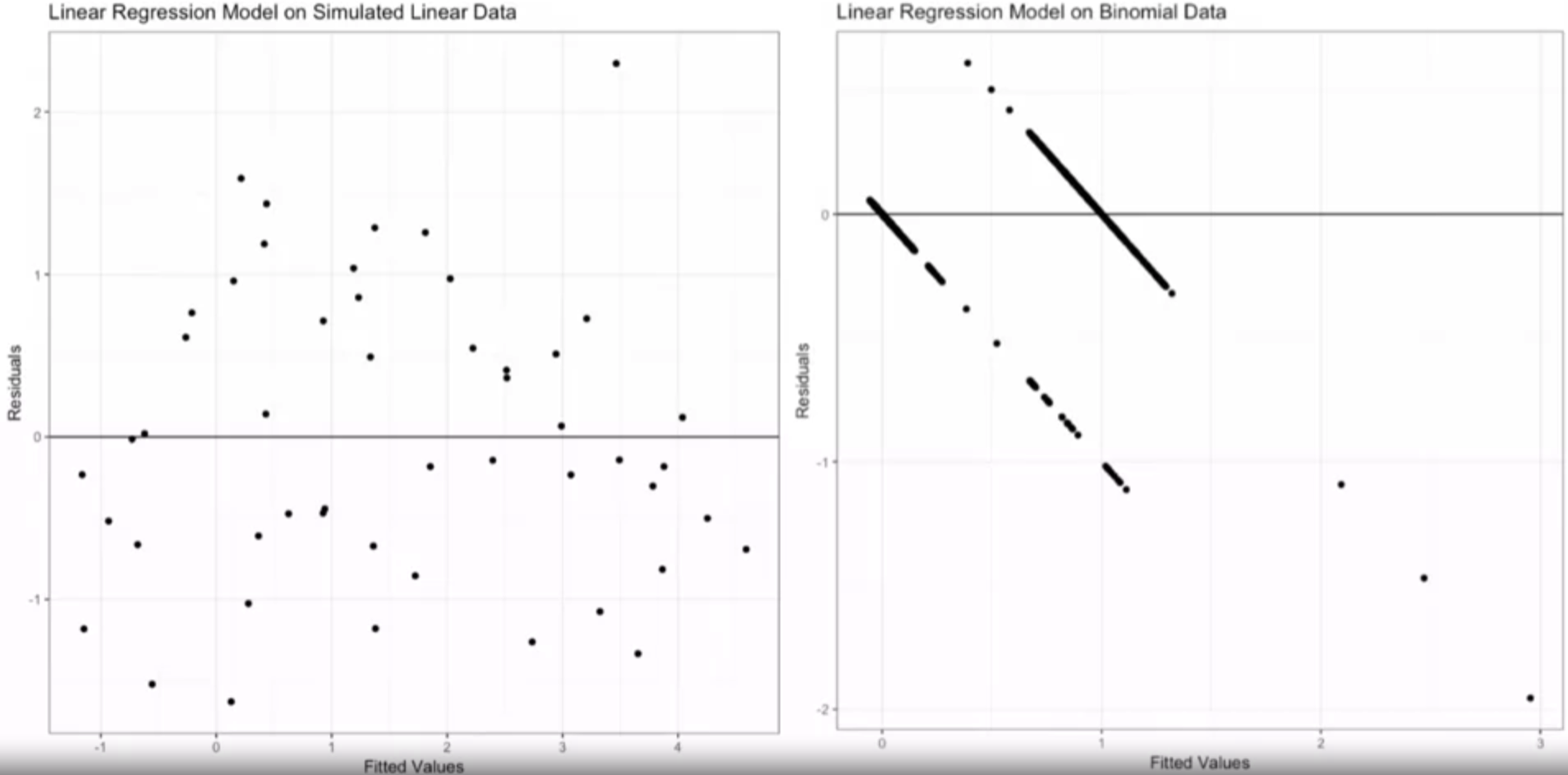
# Week 1: Introduction to Generalized Linear Models through Binomial Regression

## Introduction to Generalized Linear Models

### From Linear Models to Generalized Linear Models

1. Multiple Linear Regression Assumptions
   1. Linearity:
   2. Independence: Errors are uncorrelated ( for all )
   3. Homoscedasticity: variance is constant across all measurments ( for every value of
   4. Normality:
      1. If have normality and independence -> Get uncorrelated
2. Example Binomial Response: Whether a Political Candidate Wins or Not
   1. Response Variable: Politician wins or loses (binary)
   2. Features:
      1. Money spent on campaign
      2. Amount of time spent campaigning
      3. Whether or not the candidate is an incumbant
   3. Let ; then:
      1. Probability mass function:
      2. Mean:
      3. Variance:
   4. Assumption Violations:
      1. Continuity/Normality of the response
      2. Homoscedasticity:
      3. 
      4. Residual plot does not look like what we’d want
   5. Solution:
      1. Use Binomial Regression (in particular logistic regression). Binomial Regression is a specific type of *generalized linear model* (GLM). Other GLMs allow for other types of data (e.g. Poisson regression for count data)
3. GLMs extend the linear model framework to non-normal responses (e.g., counts). We can also extend linar models in other ways:
   1. Nonlinear structure. Nonparametric and semiparametric models allow us to capture data where the predictors are nonlinearly related to the response
   2. Correlation Structure. Mixed effects models accound for data with a grouped, nested, or hierarchical structure (In this case it can be helpful to use time-series models, spatial, and mixed effect models (not covered in this course).

### The Components of a GLM

1. Response distributions not suitable for linear regression:
   1. Examples:
      1. Binomial
      2. Poisson
      3. Multinomial
      4. Gamma, etc.
   2. When we say “model the response,” we mean “model the mean or some function of the mean”
   3. For other generalized linear models:
2. 3 Components of a GLM.
   1. Definitions. Let:
      1. be a set of predictors (p is # of predictors in a model)
      2. be a response
      3. be a vector of parameters
   2. Components:
      1. Random component: The response conditioned on the predictors must be a random variable from the exponential family of distributions and this is not to be confused with the exponential distribution itself
      2. Systematic component: Linear combination of predictor methods and parameters
      3. Link function: Related the random and systematic components together. In particular, it relates the mean of the response, the random component, to the linear systematic component
         1. Systematic Component:
         2. Describes how the mean response, is linked to
         3. In case of normal distribution GLM reduces to standard linear regression
         4. For GLM, we think the explanation connecting them is linear but requires a function.

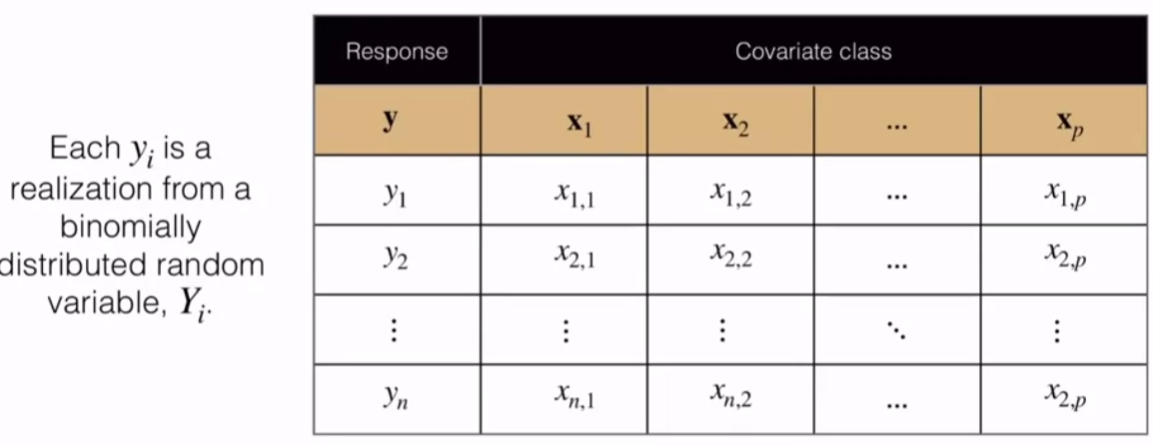
### The Exponential Family of Distributions

1. Definition: is a random variable from the exponential family if the distribution (pdf or pmf) can be written as:

   2. Dispersion parameter; represents scale
   3. Canonical parameter; represents location of distribution
2. Example: Show that the binomial distribution is a member of the exponential family
   1. ; where
   2. : implies that
3. Mean and Variance of Exponential Family:
4. Takeaways:
   1. Response must be from the exponential family
   2. Why? The canonical parameter plays a role (more details to come)
   3. Many of the common distributions e.g. binomial, Poisson, normal, exponential are all from th exponential family

## Binomial Regression

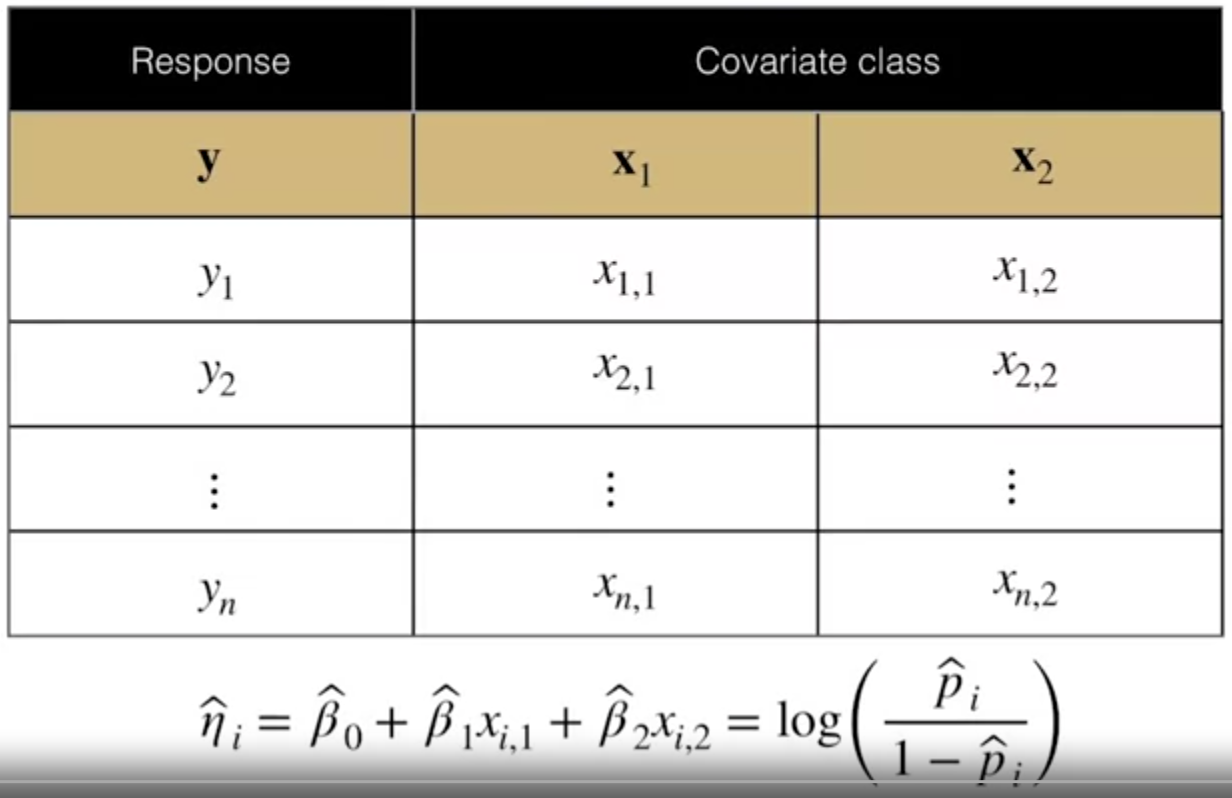
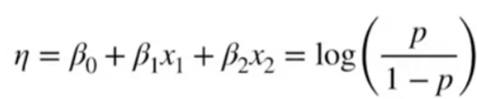
### Introduction to Binomial Regression

1. Recall 3 components:
   1. Random – response
   2. Systematic – set of predictors and parameters
   3. Link function
   4. Dataframe visualization
   5. For binomial; the response vector is expected to be a vector of realizations from the binomial distribution. Generally, each measurement of the response could be from a different binomial distribution (
   6. Particulaly popular: This is then a Bernoulli random variable. This is the case of a binary classification problem and referred to as logistic regression
2. Random component. Each is a realization of
3. Goals:
   1. Predict or explain Using a covariate class
   2. Predict or explain using a covariate class
4. Systematic Component:
5. Link function, , describes how the mean response is linked to . For binomial:
   1. Logit (most popular/default)
      1. Note: inverse of g exists between 0 and 1
   2. Probit
      1. ; inverse of the cdf of the standard 0,1 distribution; inverse of this g is also between 0 and 1

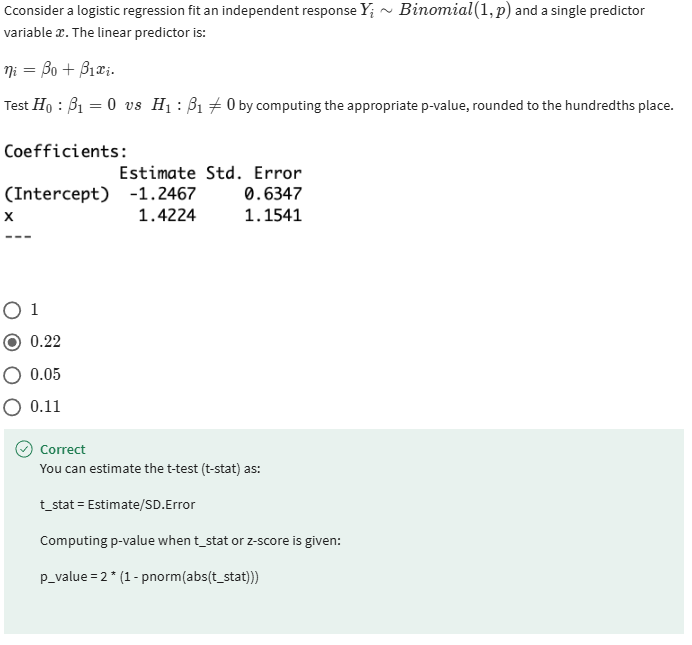
### Binomial Regression Parameter Estimation

1. Parameter estimation – maximum liklihood estimation
   1. Marginal pmf:
      1. Binomial distribution f
   2. Joint pmf:
      1. with independence assumption can multiply these together
   3. Likelihood function:
      1. Joint pmf interpretted as function of the parameters with the data fixed
   4. Log-likelihood function
   5. Maximize
      1. get a parameter estimate
      2. Since, the parameters show up with a non-linear relation; we cannot use an analytical method to solve this -> an iterative method is used

### Interpretation of Binomial Regression

1. 
2. **Definition:** Let event have probability of occurrence. Then the odds in favor of is defined as
3. **Example:** If a biased coin is such that and , then the odds of event the coin lands on heads twice in two flips is:
   1. Since >1, more probable than not of occurring
4. Binomial/Logistic Regression
   1. Interpretation: Note that the logistic regression model is formulated in terms of odds:
   2. So the parameters of the model can be interpreted as follows:
      1. represents the log odds of success when all predictors are equal to 0
      2. with all other predictors held constant increases the log-odds of success by . Or, a unit increase in with all other predictors held constant increases the odds of success by

## Binomial Regression and Goodness of Fit



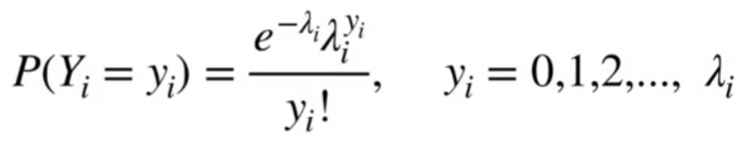
# Week 2: Models for Count Data

## Poisson Regression Basics

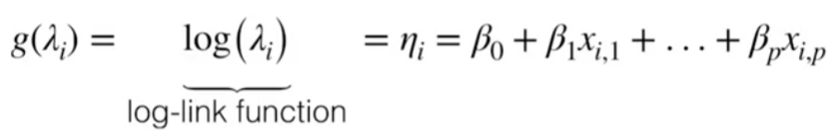
### Poisson Regression: A New Model for Count Data

1. General Scenario:
   1. 
   2. Given a Dataframe where each is a realization from a Poisson distributed random variable
   3. : is a count statistic
   4. : the rate parameter; potentially different for every measurement
   5. Poisson regression is really a way to try to model this rate parameter based on other information
2. Random Component:
   1. Conditioned on the predictors:
   2. Goal: predict or explain using a covariate class,
   3. Mean:
   4. Variance:
   5. Canonical Parameter:
3. Systematic Component
   1. No reason to think that would be positive, which is required of a count/rate (link function solves this)
4. Link Function:
   1. , describes how the mean response is linked to
5. Example: Rate response and the offset term.
   1. Construct a model to predict the number of times an individual would be admitted to a hospital. The covariant class, might include things like age, gender, other health conditions, heart condition, diabetes etc.
   2. Response measurement: length of observation should matter. Can modify model to include the exposure (i.e. length of time the individuals were observed)
   3. Herein: the mean
   4. Using the log-link function, we see that:
   5. Offset Term: (Can use this by adding to the opposite side
   6. i.e. ; family=’poisson’ here

### Poisson Regression Parameter Estimation

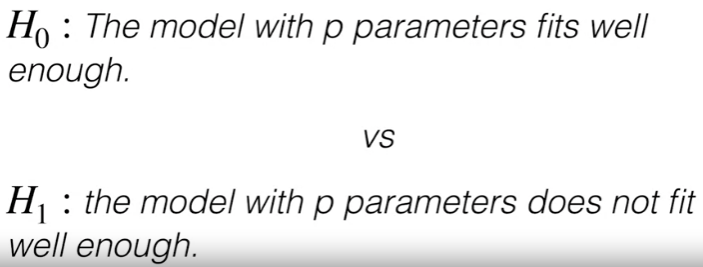
1. Marginal pmf
   1. 
2. Joint pmf
3. Likelihood function
4. Log-likelihood function
   1. Once again, this is a non-linear relation which requires numeric solving
5. Maximize!

### Poisson Regression on Real Data in R

1. 
2. can be interpreted as the mean of the response when each predictor is set to zero
3. can be interpreted as the multiplicative increase in the mean of the response for a one unit increase in holding all other predictors constant

## Poisson Regression Inference and Goodness of Fit

### Goodness of Fit for Poisson Regression I

1. Goodness of Fit definitions:
   1. Deviance:
      1. Smaller deviance means better fit
   2. Null Deviance: deviance for the model with just an intercept term (
   3. Saturated Deviance
      1. Each sample gets the best possible parameter
   4. Residual Deviance
      1. : deviance of model fit with p parameters
      2. Residual deviance follows a chi-squared distribution
      3. Under the hypothesis that the p model fits the data, the degrees of freedom for the chi-squared here will be the degrees of freedom for the p model minus the degrees of freedom for the saturated model. Think about what the degrees of freedom are for that saturated model, they'll actually be zero. Here we'll see that this quantity will have n minus p plus one degrees of freedom.
      4. 
   5. Deviance residuals
      1. Through some maneuvering can get something that can be squared and summed to act like residuals
      2. Knowing that we construct Chi-Squared distributions from the sums of squares of standard normals, it turns out that under the right assumptions, this quantity will be approximately normal 0, 1. That's not always the case, but under the assumption that the model fits well and that we have a high number of counts

### Goodness of Fit for Poisson Regression II

1. Pearson’s

   2. : Number of observations
   3. : Expected mean number of observations under some assumed distribution
   4. A large value is evidence against the null that the model fit is sufficient
   5. Interpretation: can interpret as the sum over square standardized residuals
   6. Here the residual is

### Overdispersion

1. Reminder of facts about the Poisson distribution:
2. Definitions:
   1. Count response data are **overdispersed** when the variance of the response is greater than the mean:
   2. **Underdispersion** occurs when the mean is greater than the variance of the response
3. Real overdispersion vs Apparent overdispersion
   1. Real Overdispersion:
      1. Zero inflation: the real response contains far more 0s than what one would expect under the Poisson distribution. To help address this, one might consider (Common in # of insurance claims)
         1. Hurdle Models
         2. Mixture Models
      2. Correlation: through space or time
         1. Complex correlation structure arising from hierarchical, clustered or spatial data
         2. Ex: researchers might be interested in modeling the # of cases of a particular diseas in a particular country. Without including spatial information in the model, a standard Poisson regression model would likely show evidence of overdispersion.
   2. Apparent Overdispersion
      1. Outliers
         1. Included in dataset when shouldn’t be
      2. Missing predictors
         1. When additional predictors are available, does adding them help address the overdispersion
      3. Bad link function
4. How to detect overdispersion
   1. Generalize to 2-parameter model
   2. Statistically significant test would be evidence of overdispersion
5. Solutions
   1. Quasilikelihood
      1. Adjust the standard errors by multiplying the variance by the above
      2. Standard error for :
      3. So this method is called Quasilikelihood, because the addition of the dispersion parameter means that the model is strictly speaking no longer Poisson, but it's using some other fitting procedure. So I think it's important to note that the presence of overdispersion does not impact the validity of maximum likelihood estimation, as a method for estimating parameters. Infact the MLS themselves are unchanged, whether we account for overdispersion using Quasilikelihood methods or not. So what does change when accounting for overdispersion is the estimated variance of the parameter estimates.
      4. typically, these variances will be severely underestimated if over dispersion is not accounted for. And that can result in statistically insignificant predictors being treated as statistically significant.
   2. Negative binomial regression
      1. Introduces a second paramter for flexibility in accounting for overdispersion
      2. It’s a true likelihood based method as opposed to Quasilikelihood method

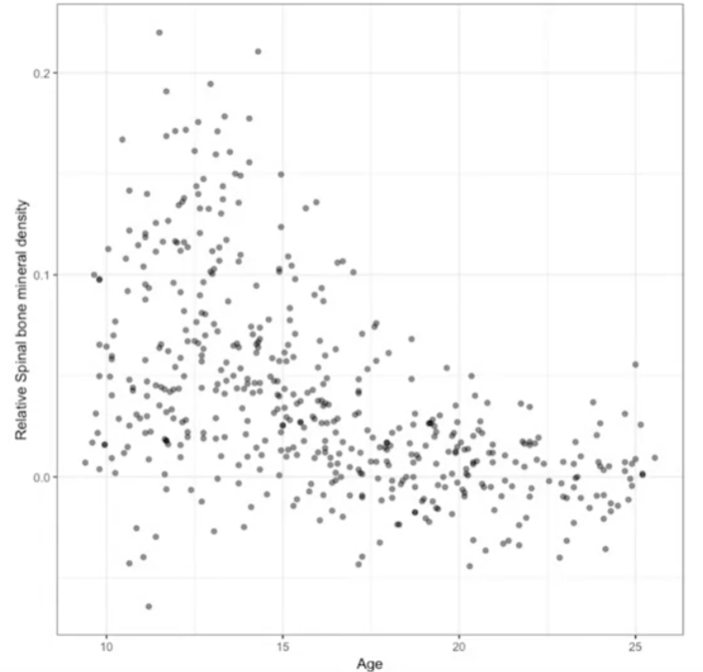
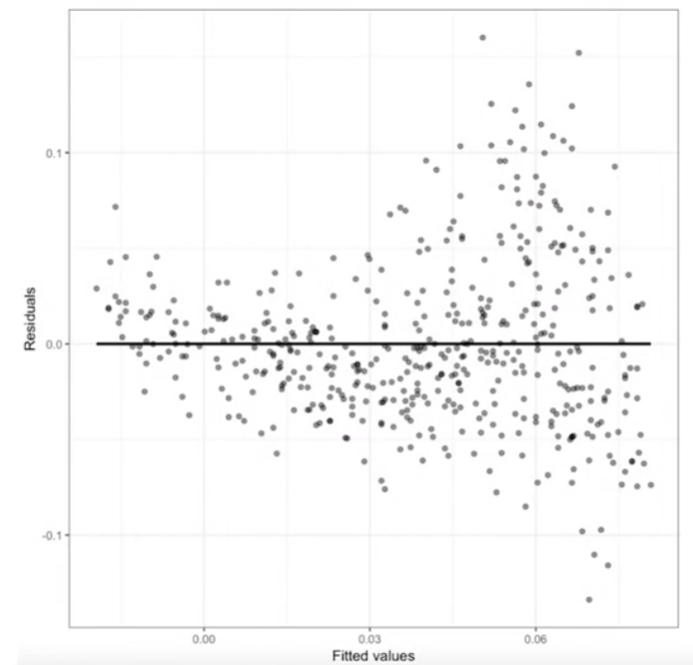
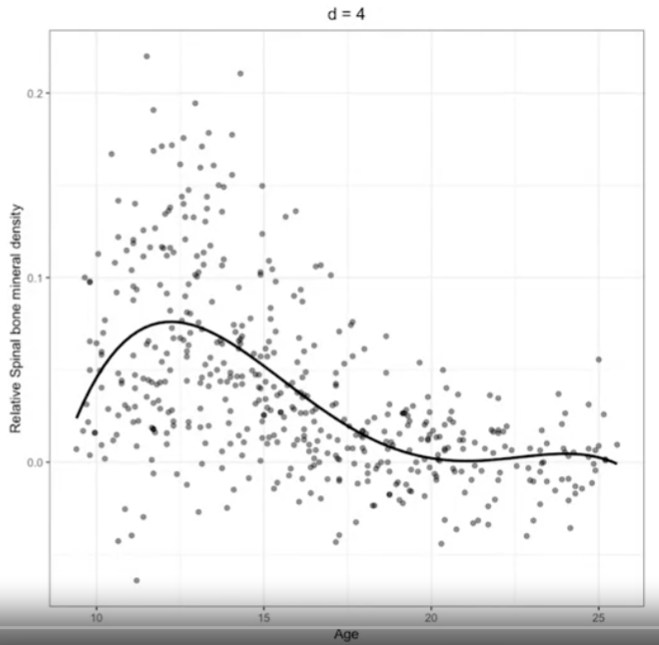
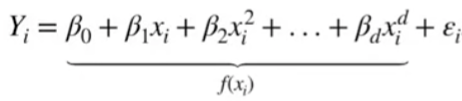
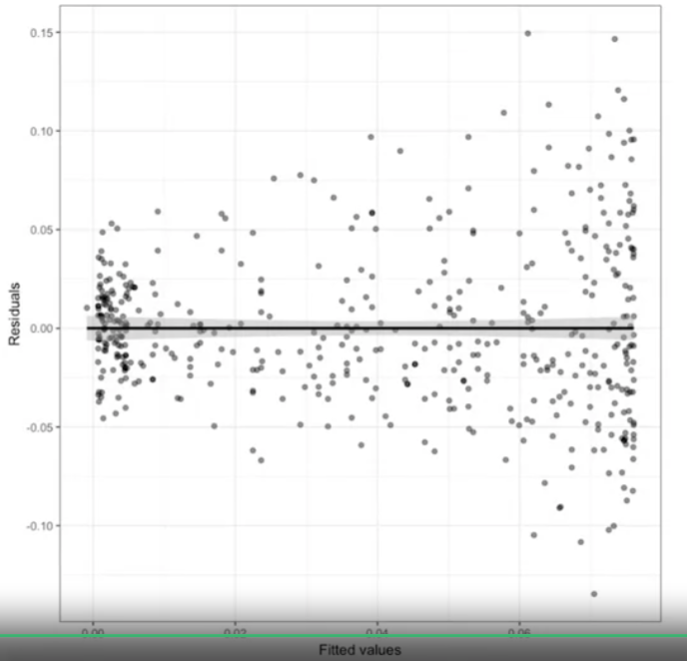
# Week 3: Introduction to Nonparametric Regression

## Nonparametric Regression: Theory

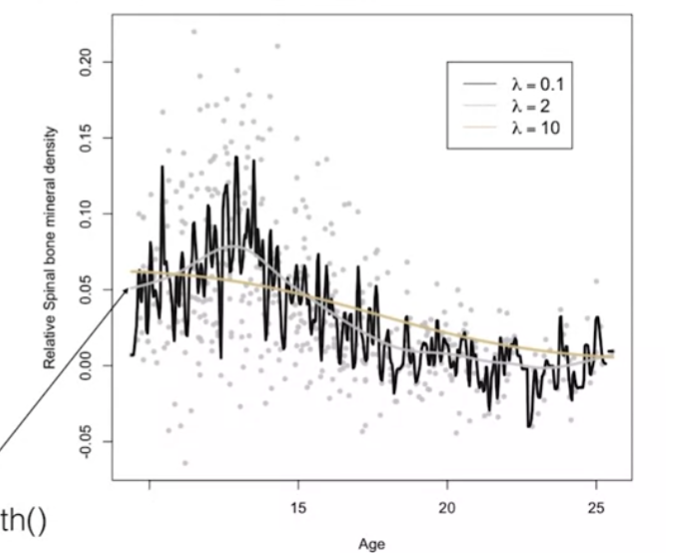
### Introduction to Nonparametric Regression Models

1. Motivation/ Definitions:
   1. A **parametric**  statistical model is a family of probability distribution with a finite set of parameters
      1. Example: Normal regression model:
         1. Response (***Y***), is normally distributed with a mean ( and variance covariance matrix (
         2. The mean has p+1 parameters, and the variance/covariance matrix has a single parameter
   2. A **nonparametric** statistical model is a family of probability distributions with infinitely many parameters
      1. Example:
      2. is an arbitrary function where .
      3. Since the function is arbitrary on this interval, no finite set of parameters could specify its form. Would need an infinite number of parameters to specify exactly what this function would be. To truly specify, we would need infinite data.
   3. A relatively general form of a statistical modeling problem is:
      1. In normal regression:
      2. In Poisson regression:
   4. In nonparametric, we try to learn f
2. How do we attempt to learn ?
   1. Assume it comes from smooth family of functions. In this case, the set of potential fits to the data is much larger than the parametric approach
3. Advantages and Disadvantages to nonparametric approach
   1. Advantages:
      1. Flexibility
      2. Fewer distributional assumptions
   2. Disadvantages
      1. Less efficient when structure of the relationship is available
      2. Interpretation difficulties

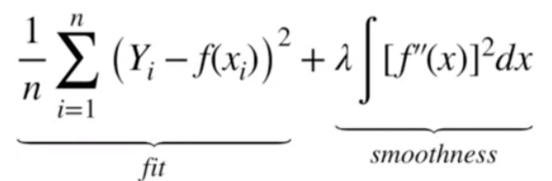
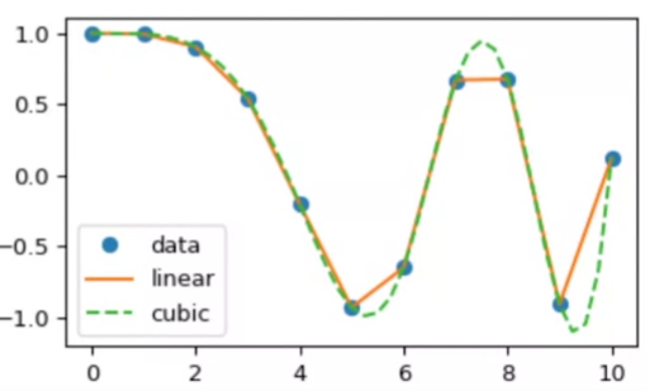
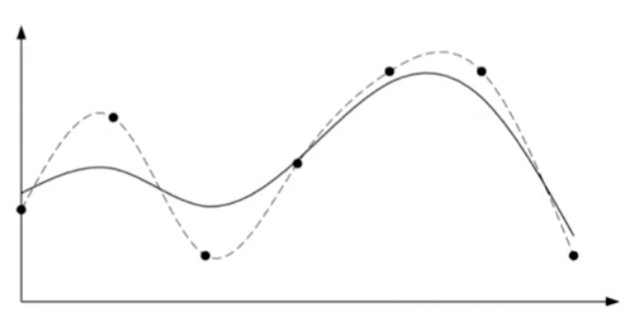
### Motivating Kernel Estimators

1. Scatter Plot relationship between response and Age
   1. 
   2. Seems to be some underlying nonlinear structure
2. Residual Plot:
   1. 
   2. By looking at the residual plot, we see that there is a non-constant variance which implies a nonlinear structure that’s not being captured and that’s showing up in the data
3. Attempt to add polynomial terms:
   1. 
   2. 
   3. How do we choose d? Heuristics:
      1. Option A: Start with d=1 and add terms until they are no longer statistically significant
      2. Option B: Start with large d and eliminate terms that are not statistically significant, starting with highest order terms
      3. Note: generally a bad idea to eliminate a lower d than
   4. 
   5. Even still, this shows signs of misfit and non-constant variance
   6. As one adds predictors/features, the process of selecting d becomes incrementally messy and the heuristics don’t behave consistently.
4. To address this, **kernel smoothers** are used as a process to define the underlying function
   1. Nonparametric method for choosing the nonlinear structure that best fits the data

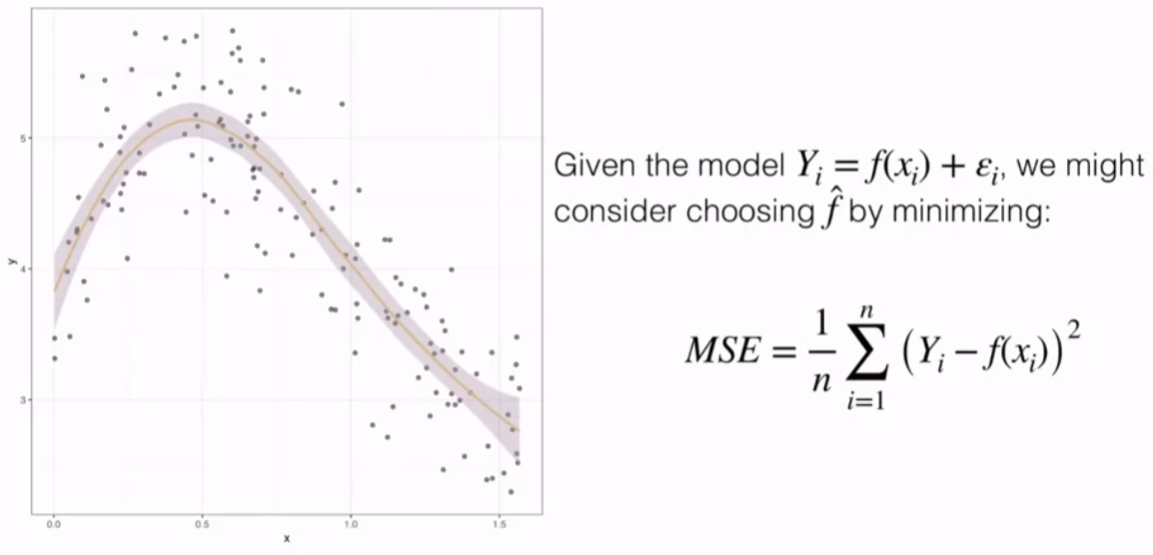
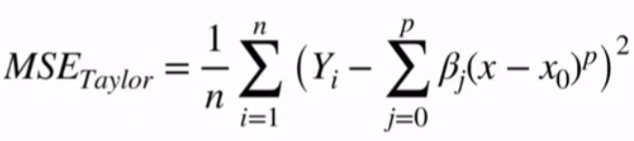
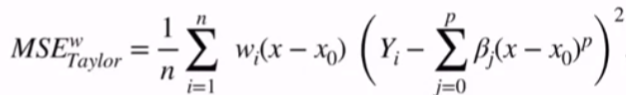
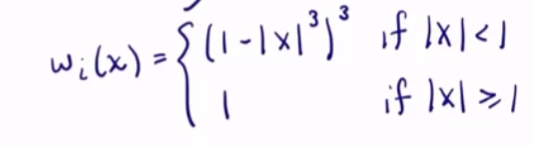
### Kernel Estimators

1. Definitions:
   1. Kernel estimate can be understood as simple, weighted local moving average to the response
   2. Model:
      1. Assume,
      2. is not specified and is to be estimated from the data
      3. ; where these are the weights
   3. Kernel (): Nonnegative, real-valued function such that for all values of x and (normalization)
      1. These are defining features of pdfs
      2. Commonly used kernels:
         1. Uniform/rectangular:
         2. Gaussian/Normal:
            1. Note this is not local, but decays very quickly
         3. Epanechnikov:
      3. Estimation procedure is not too sensitive to the choice of kernel
   4. Bandwidth (:
      1. 
      2. The bandwidth or smoothing parameter controls the smoothness or the bumpiness of the estimate
      3. Larger smoother the fit
      4. How do we choose?
         1. Rougher fits are less plausible, but oversmoothing
         2. Try to find “Least smooth fit that does not show any implausible fluctuations”
         3. Here we are trying to find something like the grey curve
         4. Note: if will be used in making predictions of future values, the choice of lambda is consequential. Faraway goes on to defend the subjective approach to choosing lambda: “If the method of selecting the amount of smoothing seems disturbingly subjective, we should also understand that selecting a family of parametric models, for example, a standard normal regression model for the same data would also involve a great deal of subjective choice”
      5. There are some automated ways of selecting . For instance, cross-validation procedure. But sometimes this doesn’t result in a plausible lambda.

### Smoothing Splines

1. Background/Motivation:
   1. Given the model , we might consider choosing by minimizing
      1. Note without specifying the function, , the solution by minimizing MSE would be . Here, we would just be interpolating and extrapolating instead of finding some sort of mean curvature
   2. Solution: minimize:
      1. This offers a roughness penalty
      2. smoothing parameter
      3. 
      4. Small squared second derivative provides low curvature. controls the tradeoff
2. Definitions:
   1. 
   2. **Spline**: piecewise function, where each segment is a polynomial
   3. **Cubic spline:** is a spline where the segment polynomials are each of degree three.
   4. Splines are meant to be continuous and have continuous derivatives
   5. **Smoothing spline**: spline designed to balance fit with smoothness
      1. 
      2. Dotted – cubic spline
      3. Solid – cubic smoothing spline
3. Since we know the form of the solution, the estimation problem is reduced to a parametric problem of estimating coefficients of the piecewise polynomials

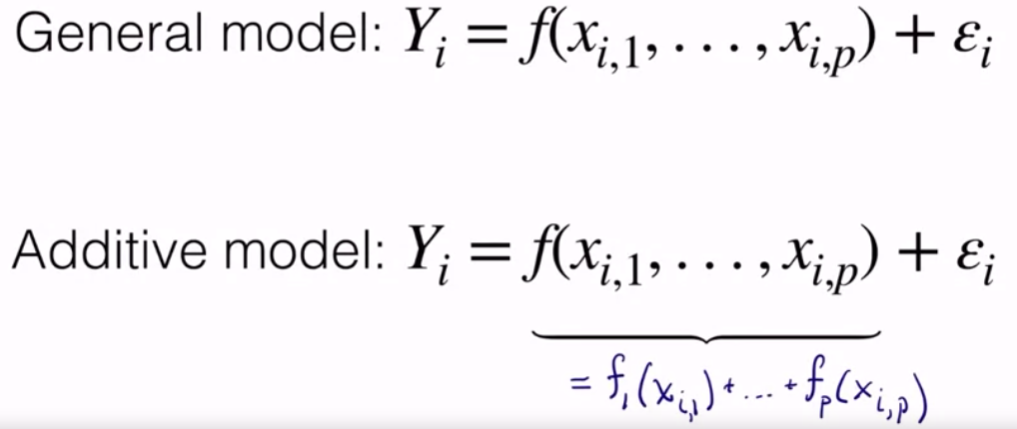
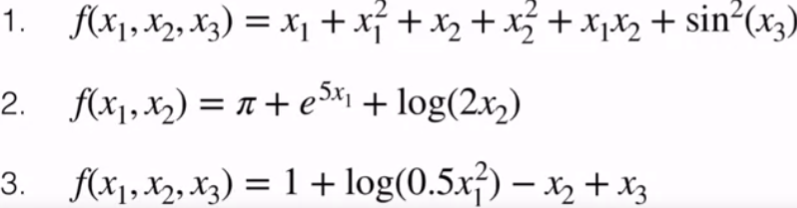
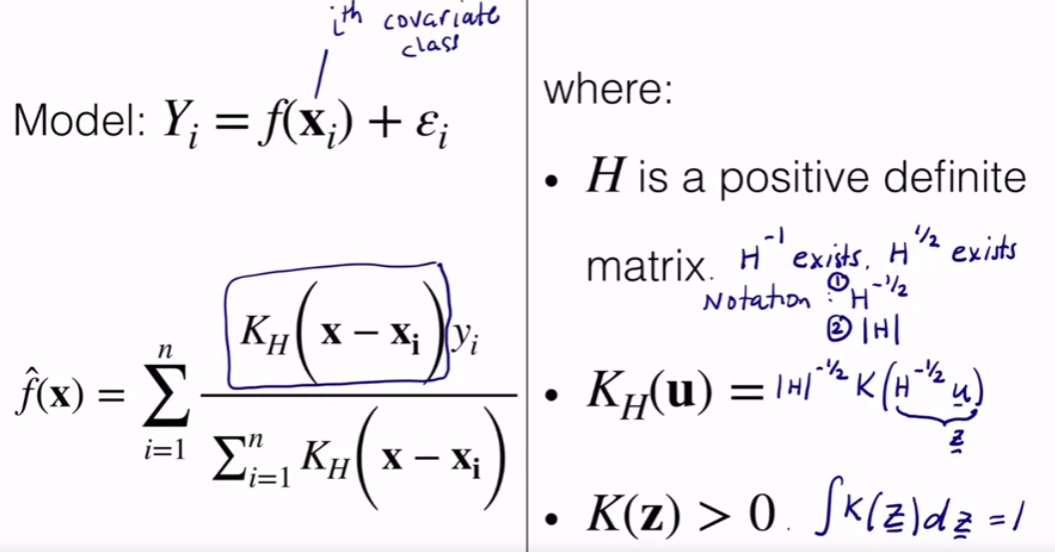
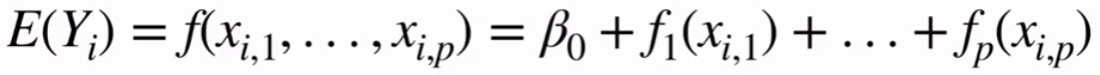
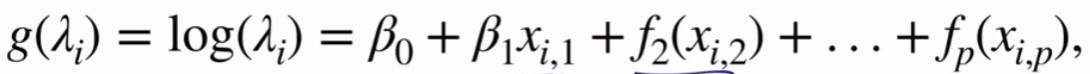
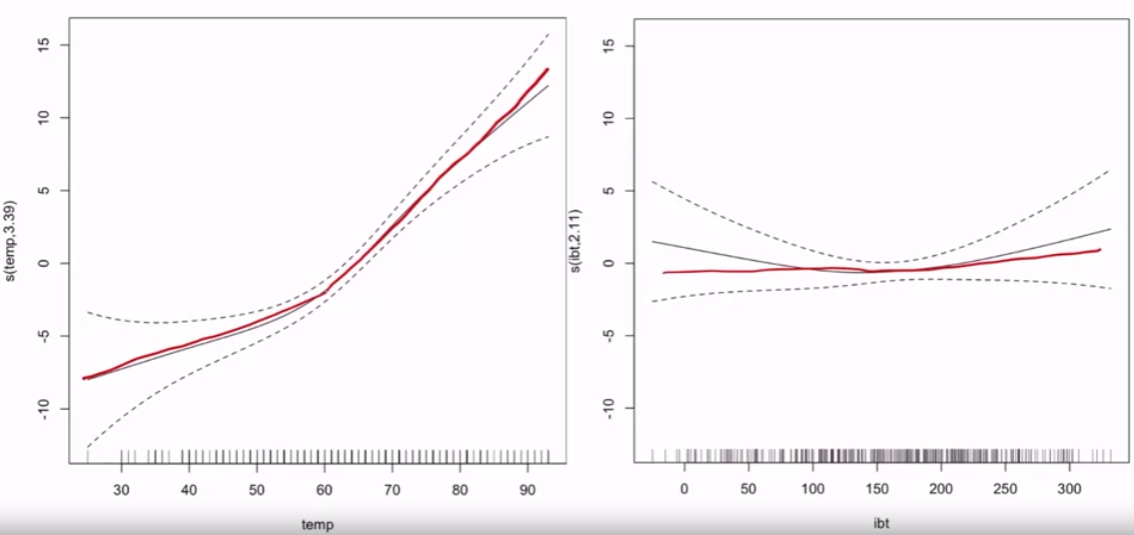
### Loess: Locally Estimated Scatterplot Smoothing

1. Combines parametric linear regression techniques with local fitting ideas that we learned about in Kernel estimation
2. General approach:
   1. 
   2. Need to put constraints on to reduce the number of parameters, thereby helping address
3. Can construct a Taylor approximation/interpolation of the data via:
4. For Loess, often doing this for p=1, local linear/low order around the window and move across the x domain
5. Can address the issue of using all points instead of those locally proximal to x\_0 with a weight function
   1. 
   2. Common weight function: 
6. Advantages:
   1. Flexible fit, can account for nonlinear where the nonlinear form is not known
   2. Relatively simple to implement
   3. Uncertainty quantification, since it relies on theory for weighted regression in the same way one would with linear regression
7. Disadvantages:
   1. Can be computationally expensive
   2. Interpretation can be more difficult
   3. Requires relatively large, densely camples data

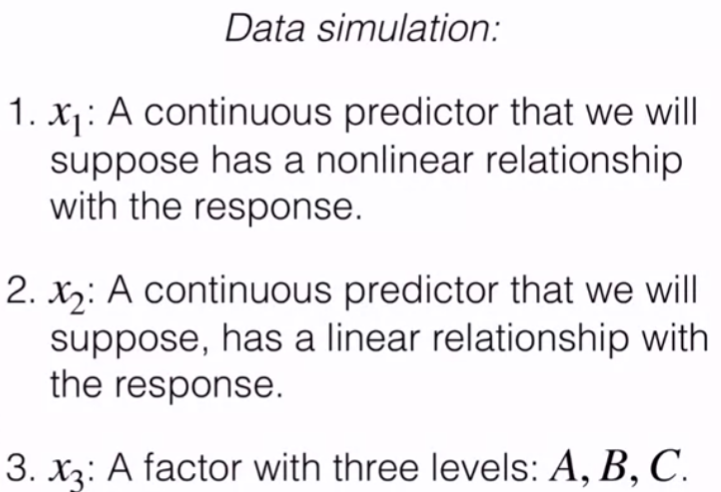
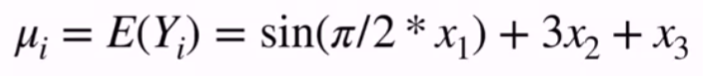
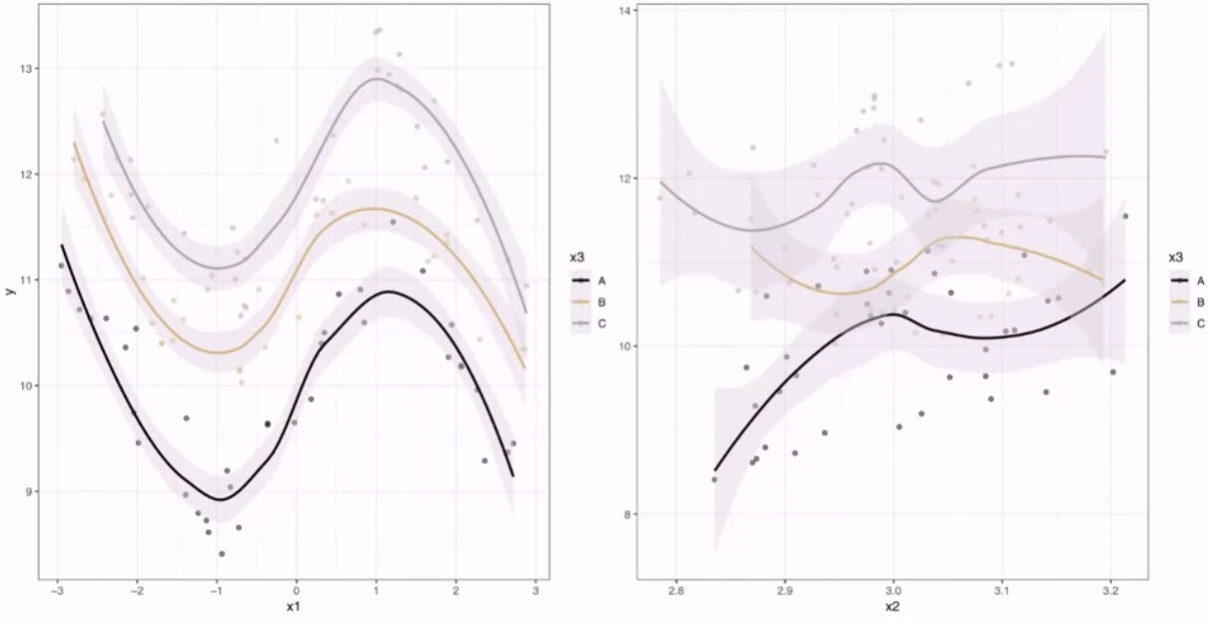
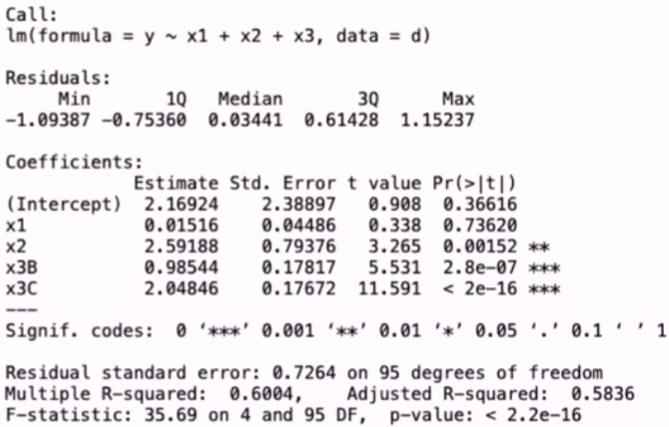
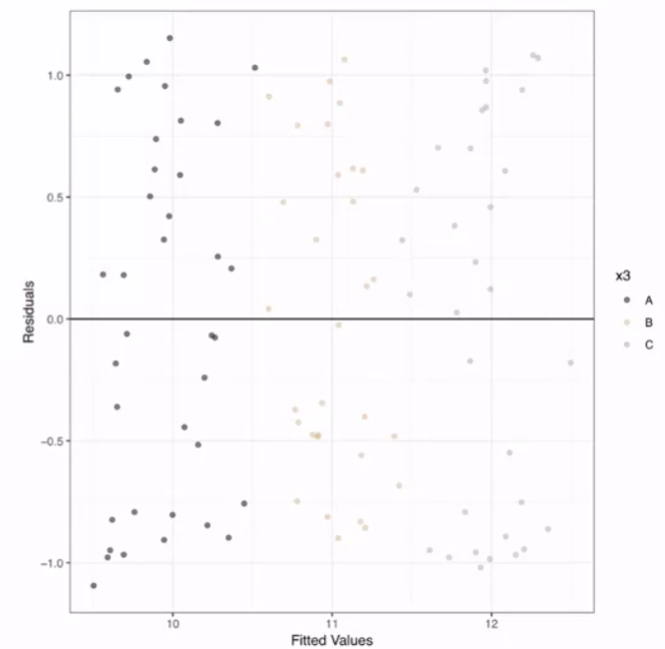
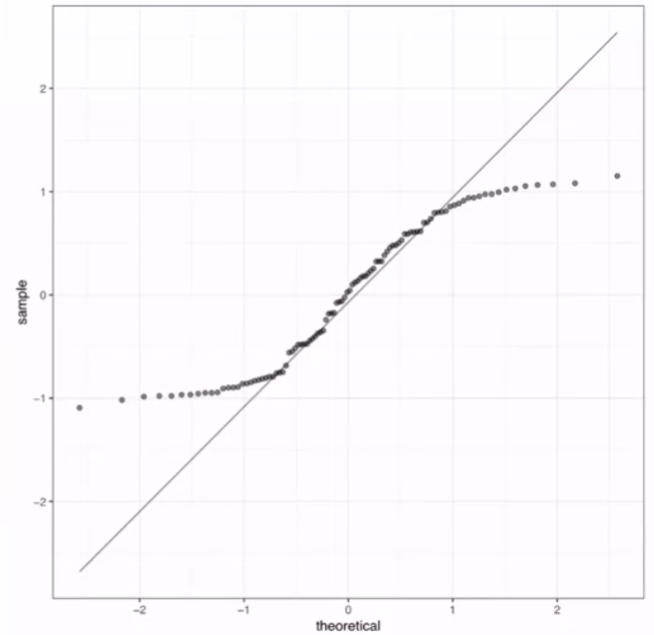
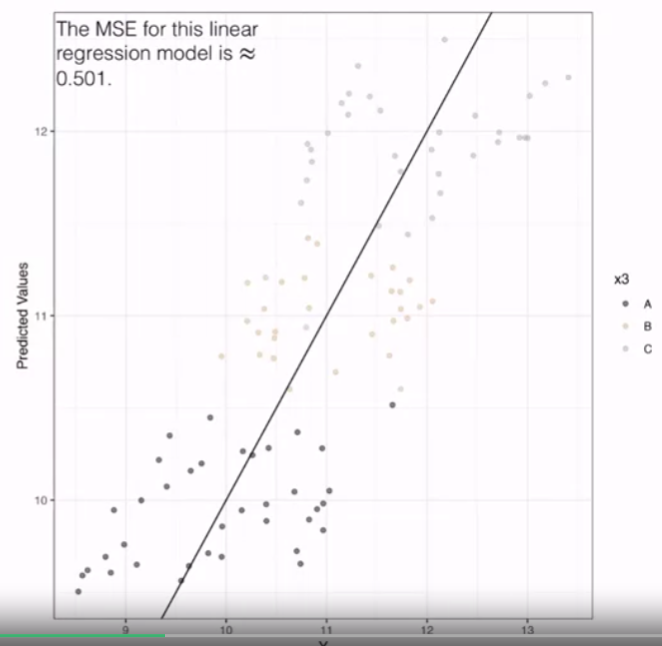
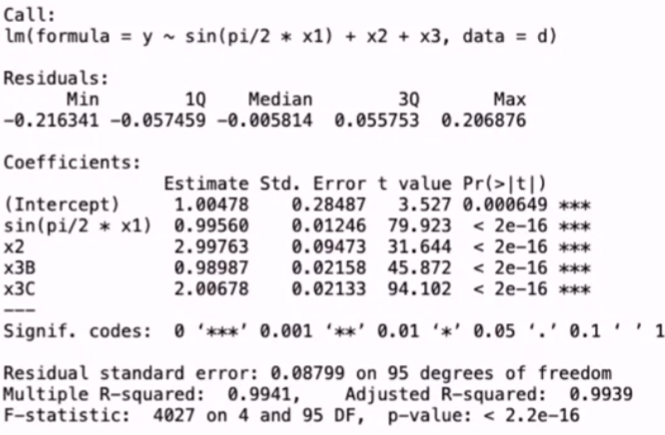
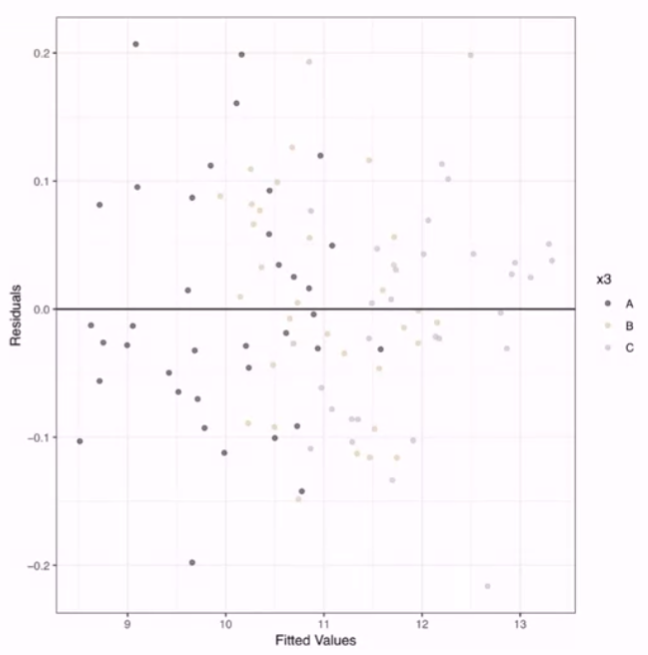
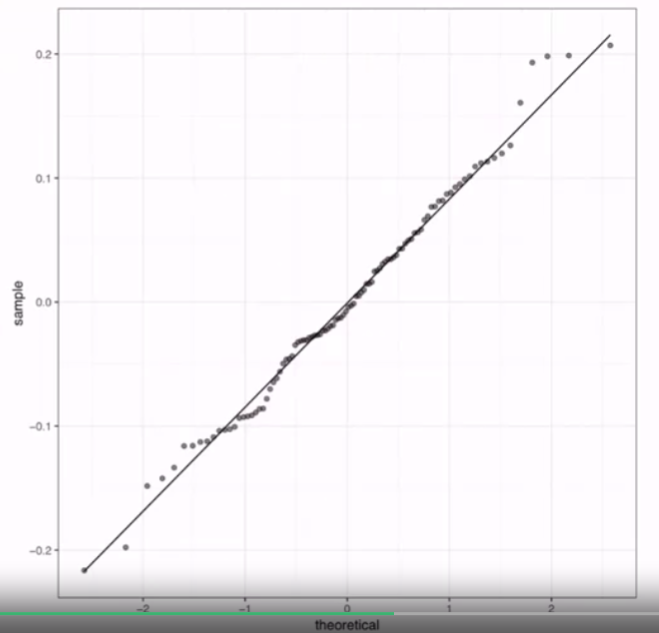
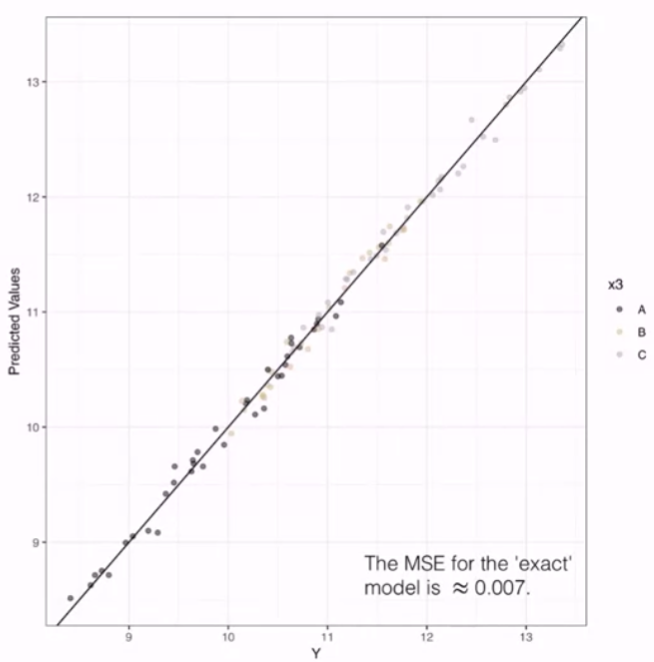
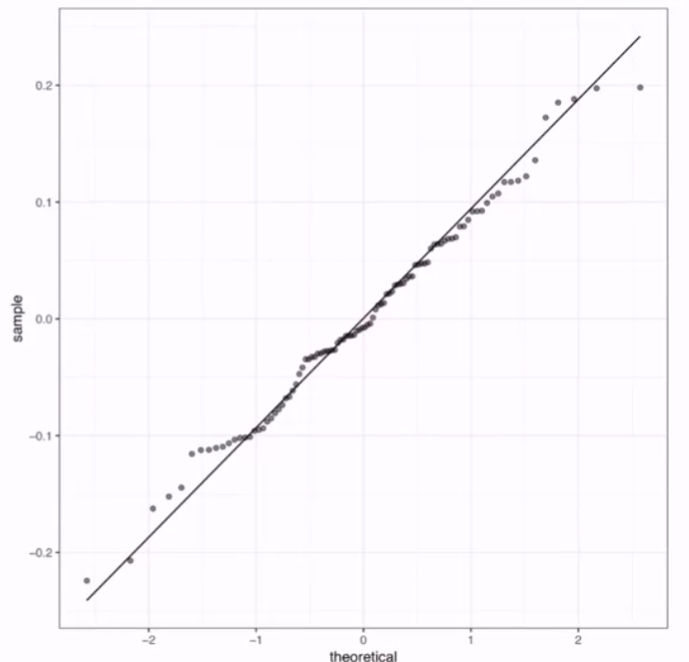
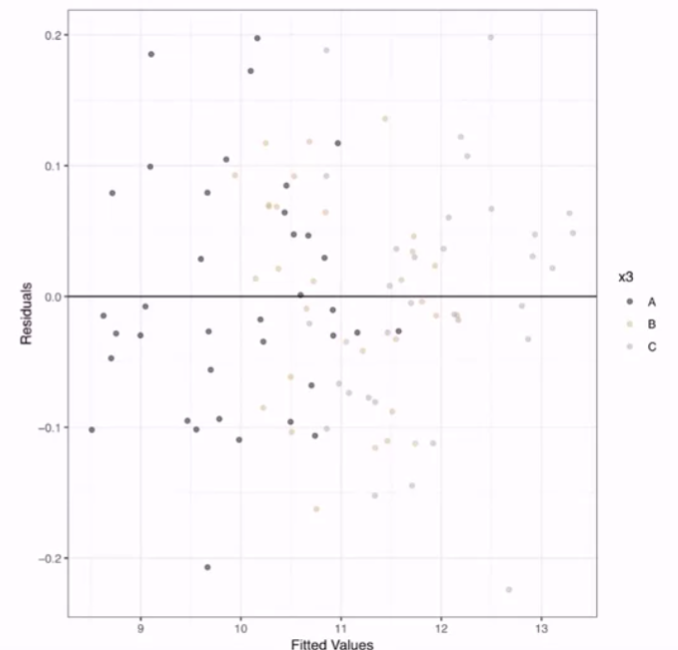
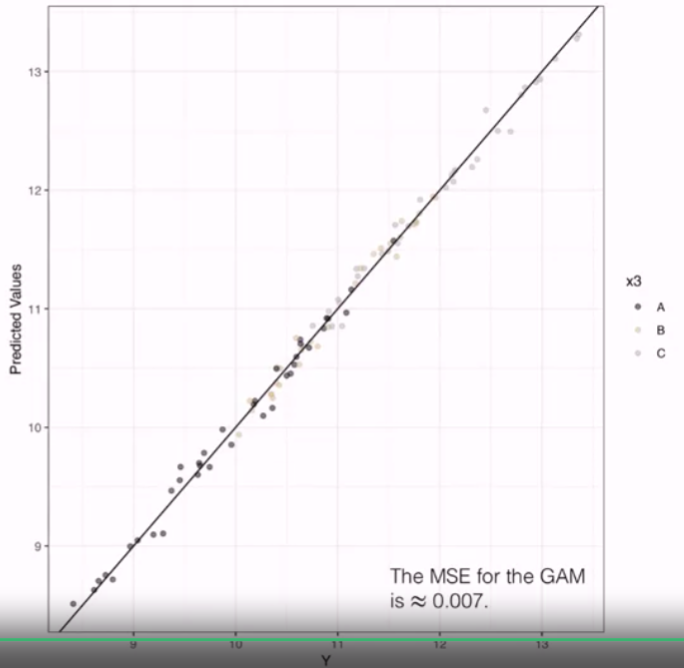
# Week 4: Introduction to Generalized Additive Models

## Generalized Additive Models: Basics

### Motivating Generalized Additive Models

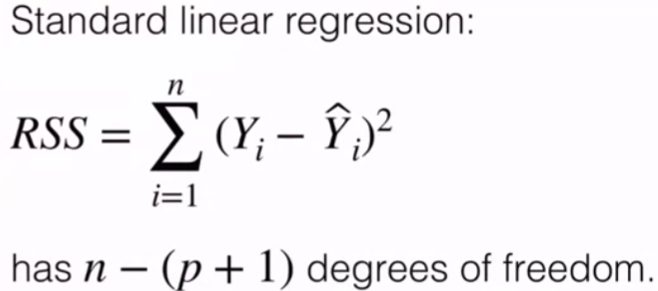
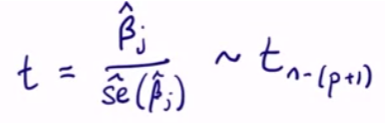
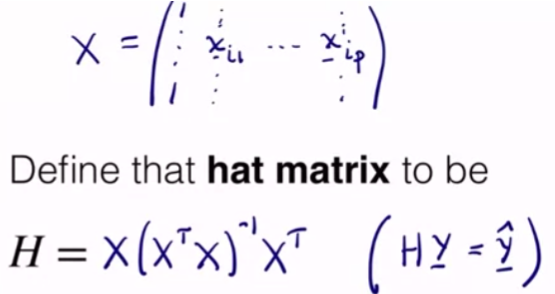
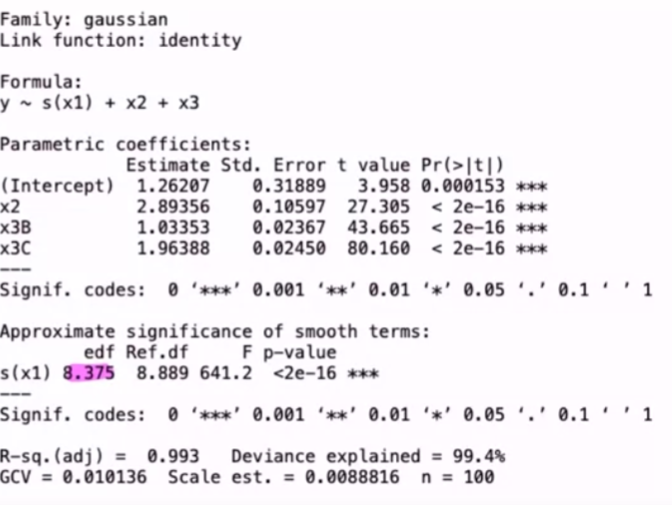
1. Background
   1. 
   2. Examples:
      1. 
      2. (1) is not additive as it cannot separate the product
      3. (2) and (3) are additive
2. How to fit non-parametric additive model:
   1. Definition
      1. 
      2. Can we simple to constrict it to be an additive model
   2. Additive Formulation:
      1. 
      2. Function of individual parameters
      3. Many options, but a fair amount of restrictions
      4. Here, we are not imposing a transformation term, can be any
      5. Use non-parametric to learn each
      6. If we know the response with a given predicter is linear, we can explicitly state it as such. IF is categorical feature, we can ecode it as such
      7. Can consider adding interaction terms, but it does add complexity
      8. Can be easily extended to non-linear methods
      9. 
   3. Example:
      1. 
      2. Since linear can fit through ibt confidence window, it’s reasonable to attempt fitting with a linear term
      3. For temp, there is a clear change in slope, which suggests using a nonlinear functional formulation

### Generalized Additive Models in R

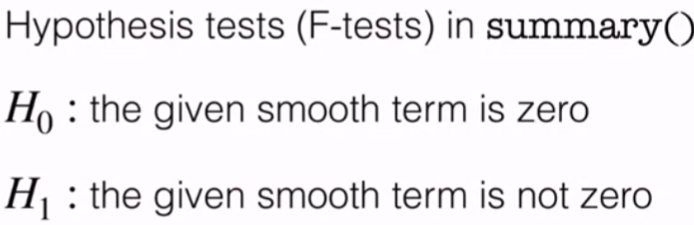
1. 
2. Construct the response formulation for the simulated data to be instructional:
   1. 
   2. Normal noise component added
3. For each model, fit a loess model (univariate smoothing situation)
   1. 
      1. Plotted for each level of x3 (A, B, and C) (factor variable/multiclass variable (3 values)
      2. Confidence bands for these models is pretty wide, for x2, no real evidence for a linear relationship. Need a multivariate model to build the prediction
   2. Modeled with linear:
      1. 
      2. Does not fit well, it is not picking up the fit well p-value for x1 and x2 is high
   3. Residual Plot
      1. 
         1. No clear issue displayes in the residuals plot
      2. QQ Plot
         1. 
         2. Strong deviation from normality in qq plot at the tails, suggesting need to account for nonlinear relation
      3. Predicted vs Actual:
         1. 
         2. Not making great predictions,
         3. Perhaps underpredict for low values and overpredict for high values
   4. “Cheat” and model with actual relationship to see how the exact parametric model does
      1. 
      2. All predicted coefficients are close to true value
      3. 
      4. 
      5. 
   5. GAM (Generalized Additive Model)
      1. 
      2. Fitting a smoothing function for x1 variable
      3. Fitting parametrically linearly for x2 and x3 based on prior exploration
      4. Family allows us to specify the link (gaussian, binomial, poisson), depending on the response variable
      5. 
      6. 
      7. GAM does much closer to true model and performs well on all three evaluation tests

## Generalized Additive Models: Inference and Data Analysis

### Inference with Generalized Additive Models: Effective Degrees of Freedom

1. 
   1. For t-test of individual feature or f-test for the model as a whole, rely on the degrees of freedom
   2. T-stat: , belonged to t-dist with (n-(p+1)) dof
2. Is there a dof analog for GAM
   1. No direct analog, but can have a proxy
   2. Need some measure of complexity in the additive model
   3. X is model matrix, can define a hat matrix (H)
      1. 
      2. tr(H) = p+1; if we think about hat matrix get model degrees of freedom
   4. Analog is **effective degrees of freedom (edf)**. Can be gained by calculating the trace of the Hat matrix. These are reported in the output of the test in R
      1. Interpretation: if a smooth term has an edf “close to” 1, then that term should enter linearly into the model.
      2. Plot of the smooth is helpful as well
   5. Prior example:
      1. 
      2. Here the edf for the smoothing of the first term is 8.375>>1
      3. Marginal relationship provides further evidence
      4. Overall, not strictly reliable, should be somewhat skeptical, but can be another useful tool to validate

### Inference with Generalized Additive Models: Tests

1. R also provides hypothesis tests, these are rough ways of setting these tests and should be used skeptically as well
2. 
3. If p very small, then smoothing term should be included
4. Deviance explained and
   1. . (residual deviance/ null deviance). Number between 0 and 1, can tell you something about the percentage of variability in the response explained by the model
   2. : adjusted r-squared; version of deviance explained. Adjusted to account for either a small sample size or model complexity (like the number of predictors). Really trying to penalize a bit for more complex models. It’s reasonable to use this stat for some model comparisons across different models