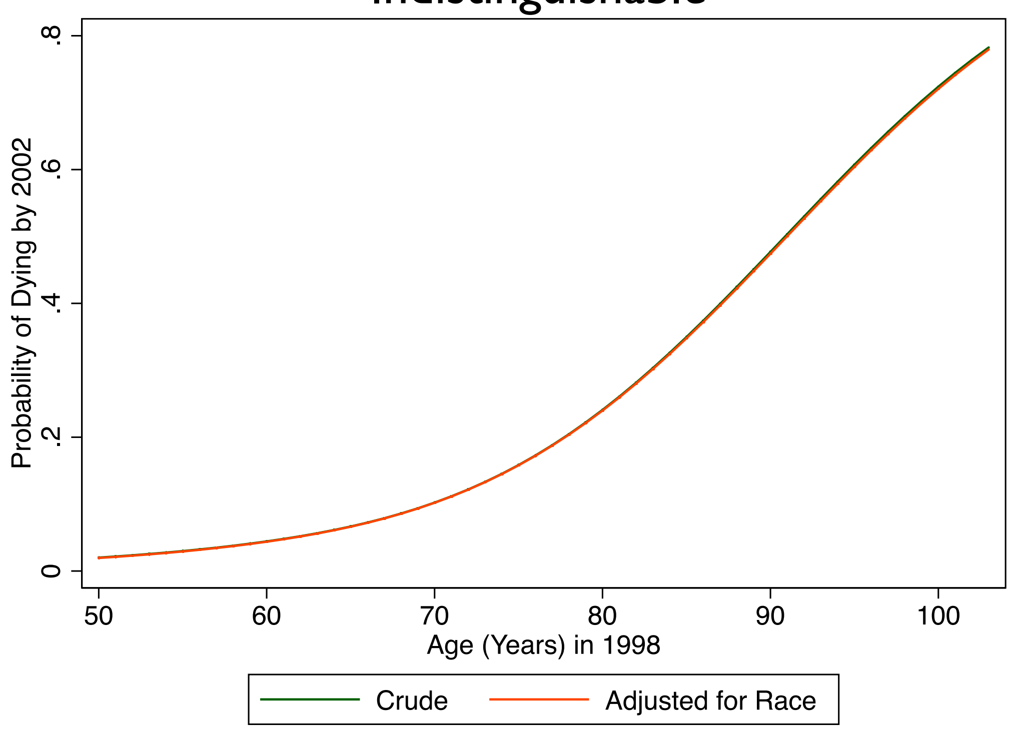
Calculates Logs Odds for the Reference Population

```
gen adjprob = exp( adjlogodds ) /(1+ exp( adjlogodds ))
```

Converts to a Predicted Probability

Crude v. Adjusted

Indistinguishable



Cutpoint Approach

- Binary cutpoint approach always inferior.
  Can always do better
- Multiple cutpoints: better
- Dilemma: flexibility v. degrees of freedom
- Discontinuous: difficult for statisticians
- Groups: appeals to clinicians
- Choice of cutpoint: difficult and important
- Residual confounding is an issue

Transformations:

Polynomial

- Two df but only moderate flexibility
- Very susceptible to outliers
- Outlier can affect curve far away
- Mathematically complex: need to graph
- Statistical v. practical significance

Transformations:
Log Transformation

- Pulls in outliers
- Can produce useful interpretation
  especially for exotic assays
- Mathematically complex
- Don't need to graph
- Can't handle negative values
- No easy test for linear v. log

Linear Fit

- Parsimonious
- No worry about overfitting
  optimizing bias/variance tradeoff
- Often "good enough"
  especially to get rid of confounding
- Need to be careful about the scale
- Usually produces good interpretation
  in the eye of the beholder

Spline: Pros

- Flexible method
- Knot placement not critical
- Can be fit in any regression model
- Allows for check on linear fit
- Potential rich information
  especially for degrees of freedom spent

Spline: Cons

- Knot number is important issue
- Too many knots can lead to overfitting
  could use cross-validation to choose
- Can only be conveyed in a plot
  difficult when adjusting other variables
- Useful when predictor is of HIGH interest

Balance

- overfitting (-)
- residual confounding (-)
- flexibility (+)
- interpretation (+)

Purpose of regression very important