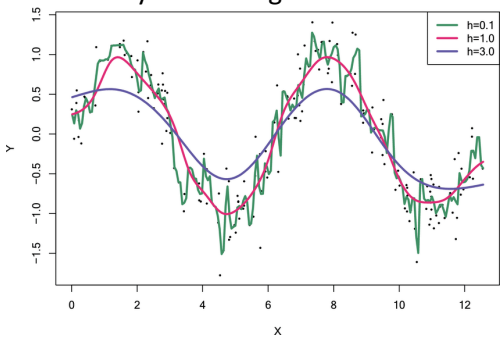
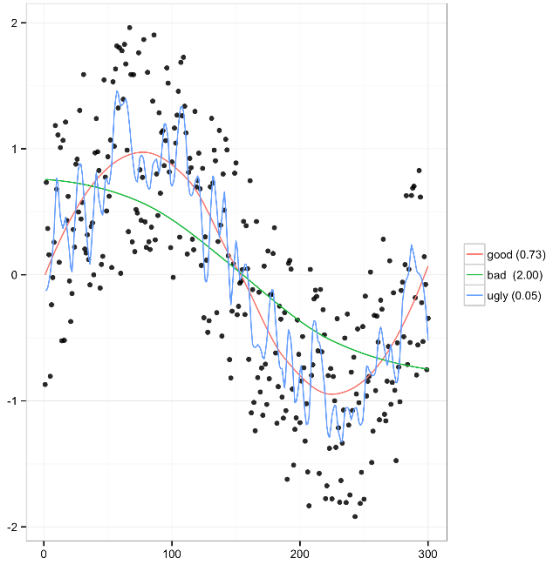
Lecture 16

Nonparametric regression methods are statistical techniques that allow us to estimate the underlying relationship between a dependent variable and one or more independent variables without making any assumptions about the functional form of the relationship. Here are some examples of nonparametric regression methods:

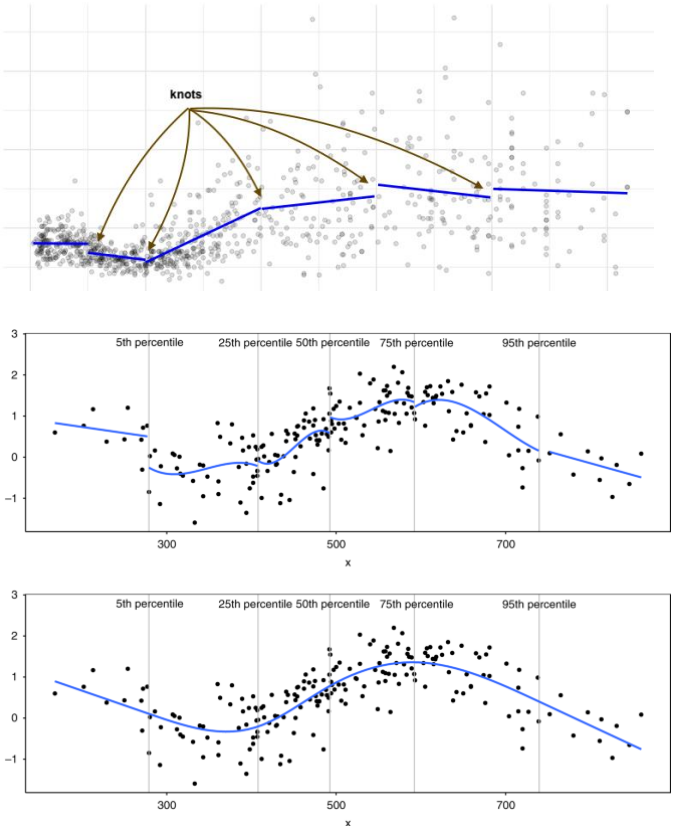
**Kernel regression**: In kernel regression, we estimate the conditional expectation of the dependent variable given the independent variables by smoothing the data with a kernel function.



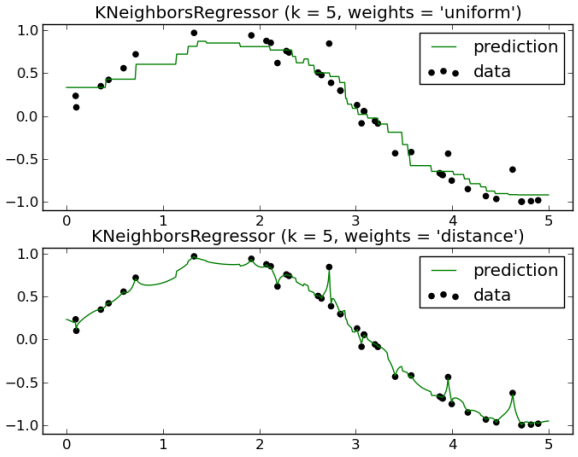

**Local regression**: Local regression, also known as LOESS (locally estimated scatterplot smoothing), involves fitting a smooth curve to the data by locally fitting a polynomial or other smooth function to a subset of the data.



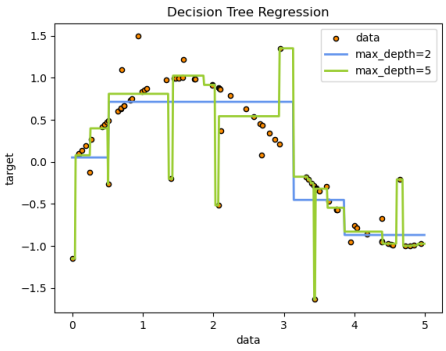

**Splines**: Splines are a flexible class of functions that can be used to fit smooth curves to the data by joining together polynomial functions of lower degree. Typically, linear or cubic splines are used. Penalties can be added to improve smoothing of the curve.



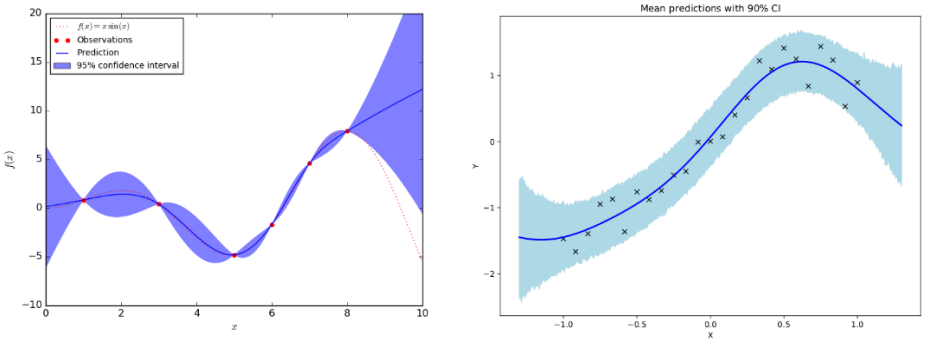

**K-nearest neighbor regression**: In k-nearest neighbor regression, we estimate the value of the dependent variable for a given value of the independent variable by taking the average of the dependent variable values of the k nearest neighbors in the training data.




**Tree-based regression**: Tree-based regression methods involve partitioning the data into smaller subsets based on the values of the independent variables and fitting a regression model to each subset.




**Gaussian processes**: Gaussian processes are a flexible nonparametric method that can be used to model complex relationships between variables by defining a prior distribution over functions and updating it based on the observed data. Models will look a little different if you assume observed points have no error, or if you assume that they do.




We aren't going to look at all of these, but we are going to briefly consider three of them: LOESS, splines, and Gaussian process regression. Besides the ones listed here, there are other kinds (for example, Support Vector Machines (SVMs) are typically used for classification, but there are methods for adapting it to regression, just as KNN and Tree-based regression are similarly adapted from classification or clustering methods.

**LOESS**

LOESS (Locally Estimated Scatterplot Smoothing) regression is a nonparametric regression method that fits a smooth curve to a scatterplot of data by locally fitting a polynomial or other smooth function to a subset of the data. Here's how LOESS regression works:

*Divide the data into subsets*: LOESS regression divides the data into a number of overlapping subsets or windows, each containing a specified proportion of the data.

*Select a degree of polynomial*: LOESS regression fits a polynomial or other smooth function to each subset of data, using a degree of polynomial that is specified by the user.

*Weigh the data*: LOESS regression weights each data point in the subset based on its distance from the point of interest, giving more weight to points that are closer and less weight to points that are farther away.

*Fit the polynomial*: LOESS regression fits the polynomial or other smooth function to the weighted data in each subset, producing a smoothed estimate of the dependent variable for each value of the independent variable.

*Combine the estimates*: LOESS regression combines the estimates from each subset to produce a single smooth curve that fits the entire dataset.

The degree of smoothing in LOESS regression can be adjusted by varying the size of the windows and the degree of the polynomial used to fit the data. By adjusting these parameters, LOESS regression can produce a curve that is more or less smooth, depending on the level of noise in the data and the desired level of precision in the estimate of the underlying function.

LOESS regression is a powerful nonparametric regression method that has several advantages over other regression methods. Here are some of the pros and cons of LOESS regression:

*Pros*:
*Flexibility*: LOESS regression can fit a wide variety of nonlinear relationships between the dependent and independent variables. It is especially useful when the relationship is complex and cannot be easily modeled using a linear regression or a simple parametric model.

*Robustness*: LOESS regression is less sensitive to outliers than some other regression methods, such as linear regression, because it assigns lower weight to outliers when fitting the local polynomial.

*Simplicity*: LOESS regression is relatively easy to understand and implement, and does not require any prior knowledge or assumptions about the functional form of the underlying relationship.

*Visualization*: LOESS regression produces a smooth curve that can be easily plotted against the data, making it a useful tool for visualizing the relationship between the dependent and independent variables.

*Cons*:
*Computationally intensive*: LOESS regression can be computationally intensive, especially for large datasets or when fitting a high-degree polynomial. This can make it impractical for some applications.

*Overfitting*: LOESS regression is prone to overfitting the data if the degree of smoothing is not set appropriately. Overfitting can lead to a model that performs well on the training data but poorly on new data.

*Choice of parameters*: LOESS regression requires the user to make decisions about the size of the windows and the degree of the polynomial to use. These choices can affect the performance of the model and may require some trial and error.

*Limited extrapolation*: LOESS regression is not well suited for extrapolation outside the range of the observed data. This is because the local polynomial is only valid within the range of the data used to fit it.

**Spline regression** is a nonparametric regression method that fits a smooth curve to the data by joining together polynomial functions of lower degree. Here's how spline regression works:

*Choose a degree of polynomial*: Spline regression chooses a degree of polynomial to fit the data, typically 1, 2, or 3.

*Divide the range of the independent variable into intervals*: Spline regression divides the range of the independent variable into a series of intervals, or knots.

*Fit a polynomial to each interval*: Spline regression fits a polynomial function of the chosen degree to the data within each interval, subject to certain constraints that ensure a smooth transition between intervals.

*Combine the polynomials*: Spline regression combines the polynomials for each interval to produce a single smooth curve that fits the entire dataset.

The resulting curve is piecewise continuous, meaning that it consists of a series of connected polynomial functions, each of which is valid only within a specific interval or knot. By using polynomials of lower degree, spline regression avoids the problem of overfitting that can occur with high-degree polynomials.

Spline regression can be used with both univariate and multivariate data, and can be extended to handle more complex relationships between the dependent and independent variables by using multiple splines, each with its own set of knots and degree of polynomial. Spline regression is a powerful tool for modeling complex nonlinear relationships between variables, and is often used in applications such as signal processing, image processing, and data smoothing.

Spline regression is a nonparametric regression method that has several advantages and disadvantages. Here are some of the pros and cons of spline regression:

*Pros*:
*Flexibility*: Spline regression can fit a wide variety of nonlinear relationships between the dependent and independent variables. It is especially useful when the relationship is complex and cannot be easily modeled using a linear regression or a simple parametric model.

*Interpolation and extrapolation*: Spline regression can interpolate between data points within the observed range of the independent variable, and can also extrapolate beyond the range of the data, provided that the underlying relationship is smooth and continuous.

*Smoothing*: Spline regression can be used to smooth noisy data by fitting a curve that captures the underlying trend without being affected by individual outliers or noise.

*No assumptions*: Spline regression does not require any prior assumptions about the functional form of the underlying relationship between the dependent and independent variables.

*Cons*:
*Choice of knots*: Spline regression requires the user to choose the number and location of the knots or intervals in the independent variable. This can affect the performance of the model and may require some trial and error.

*Computationally intensive*: Spline regression can be computationally intensive, especially for large datasets or when using high-degree polynomials. This can make it impractical for some applications.

*Overfitting:* Spline regression can overfit the data if the number of knots is too large or if the degree of polynomial is too high. Overfitting can lead to a model that performs well on the training data but poorly on new data.

*Interpretability*: Spline regression produces a piecewise continuous curve that can be difficult to interpret, especially when using multiple splines with different knots and degrees of polynomial. This can make it challenging to extract meaningful insights from the model.

**Gaussian process regression** (GPR) is a powerful and flexible nonparametric regression method that uses Bayesian inference to model complex relationships between the dependent and independent variables. Here's how it works:

*Choose a prior distribution*: GPR starts by choosing a prior distribution over the space of functions that could describe the relationship between the dependent and independent variables. The prior distribution is typically a Gaussian distribution with mean zero and a covariance function that captures the smoothness of the function.

*Compute the posterior distribution*: GPR updates the prior distribution based on the observed data to obtain a posterior distribution that reflects the uncertainty about the underlying function. The posterior distribution is also a Gaussian distribution with mean and covariance function that depend on the observed data and the prior distribution.

*Make predictions*: GPR uses the posterior distribution to make predictions for new data points by computing the conditional distribution of the dependent variable given the observed data and the new input variable. This conditional distribution is also a Gaussian distribution with mean and variance that can be used to estimate the expected value and uncertainty of the dependent variable.

*Optimize the hyperparameters*: GPR typically involves a set of hyperparameters that define the prior distribution, such as the length scale and amplitude of the covariance function. These hyperparameters can be optimized using maximum likelihood estimation or Bayesian inference to improve the fit of the model to the data.

GPR can be used with both univariate and multivariate data, and can be extended to handle more complex relationships between variables by using kernel functions that capture nonlinear or nonstationary dependencies between variables. GPR is a powerful tool for modeling complex relationships between variables and has many applications in fields such as computer vision, signal processing, and finance. However, GPR can be computationally expensive for large datasets or when using complex kernel functions.

Gaussian process regression (GPR) is a powerful nonparametric regression method that has several advantages and disadvantages. Here are some of the pros and cons of GPR:

*Pros*:
*Flexibility*: GPR can model complex nonlinear relationships between the dependent and independent variables without assuming a specific functional form of the underlying relationship. This makes it useful for a wide range of applications where the relationship is not known beforehand.

*Uncertainty quantification*: GPR provides a probabilistic framework for estimating the uncertainty of predictions, which is useful for decision making and risk management.

*Data efficiency*: GPR can be used with small datasets, making it a useful method when data is expensive or difficult to obtain.

*Hyperparameter tuning*: GPR provides a mechanism for automatically tuning hyperparameters, such as the length scale and amplitude of the covariance function, which can improve the fit of the model to the data.

*Cons*:
*Computationally intensive*: GPR can be computationally intensive, especially for large datasets or when using complex covariance functions. This can make it impractical for some applications.

*Interpretability*: The output of GPR is a probability distribution over functions, which can be difficult to interpret and communicate to non-experts.

*Scalability*: GPR is not scalable to high-dimensional problems, which can limit its applicability to some domains.

*Choice of covariance function*: The performance of GPR is sensitive to the choice of covariance function, and selecting an appropriate covariance function can require some trial and error.

Because Gaussian Process Regression is Bayesian in origin, let's consider the pros and cons of using a Bayesian perspective in the regression context.

In the context of regression analysis, a Bayesian approach provides a framework for estimating regression parameters, making predictions, and quantifying uncertainty. It allows for the incorporation of prior information, updating beliefs based on observed data, and obtaining posterior distributions for the parameters of interest. Here's a general overview of how a Bayesian approach works in regression:

Specify the Regression Model: Start by specifying the regression model, including the choice of predictors, functional forms, and error structure. For example, in simple linear regression, the model may be defined as $y = \beta_0 + \beta_1 x + \varepsilon$, where y is the dependent variable, $x$ is the independent variable, $\beta_0$ and $\beta_1$ are the regression coefficients, and $\varepsilon$ is the error term.

Specify Prior Distributions: Next, assign prior distributions to the regression coefficients and other parameters in the model. These priors capture your beliefs or knowledge about the parameter values before observing any data. The choice of priors can be informed by previous studies, expert opinions, or non-informative priors that spread the probability mass evenly. The prior distributions can be chosen from known distributions, such as normal, uniform, or exponential distributions.

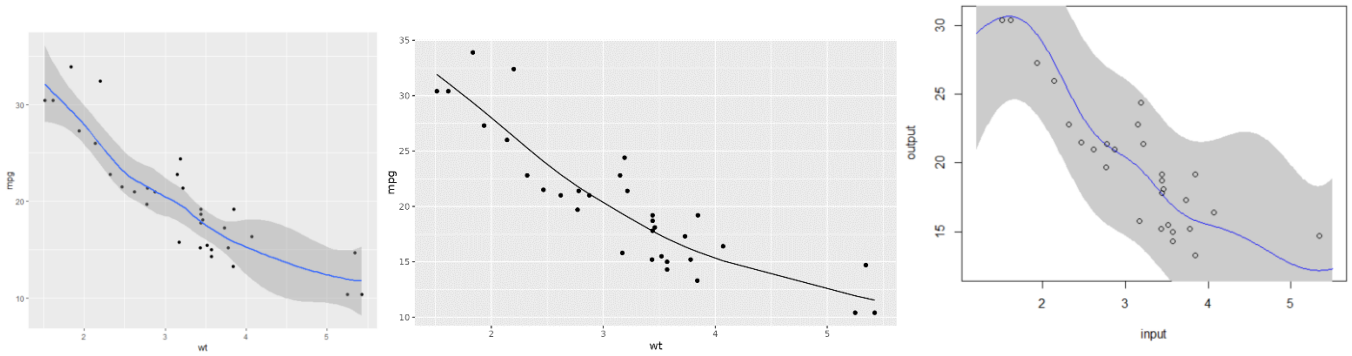
Collect and Analyze Data: Observe the data and compute the likelihood function, which quantifies the probability of observing the data given the model and parameter values. This likelihood is often derived from assumptions about the error structure, such as assuming normally distributed errors. Bayesian regression analysis involves calculating the posterior distribution of the parameters given the observed data by combining the prior distributions and the likelihood function using Bayes' theorem.

Obtain Posterior Distributions: Using Bayes' theorem, the prior distributions and the likelihood are combined to obtain the posterior distribution of the regression parameters. The posterior distribution represents the updated beliefs about the parameter values after considering the observed data. It provides a range of plausible values for each parameter, along with the associated uncertainties. Markov Chain Monte Carlo (MCMC) methods, such as Gibbs sampling or Metropolis-Hastings algorithm, are commonly used to approximate the posterior distribution when analytical solutions are not feasible.

Inference and Prediction: Once the posterior distributions are obtained, you can perform inference on the regression coefficients by examining their means, medians, or credible intervals. These intervals provide a measure of uncertainty and can be used to assess the significance of the coefficients. Bayesian regression also allows for probabilistic predictions by considering the entire posterior predictive distribution. This distribution captures the uncertainty in the predicted values, incorporating both parameter uncertainty and residual variation.

The Bayesian approach to regression analysis offers advantages such as the ability to incorporate prior knowledge, handle complex models, and provide a full probabilistic representation of the uncertainty. However, it requires careful consideration and justification of the prior distributions, and the results can be sensitive to the choice of priors. Sensitivity analyses and robustness checks are often performed to assess the influence of prior specifications.

Comparison of LOESS, spline and Gaussian process regression on the same data.




**Review for exam #2**
ANOVA/general linear model
Logistic regression
Variable transformations/non-linear models (log, polynomial, interaction terms, etc.)
Penalized regression
Encoding dummy variables

Factorial designs (general properties and motivations)
Interpret diagnostic graphs
Encoding with ordinal vs. nominal variables
Discretization
Machine learning: general types especially
Validation methods
Non-parametric models

Resources:


1. https://www.sciencedirect.com/topics/social-sciences/nonparametric-regression
2. https://www.bauer.uh.edu/rsusmel/phd/ec1-27.pdf
3. https://rcompanion.org/handbook/F_12.html
4. http://r-statistics.co/Loess-Regression-With-R.html
5. https://www.statology.org/loess-regression-in-r/
6. https://www.statology.org/spline-regression-in-r/
7. http://users.stat.umn.edu/~helwig/notes/smooth-spline-notes.html
8. https://bookdown.org/rbg/surrogates/chap5.html
9. https://rpubs.com/ncahill_stat/889056