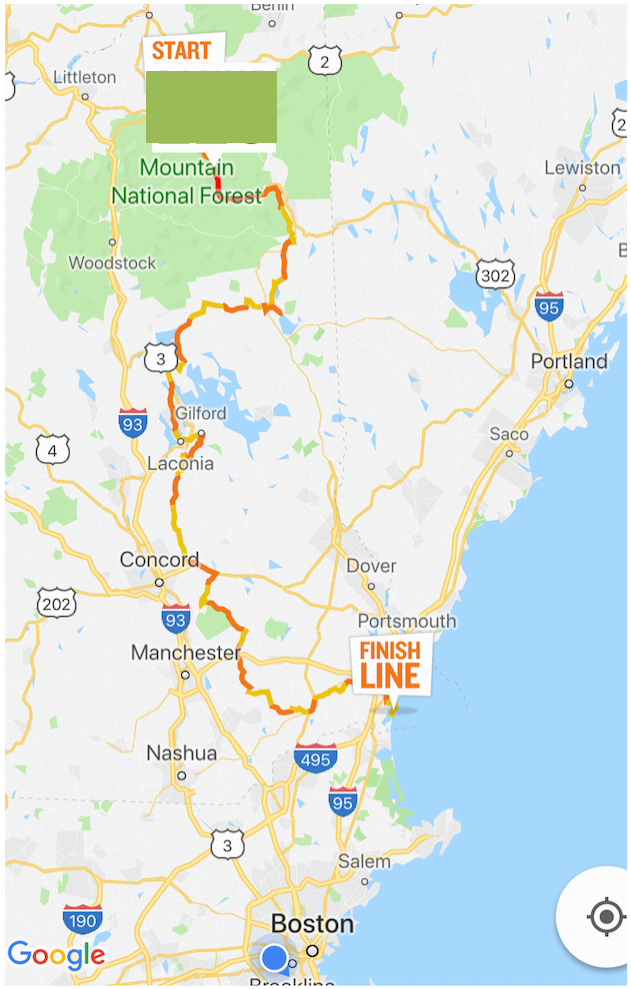
BST 210
Applied Regression Analysis



Lecture 5

Plan for Today

- Recap and Big Picture: toward Modeling
- Assessing Influence and Leverage
- Assessing Model Fit
- Flexible Modeling

Remember the Overarching Goal

- Keep in mind:
  - The simplicity of final models—we generally prefer simpler models over more complex models
  - The extent to which models meet our stated objectives/answer our research questions
  - The interpretability of models--especially if we are interested in estimating an effect
  - The cost/feasibility of actually implementing models in practice

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 $E[Y|X1,...,Xp] = \beta0 + \beta1 \cdot X1 + ... + \beta p \cdot Xp$
- Diagnostics/Checking Assumptions: LINE
  - Scatterplot, summary statistics
  - Boxplots, histograms
  - Correlations
  - Smoothing (example: 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.)?
  - Generalizations

Continue to develop framework for analysis

Then -

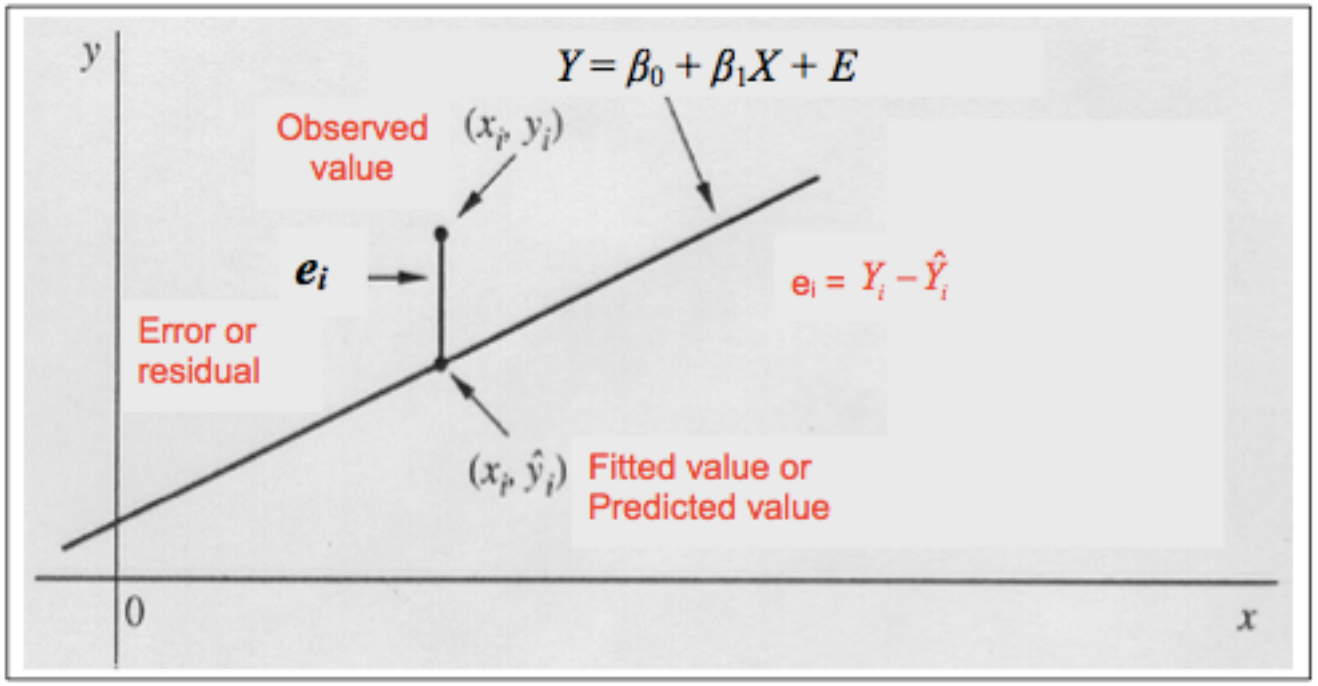
- Assess Model Fit:
  - R^2
  - MSE
  - Confidence Intervals
  - Residual Analysis: assumptions met?
  - Influence Analysis: Outliers? Leverage? Influence?
- Interpretation and inference, constant collaboration around data and meaning
- Back up and regroup as needed, delve deeper, use caution, stay organized

Recall: Residual Analysis (can contain outliers)

$$e_i = Y_i - \hat{Y}_i$$

Recall: Finding the "Best" Line

- Difference between a fitted value and an observed value is called the ***residual*** or the "***error***"
- "Best" line is the line that minimizes the error

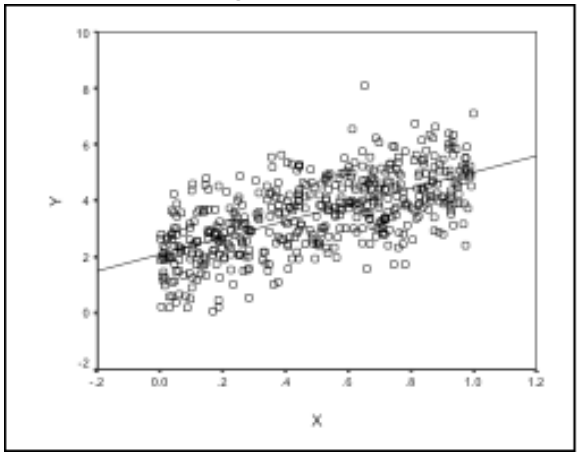


Recall: Residual Analysis

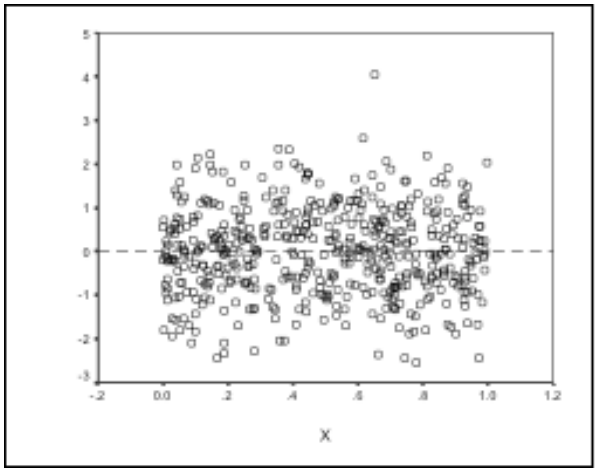
- Plots of residuals can help us to:
  - detect outliers (observations with very large residuals)
  - assess normality
  - assess homoscedasticity (equal variance assumption)
  - look for nonlinear trends (departures from linearity)

Reference – gold standard

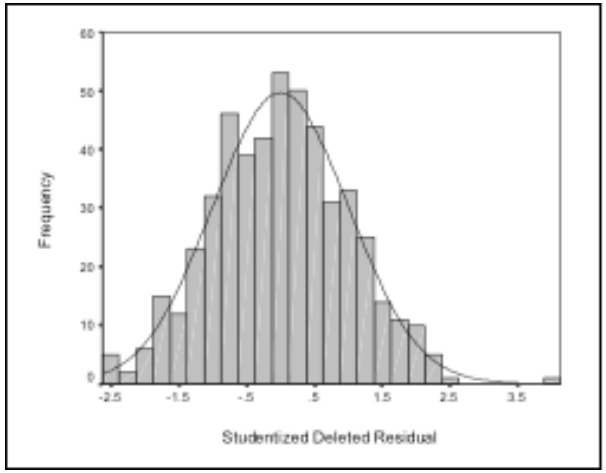
Scatterplot



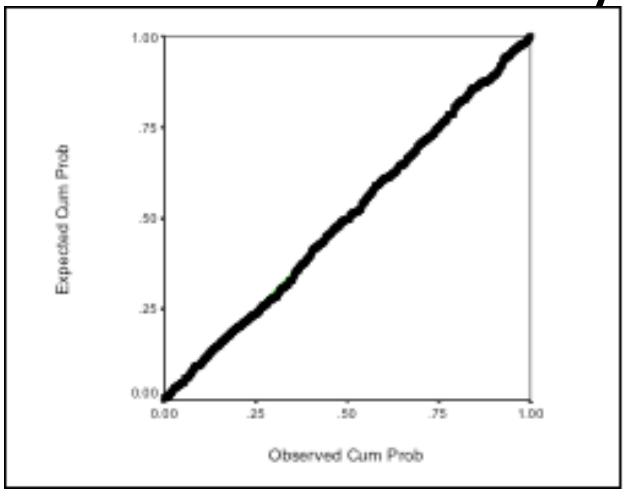
Residual Plot



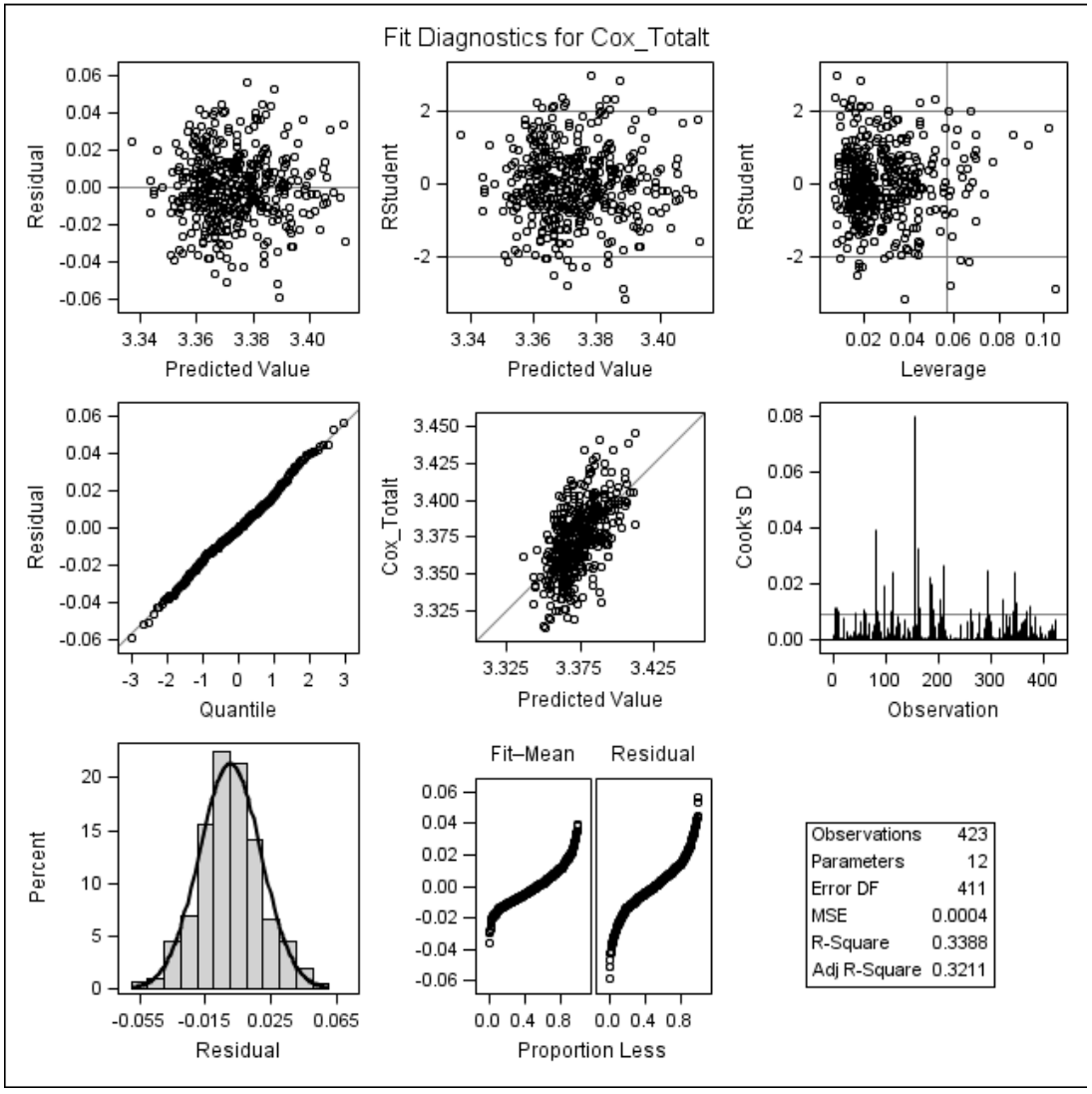
Histogram



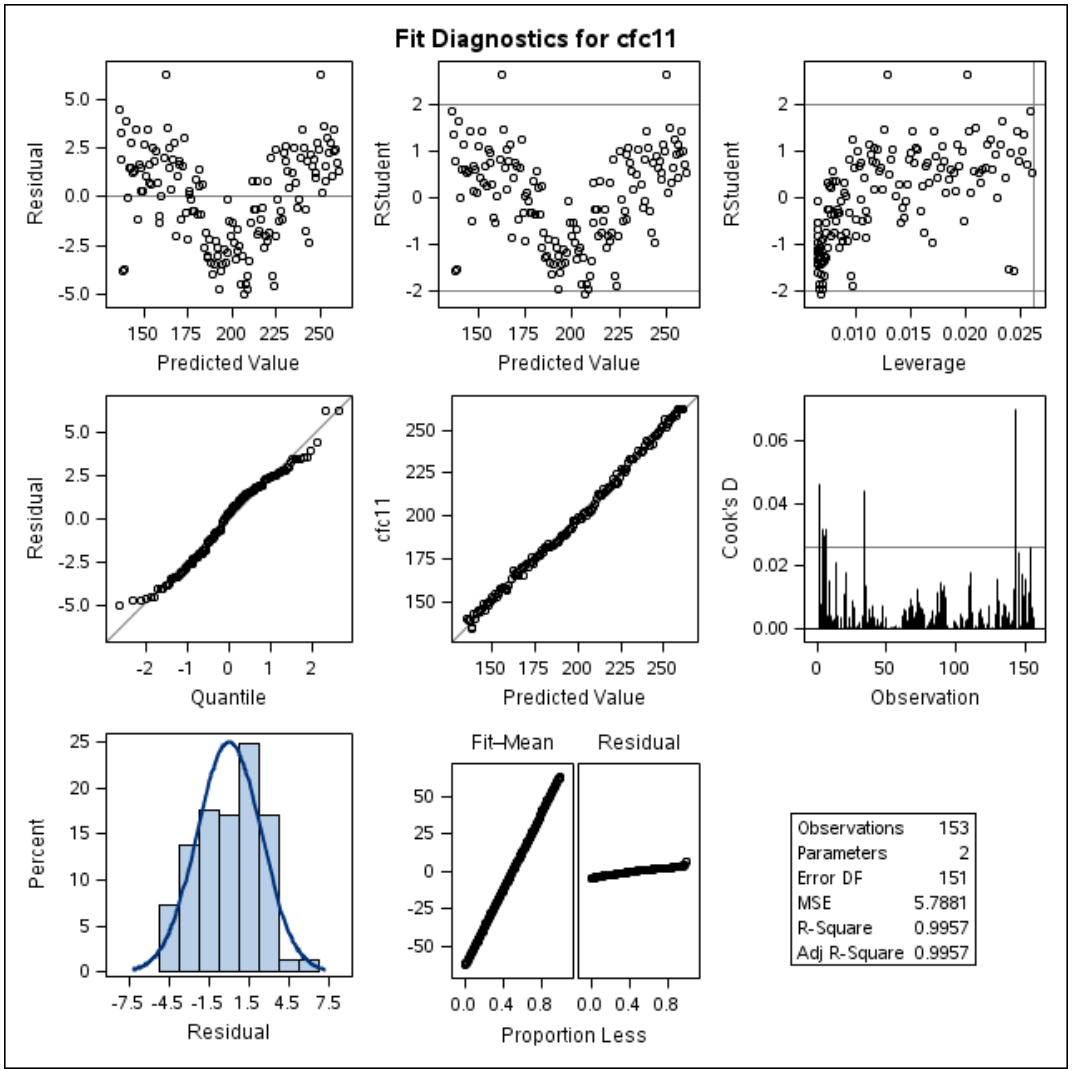
Normal Probability Plot



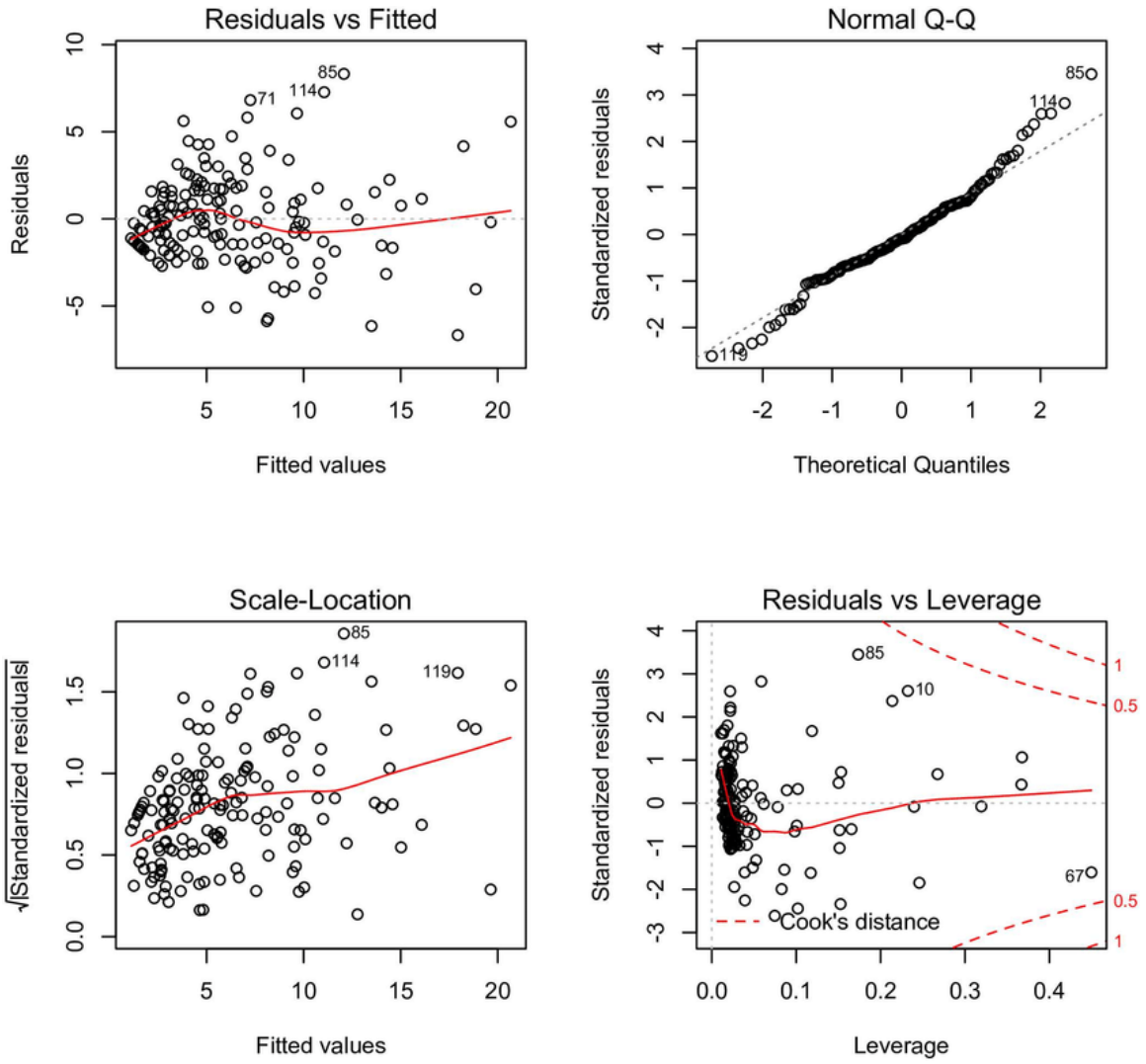
Assumptions met?



Assumptions met?

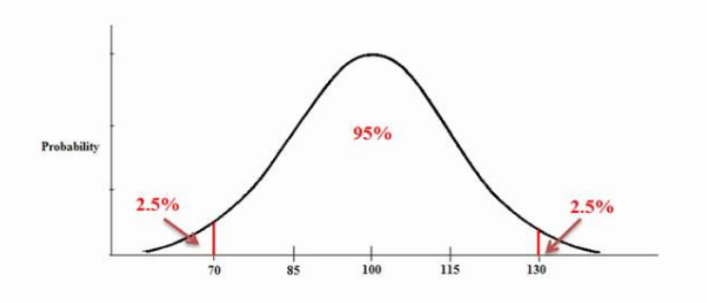


Assumptions met?



Recall: Scaling or Studentizing Residuals

- The residuals from the model will have mean zero (if an intercept is included), but will also have a variance or standard deviation
- Scaling the residuals might help in determining outlying data points—can treat such like test statistic and compare to standard normal values
- It may be easier to compare the standardized (or studentized) residuals to standard normal cutoffs to look for outlying observations, e.g., >2 (or >1.96) in absolute value
- One could compare models run first including and then excluding points with high standardized or studentized residuals

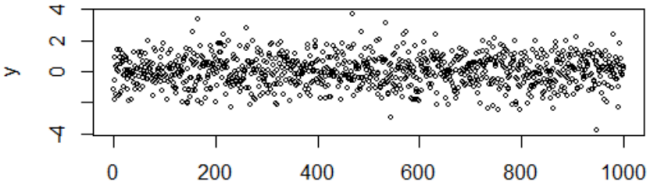


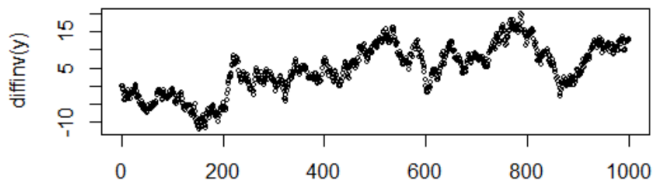
Autocorrelation in Residuals

- Autocorrelation refers to the correlation between residuals across different cases, say if the case number *i* referred to time *i*
- One could examine successive residuals to see if they were uncorrelated or not via the Durbin-Watson statistic

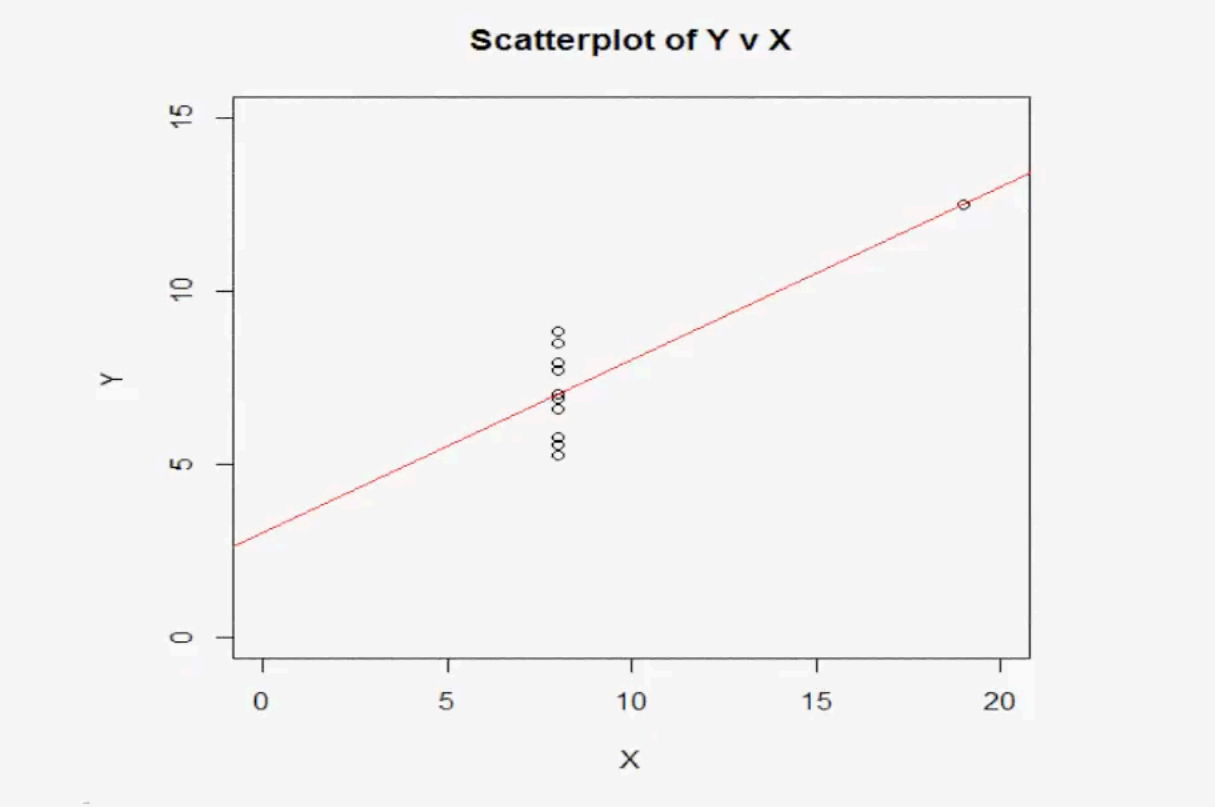
$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2}$$

- If successive residuals are uncorrelated, *d* would be equal to 2
- Positive correlation leads to $0 < d < 2$, while negative autocorrelation leads to $d > 2$
- *d* can be compared to critical values to perform a test of H_0: $d = 2$ vs. H_0: $d = 2$, with the critical values dependent on the sample size and number of covariates included





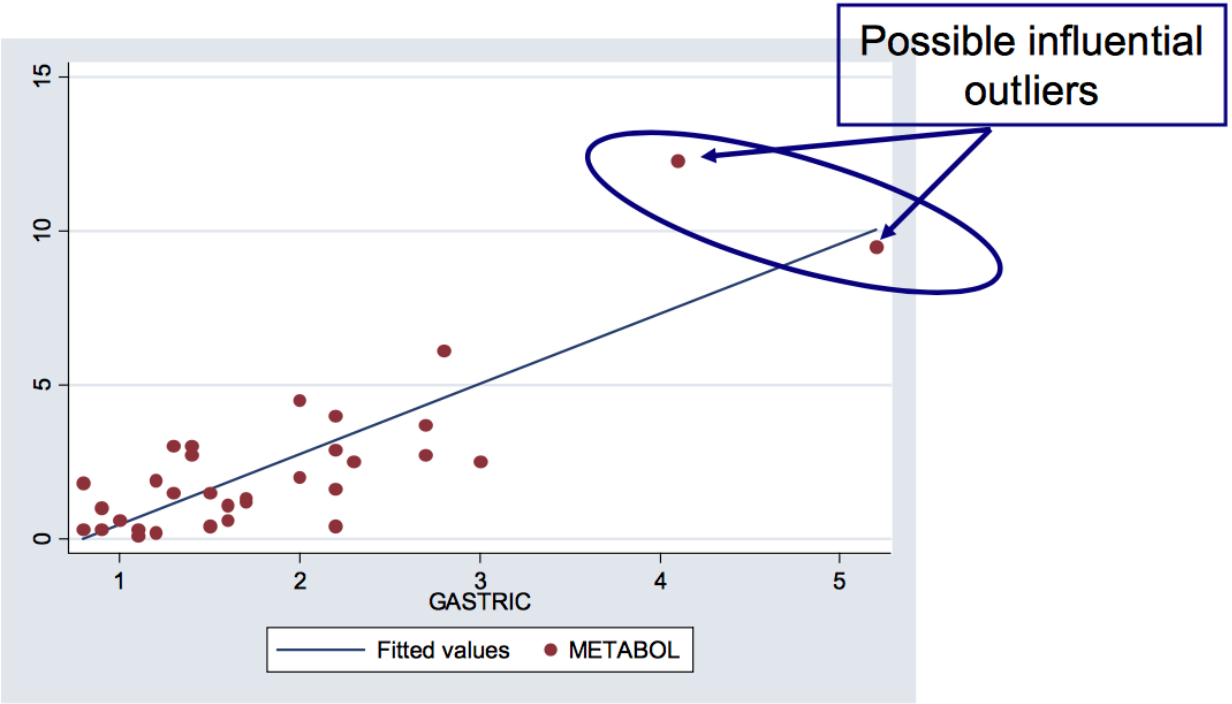
What if…



Influence Analysis

- Outliers
  - An observation with large residual.
    - An observation whose dependent-variable value is unusual given its values on the predictor variables.
    - An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.
- Leverage
  - An observation with an extreme value on a predictor variable
    - Leverage is a measure of how far an independent variable deviates from its mean.
    - These leverage points can have an effect on the estimate of regression coefficients.
- Influence
  - Influence can be thought of as the product of leverage and outlierness.
    - Removing the observation substantially changes the estimate of coefficients.

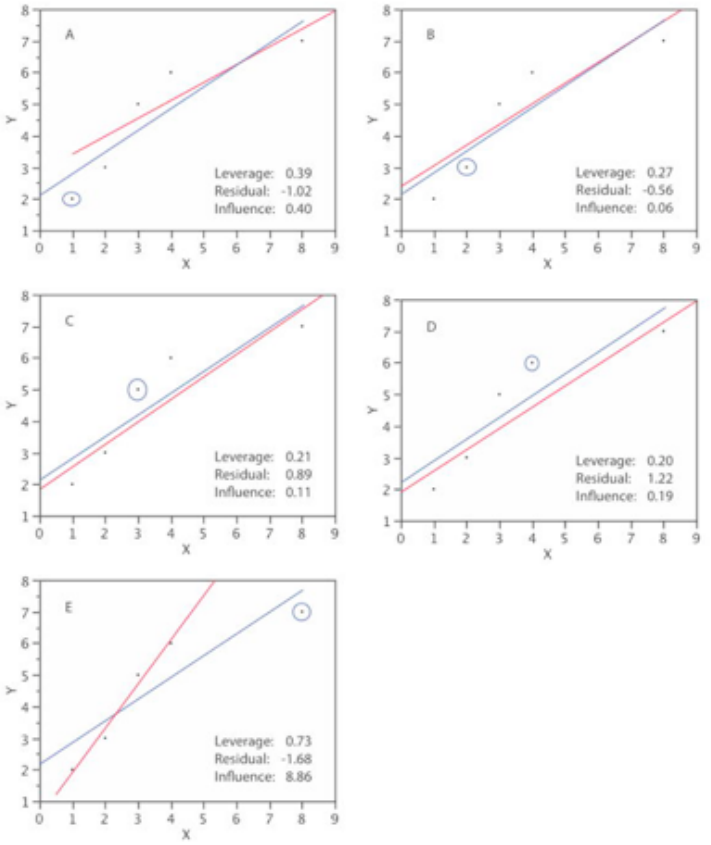
Influence Analysis



- Does the fitted regression model change when the two isolated points are removed?

Influence Analysis

- An observation is influential if its deletion substantially changes the regression results
  - High leverage => 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

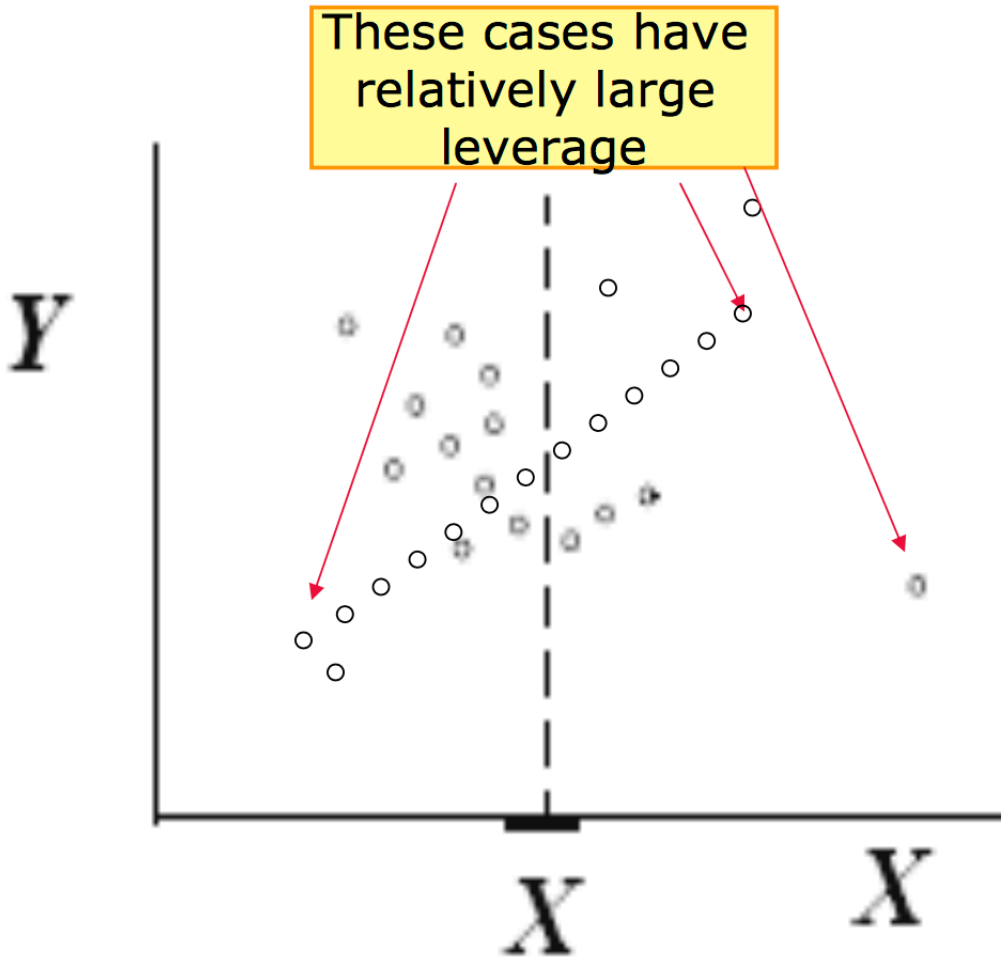


Influence Analysis

- These help identify influential observations and help to clarify the course of action.
- Use them when:
  - you suspect influence problems and
  - when graphical displays may not be adequate

- One useful set of case influence statistics:
  - D_i: Cook's Distance - for measuring influence
  - h_i: Leverage - for measuring "unusualness" of x's
  - r_i: Studentized residual - for measuring "outlierness"
    - Note: i = 1,2,..., n
- Sample use of influence statistics…

Influence Analysis: Leverage



Influence Analysis: Leverage

Leverage: h_i for the single variable case

(also called: diagonal element of the hat matrix)

- It measures the multivariate distance between the x's for case i and the average x's, accounting for the correlation structure.

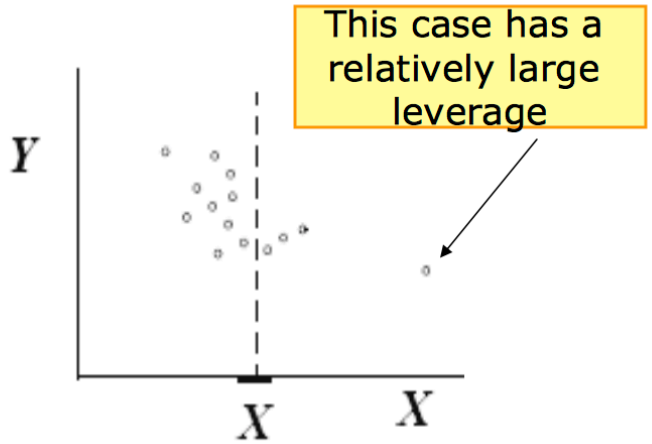
If there is only one x:

$$h_i = \frac{1}{(n-1)}\left(\frac{x_i - \bar{x}}{s_x}\right)^2 + \frac{1}{n}$$

Equivalently:

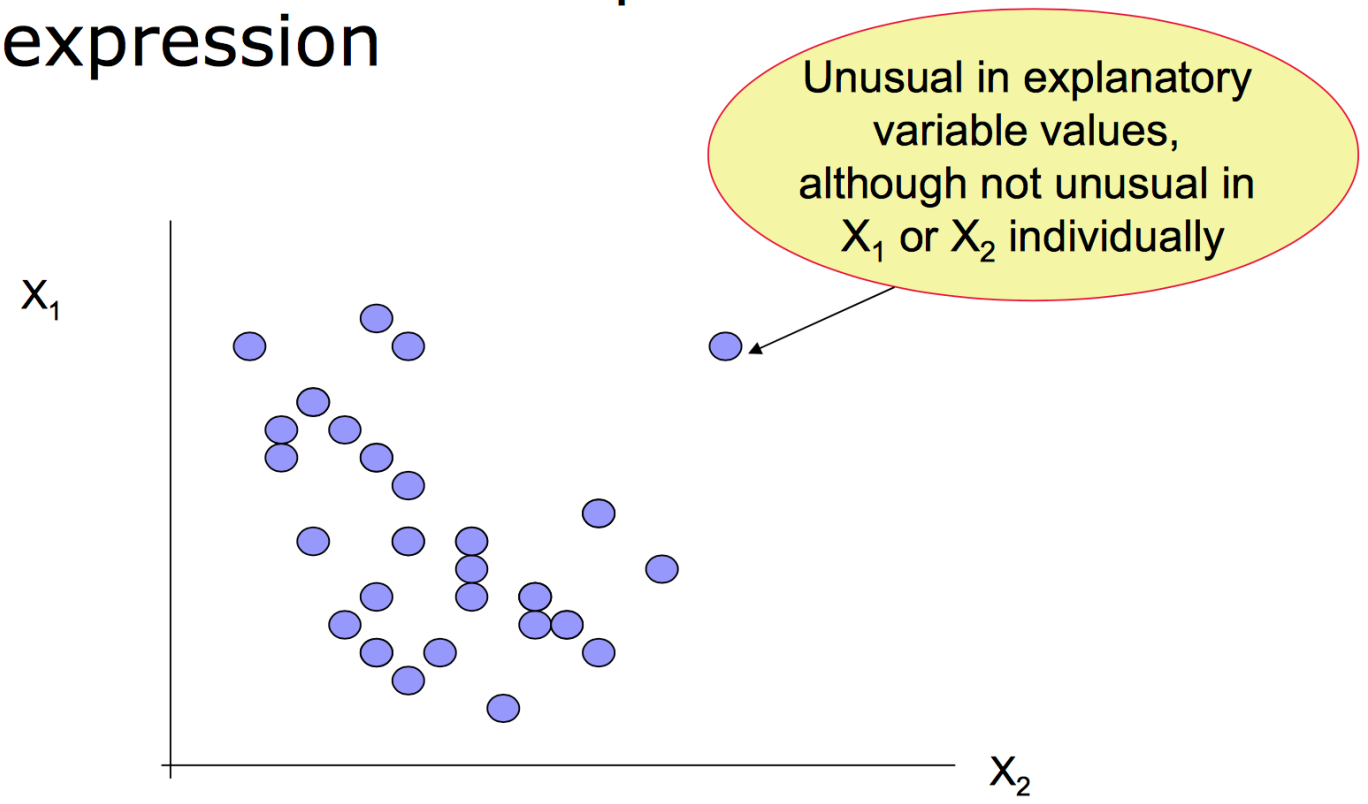
$$h_i = \frac{(x_i - \bar{x})^2}{\sum (x - \bar{x})^2} + \frac{1}{n}$$

Leverage is the proportion of the total sum of squares of the explanatory variable contributed by the i^{th} case.



Influence Analysis: Leverage

- For several x's, h_i has a matrix expression



Influence Analysis: Leverage

- One could use boxplots or histograms to assess for observations with high leverage observations
- One can show that the average of the n h_{ii} (diagonals of matrix of fitted values in multiple regression) values is $(p + 1)/n$, where p is the number of covariates (excluding the intercept) and n is the number of observations
- One rule of thumb is that $h_{ii} > 2(p + 1)/n$ signals a high leverage point (more than twice the average value)
- High leverage not always bad; just implies the observation has potential to exert strong influence on model fit.

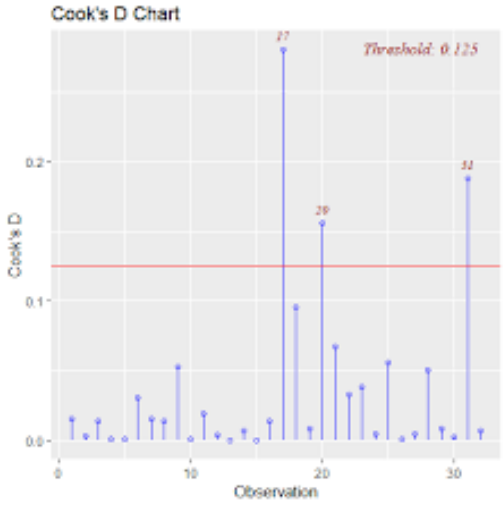
Influence Analysis: Cook's Distance

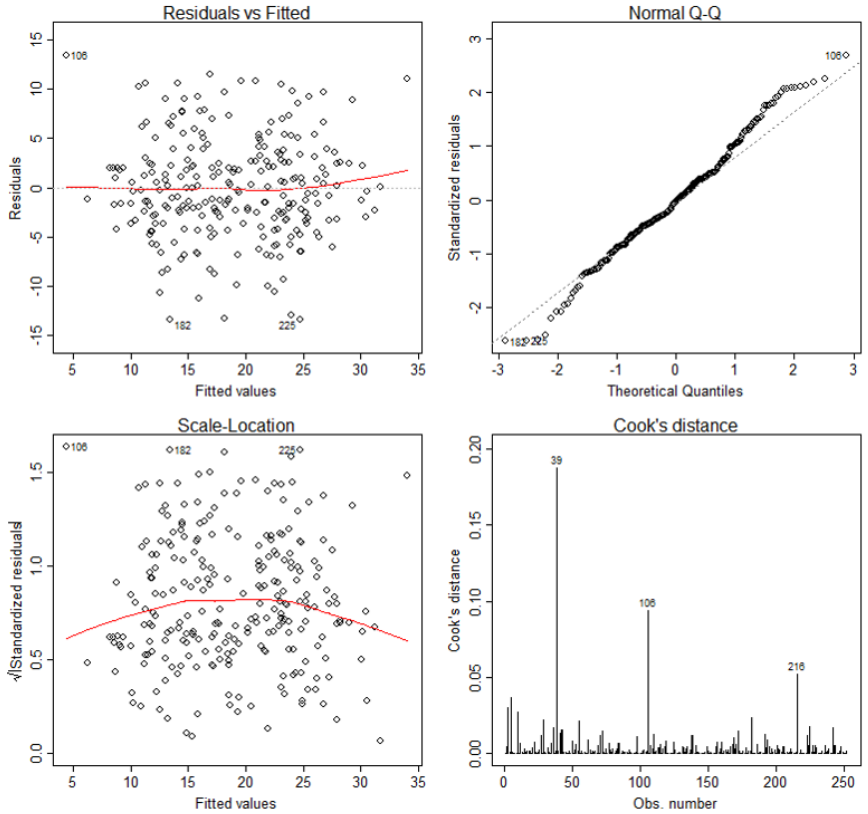
- Cook's distances are a combination of each observation's leverage and residual values; the higher these measurements, the higher Cook's distance. Measures how much the regression would change if observation i is deleted. Unlike leverage, it reflects the actual amount of influence an observation has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





Influence Analysis: *DFBETA*

- The *DFBETA* statistic quantifies how much each of the regression coefficients would change if each observation were omitted from the data set
- $DFBETA_{ik}$ 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

Influence Analysis: *DFFITS*

- A related statistic is the DFFITS measure, assessing the *i*th case's influence on $\hat{Y}_i$ and is given by

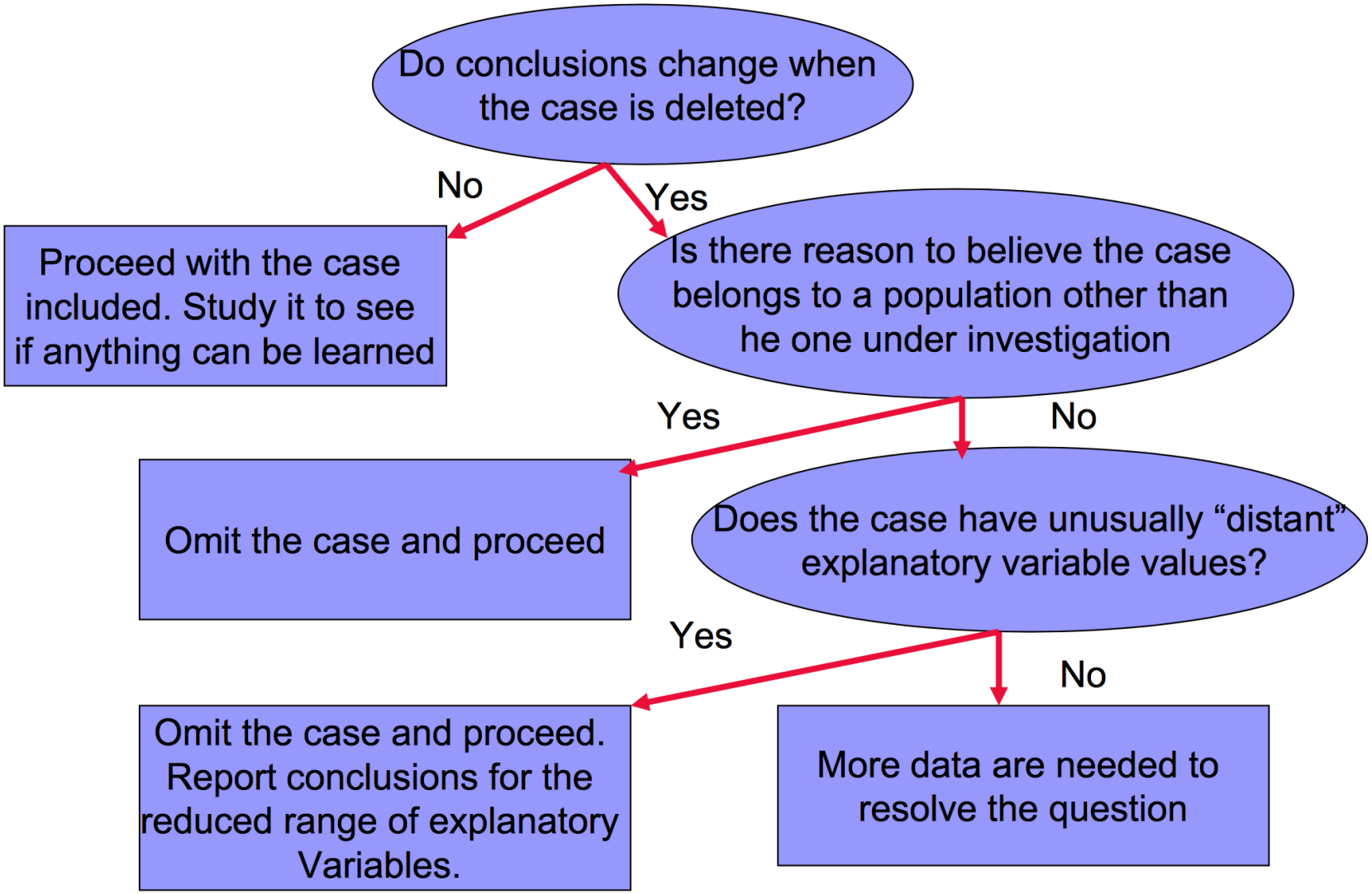
$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)i}}{s_{e(i)}/\sqrt{h_{ii}}}$$

- 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

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.

Strategy for dealing with influential cases



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!

Suppose we have a dataset with N **observations, and that we fit a regression model with** p **predictors/covariates included.** Then:

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.

- Formula: SSR = 1 − SSE SST SST
- "Better" Fitting Models Have larger R^2 values

Assessing Model Fit

- **Adjusted R2:**

addresses this issue by adding in a slight penalty that scales with the number of parameters in the model:
Formula:1− N−1 SSE N−(p+1) SST
"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:**

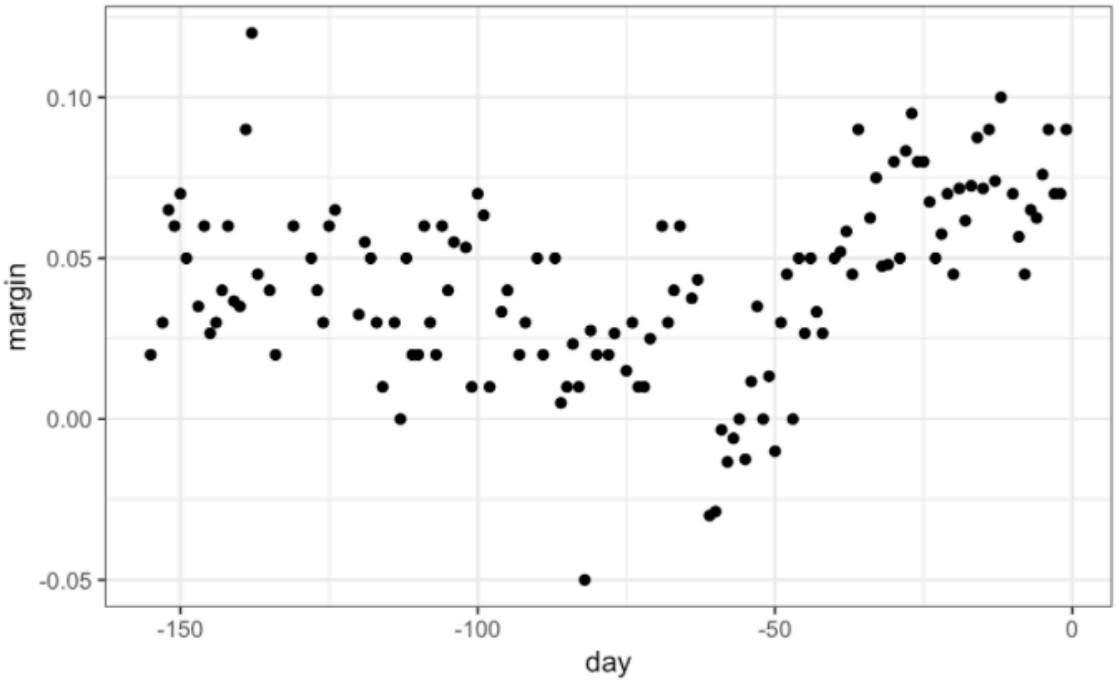
is another model selection criterion that navigates the trade-off between model complexity (which is captured by the number of parameters in the model) and model fit (which is captured by the likelihood of the model).
Formula: $2 \cdot (p + 1) - 2 \log(\hat{L})$, where $\hat{L}$ is the maximum value of the likelihood function for the model
"Better" Fitting Models Have: smaller AIC values

Next - Relaxing Assumptions of Linearity

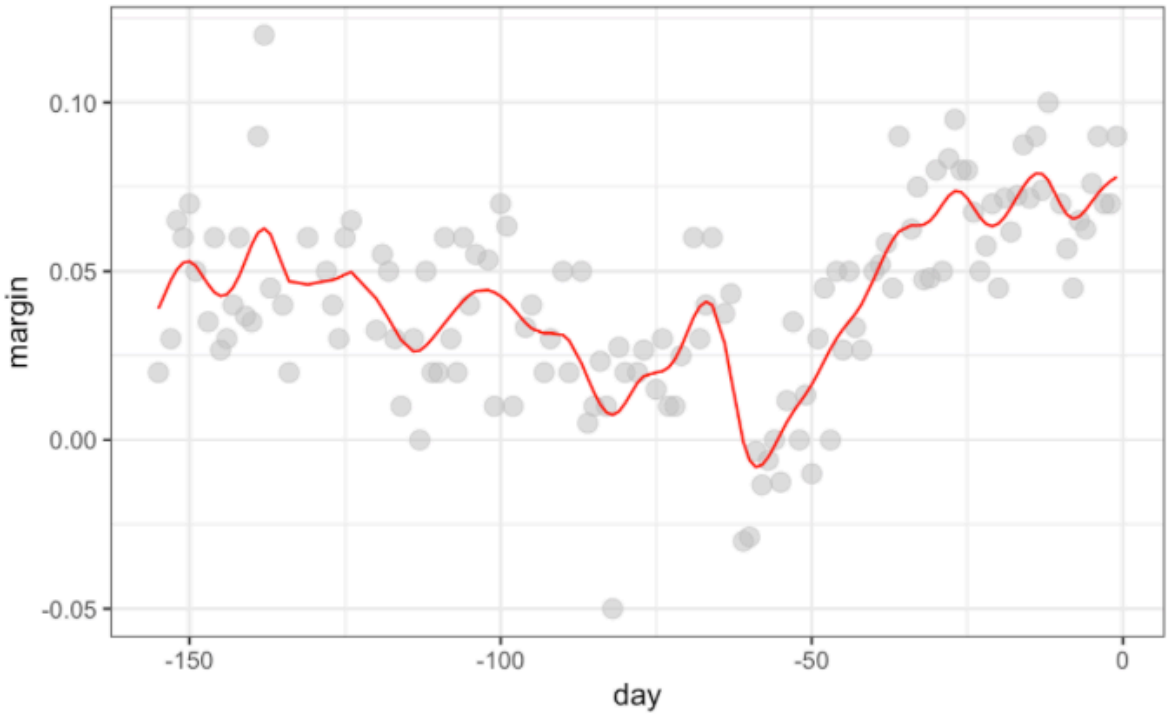
- How can I model a continuous predictor in a flexible fashion?
- Flexible modeling

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

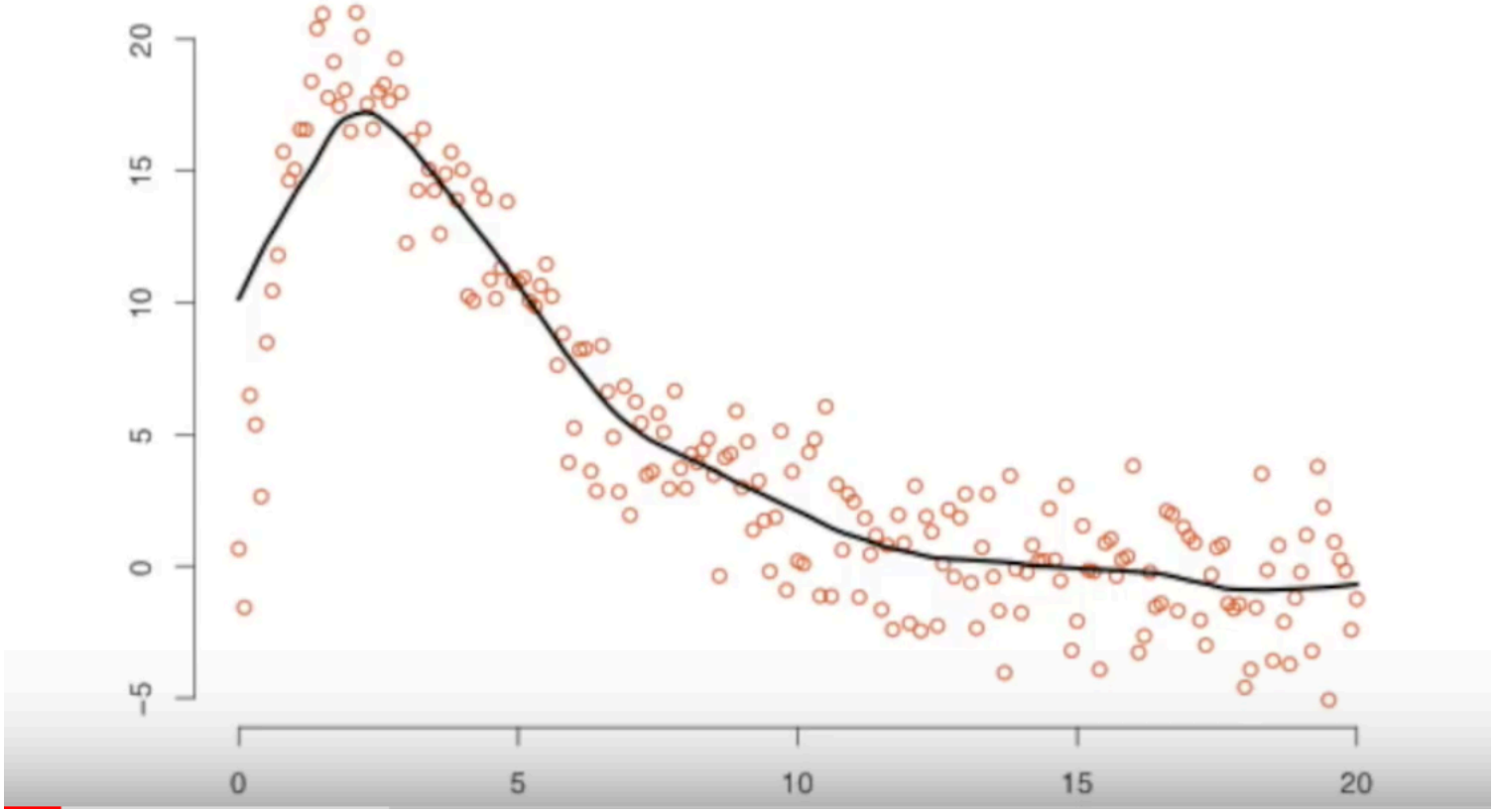
Flexible Modeling



Flexible Modeling



Flexible Modeling



Main Issue

- 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*

Main Issue

- 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 or transformations could make interpretation of β coefficients more difficult, but might make predictions better

What is Smoothing?

- Tool for summarizing the trend in a response variable as a function of one or more predictor variables
- Is less variable than the original data, hence the name smoother
- Is nonparametric – doesn't force a rigid form of the dependence (e.g., linear, quadratic)

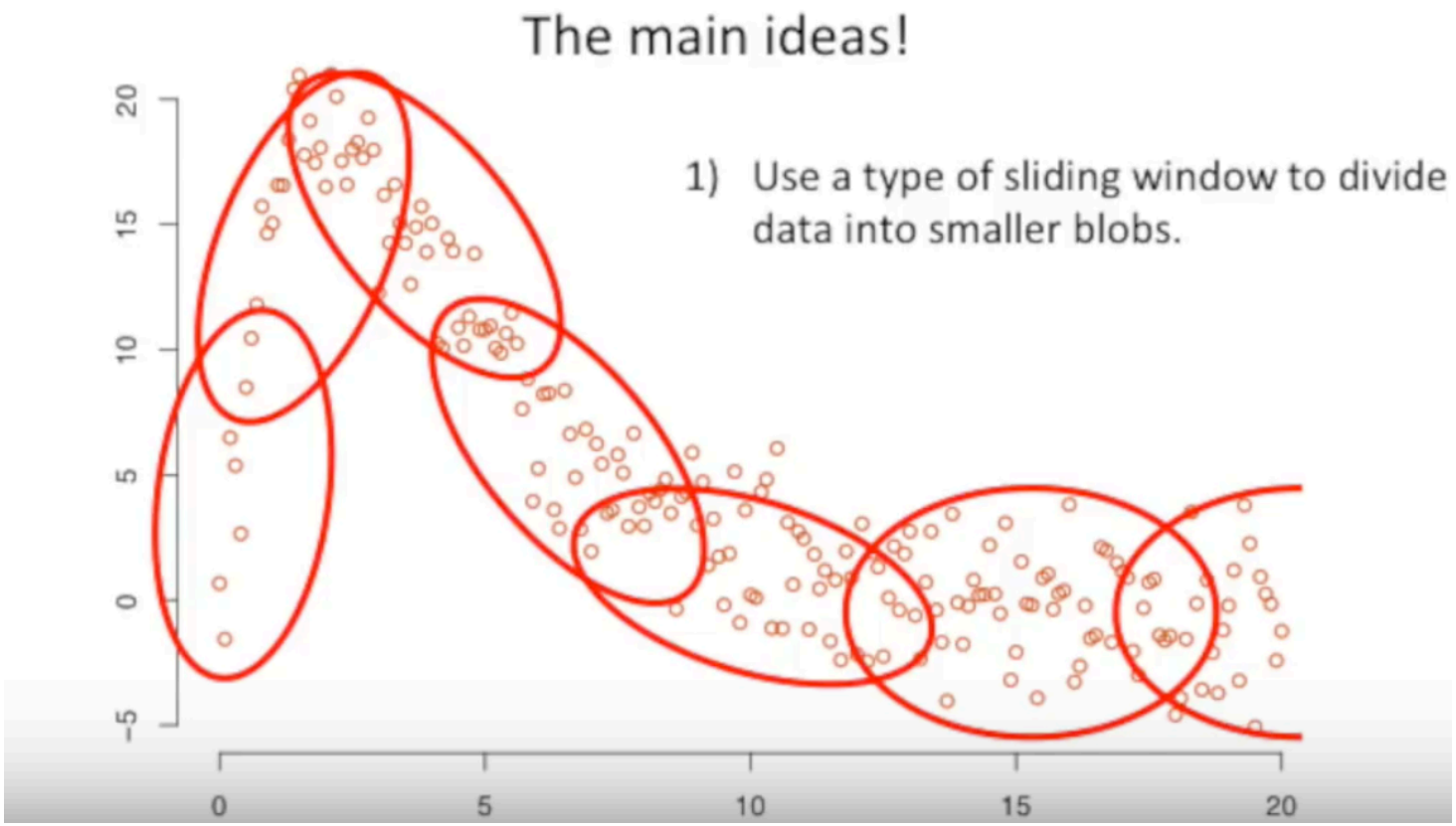
Smoothing Uses

- Description: To enhance the visual appearance of a scatterplot and assist in picking out a trend
- Prediction: To estimate the dependence of the mean of *Y* on the predictors in a less rigid fashion

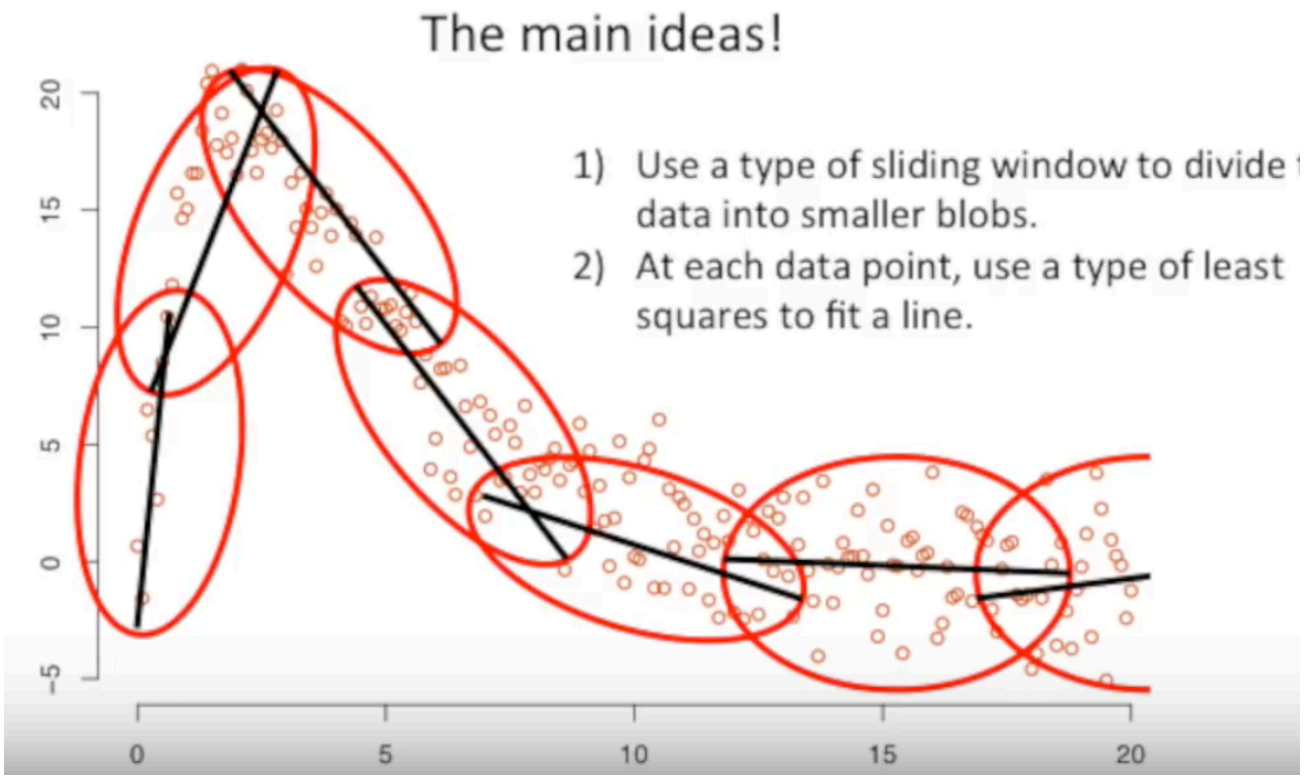
Scatterplot Smoothing

- For when there is a single predictor variable, modeling *Y* as a function of *x*
- Simple linear or quadratic regression is a special case, with a rigid form for the dependence
- Many more nonparametric methods available

Smoothing


The main ideas!
1) Use a type of sliding window to divide data into smaller blobs.
20
15
10
5
0
−5
0
5
10
15
20

Smoothing


The main ideas!
1) Use a type of sliding window to divide
data into smaller blobs.
2) At each data point, use a type of least
squares to fit a line.
20
15
10
5
0
−5
0
5
10
15
20

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*

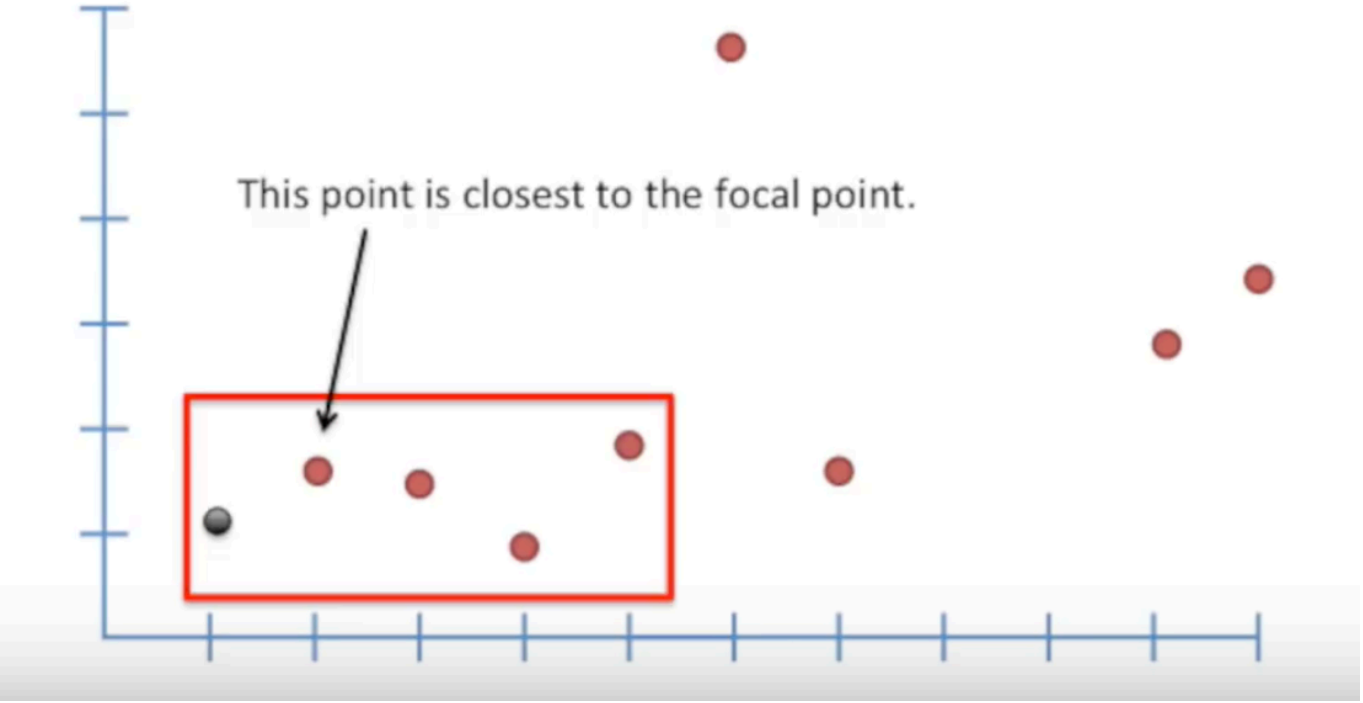
Scatterplot Smoothing

- 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)

Scatterplot Smoothing

- 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
  - Available in Stata, SAS, and R, though exact implementation may vary slightly

Smoothing



Smoothing