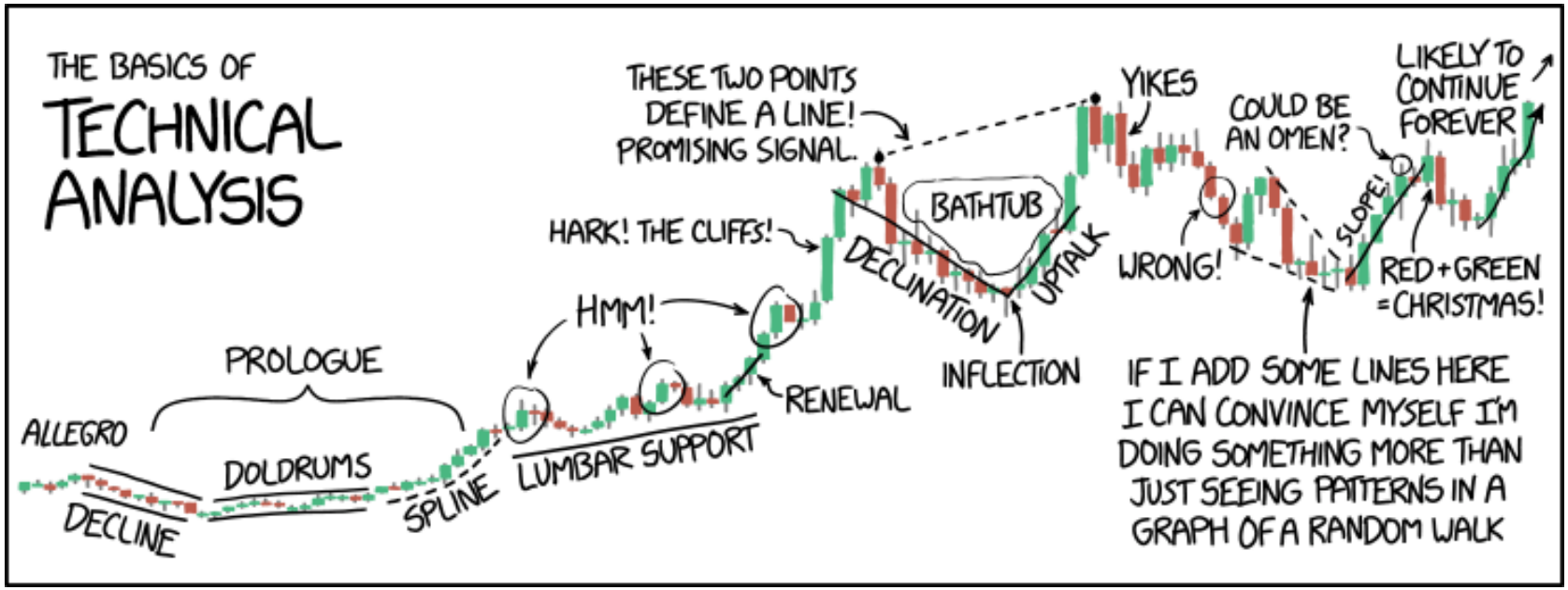
BST 210
Applied Regression Analysis



…Lowess? Cubic Spline?

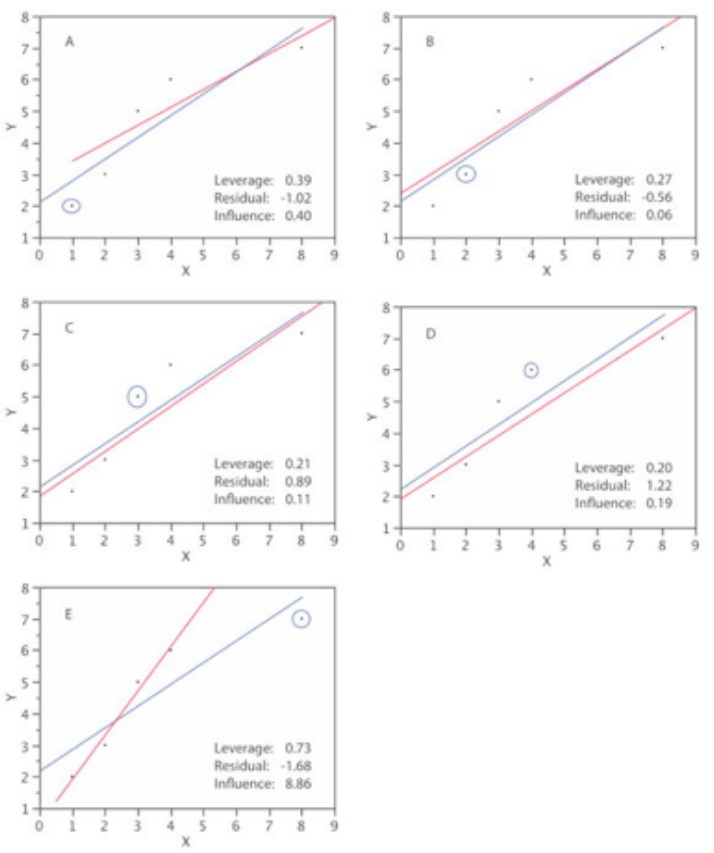
Lecture 6

Plan for Today

- Recap and Clarifications
- Framework
- Sentiments on Statistical Practice
- Game!
- Additive and more flexible models

Influence Analysis

- An observation is influential if its deletion substantially changes the regression results
  - High leverage or 'outlierness' => potential influence
- In simple linear regression, a scatterplot would usually identify influential observations, but this is more difficult in multiple linear regression
- It could be a particular combination of the covariates and outcomes that leads to an influential observation



Influence Analysis: Cook's Distance

- Cook's distances are a combination of each observation's leverage and residual values; the higher these measurements, the higher Cook's distance. Measures how much the regression would change if observation i is deleted. Unlike leverage, it reflects the actual amount of influence an observation has on a model fit.
- Calculated as

$$D_i = \left(\frac{{r_i}^2}{2}\right)\left(\frac{h_i}{1-h_i}\right)$$

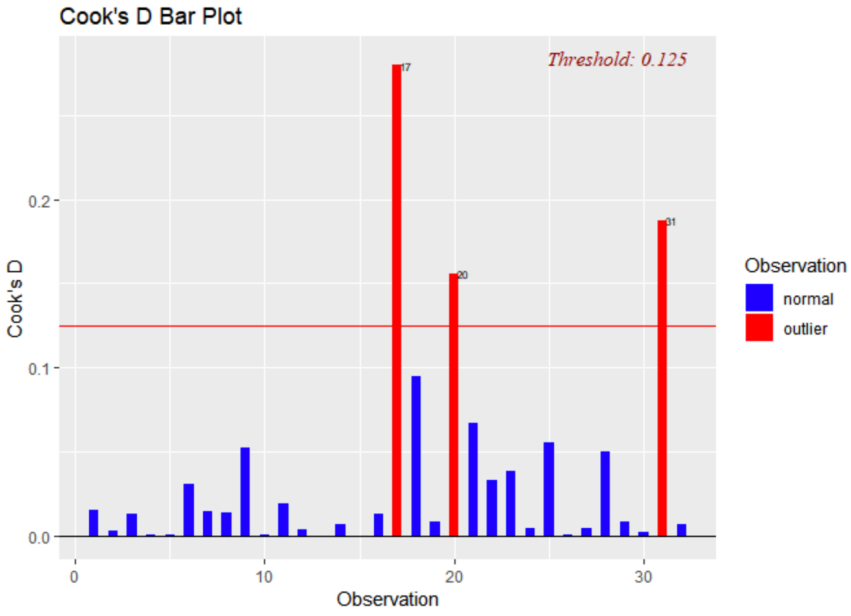
- Here r_i is the standardized residual and h_i is the hat value for the i^{th} subject. Thus, Cook's distance increases for large values of residuals, leverages, or both.
- Some authors recommend $4/(n-2)$ as a rough rule of thumb for observations with high Cook's distance; others look for gaps in the Cook's distances to find high influence points.

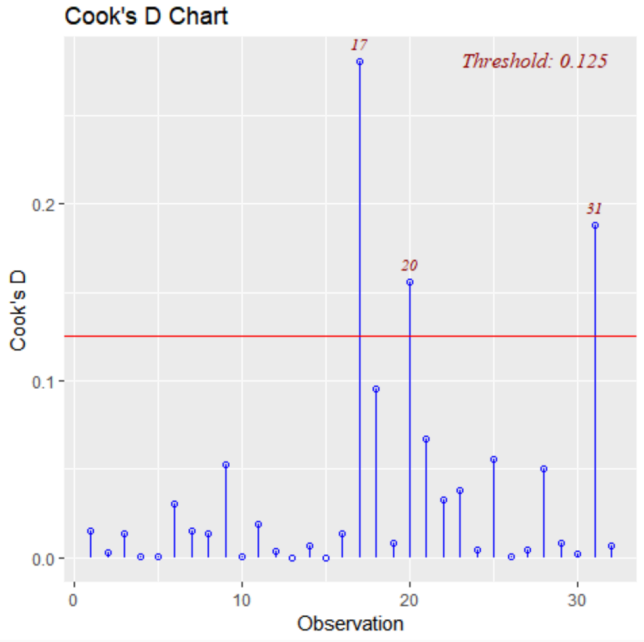
Influence Analysis: Cook's Distance

Steps to compute Cook's distance:

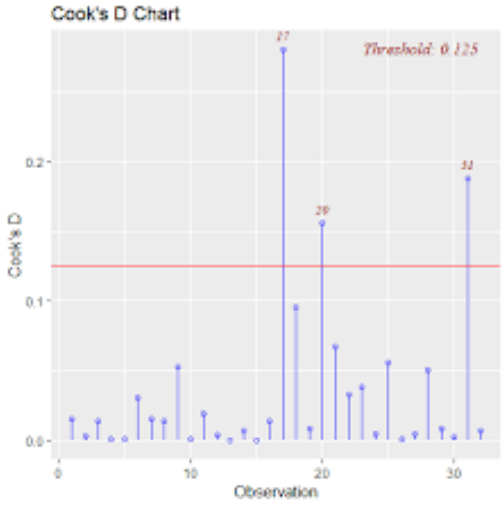
- delete observations one at a time.
- refit the regression model on remaining $(n-1)$ observations
- examine how much all of the fitted values change when the ith observation is deleted.

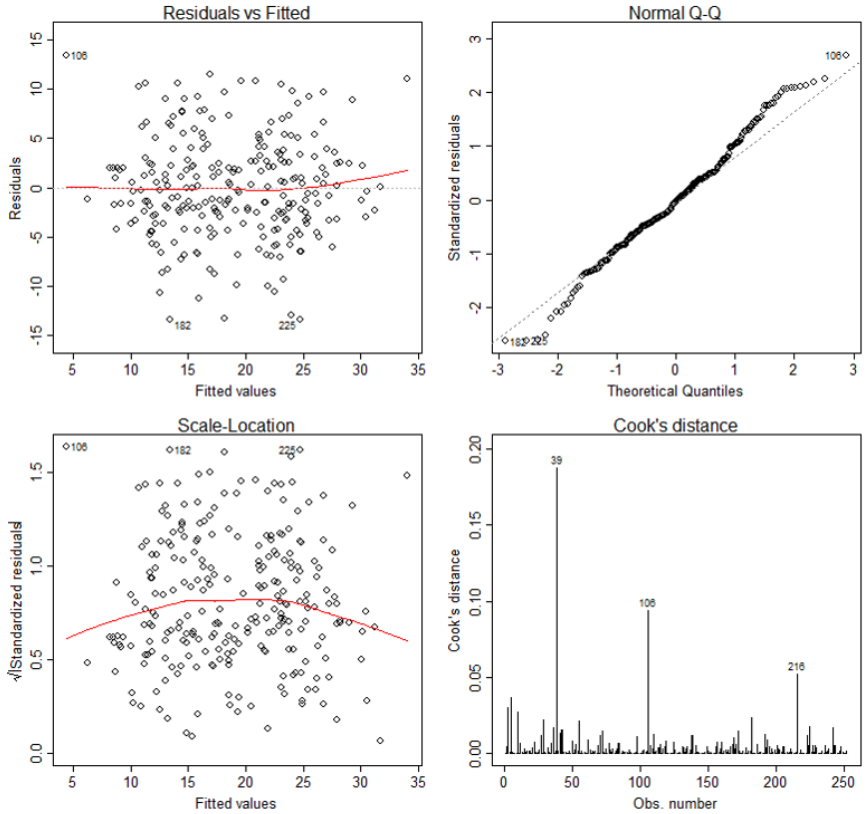
A data point having a large cook's d indicates that the data point strongly influences the fitted values.





Influence Analysis: Cook's Distance





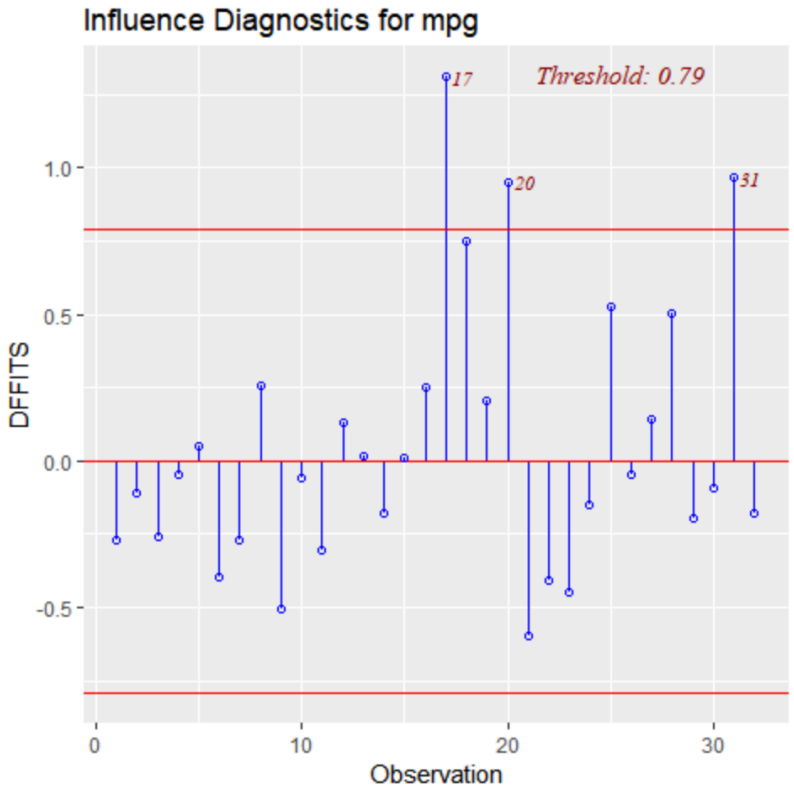
Influence Analysis: *DFFITS*

- A related statistic is the DFFITS (Difference In Fits) measure, assessing the i^{th} case's influence on $\hat{Y}_i$ and is given by

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)i}}{s_{e(i)}/\sqrt{h_{ii}}}$$

- Conceptually similar to Cook's Distance
- Scaled difference between ith fitted value obtained from the full data and the ith fitted value obtained by deleting the ith observation - quantifies the number of standard deviations that $\hat{Y}_i$ changes when the ith data point is omitted.
- Because of the scaling of $DFFITS_i$, a rule of thumb is that case *i* is relatively influential if $|DFFITS_i| > 2\sqrt{(p+1)/n}$
- One could compare models run including and then excluding points with high $DFFITS_i$ values

Influence Analysis: *DFFITS*



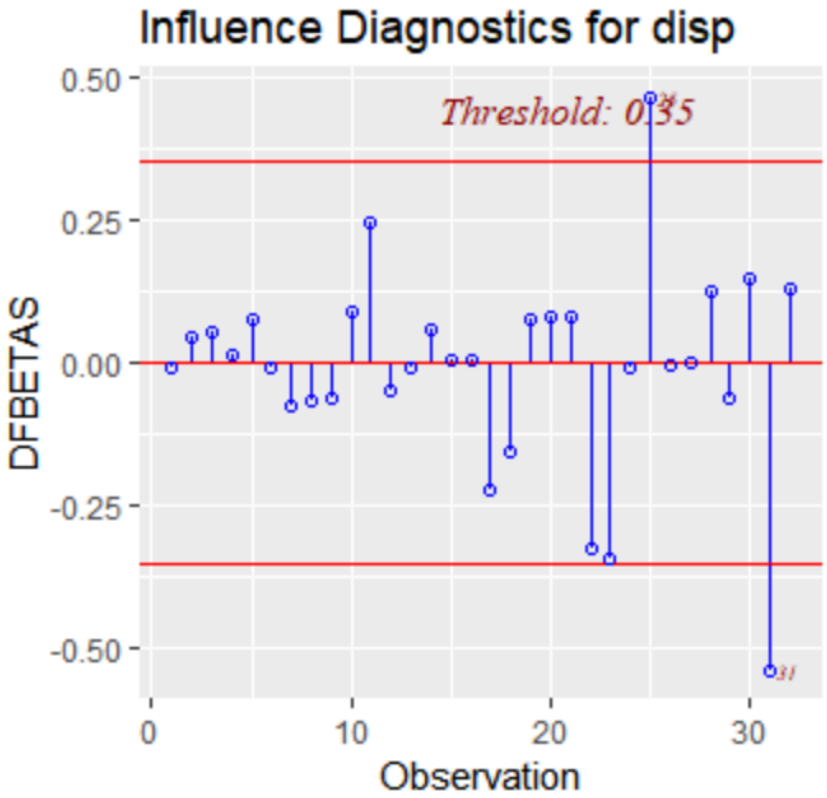
Influence Analysis: *DFBETA*

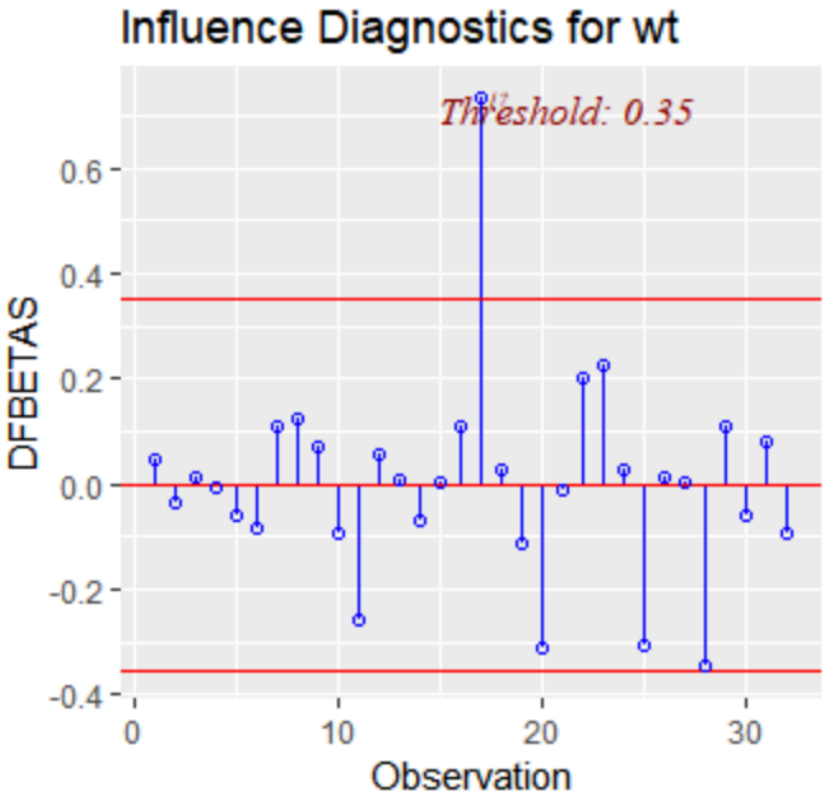
- The *DFBETA* (Difference In Beta coefficients) statistic measures the difference in each parameter estimate with and without the influential point.
- $DFBETA_{ik}$ assesses how much the k^{th} regression coefficient, β_k, changes (in standard error units) if the i^{th} observation is deleted
- The formula for $DFBETA_{ik}$ is not that important, but making boxplots or histograms of the statistics could be helpful in determining influential observations

Influence Analysis: *DFBETA*

- Because of the scaling of $DFBETA_{ik}$, a rule of thumb is that $|DFBETA_{ik}| > 2/\text{sqrt}(n)$ should detect roughly the top 5% of influential cases
- One could compare models run including and then excluding points with high $DFBETA_{ik}$ values
- There is a DFBETA for each point (for n observations and k variables there are n*k DFBETAs). Larger DFBETAs indicate observations that are influential in estimating a given parameter.

Influence Analysis: *DFBETA*

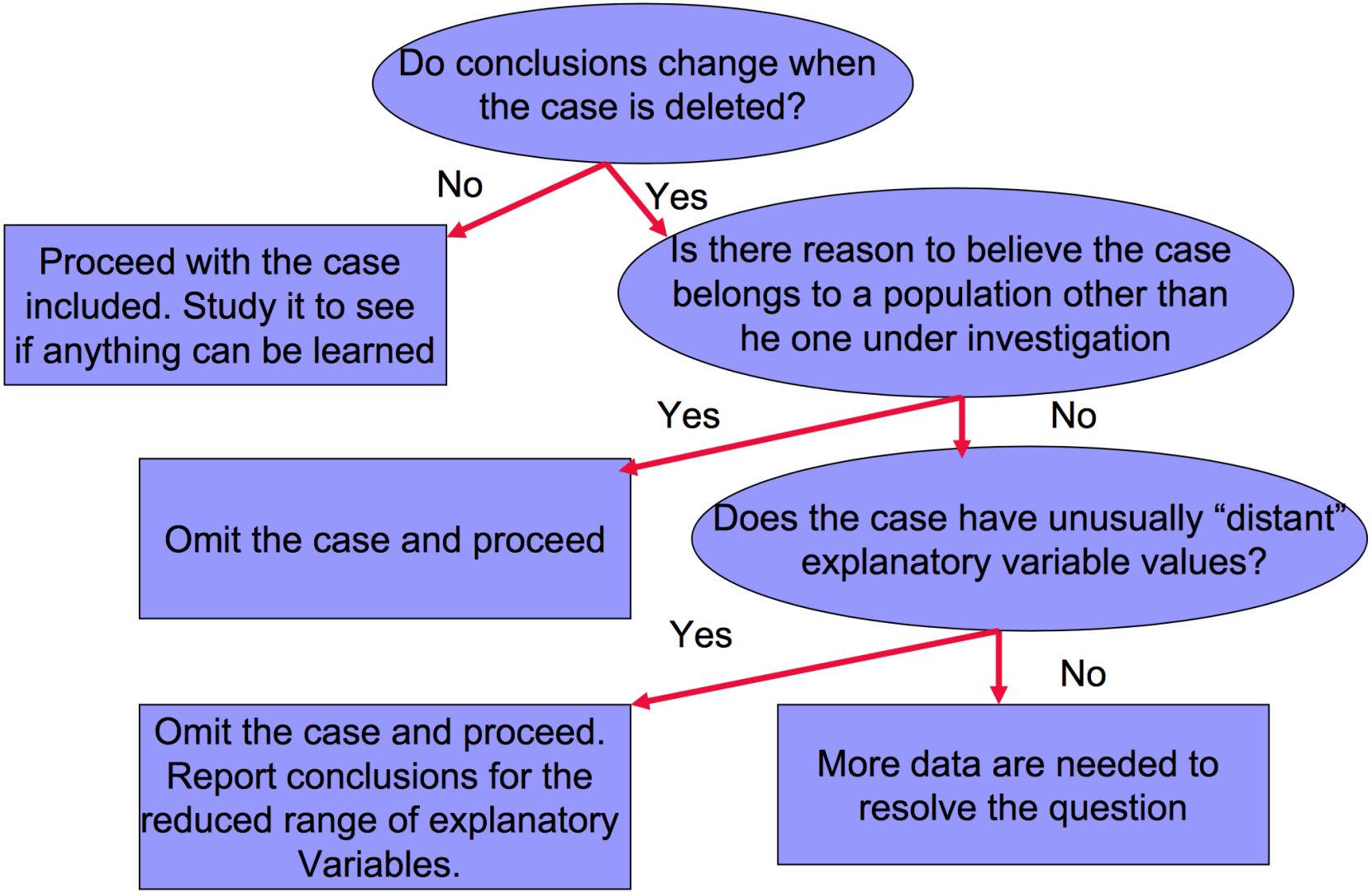




Outliers and influence: what to do?

- If outliers occur, first go back and check the original data to be sure there were no errors in data entry.
- You may want to fit the model with and without one or a small number of outliers, to see how much the model inferences change due to these outliers.
- However, be very careful of reporting results with outliers deleted.
  - This could make your model look “too good” relative to the original data.
  - Possibly you would report your inferences both with and without the outliers.
  - If things don’t change substantially, go with all of the data.

One Possible Strategy for dealing with influential cases
(there are many, and which one depends on the context and practitioner)



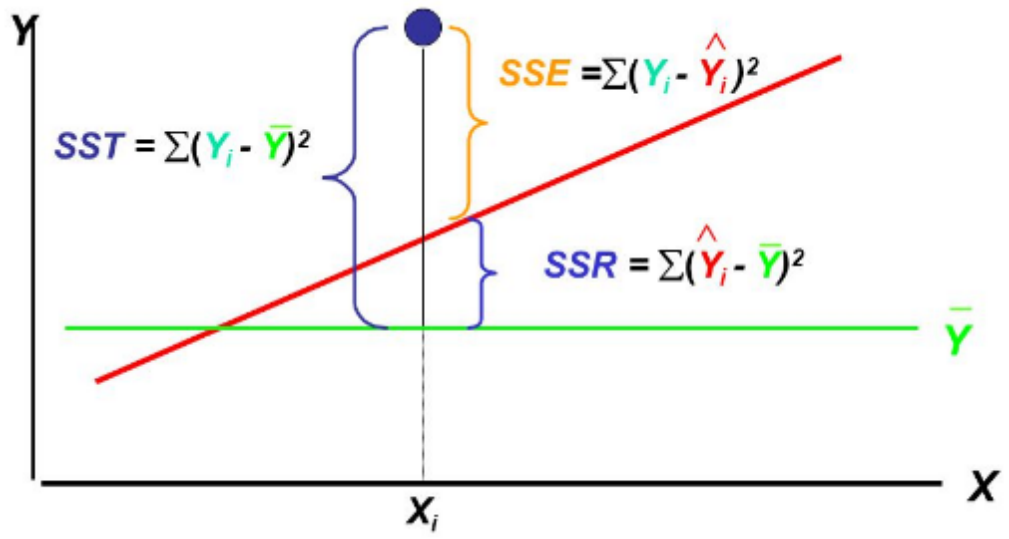
Assessing Model Fit

How Do We Decide Which Model is Best/Optimal?

If we want to go about choosing the "best" model among a collection of possible models that meet our objective, we first need to have an idea of what we mean by "best". There are many possible definitions!

Let's back up and look through a 'Sums of Squares' perspective. Suppose we have a dataset with N observations, and that we fit a regression model with p predictors/covariates included. Then:

Assessing Model Fit



SST = SSE + SSR

MSE = SSE/n

R^2 = (SST – SSE)/SST

= 1 – SSE/SST

= 1 – n*MSE/SST

- Note: R^2 is always between 0 and 1
- Note: If MSE = 0 → R^2 = 1
- Note: RMSE = sqrt (MSE)

* Why use squares?
- Differentiability needed to minimize SSR
- Keeps positive values
- Magnitude: larger differences look large

Assessing Model Fit

- R^2:

- The R^2 measures the proportion of the total variability in the outcome (Y) that our model is able to successfully explain. However, as we've discussed in both class and in the labs, the R2 isn't always a useful metric when comparing models, as it will always increase when we add additional covariates to our model—regardless of whether those covariates are actually helpful.

- Interpret and use with caution—some do not prefer use of R^2

$$R^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

- "Better" Fitting Models Have larger R^2 values

Assessing Model Fit

`Ordinary` — Ordinary (unadjusted) R-squared

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

`Adjusted` — R-squared adjusted for the number of coefficients

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST}.$$

Assessing Model Fit

- **Adjusted R2:**
- Addresses this issue by adding in a slight penalty that scales with the number of parameters in the model:
- "Better" Fitting Models Have: larger adjusted R2 values

- **Root MSE:**
- Recall that the Mean Square Error (MSE) estimates the variance of the observed outcome (Y) about our fitted regression—in other words, it estimates the amount of variability in our outcome that our model is unable to explain. So the Root MSE (RMSE) simply estimates the standard deviation instead!
- "Better" Fitting Models Have: smaller MSE/root MSE values

- **AIC:**
- **'Akaike Information'**
- Is another model selection criterion that navigates the trade-off between model complexity (which is captured by the number of parameters in the model) and model fit (which is captured by the likelihood of the model).
- Formula: $2*p - 2\log(\hat{L})$, where $\hat{L}$ is the maximum value of the likelihood function for the model (we will look at likelihood functions in upcoming lectures)
- Given a collection of models for the data, AIC estimates the quality of each model relative to each of the other models.
- "Better" Fitting Models Have: smaller AIC values

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 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 $E[Y|X1,...,Xp] = \beta 0 + \beta 1 \cdot X1 + ... + \beta p \cdot Xp$
- Diagnostics/Checking Assumptions: LINE
  - Scatterplot, summary statistics
  - Boxplots, histograms
  - Correlations
  - Smoothing
  - 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.)?
  - 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, delve deeper, use caution, stay organized

Sentiments on the ‘practice of statistics’

- Be willing to get hands dirty while ‘getting to know the data’
- Be ok with getting frustrated at times ☺
- Be open minded to learning new data import methods, cleaning methods, curation – real world data is messy
- Be not shy to ask questions and seek advice/help. Find mentors and use them (peers can be mentors too).
- Be not afraid of learning new methods or caveats to methods you know. Lean on your training and *potential knowledge*.

Advice toward your training

- Perhaps keep an ‘annotated version’ of assignments and projects which includes a running history of your personal code embedded in output and explanations (hand in only what is necessary for the homework, but keep a longer version for reference later when you cannot recall how to do something).
- This course has a bit of a ’spiraling’ approach, due to the nature of the field of statistics and learning its methods.
- Develop the ability not necessarily to have the ‘right’ answer instantly on hand, but rather a repertoire of exposures, reading, training that gives you the tools and potential to find what gets you close to viable answers.

Time for a quick game…
The Stats Are Right!
Rules:

- Room gets divided down the middle
- Class views a question at a time on the slide.
- The side of the room that answers first gets a point.
- No student can answer a question for their ‘side of the room’ more than once. It must be a different person each time.
- 6 questions total: all to do with the Circulatory Arrest peri-operative study
- 1 minute per question
- Good luck!

The Stats are Right!

- **Question 1: True of False**

- In order to evaluate effect modification of birthweight by diagnosis in modeling the outcome PDI, you need to include the variable birthweight*diagnosis, but you don't need to include the variables birthweight or diagnosis.

The Stats are Right!

- **Question 2: True of False**

- Cook's Distance incorporates mainly 'leverage' in its calculation of overall influence.

The Stats are Right!

- **Question 3: True of False**

- In order to determine the potential presence of alternatives to linearity in the association between circulatory arrest (in minutes) and PDI score, we could include a polynomial version of circulatory arrest (in minutes) in our model that already includes circulatory arrest (minutes), and if it's significant we could then infer that there is evidence of nonlinearity in the relationship with PDI.

The Stats are Right!

- **Question 4: Short Answer**

- What's another way to evaluate the question of linearity of an independent variable (predictor) in your model that is a different method than the one employed in the previous question? (ie don't use any model coefficients)

The Stats are Right!

- **Question 5: True or False**

- We would assess whether the variable ‘diagnosis’ confounds the relationship between circulatory arrest (in minutes) and PDI score by including the following variable in our model: diagnosis*circulatory arrest (minutes), then checking for a >10% change in the coefficient of circulatory arrest (in minutes).

The Stats are Right!

- **Question 6: Choose one**

- Let's say we 'binned' birthweight into 4 categories. If we are interested in determining whether or not there is a birthweight trend with outcome PDI, would we want to treat our new categorical birthweight variable as nominal or ordinal?

- Bonus point: List a downside to binning!

The Stats are Right!

- **Question 7: True or False**
- We will use ANOVA to consider an association between PDI and diagnosis.
- Bonus point: What does ANOVA stand for?

The Stats are Right!

- **Question 8: Shortish Answer**

- Why would we choose a linear model instead of ANOVA in evaluating a possible association between diagnosis and PDI score? (please list 2 reasons)

The Stats are Right!

- **Thanks for playing!**
- **Now back to flexible modeling --**

Next - Relaxing Assumptions of Linearity

- Study of 'Influence' leads us to flexible modeling so that we can avoid outlier influence
- How can I model a continuous predictor in a flexible fashion when I know I cannot force a linear association without huge influence?
- Flexible modeling!

- Scatterplot smoothing methods
- Indicator variables, polynomials, and splines
- Moving towards additive models (and generalized additive models)

→ GAMs: 'smooth' version of the GLM

More Flexible Modeling

- Linear regression imposes two key restrictions on the model: We assume the relationship between the response Y and the predictors $X_1, \ldots, X_p$ is:
  1. Linear
  2. Additive
- The truth is almost never linear; but often the linearity and additivity assumptions are *good enough*
- When we think linearity might not hold, we can try…
  - Polynomials
  - Step functions
  - Splines
  - Local regression
  - Generalized additive models
- When we think the additivity assumption doesn't hold, we can incorporate interaction terms
- These variants offer increased flexibility, while retaining much of the ease and interpretability of ordinary linear regression

More Flexible Modeling

- The goal of prediction is to estimate the true, unknown regression function, f
  - 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

- Splines are a nice way of modeling *smooth* regression functions
- To increase the *flexibility* of a spline, we increase the number of knots
- Natural cubic splines allow us *retain the model complexity of a cubic spline* while adding *two extra interior knots* at the cost of restricting our model to be *linear* outside the range of the observed data
- Smoothing splines enable us to avoid the problem of knot selection altogether, and instead specify a single parameter: the desired effective degrees of freedom for the fit
- We can put everything together into an Additive Model

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

where each f_j can be any of the fits we talked about.

Flexible Modeling

Review of Linear Models

Classical Linear Model

- Response: $Y \sim N(X\beta, \sigma^2)$
- $X\beta$ is a linear function that describes how the expected values vary based on characteristics in the data
- Linear: $\beta_0 + \beta_1 X_1^2 + sin(\beta_2 X_2)$
- Non-linear: $\beta_1 X_1 e^{\beta_2 X_2}$
- Constant Variance

Generalized Linear Model

- Response: Poisson, Gamma, Binomial, etc.
- $Y \sim F(\pi, R)$
- Expected Value:

$G(E[Y])^{-1} = X\beta$

- Variance is a function of expected value
- Responses are independent

Flexible Modeling

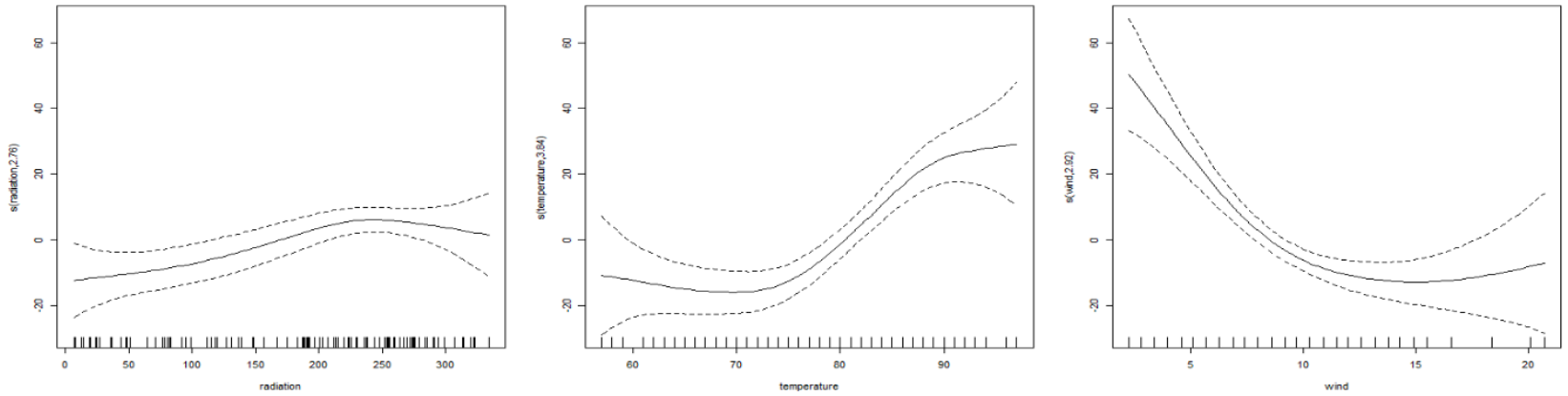
What's an Additive (or even Generalized Additive) Model?

- Linear predictor has a more general form
- $E(Y|X_1, X_2, \cdots X_p) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$

- $f_i(X_i)$ are non-parametric smoother functions
  - Smoothing Splines
  - Kernel Smoothers
  - Local Linear Regression
  - But can also be parametric functions, too

Flexible Modeling

What does this mean in real life?



- Can more easily fit models with less assumptions
- No ‘nice’ polynomial shapes are necessary
- No need for variance assumptions

Flexible Modeling

When is it appropriate to fit a GAM?

You can fit a GAM with any data where you might try fitting LMs, GLMs, and GLMMs

- GAMs are more general and with less assumptions

Common Examples

- LDF fitting
- Large data sets with complicated interaction effects
- Models with many parameters but not a lot of data per parameter
- Fitting a smoothed trend line that allows the trend to vary by year

Flexible Modeling

What are the tradeoffs?

Advantages

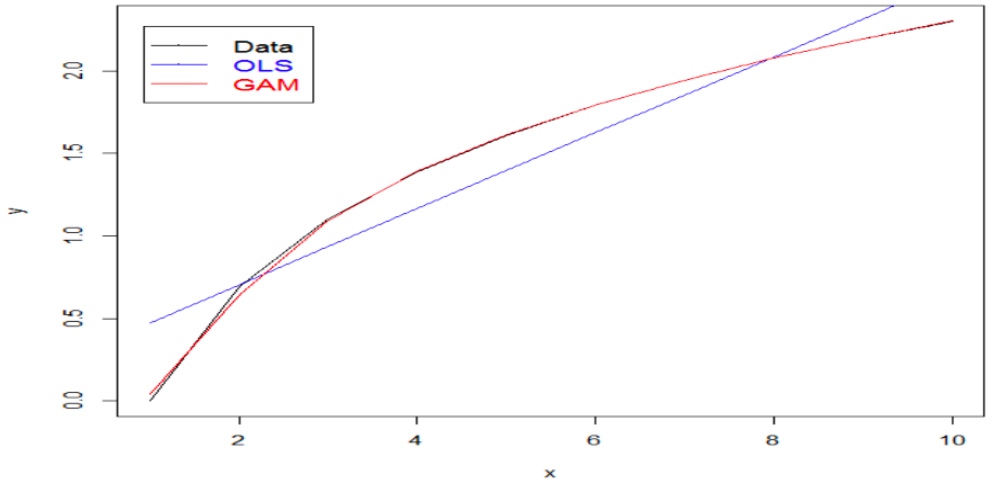
1. Useful for non-parametric and semi-parametric data
2. Useful when data doesn't fit LM/GLM assumptions
3. Can paste splines directly into Excel

Disadvantages

1. Output may be more difficult to interpret to regulators and business side
2. Must be wary of over-fitting

GAM Examples

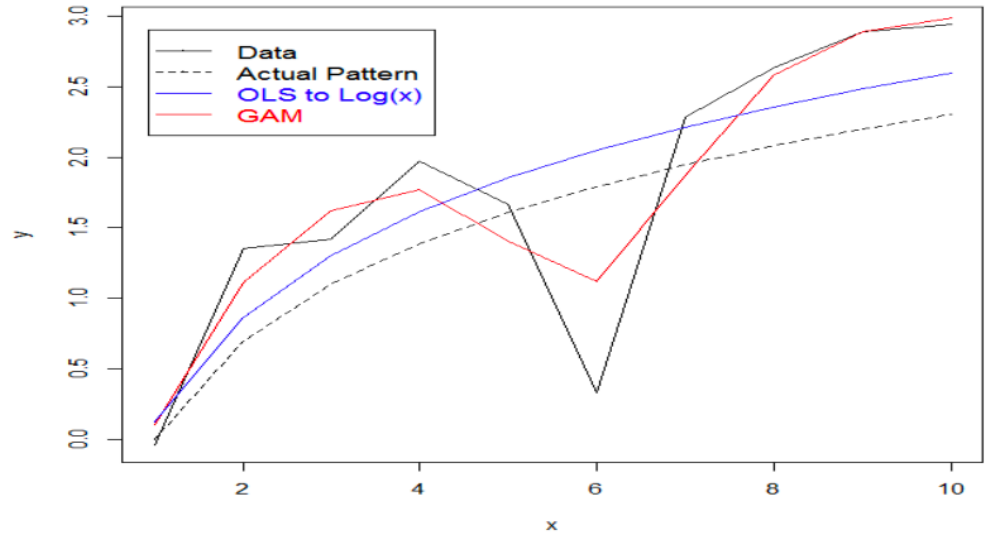
Good! →



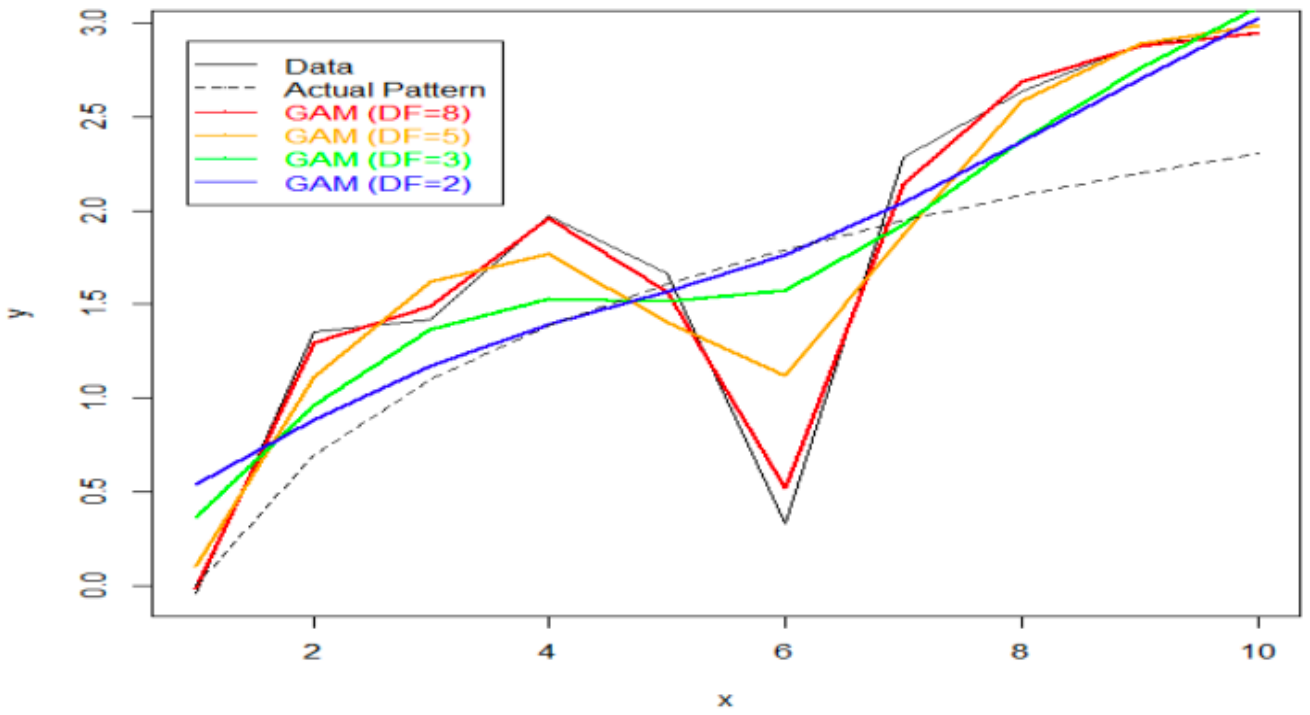
Not so good! →

Smoothing can sometimes lead to 'overfitting.'

If good parametric fit exists, use that instead.

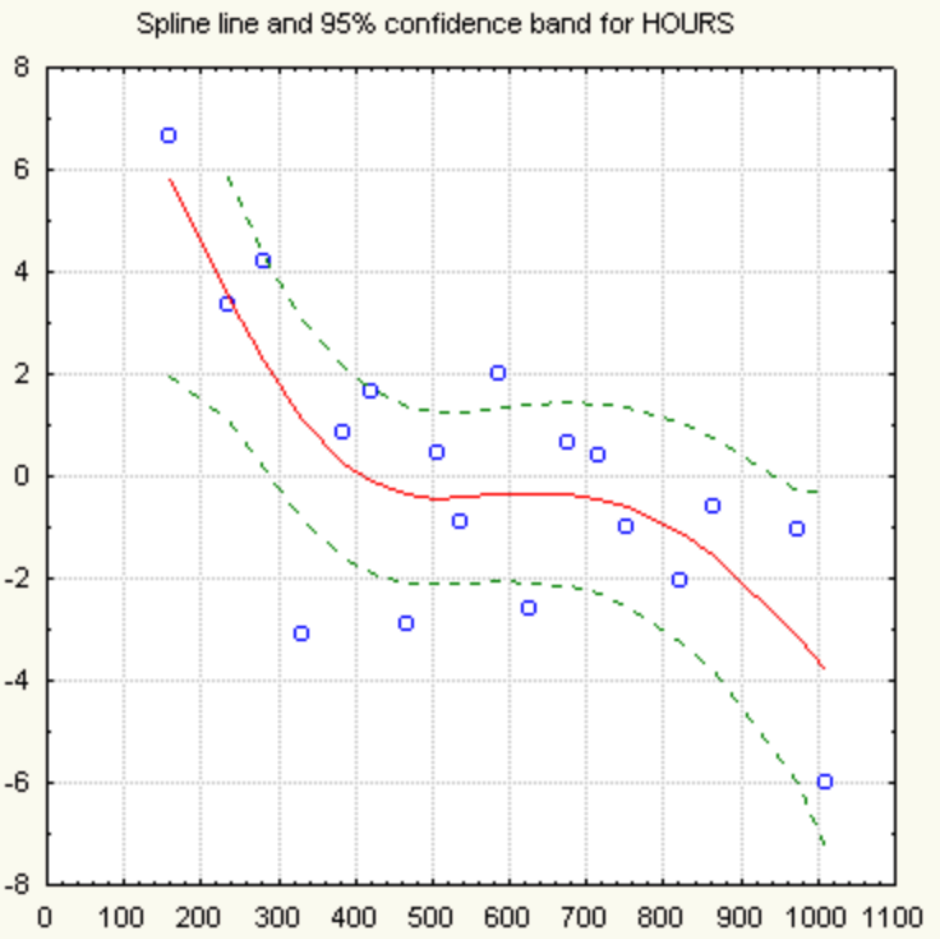


GAM Examples

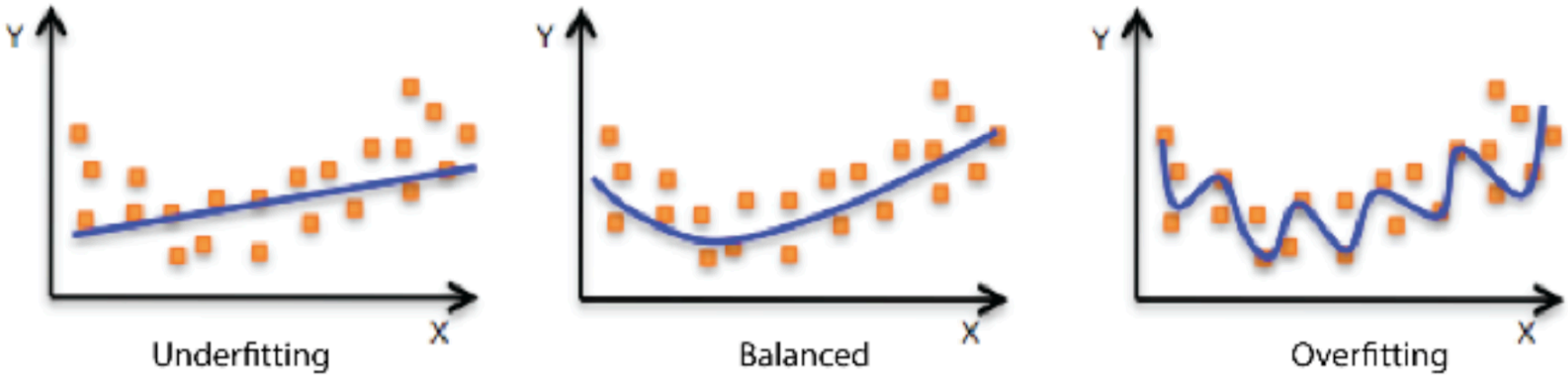


- Amount of smoothing can vary
- DF = degrees of freedom
  = number of parameters we are using to do the ‘smooth’
- A GAM ranges from a linear ‘curve’ to fitting each point exactly

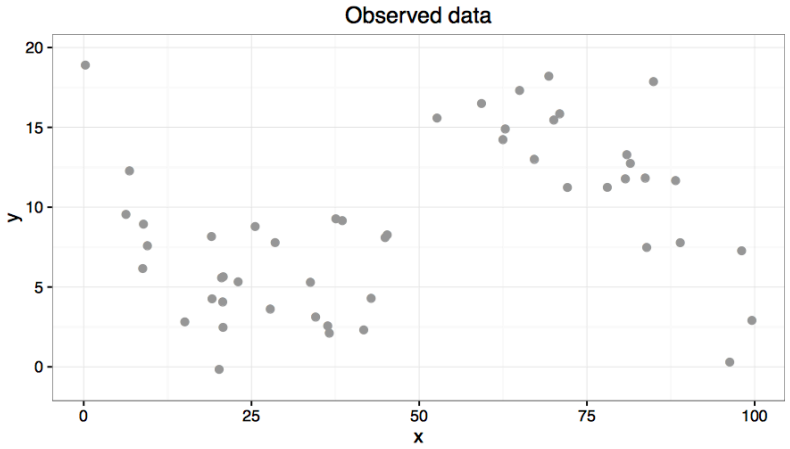
GAM Examples

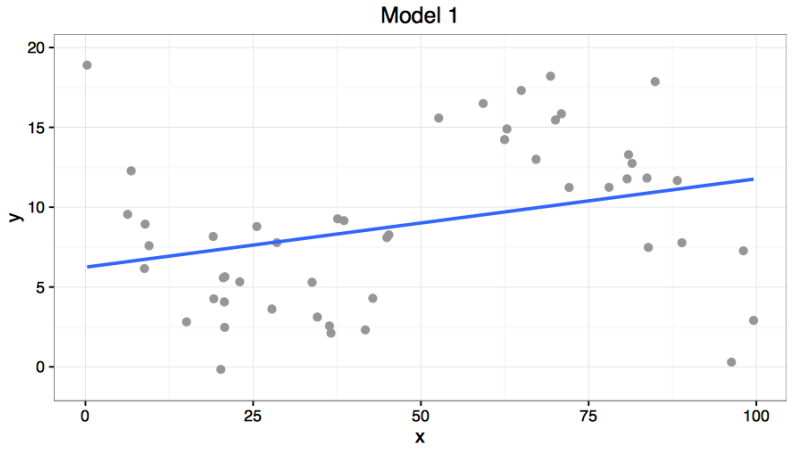


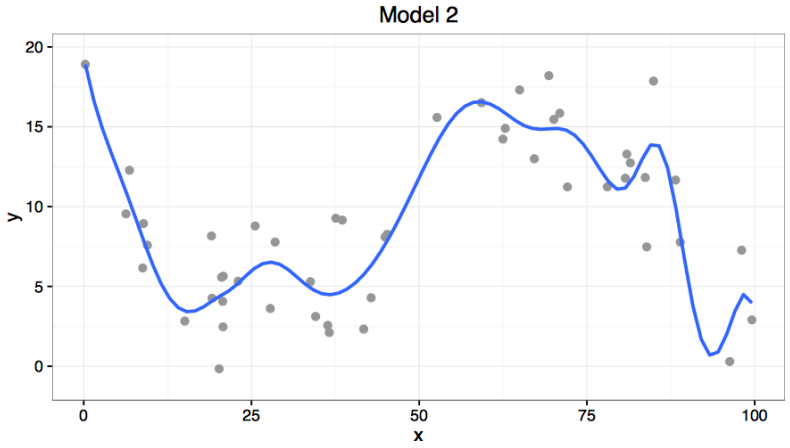
Know the gold standard

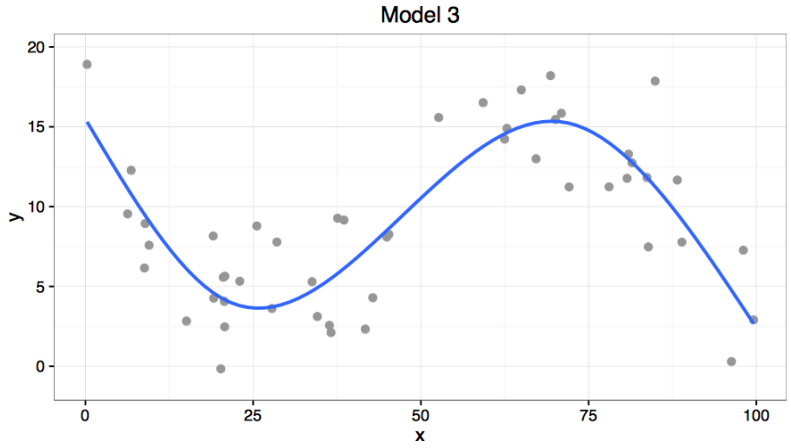


How is the fit?

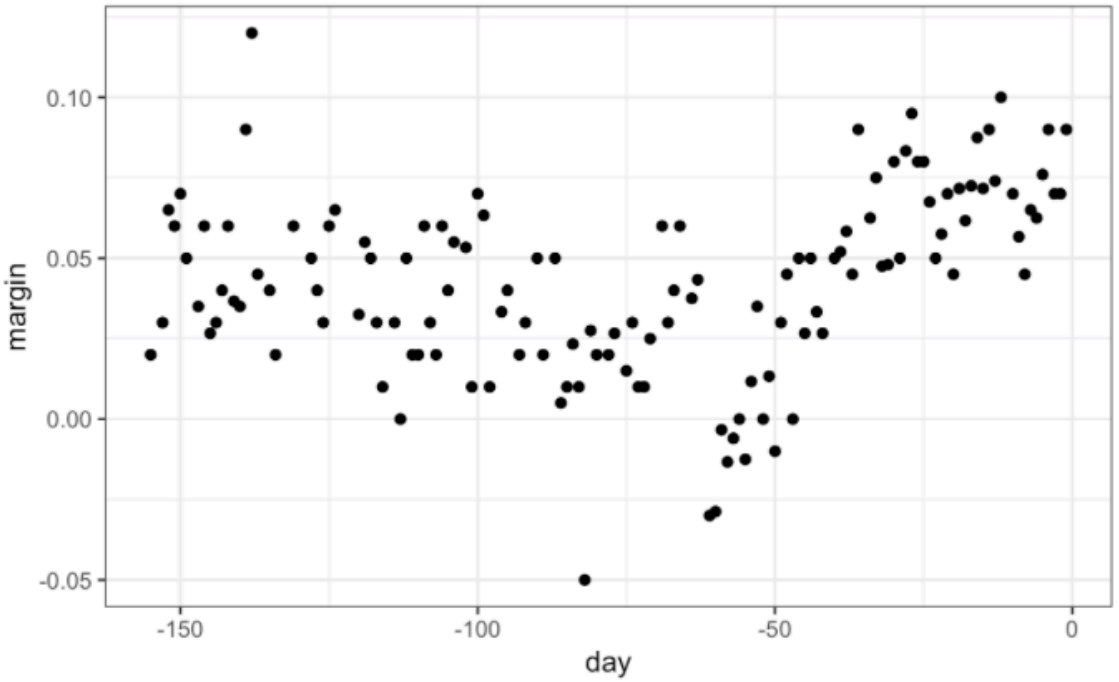




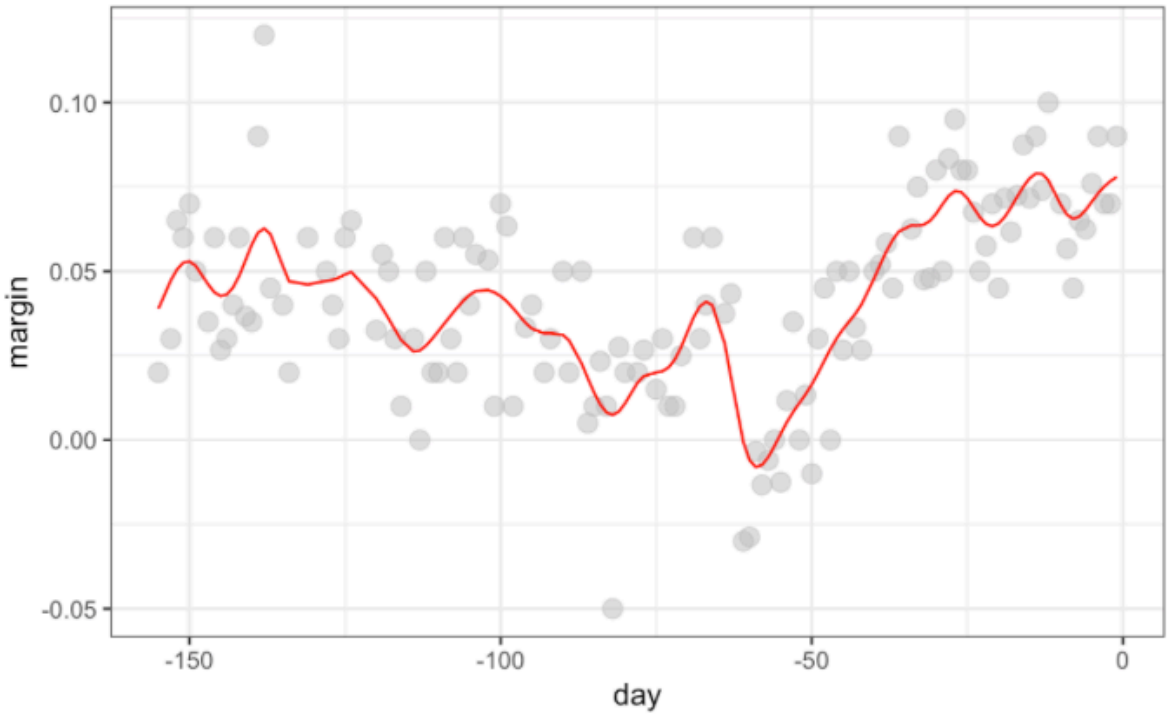




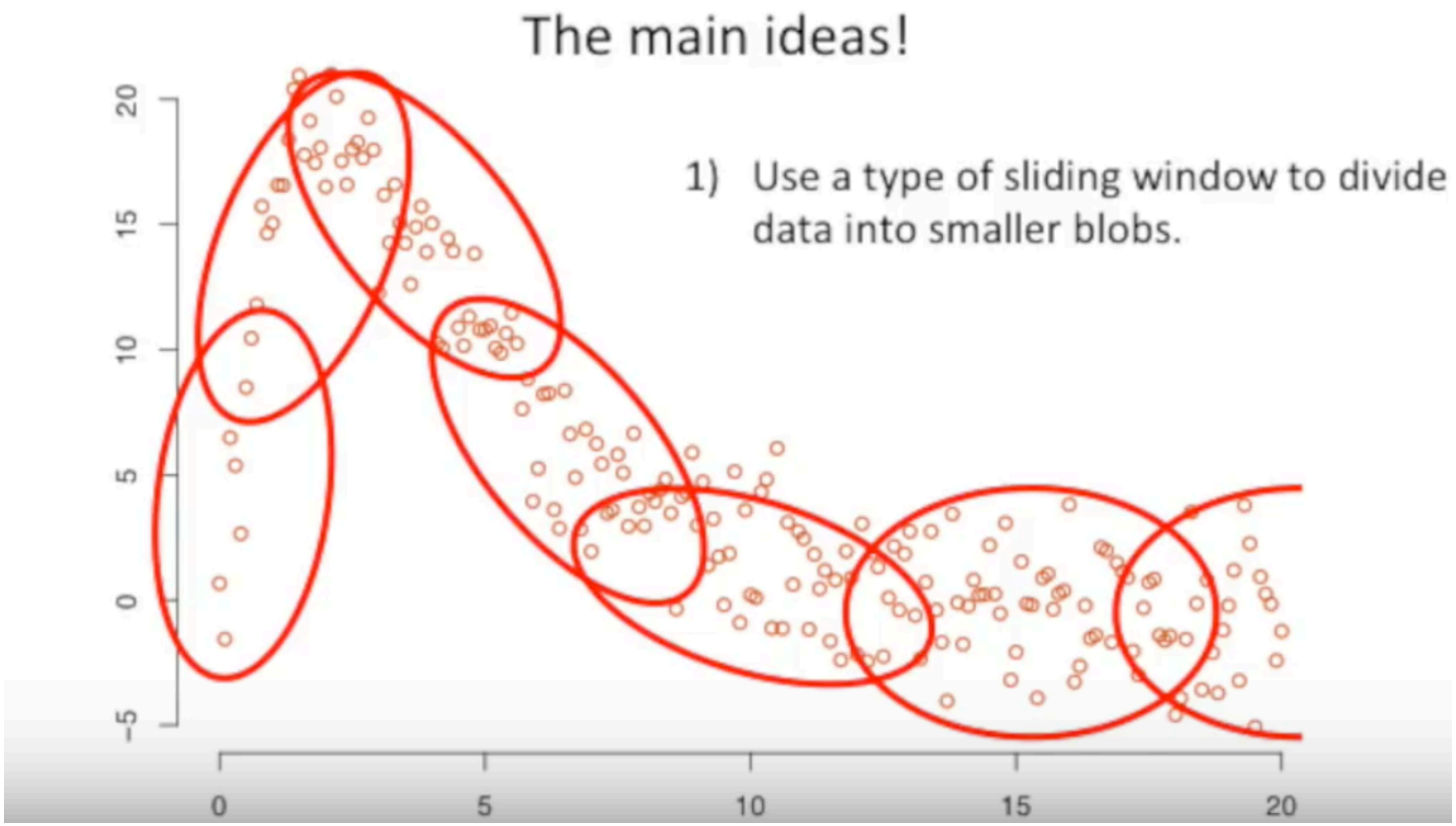
Recall this example



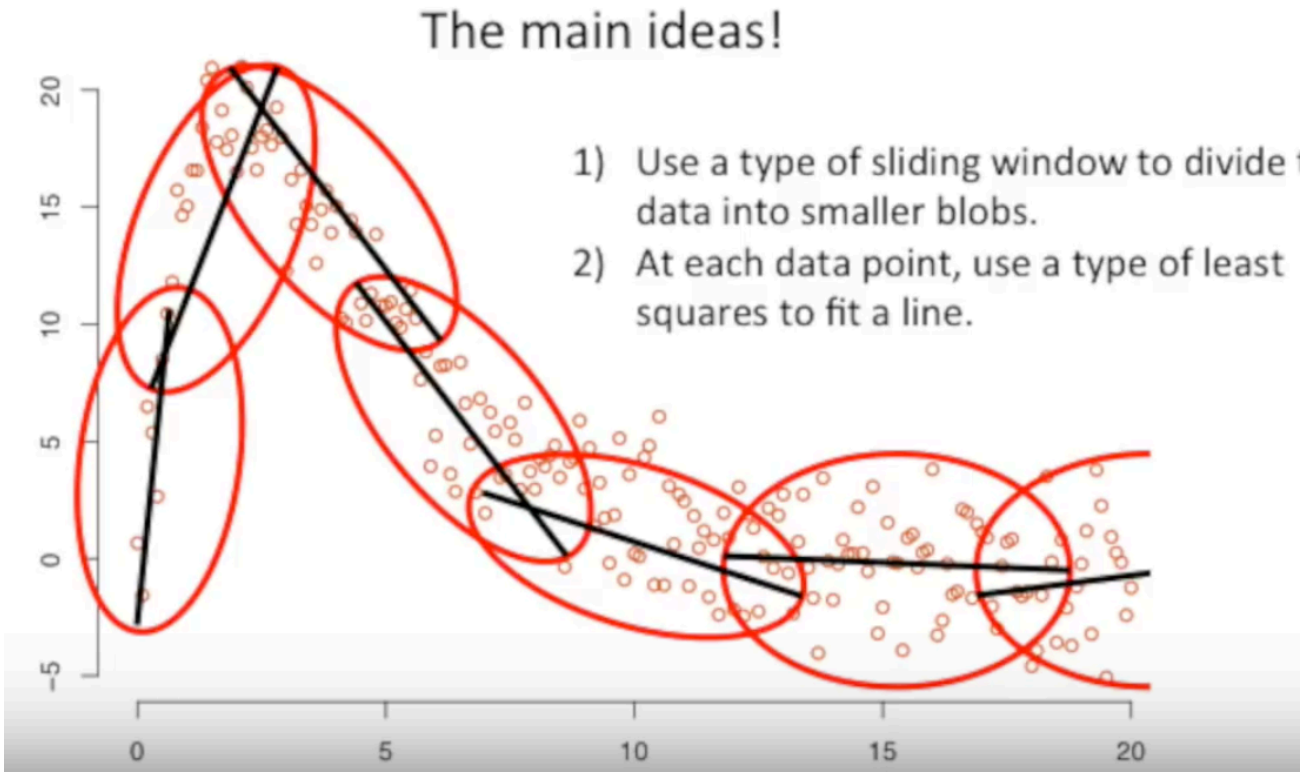
Recall this example



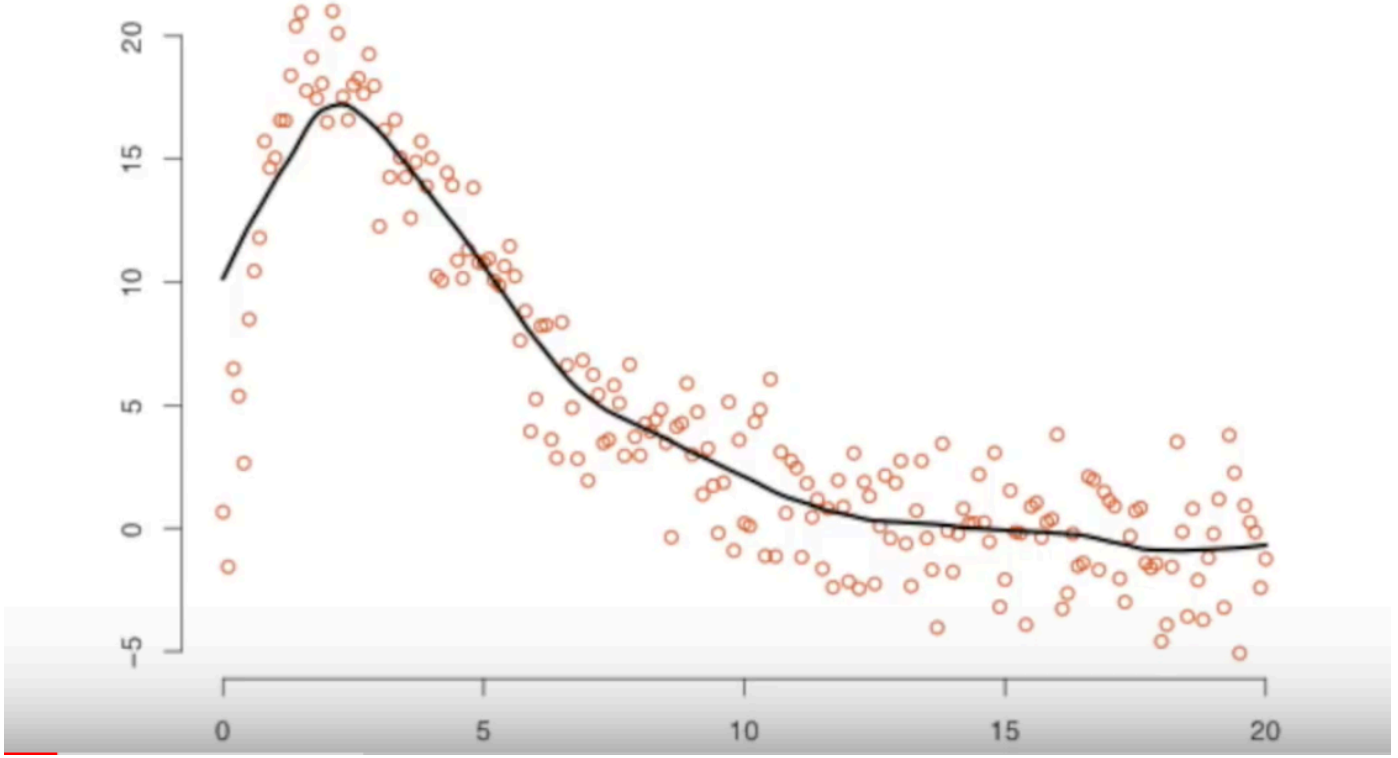
How smoothing works



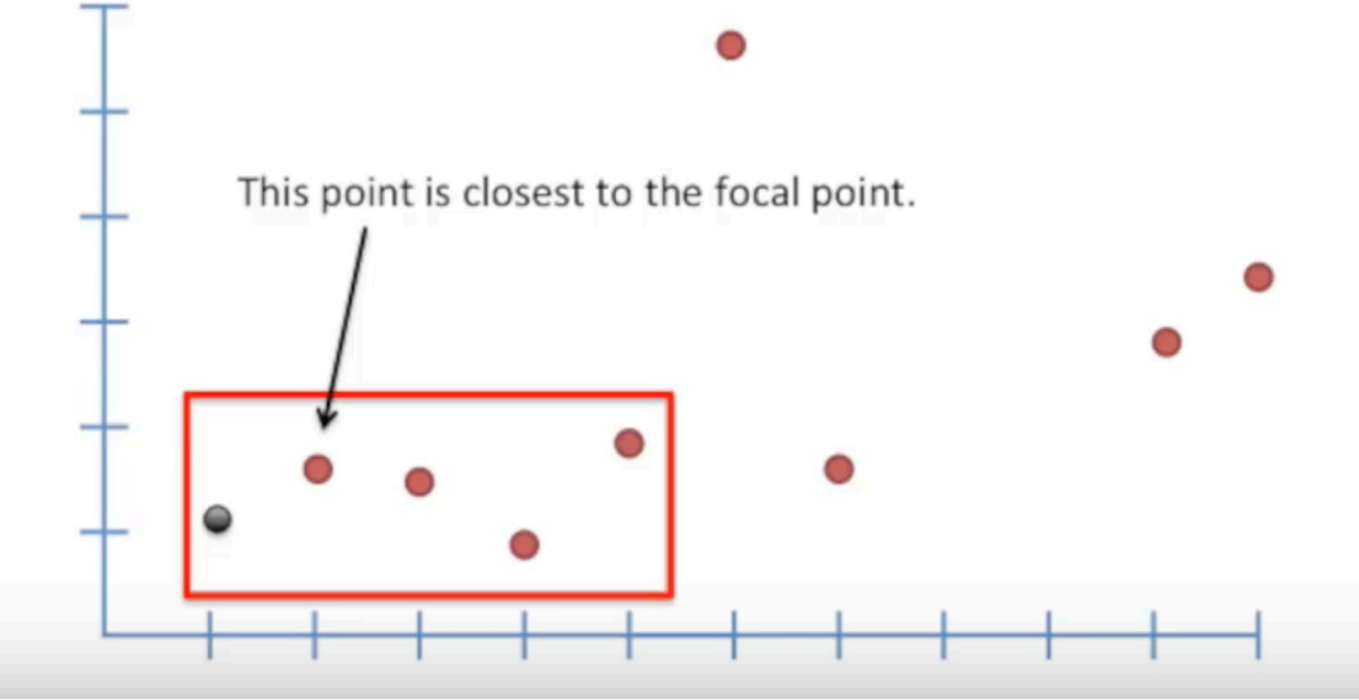
How smoothing works



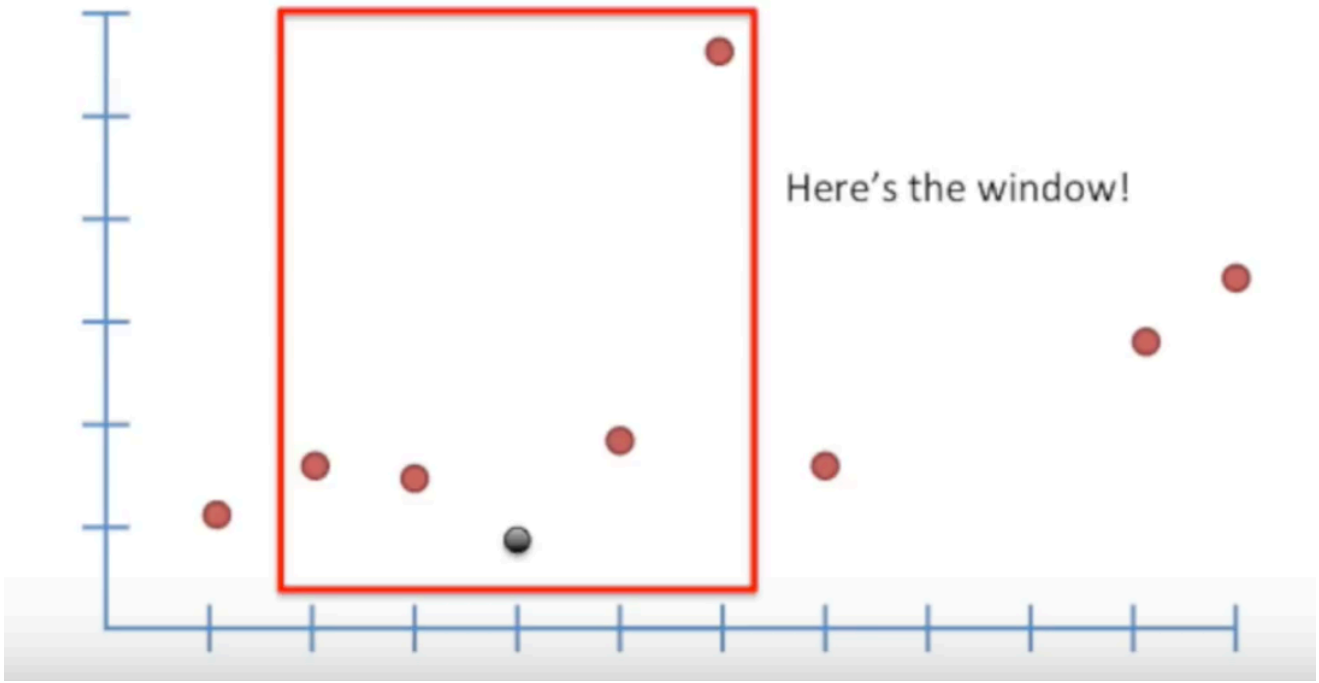
Wala!



‘Window’, ‘Kernel,…



‘Window’, ‘Kernel’



Now let’s back up…

What if uncertain about linearity of effect of a continuous covariate?

- 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

- 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

Advantages/Disadvantages of Categories

-

Use of Polynomials

- Higher order polynomials: linear, quadratic, cubic, … (or x, x^2, x^3, …)
- Tukey's power transformation: Rather than x [or Y], consider x^{λ} [or Y^{λ}] for some optimal choice of 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)