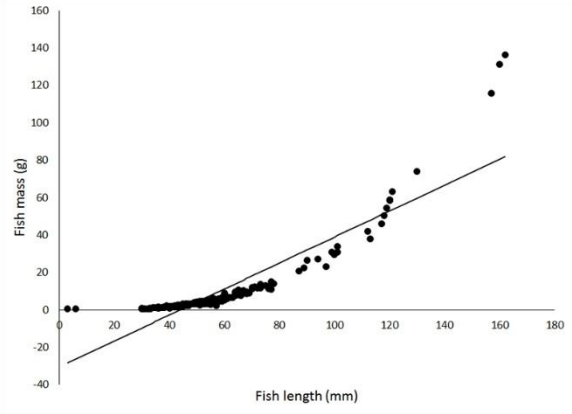
3/9/2023

Transforming variables in regression is a common technique used to improve the accuracy and interpretability of a regression model. The main reasons for transforming variables are to handle nonlinearity, heteroscedasticity, outliers, and skewed or heavy-tailed distributions in the data.

Recall from our early discussion of single variable regression, we discussed some types of regression models that were nonlinear, but considered intrinsically linear and for which we could calculate linear correlation values. We will review those models and go beyond them.



To obtain intrinsically linear models, we perform transformations of variables to obtain new relationships. For instance, we could create a reciprocal model by replacing our initial variable x with a new variable $z = \frac{1}{x}$. After transforming the variable, we can create a linear model on z.

$$y = az + b$$

This model is equivalent to

$$y = a\left(\frac{1}{x}\right) + b$$

Similarly, we can use logarithmic transformations to obtain the other intrinsically linear models.

If we apply a log transformation to x only, we get a logarithmic model.

$$z = \ln(x)$$
$$y = az + b$$

Which becomes

$$y = a(\ln x) + b$$

If we transform the y variable with log instead, we can obtain an exponential model with some algebra.

$$w = \ln y$$

$$w = ax + b$$
$$\ln y = ax + b$$

$$y = e^{ax+b}$$
$$y = e^{ax}e^{b}$$

Let $e^{b} = c$, then we get

$$y = ce^{ax}$$

Finally, if we take the log of both variables, we can obtain a power model.

$$w = \ln y, z = \ln x$$

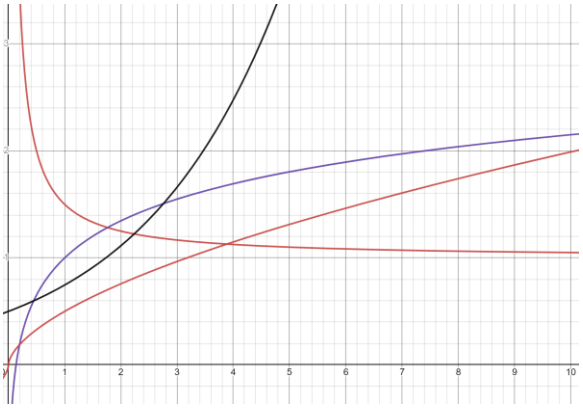
$$w = az + b$$
$$\ln y = a(\ln x) + b$$
$$y = e^{a \ln x + b}$$
$$y = e^{a \ln x}e^{b} = \left(e^{\ln x}\right)^{a}e^{b}$$

Recall that $e^{\ln x} = x$, and let $e^{b} = c$. Then we get

$$y = c(x^{a})$$

Reciprocal models without an intercept can also be considered in this category, when $a = -1$.

All four models are nonlinear, but still relatively simple curves.



We can extend these types of transformation to other kinds of algebraic manipulations, and we can extend these types of transformations to variables in multiple variable models to obtain better fits, and to repair certain types of common problems.

Here are some common methods for transforming variables in regression:

Logarithmic Transformation: This is a common method for handling skewed or highly variable data, by taking the logarithm of the variable. Logarithmic transformation can help to stabilize the variance and

reduce the impact of outliers, and can also transform nonlinear relationships into linear ones. For example, taking the logarithm of a variable that follows a power-law distribution can result in a more normally distributed variable.

Square Root Transformation: This is another method for handling highly variable data, by taking the square root of the variable. Square root transformation can also help to stabilize the variance and reduce the impact of outliers, and can transform variables that follow a quadratic relationship into a linear one.

Box-Cox Transformation: This is a general method for transforming a variable to a normal distribution by applying a power transformation, which is chosen to maximize the log-likelihood of the data. Box-Cox transformation can handle a wide range of distributions and can be used to transform variables that follow a wide range of nonlinear relationships.

- Other types of rescaling transformations of variables are possible:
    - Converting variables to z-scores
    - Rescaling variables by range to be between 0 and 1
    - Rescaling variables to values between -1 and 1
    - Applying a shift to recenter the variables without rescaling
- Some methods require some sort of adjustments because raising values to large powers or as exponentials can cause computational problems due to the size of the resulting values.

Polynomial Transformation: This is a method for transforming variables into higher-order polynomials, by adding squared, cubed, or higher-order terms to the regression model. Polynomial transformation can capture more complex nonlinear relationships between the variables, but can also lead to overfitting and instability if the degree of the polynomial is too high.

Winsorization: This is a method for handling outliers by replacing extreme values with the nearest value within a specified percentile range, such as the 1st and 99th percentiles. Winsorization can help to reduce the impact of outliers on the regression coefficients and the residuals, but can also distort the distribution of the data and reduce the precision of the estimates.

These are just a few examples of the many methods for transforming variables in regression. The choice of method depends on the specific properties of the data and the research question, and should be guided by graphical and statistical diagnostics of the model fit and residuals. The goal of variable transformation is to improve the validity and reliability of the regression model, by reducing bias, improving interpretability, and increasing the generalizability of the results.

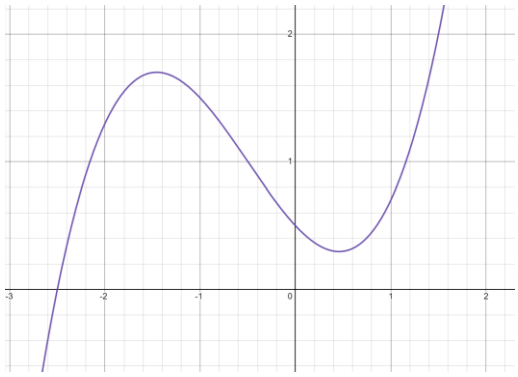
Polynomial regression models are a type of regression analysis where the relationship between the dependent variable and one or more independent variables is modeled as an nth-degree polynomial function. The polynomial function is a mathematical equation that includes terms with different powers of the independent variable(s), such as x, x^2, x^3, and so on.

Polynomial regression models are useful when the relationship between the dependent variable and the independent variable is not linear, but can be approximated by a nonlinear function (more complex than a log, exponential or power model). For example, in a simple linear regression model, the relationship between a dependent variable y and an independent variable x is modeled as $y = b_0 + b_1 x$, where

b_0 is the intercept and b_1 is the slope of the line. In contrast, in a polynomial regression model, the relationship between y and x can be modeled as

$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \ldots + b_n x^n$$

where n is the degree of the polynomial function and $b_2, b_3, \ldots, b_n$ are the coefficients of the quadratic, cubic, and higher-order terms. We can further extend the model to multiple variables. These models can capture much more complex behavior, for example, the cubic behavior in the graph below.



Polynomial regression models can be fitted using standard regression techniques, such as ordinary least squares (OLS), maximum likelihood estimation (MLE), or Bayesian inference. The choice of degree of the polynomial function depends on the specific problem and the data, and can be guided by visual inspection of the data, statistical tests of the model fit, and cross-validation procedures. Ordinary Least Squares can be fitted using the normal equation with the A matrix having 1s in the first column for the b_0 coefficient, the x values in the second column, x^2 in the third column and so on. The computational power is dependent on the number of parameters (coefficients) in the model, and much less on the number of observations.

Polynomial regression models have several advantages and limitations. One advantage is that they can capture more complex nonlinear relationships between the variables, and can provide a better fit to the data than linear models (or other nonlinear parametric models). Another advantage is that they can be easily visualized and interpreted, by plotting the polynomial function and its derivatives. However, polynomial regression models can also be prone to overfitting, especially when the degree of the polynomial is high or the sample size is small. Therefore, it is important to balance the complexity and the interpretability of the model, and to use appropriate regularization techniques, such as ridge regression or Lasso, to control the variance of the estimates.

Polynomial models, power models, logarithmic models, and exponential models are all types of nonlinear regression models that can be used to capture complex relationships between the dependent and independent variables. Here are some differences and similarities between these types of models:

Polynomial models: Polynomial models are characterized by a polynomial function that includes one or more terms with different powers of the independent variable(s). Polynomial models can capture a wide range of nonlinear relationships, including quadratic, cubic, and higher-order functions. Polynomial models are flexible and easy to fit using standard regression techniques, but can be prone to overfitting and instability when the degree of the polynomial is too high. Extrapolation can also be more

problematic since polynomials models will tend to infinity (or negative infinity) outside the range of the data.

Power models: Power models are characterized by a power-law relationship between the dependent and independent variables, such as $y = kx^p$, where k and p are parameters that determine the slope and curvature of the function. Power models are commonly used to model phenomena that exhibit scale-free behavior, such as biological growth, network dynamics, and economic systems. Power models are nonlinear and require special techniques, such as nonlinear least squares or maximum likelihood estimation, to fit the parameters.

Logarithmic models: Logarithmic models are characterized by a logarithmic transformation of the independent variable(s), such as $y = a + b\ log(x)$ or $y = a + b\ln(x)$, where a and b are parameters that determine the intercept and slope of the function. Logarithmic models can be used to model phenomena that exhibit diminishing returns or exponential growth, such as learning curves, response curves, and diffusion models. Logarithmic models are nonlinear and require special techniques, such as nonlinear regression or generalized linear models, to fit the parameters.

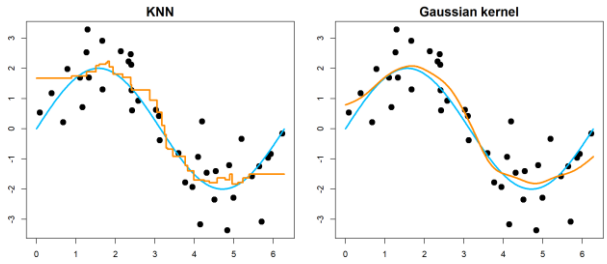
Exponential models: Exponential models are characterized by an exponential function of the independent variable(s), such as $y = ae^{bx}$, where a and b are parameters that determine the intercept and rate of change of the function. Exponential models can be used to model phenomena that exhibit exponential growth or decay, such as population growth, radioactive decay, and chemical reactions. Exponential models are nonlinear and require special techniques, such as nonlinear regression or time-series analysis, to fit the parameters.

While you may need to experiment with several models to find the best fit, there may also be physical or logical reasons to choose one type of model over another that have to be considered. For example, exponential models will reach a minimum value, while logarithmic models will not.

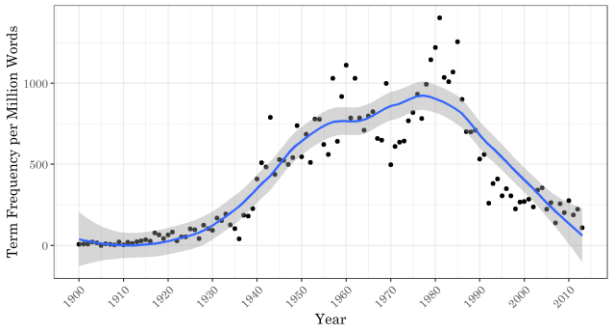
Nonparametric regression models are a class of regression models that do not make explicit assumptions about the functional form of the relationship between the dependent variable and the independent variables. Instead, nonparametric regression models estimate the relationship using flexible, data-driven techniques that are based on local or global smoothing of the data.

Here are some examples of nonparametric regression models:

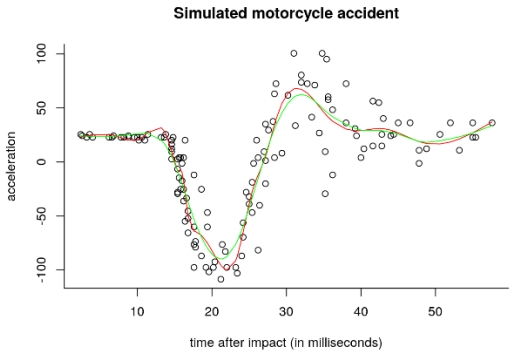
Kernel regression: Kernel regression is a nonparametric regression technique that estimates the conditional mean of the dependent variable as a weighted average of the neighboring observations, where the weights are determined by a kernel function that assigns higher weights to closer observations. Kernel regression is flexible and easy to implement, but requires the choice of a bandwidth parameter that determines the degree of smoothing.



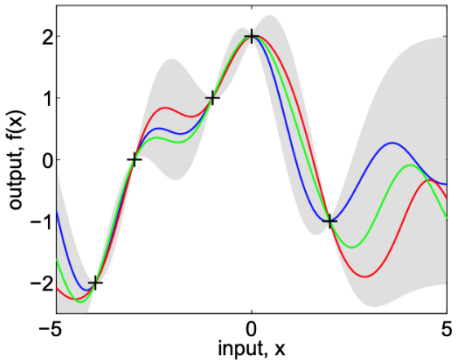
Local regression: Local regression is a nonparametric regression technique that estimates the conditional mean of the dependent variable as a weighted least squares fit of a polynomial function to the neighboring observations. Local regression can use different degrees of polynomial functions and different weighting schemes, and can provide a smooth and flexible estimate of the relationship. Loess regression is one example of this method and is the default regression method in ggplot2.



Splines regression: Splines regression is a nonparametric regression technique that estimates the conditional mean of the dependent variable as a piecewise polynomial function of the independent variable(s), where the polynomial segments are connected smoothly at a set of knots. Splines regression can use different degrees of polynomial functions and different knot locations, and can provide a flexible and interpretable estimate of the relationship.



Gaussian process regression: Gaussian process regression is a nonparametric regression technique that models the relationship between the dependent variable and the independent variable as a random function drawn from a Gaussian process prior, which is characterized by a mean function and a covariance function. Gaussian process regression can capture complex nonlinear and nonstationary relationships and can provide uncertainty estimates of the predictions.



Random forests regression: Random forests regression is a nonparametric regression technique that combines multiple decision trees that are trained on different subsets of the data and different subsets of the variables. Random forests regression can capture complex interactions and nonlinearities, and can provide important measures of the variables.

Many regression methods have a classification counterpart, and classification methods have a regression counterpart, but that is not to say that these methods are equally good.

Interaction terms: We've encountered these in ANOVA, but we can also include them in nonlinear regression models (typically polynomial models). In a regression model, interaction terms refer to the product of two or more independent variables that have a combined effect on the dependent variable. The inclusion of interaction terms in a regression model allows the model to capture the fact that the relationship between the dependent variable and each independent variable may depend on the level or value of other independent variables.

For example, suppose we have a regression model that predicts the price of a house based on its size and location, and we include an interaction term between size and location. This interaction term captures the fact that the effect of size on price may depend on the location of the house. Specifically, the coefficient of the interaction term represents the change in the effect of size on price when the location of the house changes by one unit.

Mathematically, the regression model with an interaction term can be written (for the two variable case) as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

where y is the dependent variable, x_1 and x_2 are the independent variables, $\beta_0, \beta_1, \beta_2$, and β_3 are the regression coefficients, $x_1 x_2$ is the interaction term, and ε is the error term. The coefficient β_3 measures the effect of the interaction between x_1 and x_2 on y, controlling for the main effects of x_1 and x_2. Interaction terms can be useful for modeling complex relationships between variables and improving the accuracy of the regression model. However, they can also lead to overfitting and instability if the sample size is small or the number of interaction terms is too high. It is important to carefully select the interaction terms based on domain knowledge, exploratory data analysis, and statistical significance tests. We've seen similar effects in ANOVA models involving interaction terms.

Including or excluding interaction terms in a regression model is an important decision that should be based on domain knowledge, exploratory data analysis, and statistical significance tests. Here are the general steps involved in the process of including or excluding interaction terms in a regression model: Understand the variables: Before including or excluding interaction terms in a regression model, it is important to understand the variables and their potential relationships. For example, if there is reason to believe that the effect of one variable on the dependent variable depends on the value of another variable, then an interaction term might be necessary.

Explore the data: Exploratory data analysis can help identify potential interactions between variables. This can be done by examining scatter plots, correlation matrices, and other graphical and numerical summaries of the data. If there appears to be a relationship between two variables that is not linear or additive, then an interaction term might be necessary.

Specify the model: Once potential interaction terms have been identified, a regression model can be specified that includes or excludes the interaction terms. This can be done by adding or removing terms from the regression equation, as appropriate. It is important to consider the number of interaction terms relative to the sample size, as too many interaction terms can lead to overfitting and instability.

Test the model: The specified model can be tested for statistical significance using standard methods such as hypothesis testing, model selection criteria, and goodness-of-fit tests. If the interaction terms are statistically significant and improve the fit of the model, then they can be included. If the interaction terms are not statistically significant or do not improve the fit of the model, then they can be excluded.

Validate the model: Once a final model has been selected, it is important to validate the model using methods such as cross-validation, bootstrapping, or out-of-sample testing. This can help ensure that the model is not overfitting or underfitting the data and is generalizable to new data.

In future lectures, we'll look at additional validation measures, but for now, we can still apply the variety of model diagnostic tools. Model diagnostics are important for assessing the goodness of fit and the validity of assumptions in nonlinear regression models. Here are some common diagnostic techniques for nonlinear regression models:

Residual plots: Similar to linear regression, residual plots are a useful diagnostic tool for nonlinear regression. A residual plot shows the difference between the predicted and observed values of the dependent variable, plotted against the independent variable(s). A good model will have residuals that are randomly scattered around zero, indicating that the model is fitting the data well.

Normal probability plots: A normal probability plot is used to assess the normality of the residuals in a regression model. If the residuals are normally distributed, the points on the normal probability plot will form a roughly straight line. Nonlinear regression models may require transformation of the dependent variable to achieve normality.

Cook's distance: Cook's distance is a measure of the influence of each observation on the fitted values in the regression model. Large values of Cook's distance indicate that the corresponding observation has a large impact on the model, and may need to be investigated further.
Leverages: Leverages are a measure of how extreme an observation is in terms of its predictor values. High leverage points can affect the fit of the model, and may require further investigation.

Outlier detection: Outliers are observations that have a large impact on the fitted values in the regression model. Nonlinear regression models may require outlier detection methods specific to the model, such as Mahalanobis distance or studentized residuals.

Cross-validation: Cross-validation is a method for testing the validity of the model by evaluating the performance of the model on a separate dataset. Nonlinear regression models can benefit from cross-validation techniques, such as k-fold cross-validation, leave-one-out cross-validation, or bootstrapping.

Which is to say that many of the same methods we looked at for simple linear regression are also useful here. After the break, we'll look at penalized regression methods in more detail.

Variable transformations can impact model assumptions in regression in both positive and negative ways. Here are some pros and cons to consider:

Pros:

Improved model fit: In some cases, transforming variables can improve the fit of the regression model, leading to more accurate predictions and better inference.

Better linearity: Transformations can sometimes make the relationship between the dependent and independent variables more linear, which can simplify interpretation and improve the stability of the estimates.

Better normality: Transformations can sometimes make the distribution of the dependent variable more normal, which is a key assumption for many regression models.

Cons:

Non-intuitive interpretation: When variables are transformed, the coefficients in the regression model can become harder to interpret in a meaningful way.

Loss of information: Transformations can sometimes result in a loss of information or distortions in the data, making it harder to draw accurate conclusions.

Changes in assumptions: Transformations can also impact the assumptions of the regression model, making it harder to draw accurate inferences. For example, transforming a variable can impact the assumption of homoscedasticity or normality.

Have a nice spring break!

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://book.stat420.org/transformations.html
3. https://bookdown.org/ejvanholm/Textbook/transforming-variables-in-regression.html
4. https://www.geeksforgeeks.org/polynomial-regression-in-r-programming/
5. https://datascienceplus.com/fitting-polynomial-regression-r/
6. https://www.statology.org/polynomial-regression-r/
7. https://data-flair.training/blogs/r-nonlinear-regression/
8. https://bookdown.org/anshul302/HE902-MGHIHP-Spring2020/LogitTrans.html
9. https://statisticsbyjim.com/regression/r-squared-invalid-nonlinear-regression/
10. https://statisticsbyjim.com/regression/r-squared-invalid-nonlinear-regression/
11. https://www.hiercourse.com/docs/Rnotes_nonlinear.pdf
12. https://www.pnw.edu/wp-content/uploads/2020/03/Lecture-Notes-12-8.pdf
13. https://sciences.usca.edu/biology/zelmer/305/trans/
14. https://teazrq.github.io/stat432/rlab/kernel.html
15. https://stackoverflow.com/questions/58041940/loess-confidence-intervals-excessively-narrow-in-qqplot2
16. https://lbelzile.github.io/lineaRmodels/splines.html
17. https://paperswithcode.com/method/gaussian-process