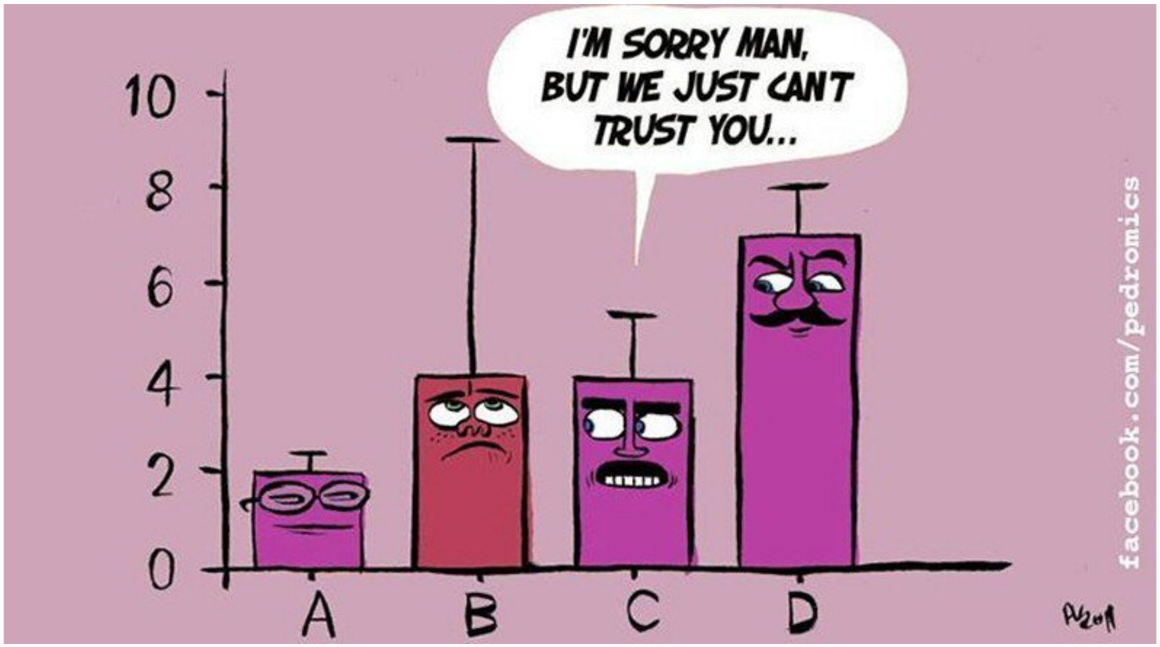
# BST 210
# Applied Regression Analysis




***Classroom treats today***
Bars are gluten free but contain peanut
Cookies are gluten/dairy/nut free

# Lecture 7

## Plan for Today

- Classroom treats and coffee!
- Notes on lecture structure and what’s important
- Framework for Analysis
- Relaxing the linearity assumption: delving into smoothing details
- Additive, more flexible models

# Continue to develop general framework for analysis

First -

- Learn the topic/study well, really well
- Collaborate to define motivating questions of interest, check PubMed, other sources
- What are your goals? (Inference: Prediction vs Interpretation)
- What techniques might help to achieve answers? Which do the data warrant? (develop intuition, read literature)
- Possible Confounding or Effect Modification to account for? Have the right model?
- Keep an open mind, and the larger picture – there is no recipe

# Continue to develop general framework for analysis

Next -

- Consider model $E[Y|X1,\ldots,Xp] = \beta 0 + \beta 1 \cdot X1 + \ldots + \beta p \cdot Xp$
- Diagnostics/Checking Assumptions: LINE
  - Scatterplot, summary statistics
  - Boxplots, histograms
  - Correlations
  - Smoothing (Lowess?)
  - Residual Analysis: assumptions met?
  - Influence Analysis: Outliers? Leverage? Influence?
- Hypothesis testing/modeling:
  - t-test?
  - Correlation (r)?
  - ANOVA useful?
  - Nonparametric approach better?
  - Linear regression or extensions (multiple reg.)?
  - More flexible modeling needed (splines, additive models)?
  - Generalizations

# Continue to develop framework for analysis

Then -

- Assess Model Fit:
  - Residual Analysis: assumptions met?
  - Influence Analysis: Outliers? Leverage? Influence?
  - Confidence Intervals
  - $R^2$, Adjusted $R^2$
  - MSE, RMSE, AIC
  - Interpretation and inference, constant collaboration around data and meaning
  - Back up and regroup as needed (flexible modeling, etc?), delve deeper, use caution, stay organized

# Flexible Modeling: the motivating challenge

- Not all continuous covariates will have a linear relationship with the outcome variable
- Transformations could be used on the outcome $Y$ or a covariate $x$
- Categorization, piecewise polynomials, additive models, or other possible changes to $x$ are possible to assess and model possible nonlinear effects of $x$ on $Y$

# Flexible Modeling: the motivating challenge

- Inclusion of a continuous covariate as linear in a final model probably should include evaluation of the appropriateness of the linear assumption
- Of course, inclusions of other forms of covariates (polynomial additive, etc) or transformations could make interpretation of β coefficients more difficult, but might make predictions better
- With above said, always go with simpler model if appropriate, but more flexible forms of modeling should always be considered, when viable

# More Flexible Modeling

- All flexible models we will discuss are simple cases of generalized additive models. Here we restate the additive model and will then delve into the special (simpler) cases first before moving our way toward the (more general) additive model itself.

  - Recall the Linear Regression Model

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

  - We can now extend this to the far more flexible Additive Model

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon$$

  - Each $f_j$ can be any of the different methods we just talked about: Linear term $(\beta_j X_j)$, Polynomial, Step Function, Piecewise Polynomial, Degree-$k$ spline, Natural cubic spline, Smoothing spline, Local linear regression fit, …
  - You can mix-and-match different kinds of terms

# Flexible Modeling Options: Smoothing

- Is less variable than the original data, hence the name smoother
- Helpful when *scatterplot smoothing* with a single predictor variable x, modeling *Y* as a function of *x*
- Possible starting point for assessing a trend—may otherwise miss it
- Simple linear or quadratic regression is a special case, having a rigid form for the dependence
- Many more nonparametric (no rigid form for dependence) methods available
- These are data dependent approaches, unlike model dependent approach such as linear regression

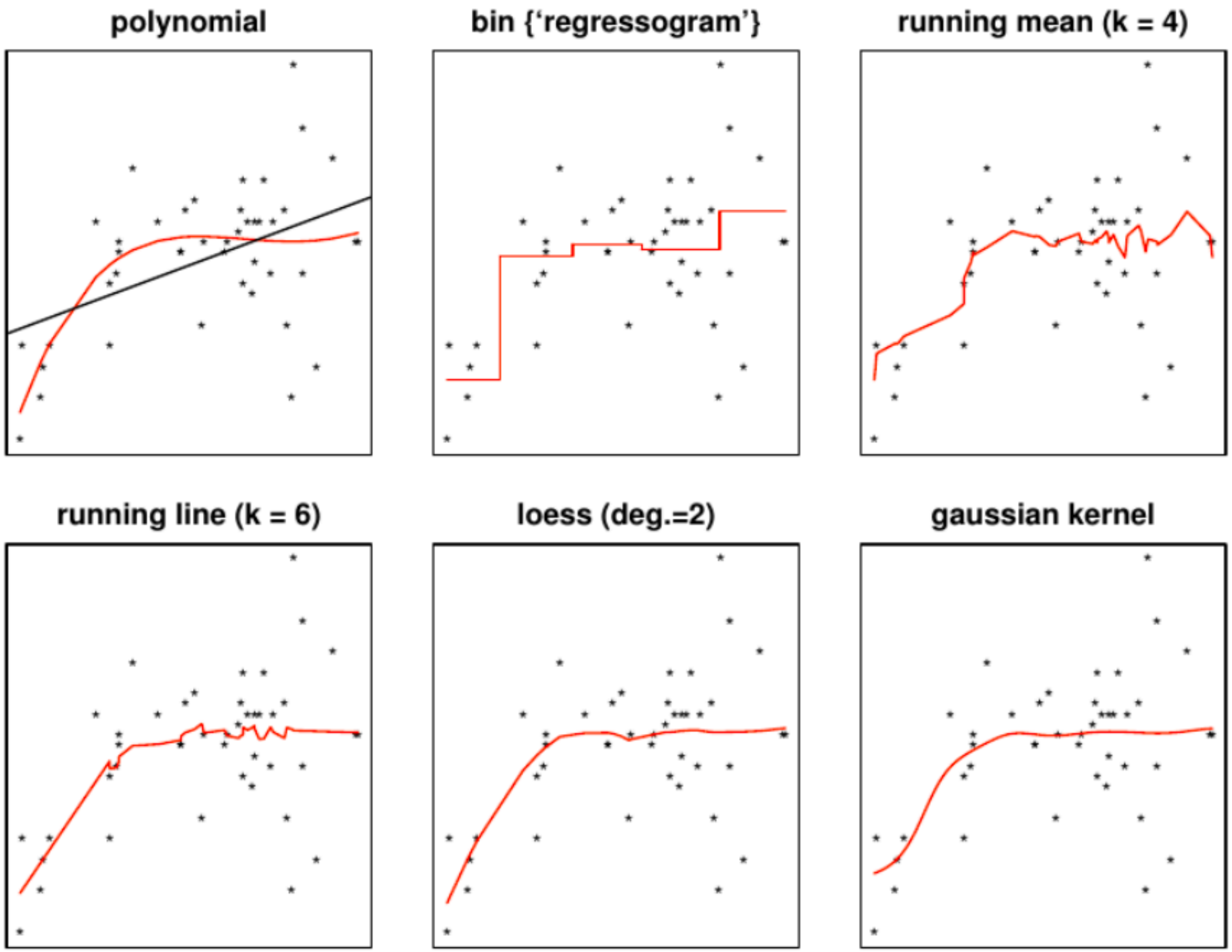
# Scatterplot Smoothing

- Bin smoother:  Create categories for *x* variable and average *Y* over each category ("piecewise constant")
- Running mean smoother:  Mean of *Y* over a (moving) neighborhood of *x*
- Piecewise linear: Over categories of the *x* variable
- Running line smoother:  Linear fit of *Y* on *x* over a (moving) neighborhood of *x*
- Kernel smoother:  Locally weighted running mean smoother of *Y* over a (moving) neighborhood of *x* (the kernel has higher weights the closer you are to the middle of the neighborhood). Kernel means 'the central or most important part of something.'

# And we've already been using…

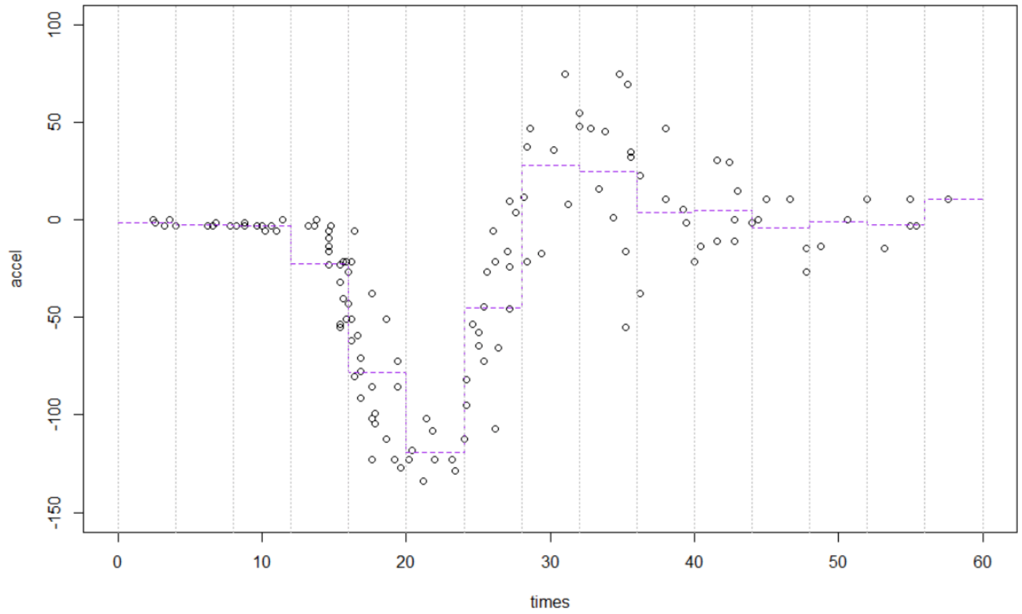
- Lowess: Locally weighted running line smoother of *Y* over a (moving) neighborhood of *x* (the kernel has higher weights the closer you are to the middle of the neighborhood)
  - A sophisticated but very useful approach to smoothing-combines best of previous concepts
  - Accommodates nonlinearity, uses overlpapping windows of data, higher weights for middle than ends, shifting windows over range
  - Available in Stata, SAS, and R, though exact implementation may vary slightly

# Scatterplot Smoothing



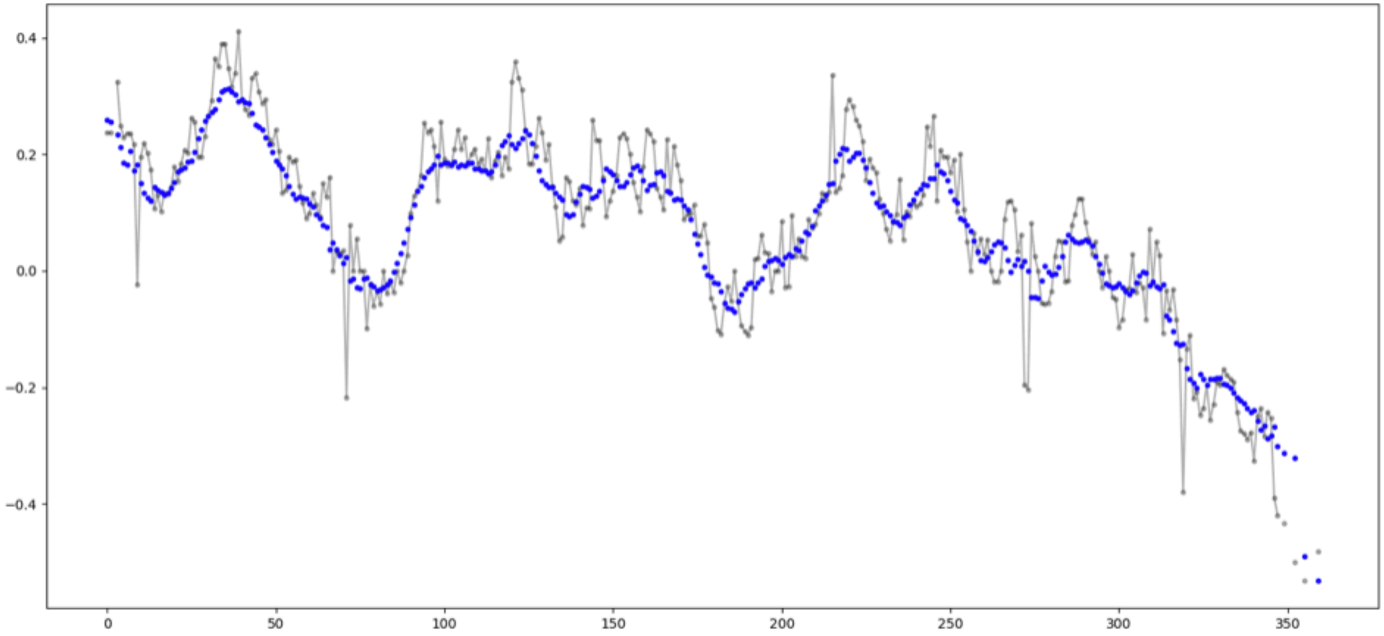
# Scatterplot Smoothing

- Bin smoother: Create categories for *x* variable and average *Y* over each category ("piecewise constant")



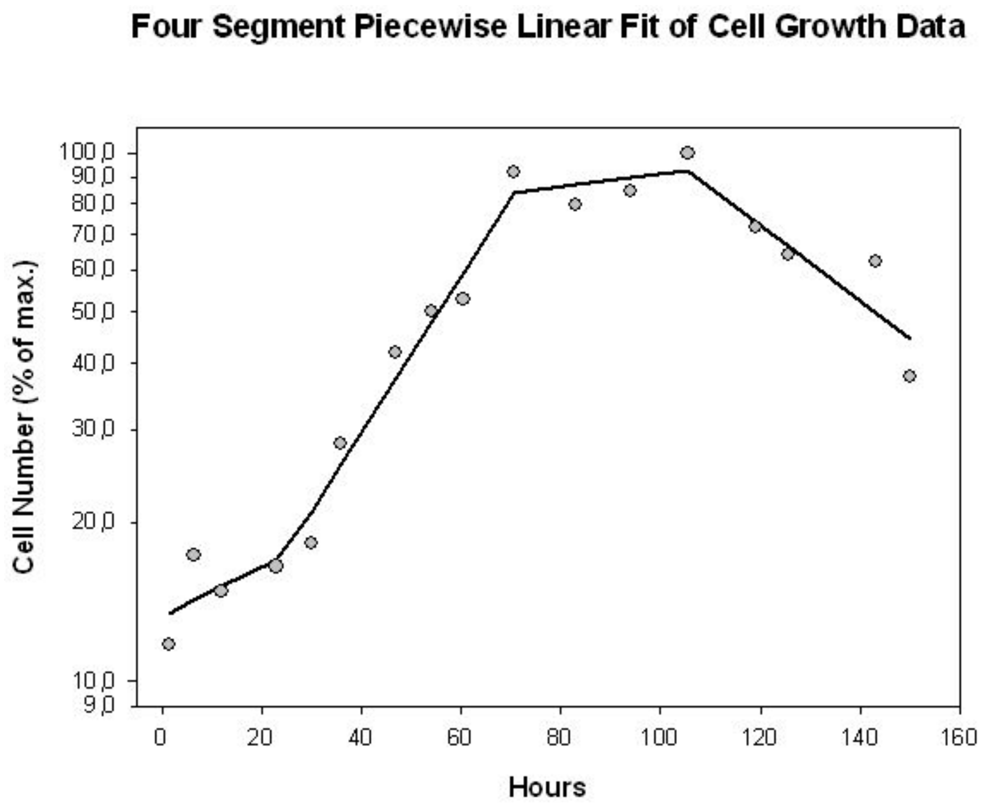
# Scatterplot Smoothing

- <u>Running mean smoother</u>: Mean of *Y* over a (moving) neighborhood of *x*



# Scatterplot Smoothing

- Piecewise linear: Over categories of the *x* variable

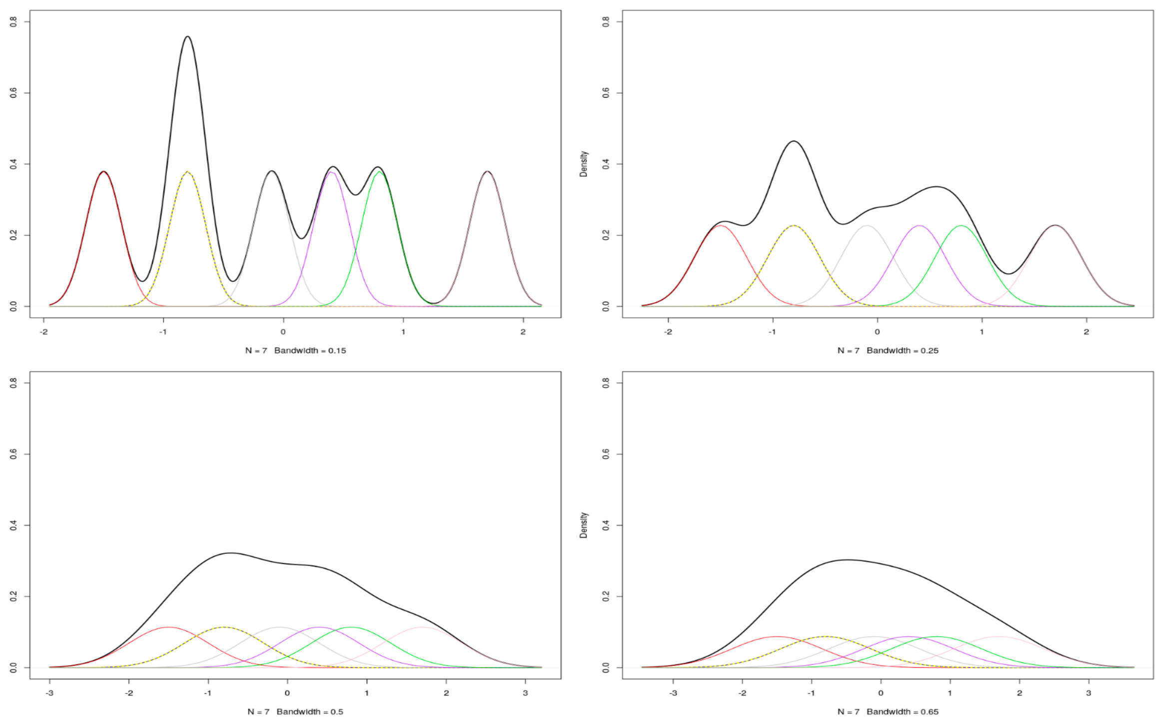


# Scatterplot Smoothing

- Running line smoother: Linear fit of $Y$ on $x$ over a (moving) neighborhood of $x$

# Scatterplot Smoothing

- <u>Kernel smoother</u>: Locally weighted running mean smoother of *Y* over a (moving) neighborhood of *x* (the kernel has higher weights the closer you are to the middle of the neighborhood) – below increasing bandwidth



# We’ve already been using…

- https:rafalab.github.io/dsbook/smoothing.html#bin-smoothing//

# Framework for Analysis: caveats

– Recall the '*no recipe*' promise!

– Let's consider an example where paying close attention, to the data and caveats around the methods, proves valuable…

# Example: Hyponatremia

- **Given 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

- 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

The NEW ENGLAND JOURNAL of MEDICINE

The Washington Post

Los Angeles Times

U.S.News & WORLD REPORT

The New York Times

Chris Almond

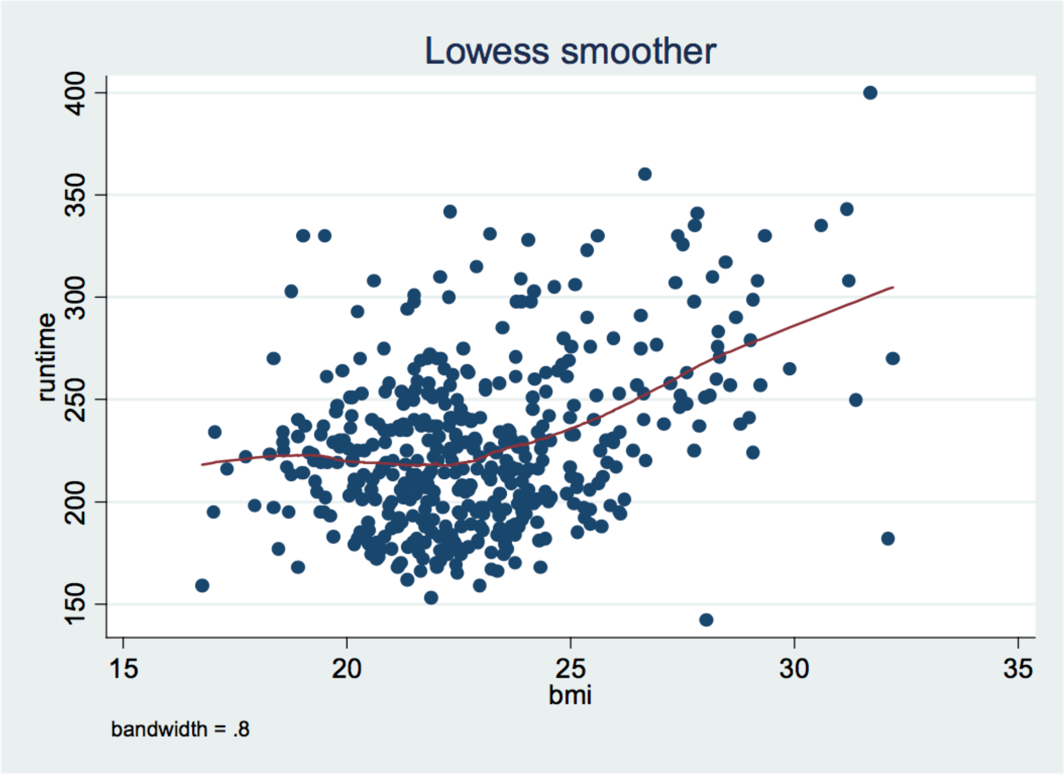
# Example: Hyponatremia

- 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
- Note: there are various ways to get into Boston Marathon

```
. 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?

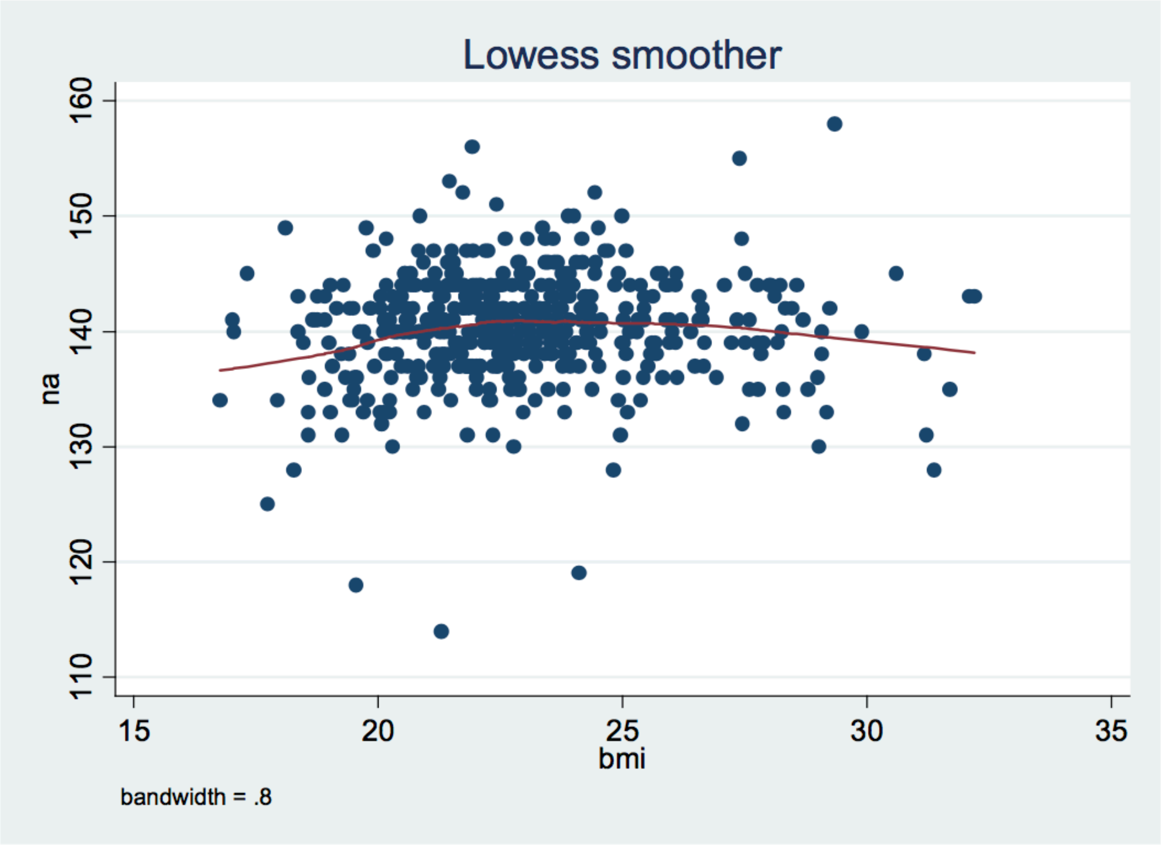
```
. regress na runtime

      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 linear regression model say?
- Ideas for next step?

```
. 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

- $x^2$ relationship?
- What does linear regression model say?

```
. 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.
- Why did we pick 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 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 $x^\lambda$ [or $Y^\lambda$] 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 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 $\lambda$, where in the limit as $\lambda \to 0$ we get $\log(x)$ [or $\log(Y)$]

# Use of Piecewise Polynomials

- (Splines are just ‘piecewise polynomials’. Need to pick the knot points.)
- Regression splines: Piecewise constant (categories), piecewise linear, piecewise cubic
- Regression splines with constraints: To force continuity, smoothness in the derivatives, linearity at the ends, or similar constraints
- Moving towards generalized additive models (don’t need to pick the knots)

# Advantages/Disadvantages of Polynomials or Transforms

-
-
-
-

# Natural Log Transforms

- Using a natural log transform (on an $x$ 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 Transforms

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

  $\beta_1$ = slope, or the change in $Y$ corresponding to a 1 unit increase in $x$
- Here, $\beta_1$ 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 Transforms

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

  $\beta_1$ = 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 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 $\beta_1$, one can calculate a 95% CI for $\beta_1 \cdot \log(1.01)$

# Natural Log 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 $\beta_1$, one can calculate a 95% CI for $\beta_1 \cdot \log(1.10)$

# Natural Log Transforms

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

  $\beta_1$ = 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 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

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

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

# Natural Log 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

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

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

# Natural Log Transforms

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

  $\beta_1$ = 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 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

- Need to choose order of spline, knot points, constraints
  - Order of spline: Constant, linear, cubic
  - Knot points: Change points
  - Constraints: On continuity or derivatives at knots or ends

# 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 $u \leq 0$

# 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

# (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 ones

# 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

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

- 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

# 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
- 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

# 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 binary 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.