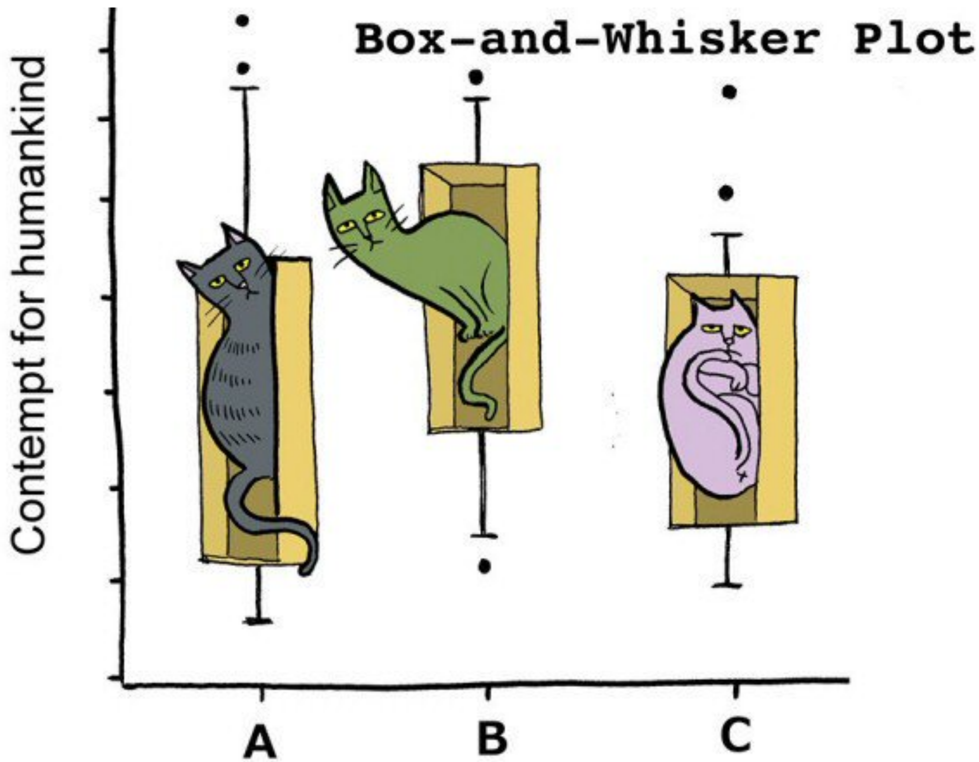
BST 210
Applied Regression Analysis



Lecture 8
Plan for Today

- A few course details
- Smoothing, Splines, Additive Models continued
- Ahead: beginnings of Model Selection- Nested Modeling in context of GAMS
- Questions since last class

Example: Hyponatremia

- **2002 Boston Marathon Results**
- **Boston, MA USA**
  April 15, 2002
- Finishers: 14400; Males - 9149 , Females - 5251
- Male Winner: 2:09:02 | Female Winner: 2:20:43
- Average Finish Time: 3:43:01 | STD: 0:34:23
- Hyponatremia paper was held for publication until April

- Source: Boston Athletic Association (host/sponsor)

Example: Hyponatremia

- Study objective was to investigate predictors of *hyponatremia* in runners of Boston Marathon
- Hyponatremia is continuous -- made binary: low sodium, ≤135 mmol/l
- Study includes 488 Boston marathon runners in 2002
- Hyponatremia is life threatening; low sodium is not good
- 1 death in 2002 Boston Marathon at ~mile 20; hyponatremia suspected

- (Chris) Almond et al. (2005), New England Journal of Medicine, 352:1550-1556. (was an MPH student here)

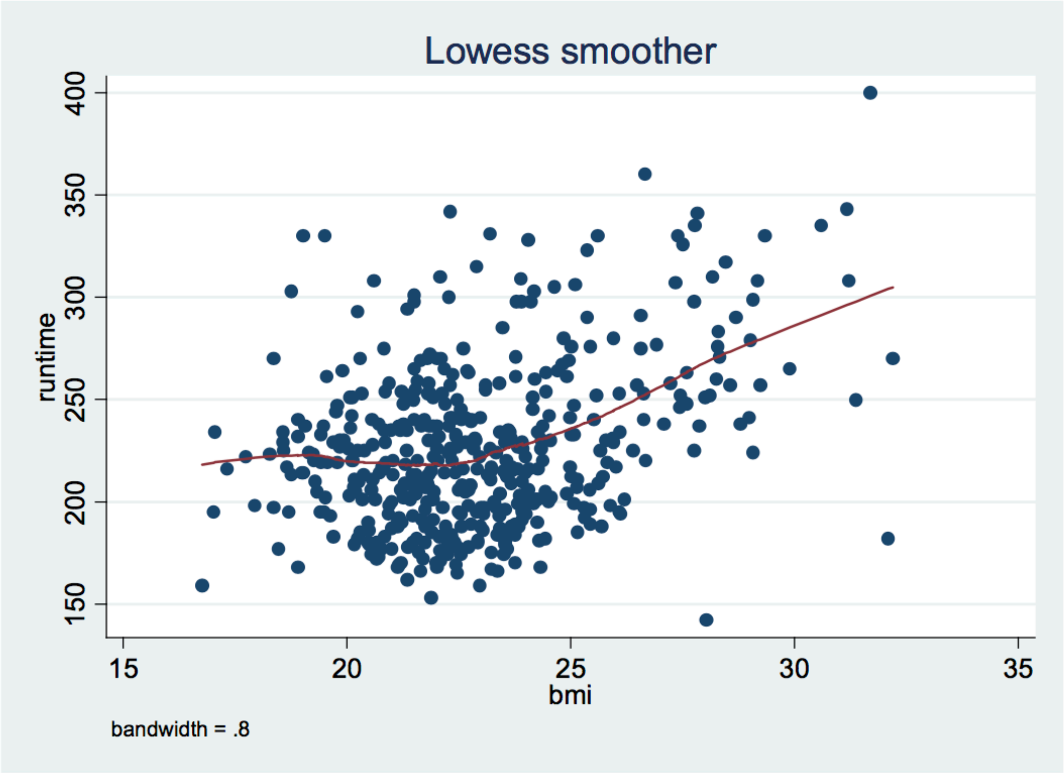
Example: Hyponatremia

- Data not collected on many (488 out of 14k+ runners); no elite runners
- Concerned about measurement error?
- *Self reported*: gender, run time, body mass index, weight gain, water consumption, urination freq, and more
- Useful to have clinically meaningful cutoff for sodium variable (≤135 mmol/l)

Example: Hyponatremia

MARATHON DATASET (Lowess Smoothing)

. lowess runtime bmi



Example: Hyponatremia

- Linear relationship?
- Relationship seems different at BMI = 24 (try piecewise linear?)
- Note: there are various ways to get into Boston Marathon
- For every 1 unit increase in BMI $\leqq$ 24, 'runtime' increases by .39 minutes.
- For every 1 unit increase in BMI > 24, 'runtime' increases by (10.97+.39) minutes, on average.

```
. generate bmi24 = max(0, bmi - 24)

. regress runtime bmi bmi24

      Source |       SS       df       MS              Number of obs =     455
-------------+------------------------------           F(  2,   452) =   49.31
       Model |   134336.02     2   67168.01            Prob > F      =  0.0000
    Residual |  615660.187   452  1362.08006           R-squared     =  0.1791
-------------+------------------------------           Adj R-squared =  0.1755
       Total |  749996.207   454  1651.97402           Root MSE      =  36.906

------------------------------------------------------------------------------
     runtime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .3900295   1.152247     0.34   0.735    -1.874397    2.654456
       bmi24 |   10.97142   2.093023     5.24   0.000     6.858159    15.08468
       _cons |   210.2836   25.36821     8.29   0.000     160.4293    260.1378
------------------------------------------------------------------------------
```

Example: Hyponatremia

. lowess na runtime

Example: Hyponatremia

- Linear relationship?
- What does linear regression say? What is N? What is SE?
- Does the direction of the association make sense?
- For every 100 minutes longer it took to complete Boston, sodium level dropped on average by 1.9 units (mmol/l).

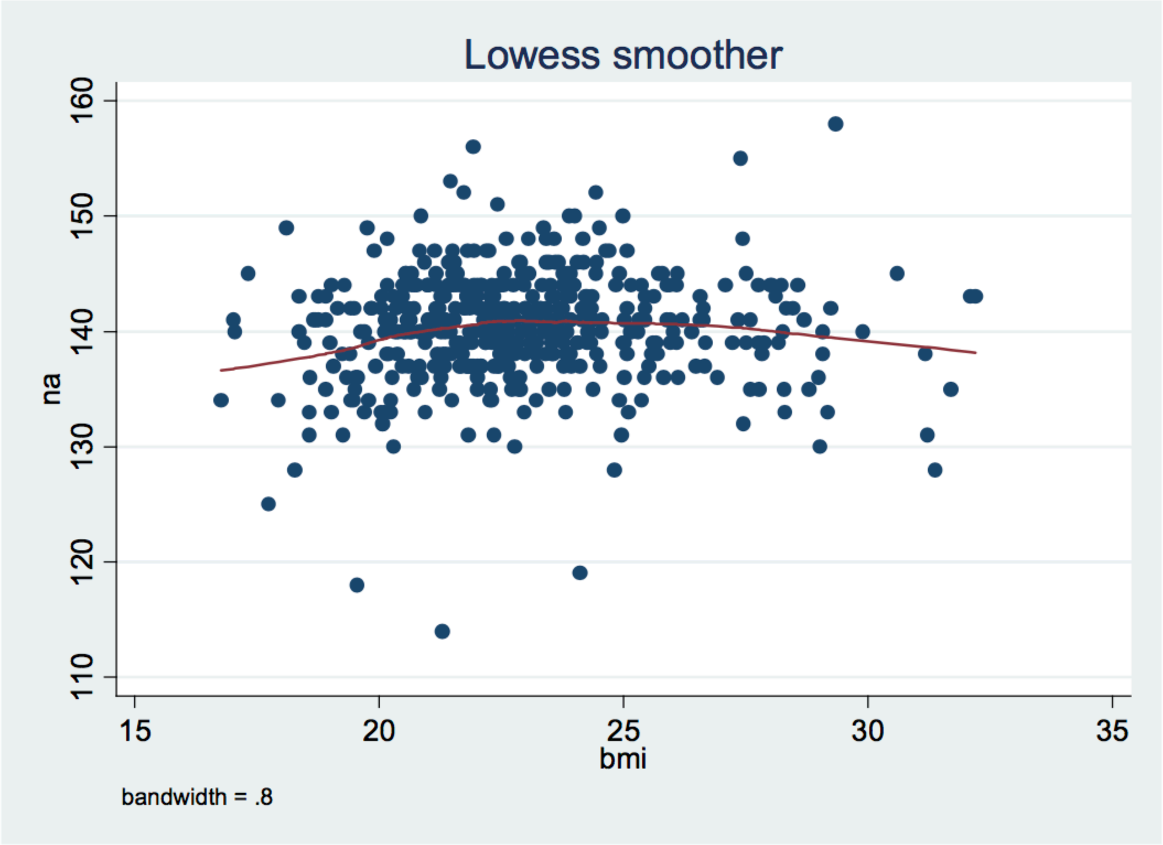
```
. regress na runtime

      Source |       SS       df       MS              Number of obs =     477
-------------+------------------------------           F(  1,   475) =   13.00
       Model |   288.49509     1   288.49509           Prob > F      =  0.0003
    Residual |  10542.2219   475  22.1941514           R-squared     =  0.0266
-------------+------------------------------           Adj R-squared =  0.0246
       Total |   10830.717   476  22.7536071           Root MSE      =  4.7111

------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     runtime |  -.0187156    .005191    -3.61   0.000    -.0289159   -.0085154
       _cons |   144.6237   1.190519   121.48   0.000     142.2844    146.9631
```

Example: Hyponatremia

. lowess na bmi



Example: Hyponatremia

- Linear relationship?
- What does estimated regression model tell us?
- Do we stop here?

```
. regress na bmi

      Source |       SS       df       MS              Number of obs =     465
-------------+------------------------------           F(  1,   463) =    1.48
       Model |  34.1804956     1  34.1804956           Prob > F      =  0.2239
    Residual |  10668.8001   463  23.0427649           R-squared     =  0.0032
-------------+------------------------------           Adj R-squared =  0.0010
       Total |  10702.9806   464  23.0667686           Root MSE      =  4.8003

------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .1005134   .0825281     1.22   0.224    -.0616627    .2626894
       _cons |   138.0882   1.905888    72.45   0.000     134.3429    141.8335
------------------------------------------------------------------------------
```

Example: Hyponatremia

- Consider x^2 relationship instead?
- What does estimated regression model tell us now?
- Those with lowest and highest BMI have on average the lowest sodium levels, and are at higher risk of hyponatremia than those with average BMI.
- What does the added curvature in the model tell us? (we can take first derivative of model equation with respect to bmi, to gain sense of concavity)

```
. generate bmisq = bmi * bmi
(23 missing values generated)

. regress na bmi bmisq

      Source |       SS       df       MS              Number of obs =     465
-------------+------------------------------           F(  2,   462) =    7.82
       Model |  350.647635     2  175.323817           Prob > F      =  0.0005
    Residual |   10352.333   462  22.4076472           R-squared     =  0.0328
-------------+------------------------------           Adj R-squared =  0.0286
       Total |  10702.9806   464  23.0667686           Root MSE      =  4.7337

------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   3.785527   .9839287     3.85   0.000     1.851997    5.719057
       bmisq |  -.0770051   .0204905    -3.76   0.000    -.1172713   -.0367389
       _cons |   94.63835   11.71347     8.08   0.000     71.62006    117.6566
------------------------------------------------------------------------------
```

Example: Hyponatremia

- If we hadn't explored with Lowess, we may have only modeled BMI as linear, and then would have likely deemed it not significant for inclusion in overall model of 'na' (sodium), and we shouldn't be excluding BMI from the model.
- If we hadn't tried a linear regression we wouldn't have appreciated the statistical significance of the effect of 'runtime' on sodium level, despite the 'flat' appearance of the Lowess smooth.
- Lowess aided in developing 'piecewise model' (ie change at 24) for the effect of BMI on 'runtime.'

- Options for picking cutoff of 24?
- Pros and cons of categorizing BMI
- Pros and cons of polynomial BMI
- Concern around the standard categorization of BMI for this particular hyponatremia study?

So, what if uncertain about linearity of effect of a continuous covariate? (from most to least simple)

- Use a categorized version of the covariate
- Use polynomial or nonlinear terms for the covariate
- Use a transformation of the covariate (or outcome) variable
- Use piecewise linear (or cubic) terms for the covariate (also known as splines)
- Use Generalized Additive Models (GAMs)

Use of Categories

- Usually decided in advance
- Based on external data (e.g., WHO guidelines for age groups in malaria)
- Based on equal spaced intervals (e.g., age decades in cancer epidemiology)
- Based on equal sized intervals (e.g., quintiles of air pollution in environmental research)
- Based on your own data snooping – but this may be harder to justify (try cutoff=24, then 25, etc), also what use for criteria?

Advantages/Disadvantages of Categories

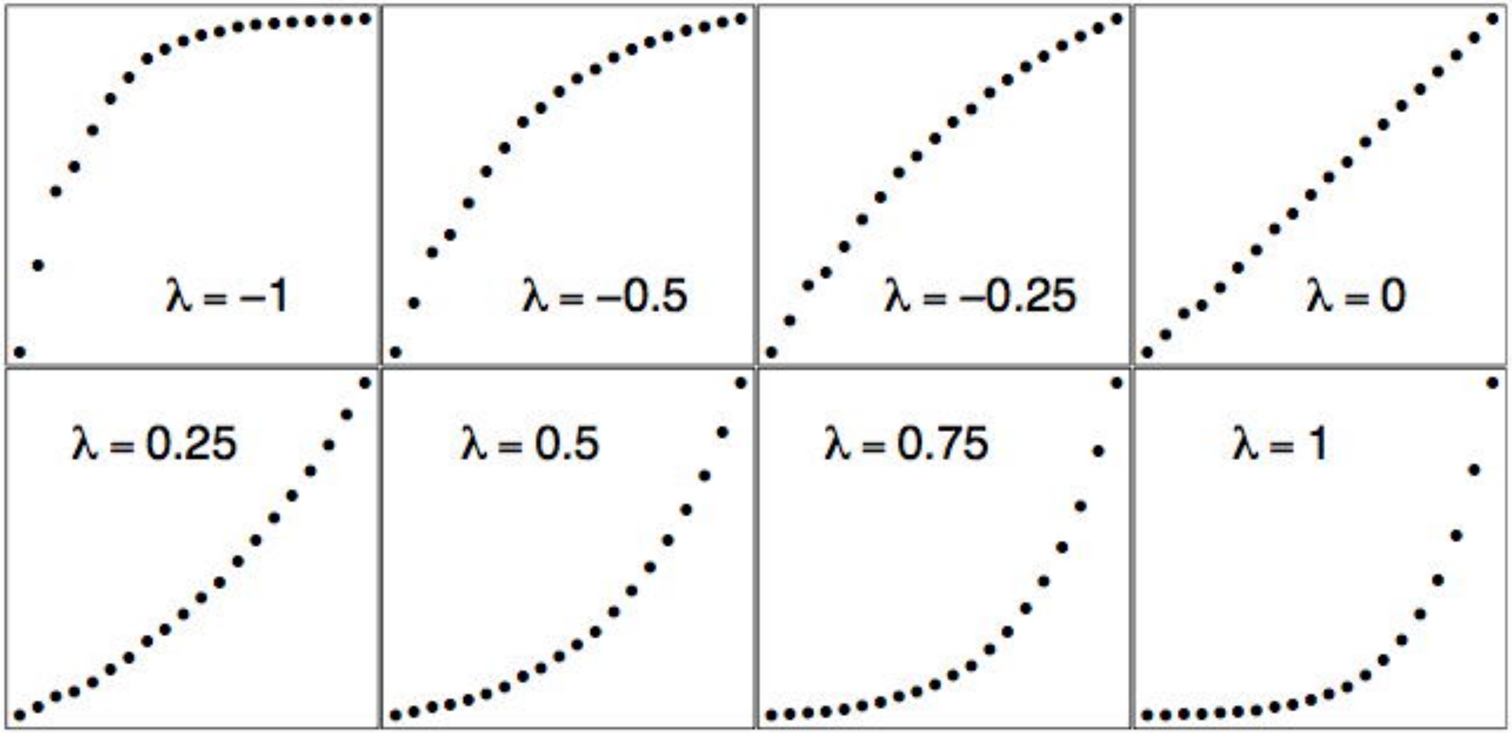
•

Use of Polynomials

- Higher order polynomials: linear, quadratic, cubic, … (or x, x^2, x^3, …) (where to stop—higher for prediction models; terms can be highly correlated)
- Tukey's power transformation: Rather than x [or Y], consider x^λ [or Y^λ] for some optimal choice of λ. Need to choose λ.
- Fractional polynomials (Tukey's ladder of transformations): …, x^{-2}, $x^{-3/2}$, $x^{-1} = 1/x$, $x^{-1/2}$, $\log(x)$, $x^{1/2}$, x, $x^{3/2}$, x^2, … (might include more than one of these in the model) Not clear how to pick the ladder.

Use of Polynomials

Tukey's transforms



Use of Transforms

- Log transforms: $\log(x)$ [or $\log(Y)$]
- $x \log(x)$: (the Box-Tidwell transform, added to a model with x in it, to assess possible nonlinearity not as strong as quadratic
- Box-Cox transform: Rather than x [or Y], consider $(x^\lambda - 1)/\lambda$ [or $(Y^\lambda - 1)/\lambda$] for some optimal choice of λ, where in the limit as $\lambda \to 0$ we get $\log(x)$ [or $\log(Y)$]

Natural Log ($\log_e$ or ln) Transforms

- Recall our original, easy to interpret linear regression model…
- Suppose $E(Y_i) = \beta_0 + \beta_1 \cdot x_i$

  β_1 = slope, or the change in Y corresponding to a 1 unit increase in x
- Here, β_1 is interpretable as the change in the average value of Y for every unit increase in x – *nice, convenient interpretation!*
- Change in both the predictor and outcome is on the measured, or absolute, scale

Natural Log ($\log_e$ or ln) Transforms

- Using a natural log transform (log base e, or ln) on *x* or *Y* can sometimes help achieve linearity of effect, or normalize the residuals (can potentially pull in residuals), making our LINE assumptions be better satisfied
- Sometimes, it might be natural to use a natural log transform on substantive grounds
- How does this affect the interpretation of β coefficients? (percentage change)
- Can you take ln(any value)? ($0<x<\infty$)

Natural Log ($\log_e$ or ln) Transforms

- Suppose instead $E(Y_i) = \beta_0 + \beta_1 \cdot \log(x_i)$

  β_1 = slope, or the change in *Y* corresponding to a 1 unit increase in $\log(x)$

- Here, $\beta_1 \cdot \log(1.01)$ can be interpreted as the change in the average value of *Y* for every 1% increase in *x*

  $$\beta_0 + \beta_1 \cdot \log(1.01 \cdot x_i) = \beta_0 + \beta_1 \cdot \log(1.01) + \beta_1 \cdot \log(x_i)$$

Natural Log ($\log_e$ or ln) Transforms

- For $E(Y_i) = \beta_0 + \beta_1 \cdot \log(x_i)$, $\beta_1 \cdot \log(1.01)$ can be interpreted as the change in the average value of Y for every 1% increase in x
- Using natural logs, $\log(1.01) = 0.00995 \approx 0.01$
- Given a 95% CI for β_1, one can calculate a 95% CI for $\beta_1 \cdot \log(1.01)$
- Estimated effect for a 1% increase in x
- (or percentage change in x, and how does that effect y)

Natural Log ($\log_e$ or ln) Transforms

- For $E(Y_i) = \beta_0 + \beta_1 \cdot \log(x_i)$, $\beta_1 \cdot \log(1.10)$ can be interpreted as the change in the average value of *Y* for every 10% increase in *x*
- Using natural logs, $\log(1.10) = 0.09531$
- Given a 95% CI for β_1, one can calculate a 95% CI for $\beta_1 \cdot \log(1.10)$

Natural Log ($\log_e$ or ln) Transforms

- Instead, suppose $E(\log(Y_i)) = \beta_0 + \beta_1 \cdot x_i$

  β_1 = slope, or the change in log(*Y)* corresponding to a 1 unit increase in *x*

- Here, after some algebra, $100(\exp(\beta_1) - 1)$ can be interpreted as the percentage change in the average value of *Y* for every 1 unit increase in *x*

Natural Log ($\log_e$ or ln) Transforms

- For $E(\log(Y_i)) = \beta_0 + \beta_1 \cdot x_i$, the percent change on the Y scale for a one unit increase in x is given by

$$100 \cdot \frac{\exp\left(\beta_0+\beta_1(x+1)\right)-\exp(\beta_0+\beta_1 x)}{\exp(\beta_0+\beta_1 x)} = 100(e^{\beta_1} - 1)$$

- Given a 95% CI for β_1, one can calculate a 95% CI for $100(\exp(\beta_1) – 1)$

Natural Log ($\log_e$ or ln) Transforms

- For $E(\log(Y_i)) = \beta_0 + \beta_1 \cdot x_i$, the percent change on the *Y* scale for a ten unit increase in *x* is given by

$$100 \cdot \frac{\exp\big(\beta_0+\beta_1(x+10)\big)-\exp(\beta_0+\beta_1 x)}{\exp(\beta_0+\beta_1 x)} = 100(e^{10\beta_1} - 1)$$

- Given a 95% CI for β_1, one can calculate a 95% CI for $100(\exp(10 \cdot \beta_1) - 1)$

Natural Log ($\log_e$ or ln) Transforms

- Finally, suppose $E(\log(Y_i)) = \beta_0 + \beta_1 \cdot \log(x_i)$

  β_1 = slope, or the change in log(*Y*) corresponding to a 1 unit increase in log(*x*)

- Here, $100(\exp(\beta_1 \cdot \log(1.01)) - 1)$ can be interpreted as the percentage change in the average value of *Y* for every 1% increase in *x*
- Here, $100(\exp(\beta_1 \cdot \log(1.10)) - 1)$ can be interpreted as the percentage change in the average value of *Y* for every 10% increase in *x*

Natural Log ($\log_e$ or ln) Transforms

- These interpretations can be made even when adjusting for other factors in the model
- Which log transform to make (if any) depends on the data set you are modelling

Now on to Regression Splines

- Keep in mind throughout our discussion of the many smoothing methods:

Essentially, a smooth just finds an estimate of f in the nonparametric regression function $Y = f(x) + \epsilon$.

Regression Splines

If one estimates f by minimizing the equation that balances least squares fit with a roughness roughness penalty, e.g.,

$$\min_{f \in \mathcal{F}} \sum_{I=1}^{n} [y_i - f(\boldsymbol{x}_i)]^2 + \lambda \int [f^{(k)}(\boldsymbol{x})]^2 \, d\boldsymbol{x} \tag{1}$$

over an appropriate set of functions (e.g., the usual Hilbert space of square-integrable functions), then the solution one obtains are smoothing splines.

* You are not responsible for this, but it's important too see the level of theory that goes into this approach, which is actually a very powerful and commonly used, intuitive method!

Regression Splines

- Splines: just piecewise polynomials with 'pieces' divided at sample values x_i
- Need to choose order of spline, knot points, constraints
- Order of spline: Constant, linear, cubic
  - Knot points: Change points, X values that divide the fit into polynomial portions
  - Constraints: On continuity or derivatives at knots or ends—want smoothness at knots
- Splines are computationally fast, enjoy strong theory, work well, are widely used

Regression Splines

- Piecewise constant is the same as a bin smoother, discontinuous at knots
- Piecewise linear can be made continuous at knots
- Piecewise cubic can match 1^{st} and 2^{nd} derivatives at knots and/or have boundary conditions at ends (e.g., flat, linear)

Piecewise Linear Splines

- Suppose you select knot points (cut points) at $a < b < c$ for a continuous covariate x
- The linear spline function is then given by

$$E(Y) = \beta_0 + \beta_1 x + \beta_2 (x - a)_+ + \beta_3 (x - b)_+ + \beta_4 (x - c)_+$$

where $(u)_+ = u$ for $u > 0$ and 0 for $u \leq 0$

- Interpretation:
- β_2 is change in slope from β_1;
- β_3 is change in slope from $\beta_1 + \beta_2$;
- β_4 is the change in slope from $\beta_1+\beta_2+\beta_3$.

Piecewise Linear Splines

$$E(Y) = \beta_0 + \beta_1 x + \beta_2 (x - a)_+ \\ + \beta_3 (x - b)_+ + \beta_4 (x - c)_+$$

- Thus, *k* knot points would lead to a model with $k + 2$ parameters (including the intercept)
- This linear spline is continuous with slopes changing at each knot point

Piecewise Cubic Splines

- A cubic spline is a spline constructed of piecewise cubic polynomials which pass through a set of knot (or ‘control’) points.
- The second derivative of each polynomial is commonly set to zero at the endpoints, since this provides a boundary condition that completes the system of equations.

Piecewise Cubic Splines

- For knot points at $a < b < c$ for a continuous covariate x, the cubic spline function is given by

$$
\begin{aligned}
E(Y) = {} & \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \\
& + \beta_4 (x - a)_+^3 + \beta_5 (x - b)_+^3 \\
& + \beta_6 (x - c)_+^3
\end{aligned}
$$

- Thus, k knot points would lead to a model with $k + 4$ parameters (including the intercept)

Piecewise Cubic Splines

$$
\begin{aligned}
E(Y) = \beta_0 &+ \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \\
&+ \beta_4 (x - a)_+^3 + \beta_5 (x - b)_+^3 \\
&+ \beta_6 (x - c)_+^3
\end{aligned}
$$

- This cubic spline is continuous and also has continuous first and second derivatives at the knots, so the fit looks smooth to the eye
- Sometimes cubic splines can get wavy near the tails, so placing restrictions can be helpful

(Continuous) Piecewise Polynomials

- Piecewise linear with knot points at 50 and 100 can be modeled continuously as x, $(x-50)_+$, and $(x-100)_+$
- Piecewise cubic with knot points at 50 and 100 can be modeled continuously as x, x^2, x^3, $(x-50)_+^3$, and $(x-100)_+^3$
- Here $(\ )_+$ denotes the positive part, so $(0)_+ = 0$, $(5)_+ = 5$, $(-5)_+ = 0$
- Thus, continuous piecewise polynomials contain fewer parameters than discontinuous smooths

Restricted Cubic Splines

- Piecewise cubic splines with *k* knots, often placed as percentiles of the predictor distribution (e.g., 20^{th}, 40^{th}, 60^{th}, and 80^{th} percentiles of *x*) or at pre-selected values
- Flexible association of *x* with *Y*
- Smooth at each knot point (avoiding unrealistic sharp bends)
- Constrained to be linear beyond the extreme knots (improving behavior in the tails)

Restricted Cubic Splines

- For k knots, there are $k - 1$ spline variables needed (so number of knots is reduced)
- An advantage is that the first variable is the "linear" term in x, so we can assess nonlinear effects by considering the other spline terms created
- A disadvantage is that the coefficients are not easily interpretable, the fit can only be assessed graphically

Restricted Cubic Splines

- Sometimes these are called natural cubic splines
- Spline fits can be sensitive to the number and placement of the knots
- But they can be a relatively easy way to assess “nonlinearity” of *x* effects

Basis Functions

- We’ve written out certain piecewise linear and piecewise cubic models, but note that these parameterizations often have major collinearity between the covariates
- Many software packages use some other set of covariates (“basis functions”) to model the spline that have less collinearity between the covariates
- The number of parameters and fitted values should be the same, however

Cubic Smoothing Splines

- To be more flexible, may want to choose a lot of knot points (even every *x* value!)
- As this can lead to too much "wiggles" need to penalize the fit for discontinuities in the 2^{nd} derivative
- Mathematically complex, but very useful in practice (related to GAMs)

Advantages/Disadvantages of Splines

-
-
-
-

Basis Functions

- We've written out certain piecewise linear and piecewise cubic models, but note that these parameterizations often have major collinearity between the covariates
- Many software packages use some other set of covariates ("basis functions") to model the spline that have less collinearity between the covariates
- The number of parameters and fitted values should be the same, however

Smoothers with Multiple Predictors

- Most scatterplot smoothers can be generalized to multiple predictors
- Special case is that of generalized additive models (GAMs)
- Mathematically complex, but very useful in practice

Multiple predictor approaches

- Linear regression:

  $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

- Nonparametric/Smooth regression:

  $E(Y) = f(x_1, x_2, x_3)$

- Generalized Additive Model (GAM):

  $E(Y) = f_1(x_1) + f_2(x_2) + f_3(x_3)$

  or

  $E(Y) = f_1(x_1) + f_2(x_2) + \beta_3 x_3$, etc.

Smoothing Issues in GAMs

- Choice of smoothing method
- Amount of smoothing (approximate degrees of freedom, weighting)
- Bias/variance trade-off
  - Too much smoothing leads to small bias but higher variance
  - Not enough smoothing leads to larger bias but possibly smaller variance

Smoothing Issues in GAMs

- GAMs are often called semiparametric methods
  - Parametric in the error distribution (e.g., normality of residuals for continuous outcomes, logit assumption for binary outcomes)
  - But nonparametric in the assumptions of the smoothed covariates (not assumed linear, quadratic, etc.)

Smoothing Issues in GAMs

- Generally, GAM procedures maximize a penalized log likelihood function based on cubic smoothing splines for each smoothed variable
- The smoothness is determined by an “equivalent degrees of freedom” (higher means more wiggles, lower means fewer wiggles)

Computing of GAMs

- Stata has a procedure that can be loaded in, but it is old code (written in Fortran), is sometimes a bit fussy to run or difficult to interpret, and does not run on Macs (so probably go with cubic splines instead)
- SAS and R have nice GAM code, easy to use, though sometimes can be a bit fussy to run or difficult to interpret

Advantages/Disadvantages of GAMs

GAMs/Splines: Final thoughts

- For virtually any data set having continuous covariates, use of smoothing, splines, and GAMs offers a good way to assess "linearity" of effect
  - Many studies would benefit by appropriate assessment of linearity assumptions

GAMs/Splines: Final thoughts

- GAMs/splines are getting more attention in the medical and epidemiologic literature
  - In some cases, smoothing or spline or GAM plots are presented in papers, often with confidence bands
  - In other cases, the spline models or GAMs help support an assumption of linearity or 'quadraticness' of effect, or piecewise linearity, or threshold effects, etc., expressed in a parametric way

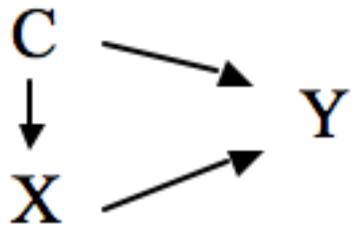
GAMs/Splines: Final thoughts

- Given software availability, use of smoothing, splines, or GAMs is relatively easy
  - We seldom plot a scatterplot without an overlaid Lowess smooth
  - We seldom finalize a regression analysis without trying cubic splines or GAMs for continuous covariates under consideration

A few questions since last class

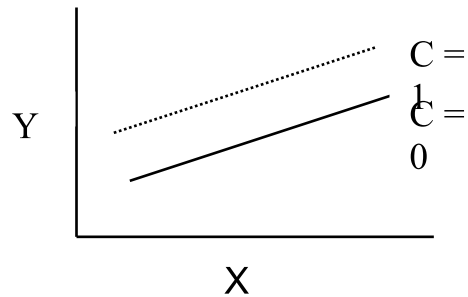
Confounding Review

- A variable is a **confounding variable** if it satisfies two conditions (classical definition of confounding):
  - It is a risk factor for the outcome
  - It is associated with exposure, but not a consequence of exposure
- Failure to control for confounding can lead to **bias**

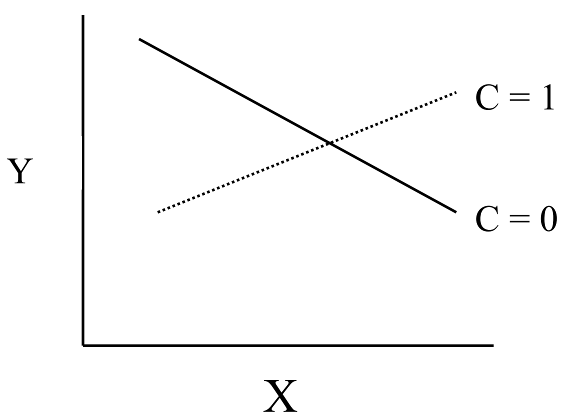


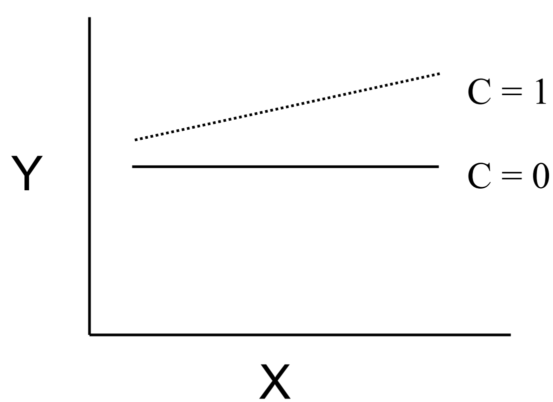
Effect Modification Review

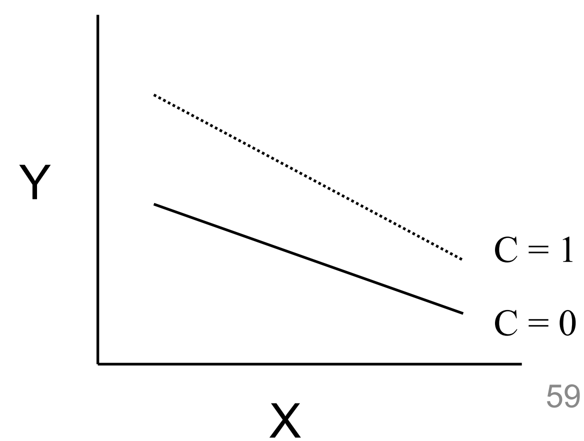
- Relationship between variable (X) and outcome (Y) differs by level of third variable (C)
- Example: No effect modification (parallel slopes)



- Example: effect modification (NOT parallel slopes)







‘Adjustment for gender’

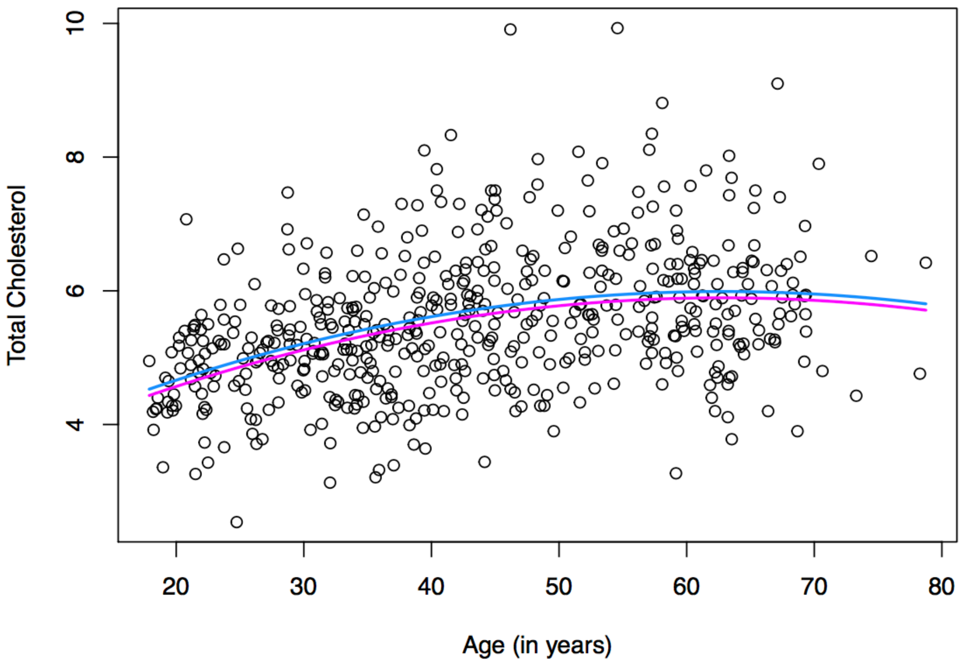


Figure 4: Fitted regression lines from Question 2 (b), with seperate lines for males (pink) and females (blue).

'Effect Modification by Gender'

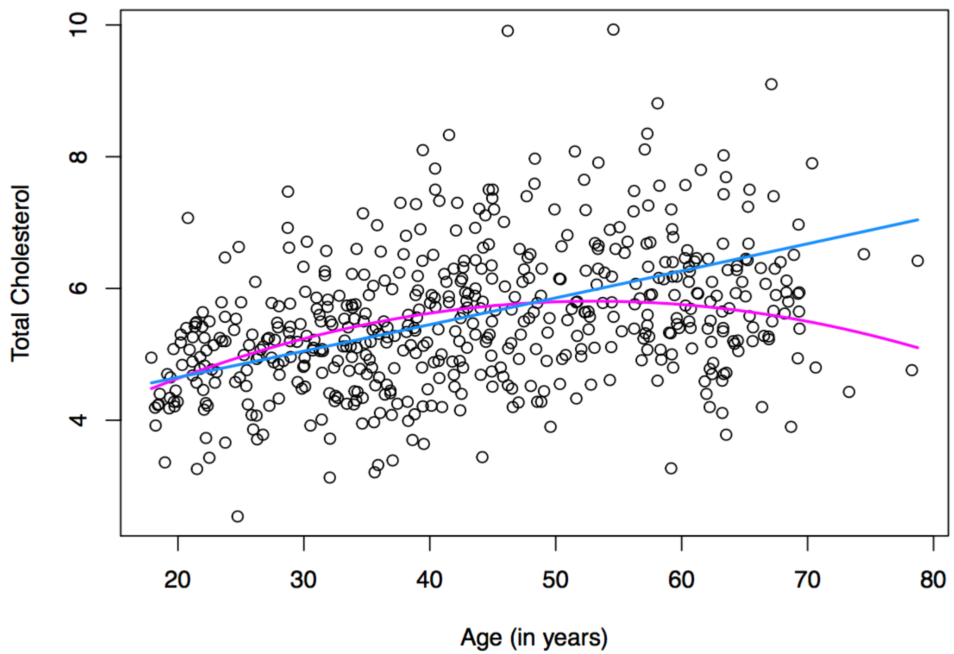


Figure 5: Fitted regression lines from Question 2 (d), with seperate lines for males (pink) and females (blue).