

Introduction to MARS, Logistic Regression, & Survival Analysis

Presented by: Derek Kane



Overview of Topics

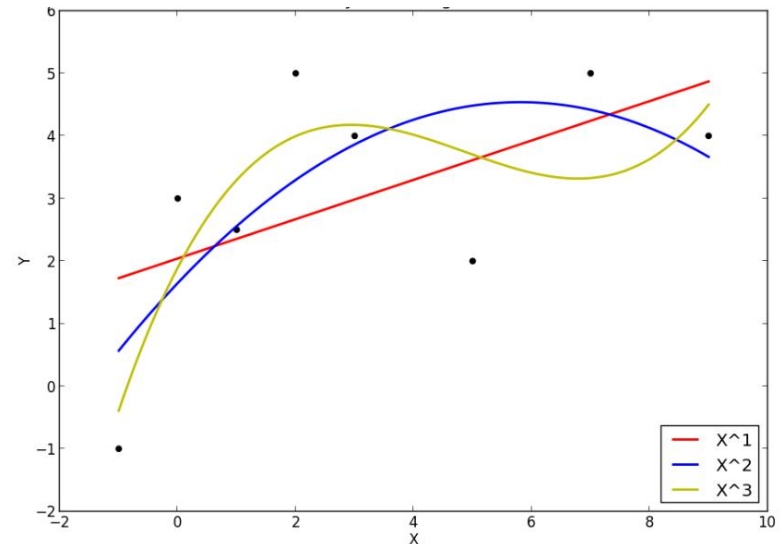
- ❖ Extending Regression Techniques
- ❖ MARS
- ❖ Logistic Regression
- ❖ Survival Analysis
- ❖ Practical Application Example
 - ❖ Crime Prediction in US



Extending Regression Techniques

- ❖ In previous lectures, we have extensively reviewed the mechanics of multiple linear regression, including the various OLS assumptions.
- ❖ We have also spent some time building on top of this framework and extending regression into Ridge Regression, LASSO, and Elastic Net techniques.
- ❖ There are even more variations of linear regression such as polynomial representations and other variants which we have not fully explored.
- ❖ The purpose of this presentation is to further expose us to advanced regression topics (linear/nonlinear) and continue to build off of our regression knowledge base.

Polynomial Regression



$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

Introduction to MARS

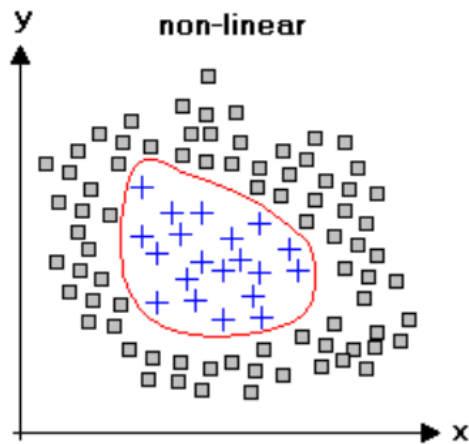
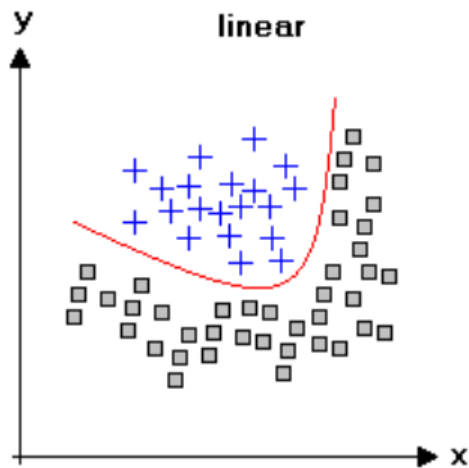
- ❖ In statistics, Multivariate Adaptive Regression Splines (MARS) is a form of regression analysis introduced by Jerome H. Friedman in 1991.
- ❖ MARS is a non-parametric regression technique and can be seen as an extension of linear models that automatically models non-linearities and interactions between variables.
- ❖ The term "MARS" is trademarked and licensed to Salford Systems.
- ❖ In order to avoid trademark infringements, many open source implementations of MARS are called "Earth". (Ex. R Package "earth")



Jerome Friedman



Introduction to MARS



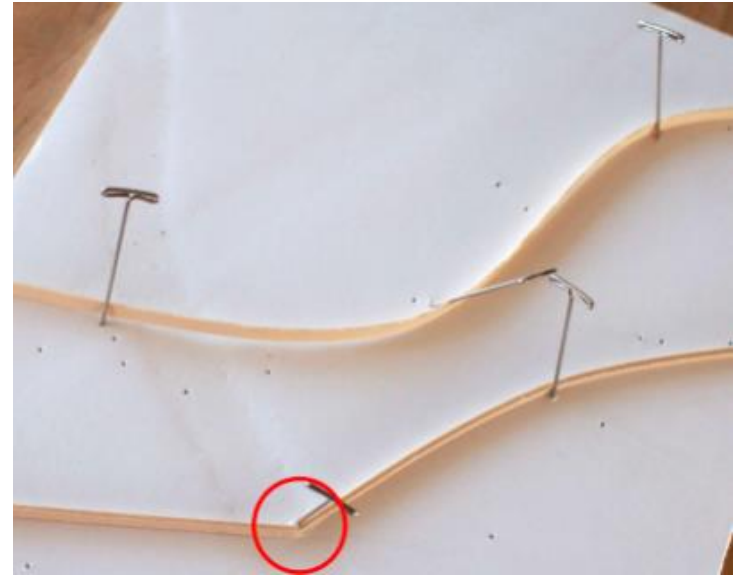
Why use MARS models?

- ❖ MARS software is ideal for users who prefer results in a form similar to traditional regression while capturing essential nonlinearities and interactions.
- ❖ The MARS approach to regression modeling effectively uncovers important data patterns and relationships that are difficult, if not impossible, for other regression methods to reveal.
- ❖ MARS builds its model by piecing together a series of straight lines with each allowed its own slope.
- ❖ This permits MARS to trace out any pattern detected in the data.

Introduction to MARS

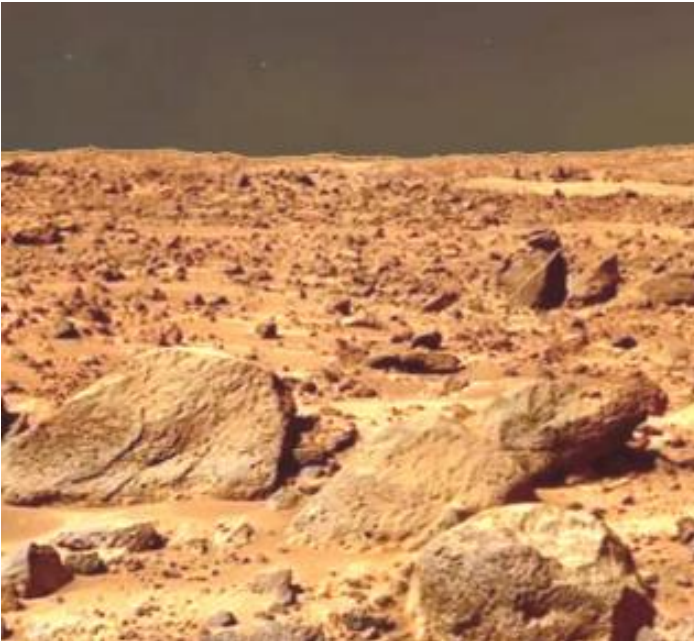
Why use MARS models?

- ❖ The MARS model is designed to predict continuous numeric outcomes such as the average monthly bill of a mobile phone customer or the amount that a shopper is expected to spend in a web site visit.
- ❖ MARS is also capable of producing high quality probability models for a yes/no outcome.
- ❖ MARS performs variable selection, variable transformation, interaction detection, and self-testing, all automatically and at high speed.



Splines in Wood

Introduction to MARS



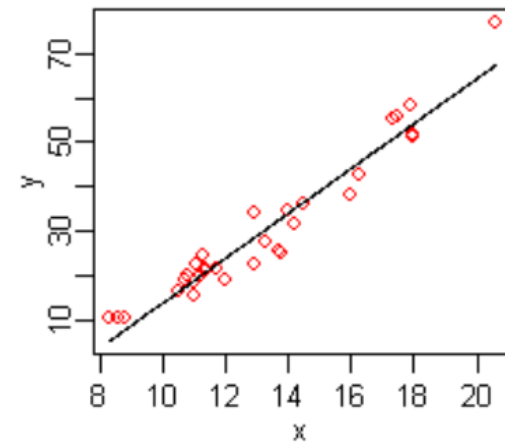
Areas where MARS has exhibited very high-performance results include:

- ❖ Forecasting electricity demand for power generating companies.
- ❖ Relating customer satisfaction scores to the engineering specifications of products.
- ❖ Presence/absence modeling in geographical information systems (GIS).
- ❖ MARS is a highly versatile regression technique and an indispensable tool in our data science toolkit.

MARS Mechanics

- ❖ This section introduces MARS using a few examples. We start with a set of data: a matrix of variables x , and a vector of the responses, y , with a response for each row in x .
- ❖ For example, the data could be:

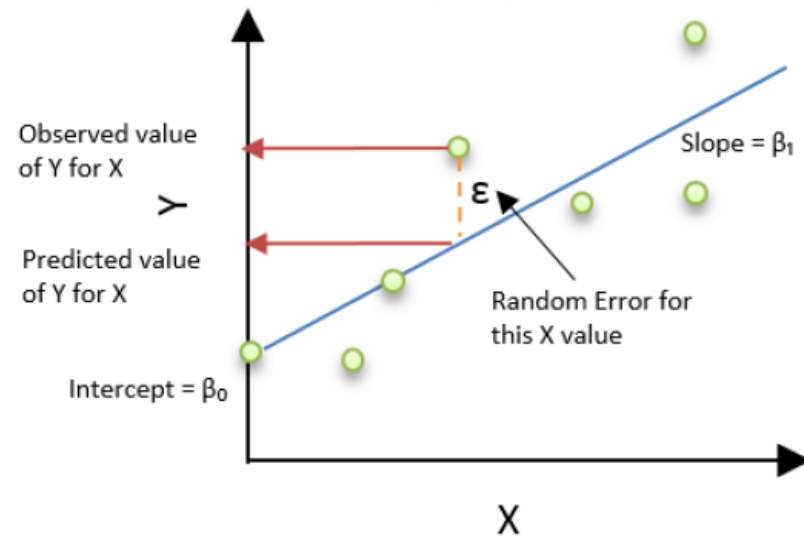
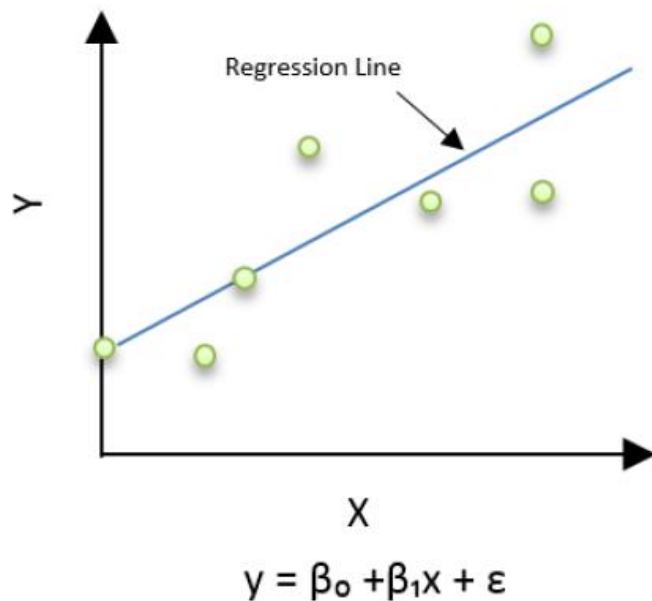
X	Y
10.5	16.4
10.7	18.8
10.8	19.7
...	...
20.6	77



- ❖ Here there is only one independent variable, so the x matrix is just a single column. Given these measurements, we would like to build a model which predicts the expected y for a given x .
- ❖ A linear model for the above data is: $\hat{y} = -37 + 5.1x$

OLS Regression Mechanics

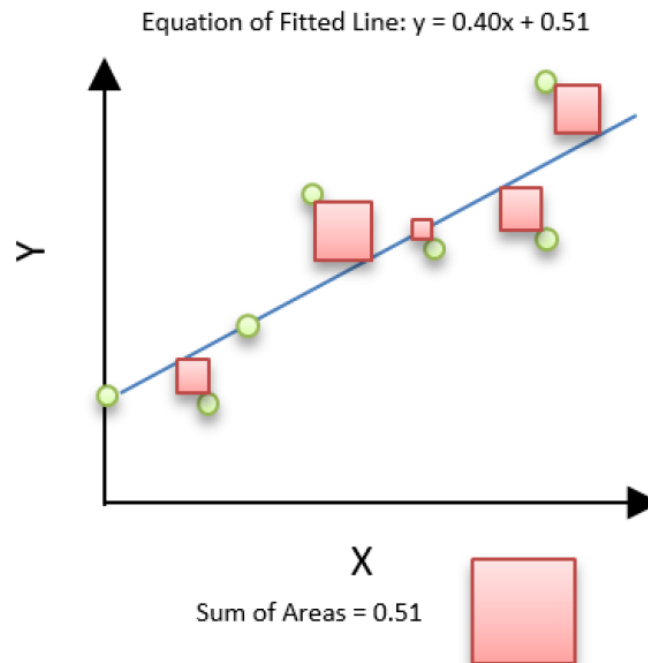
- ❖ Here is a quick refresher slide on OLS regression.



- ❖ For a more thorough breakdown of linear regression techniques and diagnostics, please refer back to some of my previous tutorials.

OLS Regression Mechanics

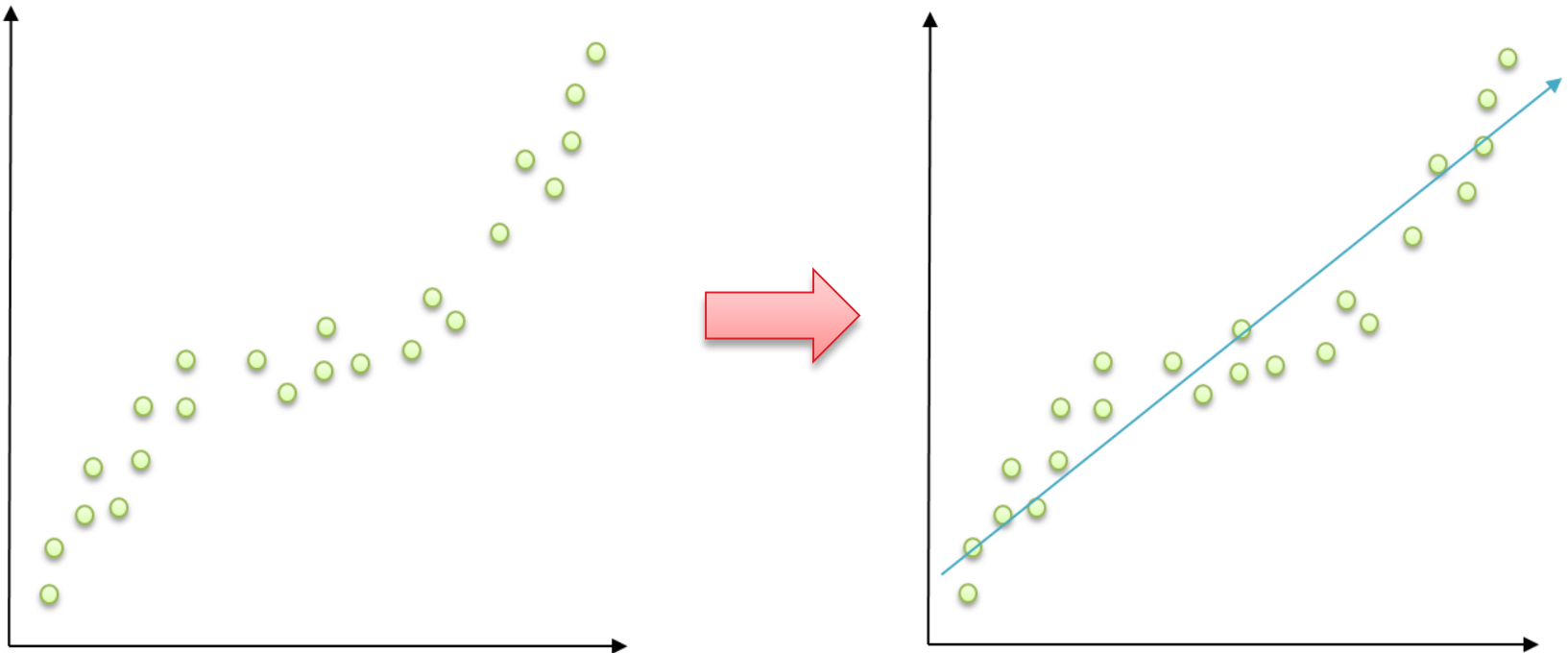
- ❖ We take the square of the error value and then sum the total errors to reach the Sum of Square Error.



- ❖ The OLS based approach is trying to minimize this error by producing a line with the lowest total sum of the error.

MARS Mechanics

- ❖ The basic idea is similar with the MARS algorithm. Lets take the following chart as an example:



- ❖ If we are trying to create a line with the smallest square error, it may look something like this chart.

MARS Mechanics

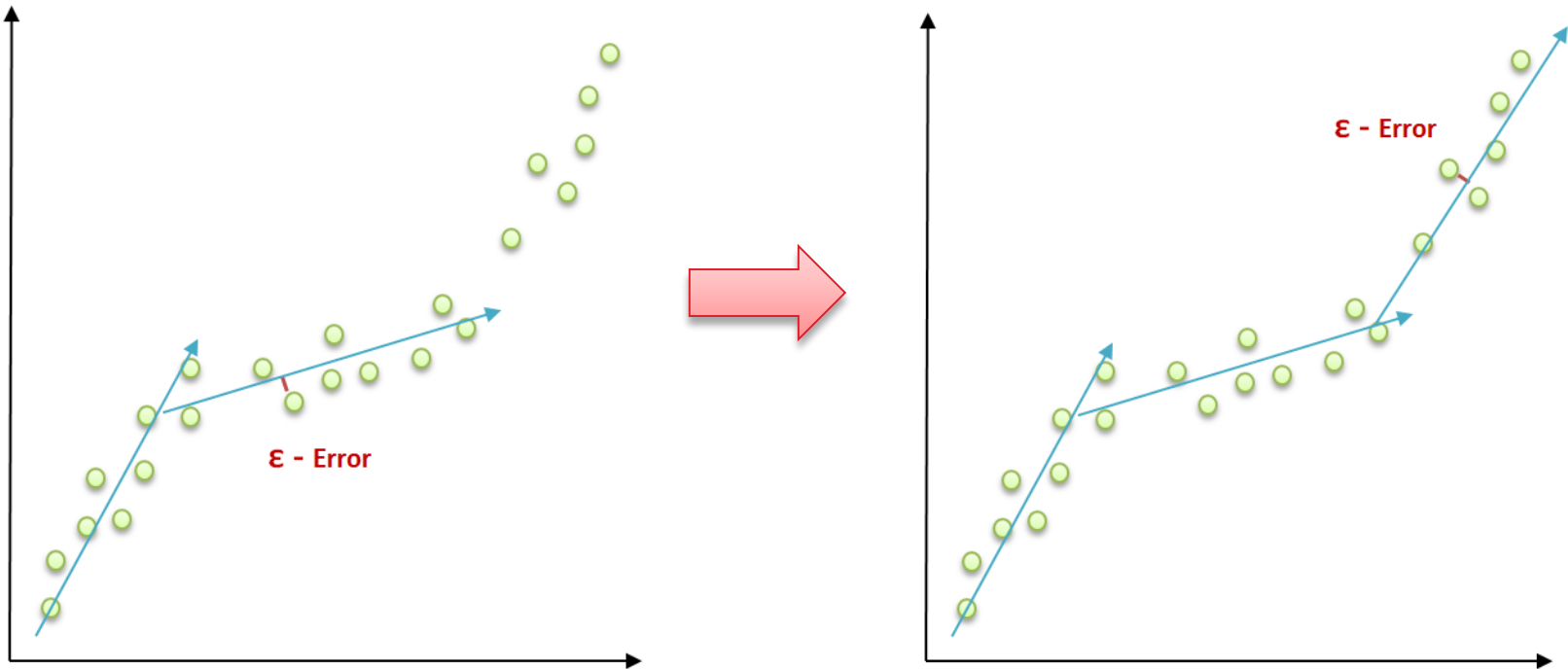
- ❖ The basic error can be seen in the chart below.



- ❖ However, if we fit a linear line on only a small part of the line (spline) we can measure the error related to that specific spline and that portion of the data.

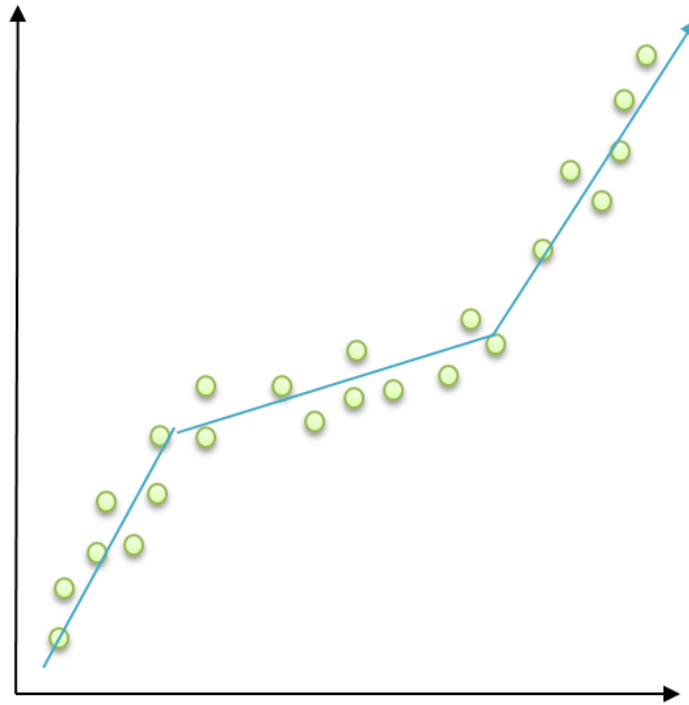
MARS Mechanics

- ❖ We then create a pivot point and then add a second linear line connecting to this line to create a “knot” in the model.



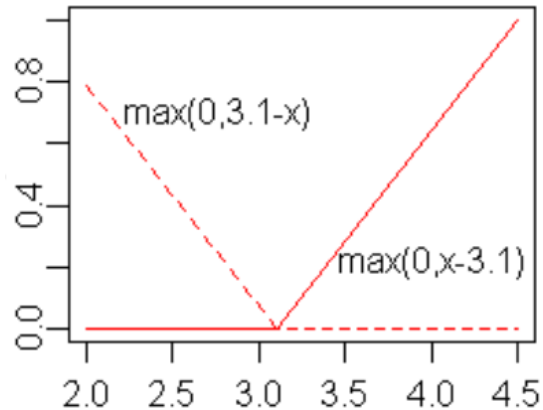
MARS Mechanics

- ❖ The resulting model will have a lower overall error term and can match the pattern of the data in a much more eloquent manner.



- ❖ This MARS model contains 3 separate splines and 2 hinge functions.

MARS- Hinge Functions



Hinge functions are a key part of MARS models:

- ❖ A hinge function is the point where a linear regression models line is shifted into a different linear regression line.
- ❖ There are 2 functions with one being the hinge we are looking at utilizing and the reciprocal.
- ❖ The hinge functions are the expressions starting with max (where $\max(a,b)$ is a if $a > b$, else b).
- ❖ Hinge functions are also called hockey stick or rectifier functions.
- ❖ This is primarily due to the characteristic shape of a hockey stick that the hinge functions sometime take.

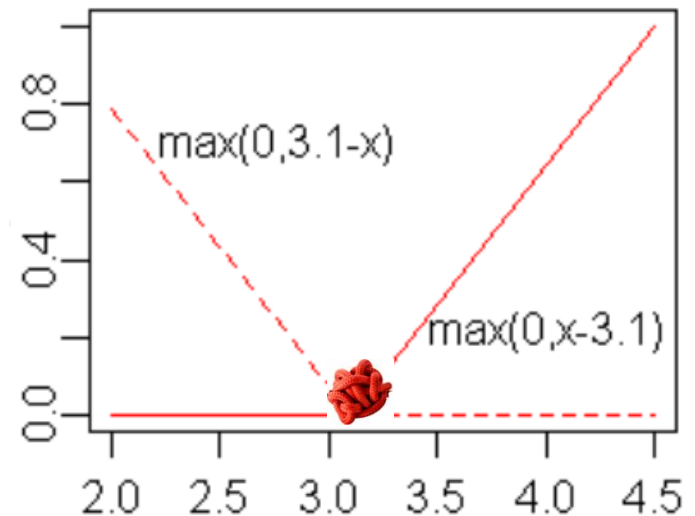
MARS- Hinge Functions

A hinge function takes the form:

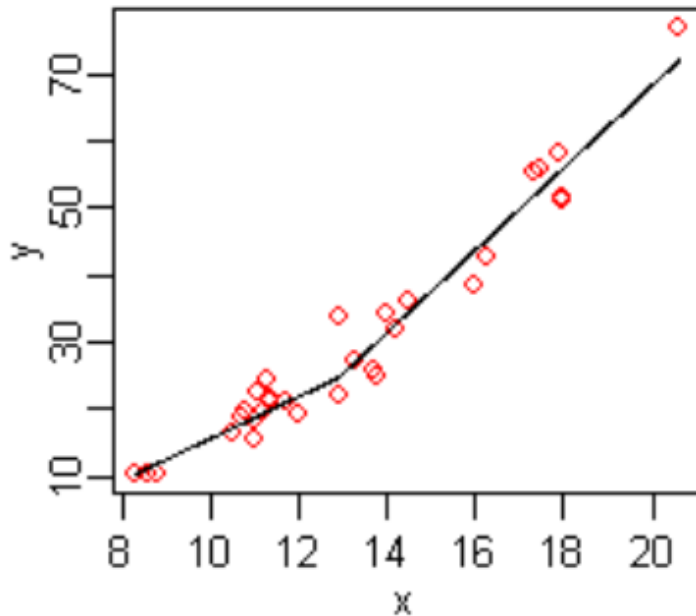
$$\max(0, x - c) \text{ or } \max(0, c - x)$$

- ❖ where c is a constant, called the knot.
- ❖ The figure on the right shows a mirrored pair of hinge functions with a knot at 3.1.

Note: One might assume that only piecewise linear functions can be formed from hinge functions, but hinge functions can be multiplied together to form non-linear functions.



MARS- Hinge Functions



- ❖ A hinge function is zero for part of its range, so can be used to partition the data into disjoint regions, each of which can be treated independently.

- ❖ For our example a mirrored pair of hinge functions in the expression:

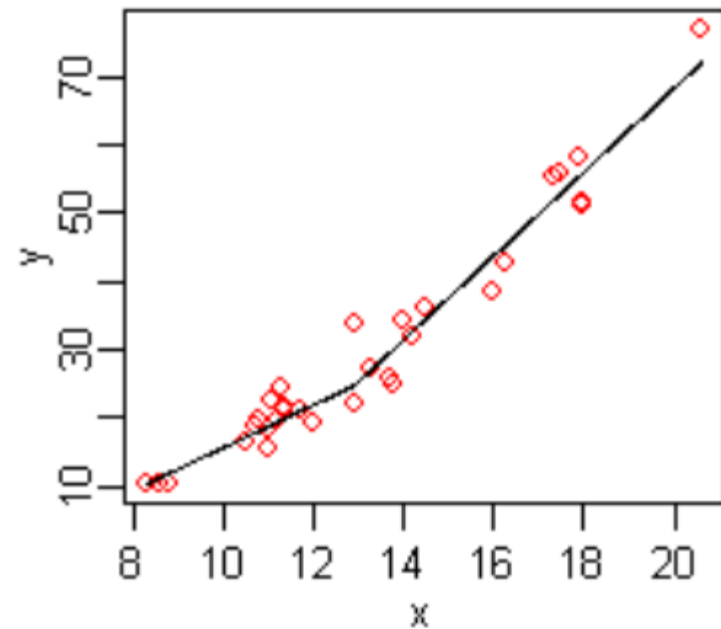
$$6.1 \max(0, x - 13) - 3.1 \max(0, 13 - x)$$

- ❖ This creates the piecewise linear graph shown for the simple MARS model on the left hand side.

MARS Model

- ❖ When we turn to MARS to automatically build a model taking into account non-linearities.
- ❖ The MARS software constructs a model from the given x and y as follows:

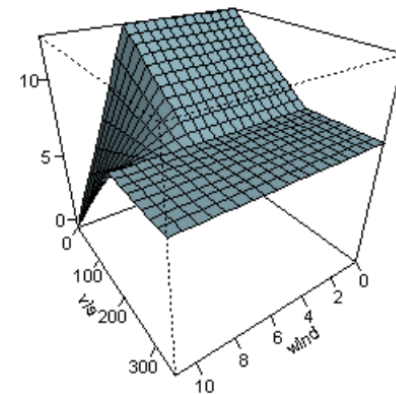
$$\begin{aligned}\hat{y} = & 25 \\ & + 6.1 \max(0, x - 13) \\ & - 3.1 \max(0, 13 - x)\end{aligned}$$



MARS Mechanics

- ❖ In general there will be multiple independent variables, and the relationship between y and these variables will be unclear and not easily visible by plotting.
- ❖ We can use MARS to discover that non-linear relationship. An example MARS expression with multiple variables is:

$$\begin{aligned}\text{ozone} = & 5.2 \\ & + 0.93 \max(0, \text{temp} - 58) \\ & - 0.64 \max(0, \text{temp} - 68) \\ & - 0.046 \max(0, 234 - \text{ibt}) \\ & - 0.016 \max(0, \text{wind} - 7) \max(0, 200 - \text{vis})\end{aligned}$$

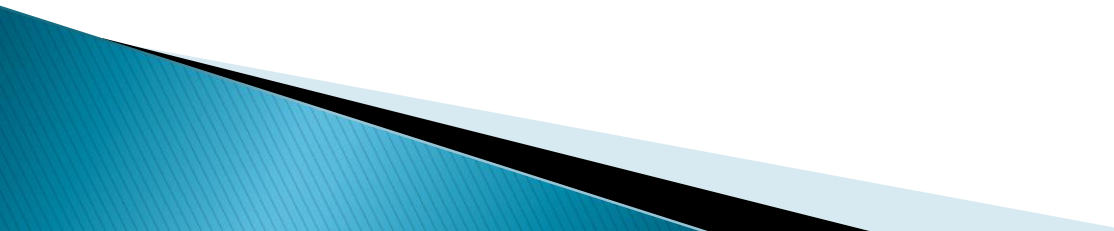


- ❖ This expression models air pollution (the ozone level) as a function of the temperature and a few other variables.
- ❖ The figure on the right plots the predicted ozone as wind and vis vary, with the other variables fixed at their median values.
- ❖ The figure shows that wind does not affect the ozone level unless visibility is low. We see that MARS can build quite flexible regression surfaces by combining hinge functions.

MARS – Pros and Cons

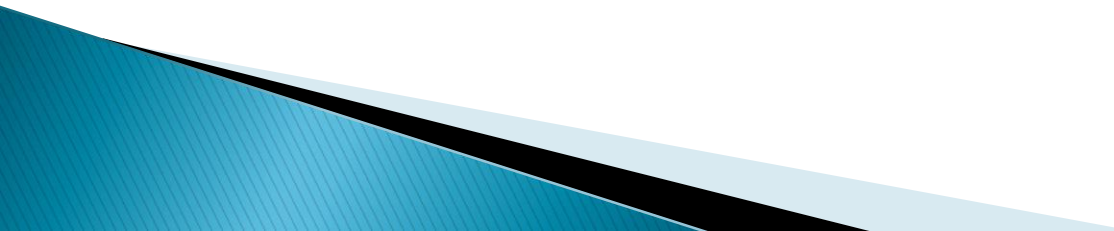
- ❖ It is useful to compare MARS to recursive partitioning. (Recursive partitioning is also commonly called regression trees, decision trees, or CART)

Pros:

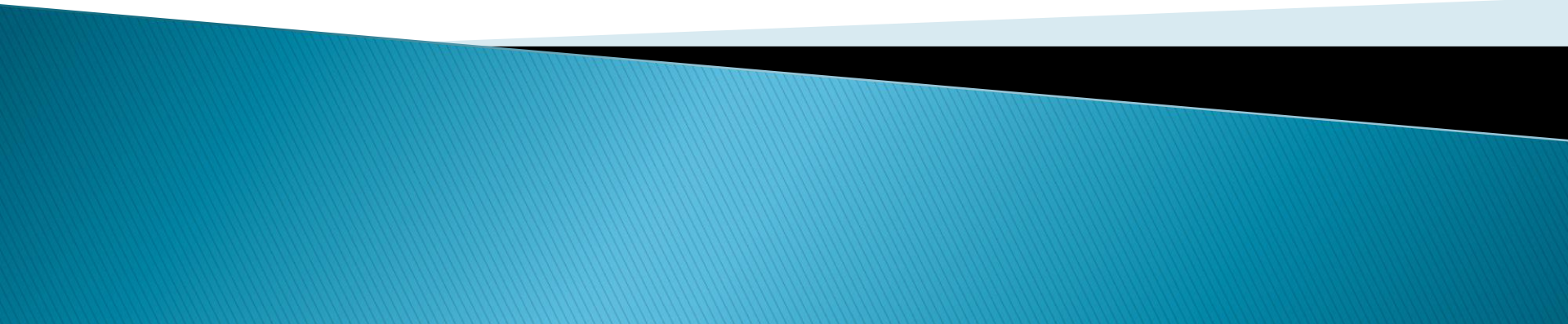
- ❖ MARS models are more flexible than linear regression models.
 - ❖ MARS models are simple to understand and interpret. Compare the equation for ozone concentration example to, say, the innards of a trained neural network or a random forest.
 - ❖ MARS can handle both continuous and categorical data.
 - ❖ MARS tends to be better than recursive partitioning for numeric data because hinges are more appropriate for numeric variables than the piecewise constant segmentation used by recursive partitioning.
 - ❖ Building MARS models often requires little or no data preparation.
 - ❖ The hinge functions automatically partition the input data, so the effect of outliers is contained.
 - ❖ MARS models tend to have a good bias-variance trade-off.
- 

MARS – Pros and Cons

Cons:

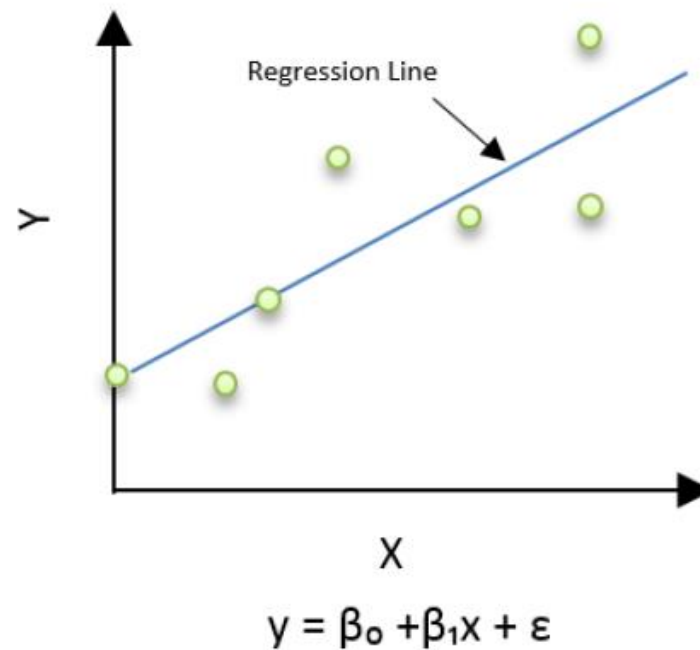
- ❖ Recursive partitioning is much faster than MARS.
 - ❖ With MARS models, as with any non-parametric regression, parameter confidence intervals and other checks on the model cannot be calculated directly (unlike linear regression models). Cross-validation and related techniques must be used for validating the model instead.
 - ❖ MARS models do not give as good fits as boosted trees, but can be built much more quickly and are more interpretable.
 - ❖ The earth, mda, and polyspline implementations do not allow missing values in predictors, but free implementations of regression trees (such as rpart and party) do allow missing values using a technique called surrogate splits.
- 

Introduction to Logistic Regression



Introduction to Logistic Regression

- ❖ When we revisit the classic OLS regression model example from before, we see that the value of the regression line is continuous in nature.



- ❖ In other words, the regression line can range in value from $-\infty$ to $+\infty$ and is unbounded.

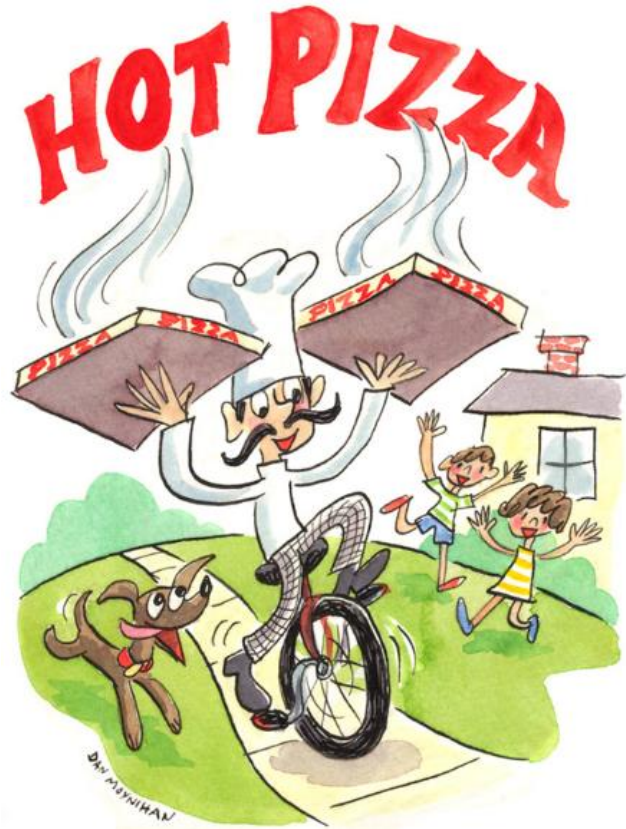
Introduction to Logistic Regression

- ❖ What if we wanted to utilize a linear regression model on a variable that is not continuous in nature?
- ❖ Lets say that we want to predict a “yes or no” variable that we have encoded into a binary response. (Ex. 0 = no and 1 = yes).
- ❖ When I personally think about binary response variables like “yes or no” I immediately think about probabilities, even though this is subconscious.

For example:

Q: Would I like to get some pizza for lunch?

A: I really would like some pizza...



Introduction to Logistic Regression



- ❖ I then ponder “how badly” I would like pizza and begin to associate a probability that I will order a slice for lunch.
- ❖ If I am really hungry and craving mozzarella, the probability will be much higher to saying “yes” (or the number 1).
- ❖ If I just ate a large sandwich, I will be full and the probability will be much lower.
- ❖ Ex. The probability I will get pizza is less than 10% now...

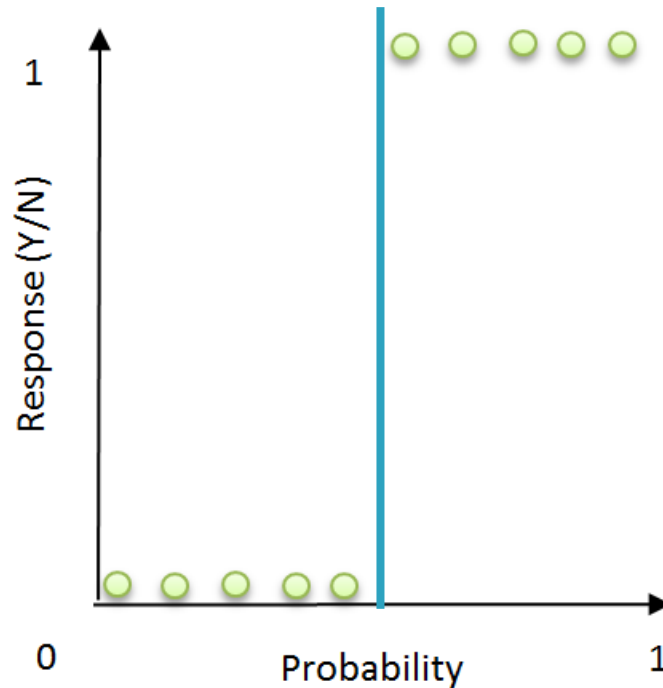
Introduction to Logistic Regression

- ❖ At some point we have to say whether or not we will get up and purchase the pizza.
- ❖ Another way to put it is "at what probability threshold will I go from not getting the pizza (a value of 0) to actually getting the pizza (a value of 1)".
- ❖ The general idea is that we state that if the probability is 0.5 or below, we won't get the pizza. If it is greater than 0.5, then let's get the pizza. (Ex. Two Face's coin)
- ❖ This cutoff probability point is a very interesting idea and the basis for a lot of statistical theory (Ex. Logit vs. Probit models).



Introduction to Logistic Regression

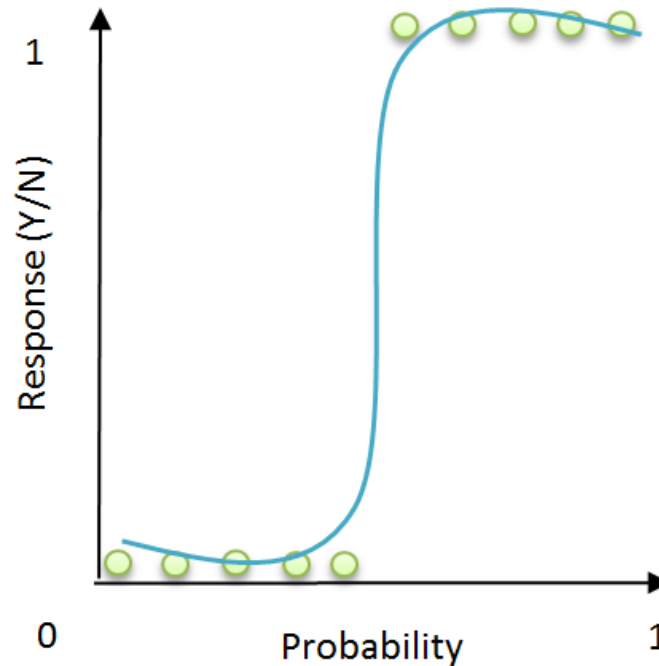
- ❖ If we were to create a sample plot with the outcome variable (yes/no) on the Y Axis and the probabilities on the X axis, we would see something like this chart:



- ❖ The blue line represents probability of 0.50 and the exact point where we shift from "no" to "yes". This point is sometimes referred to as an activation function.

Introduction to Logistic Regression

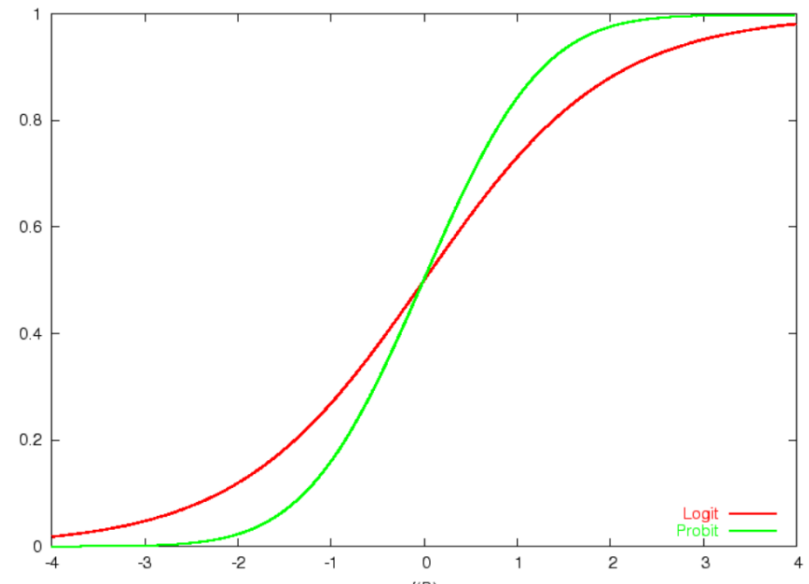
- ❖ If we were to construct a line to encapsulate this idea, it would probably look something like this:



- ❖ This shape is generally referred to as a sigmoid curve which resembles an S. This is a central concept to logistic regression.

Introduction to Logistic Regression

- ❖ There are a couple different forms of the sigmoid curve in logistic regression which are used called "logit" and "probit" based models (although there are other log-link functions).
- ❖ We can see the subtle difference in the sigmoid shape here.
- ❖ The difference in how they function is fairly minor and a skilled statistician will know when to use one variant over the other.
- ❖ We will focus our efforts on the logit model in this presentation.

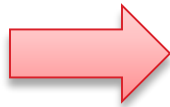


Logistic Regression

- ❖ Logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors).
- ❖ Because of the dichotomous nature (0 or 1) of the dependent variable, y , a multiple linear regression model must be transformed in order to avoid violating statistical modeling assumptions.

Linear Regression

- ❖ $Y = \beta_0 + \beta_1 X_1 + \dots + \varepsilon$



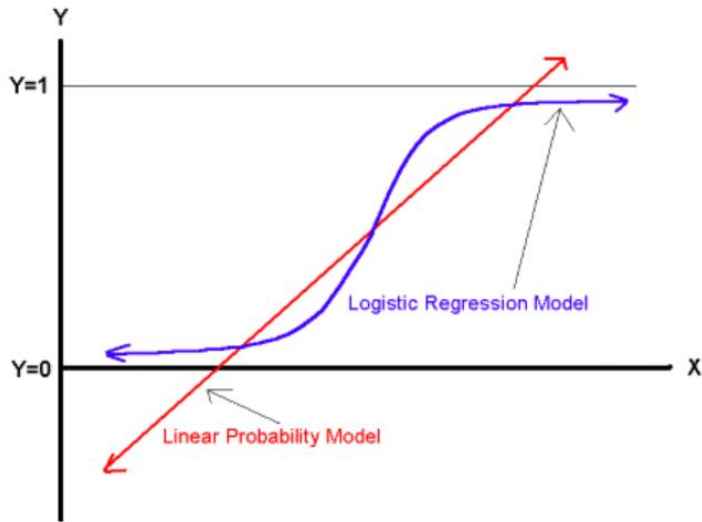
Logistic Regression

$$\ln \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + \dots + \varepsilon$$

- ❖ It is necessary that a logistic regression take the natural logarithm of the odds of the dependent variable being a case (referred to as the logit or log-odds) to create a continuous criterion as a transformed version of the dependent variable.

Logistic Regression

Comparing the LP and Logit Models



- ❖ Thus, the logit transformation is referred to as the link function in logistic regression.
- ❖ Although the dependent variable in logistic regression is binomial, the logit is the continuous criterion upon which linear regression is conducted.
- ❖ The logit of success is then fit to the predictors using linear regression analysis.
- ❖ The predicted value of the logit is converted back into predicted odds via the inverse of the natural logarithm, namely the exponential function.

Odds & Odds Ratios

- ❖ Therefore, although the observed dependent variable in logistic regression is a zero-or-one variable, the logistic regression estimates the odds, as a continuous variable, that the dependent variable is a success (a case).
- ❖ In some applications the odds are all that is needed. The basic approach is to use the following regression model:

$$\ln Odds(E) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- ❖ Odds(E) is the odds that event E occurs, namely:

$$Odds(E) = \frac{P(E)}{P(E')} = \frac{P(E)}{1 - P(E)}$$

- ❖ Where p has a value $0 \leq p \leq 1$ (i.e. p is a probability value), we can define the odds function as:

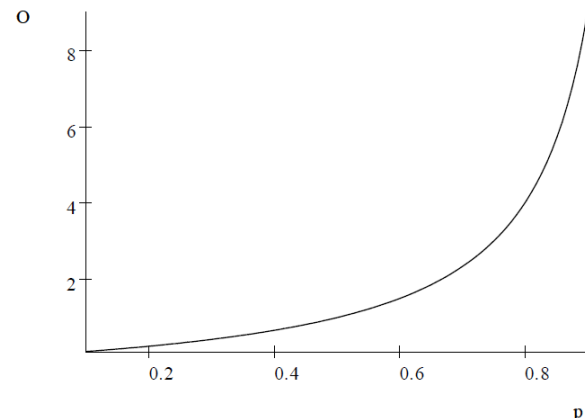
$$Odds(p) = \frac{p}{1 - p}$$

Odds & Odds Ratios

- ❖ The logit function indicates a mathematical relationship between the probability and the odd's ratio as depicted on the right.
- ❖ The concept of the odds and odds ratios can sometimes be misinterpreted by non-statisticians.
- ❖ To provide additional clarity we will use the following definitions:
 - ❖ **Odds:** the ratio of the expected number of times an event would occur to the expected number of times it will not occur.
 - ❖ **Odds Ratio:** For a binary variable (0 to 1), it is the ratio of the odds for the outcome = 1 divided by the odds of the outcome = 0.

Probability	Odds
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

$$Odds(p) = \frac{p}{1-p}$$



Odds & Odds Ratios

- ❖ The logit function is the log of the odds function, namely $\text{logit}(E) = \ln \text{Odds}(E)$, or:

$$\text{logit}(p) = \ln \text{Odds}(p) = \ln \frac{p}{1-p} = \ln p - \ln(1-p)$$

- ❖ Based on the logistic model as described before, we have:

$$\text{logit}(\pi) = \ln \text{Odds}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

- ❖ where $\pi = P(E)$. It now follows that:

$$\frac{P(E)}{1 - P(E)} = \text{Odds}(E) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon}$$

- ❖ and so:

$$p = P(E) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k}} = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k)}} = \frac{1}{1 + e^{-b_0 - \sum_{j=1}^k b_j x_j}}$$

Odds & Odds Ratios

- ❖ For our purposes we take the event E to be that the dependent variable y has value 1. If y takes only the values 0 or 1, we can think of E as success and the complement E' of E as failure.
- ❖ The odds ratio between two data elements in the sample is defined as follows:

$$R_{x_i, x_j} = \frac{\text{Odds}(x_{i1}, \dots, x_{ik})}{\text{Odds}(x_{j1}, \dots, x_{jk})} = \frac{e^{b_0 + \sum_{m=1}^k b_m x_{im}}}{e^{b_0 + \sum_{m=1}^k b_m x_{jm}}} = e^{\sum_{m=1}^k b_m (x_{im} - x_{jm})}$$

- ❖ Using the notation $p_x = P(x)$, the log odds ratio of the estimates is defined as:

$$\text{logit}(p_{x+1}/p_x)$$

Maximum Likelihood Estimates

- ❖ Although logistic regression model, $\text{logit}(y) = \alpha + \beta x$ looks similar to a simple linear regression model, the underlying distribution is binomial and the parameters, α and β cannot be estimated in the same way as for simple linear regression.
- ❖ Instead, the parameters are usually estimated using the method of maximum likelihood of observing the sample values. Maximum likelihood will provide values of α and β which maximize the probability of obtaining the data set.
- ❖ The maximum likelihood estimate is that value of the parameter that makes the observed data most likely. Define p_i as the probability of observing whatever value of y was actually observed for a given observation:

$$p_i = \begin{cases} \Pr(y_i = 1 | x_i) & \text{if } y_i = 1 \text{ is observed} \\ 1 - \Pr(y_i = 1 | x_i) & \text{if } y_i = 0 \text{ is observed} \end{cases}$$

- ❖ So, for example, if the predicted probability of the event occurring for case i was .7, and the event did occur, then $p_i = 0.7$. If, on the other hand, the event did not occur, then $p_i = 0.30$.

Maximum Likelihood Estimates

- ❖ If the observations are independent, the likelihood equation is:

$$L(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = \prod_{i=1}^N p_i$$

- ❖ The likelihood tends to be an incredibly small number, and it is generally easier to work with the log likelihood.
- ❖ Ergo, taking logs, we obtain the log likelihood equation:

$$\ln L(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = \sum_{i=1}^N \ln p_i$$

- ❖ The maximum likelihood estimates are those values of the parameters that make the observed data most likely. That is, the maximum likelihood estimates will be those values which produce the largest value for the likelihood equation (i.e. get it as close to 1 as possible; which is equivalent to getting the log likelihood equation as close to 0 as possible).

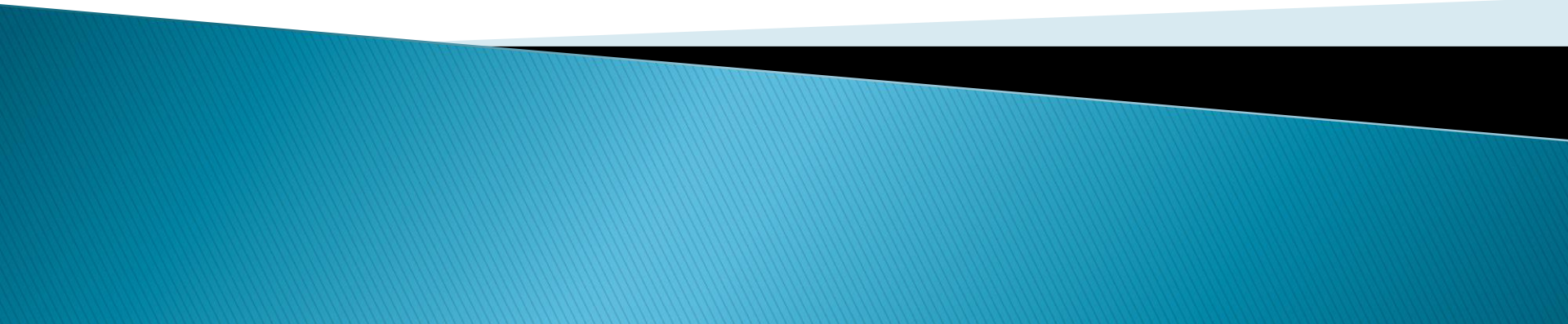
Properties of MLE

- ❖ The ML estimator is consistent. As the sample size grows large, the probability that the ML estimator differs from the true parameter by an arbitrarily small amount tends toward 0.
- ❖ The ML estimator is asymptotically efficient, which means that the variance of the ML estimator is the smallest possible among consistent estimators.
- ❖ The ML estimator is asymptotically normally distributed, which justifies various statistical tests.



In 1922, Ronald Fisher introduced the method of maximum likelihood.

Introduction to Survival Analysis



Introduction to Survival Analysis



- ❖ Survival Analysis is a set of techniques that study of “events of interest” where the outcome variable is the time until the occurrence of an event of interest.
- ❖ Survival Analysis is the study of **TIME**

Survival Analysis attempts to answer questions such as:

- ❖ What is the proportion of a population which will survive past a certain time?
- ❖ Of those that survive, at what rate will they die or fail?
- ❖ Can multiple causes of death or failure be taken into account?
- ❖ How do particular circumstances or characteristics increase or decrease the probability of survival?

Brief History of Survival Analysis

- ❖ These “events of interest” in the medical field typically represents the mortality rate for experimental drugs and medicines.
- ❖ The analysis generally produces a timeframe until death which is why the technique is referred to as “Survival Analysis”.
- ❖ An experimental medicine (amongst other considerations) is generally seen to be effective when the survival rate is extended for the experiment beyond the control.
- ❖ Survival Analysis is the technique which allows for this determination to be made.



Mortality and Immortality by William Michael Harnett

Survival Analysis Applications

Here are some applications of Survival Analysis:

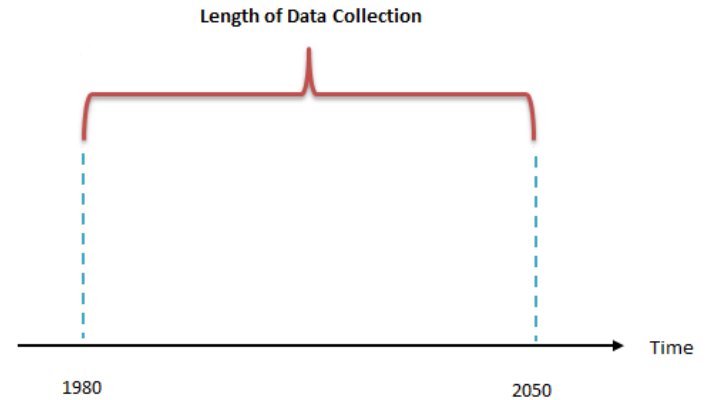
- ❖ Medical Drug Testing
- ❖ Reliability Analysis in Engineering
- ❖ Duration Modeling in Economics
- ❖ Event History Analysis in Sociology
- ❖ Criminological Analysis
- ❖ Customer Lifetime Modeling
- ❖ Actuarial Science / Risk Modeling
- ❖ Botany / Zoology



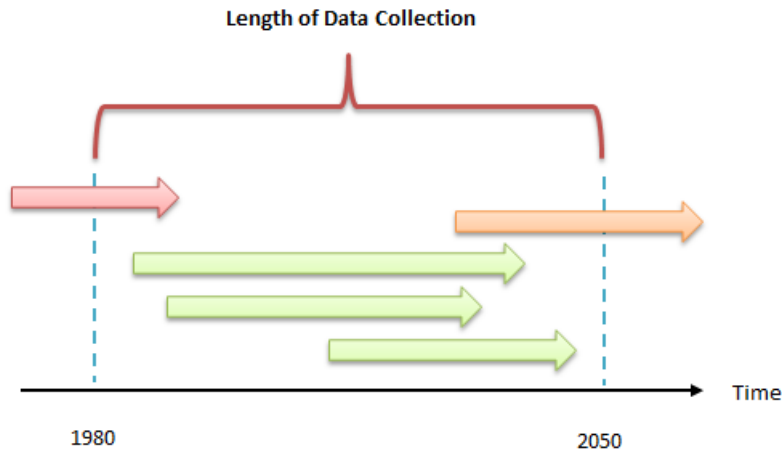
Censoring

In order to successfully prepare a survival analysis, we must first introduce the concept of censoring.

- ❖ Lets imagine that we are trying to understand factors which can cause the lifespan of a human to be shortened. We will collect the data from 1990 – 2050.
- ❖ When imagining our dataset, we might think about variables such as diet, exercise, socioeconomic conditions, etc... however we will also think about the whether we have data related to full lifespan of the person.
- ❖ Are we collecting data from the date of birth to the date of death for a particular subject?
- ❖ What if the subject is still alive after the study? This is called censoring.



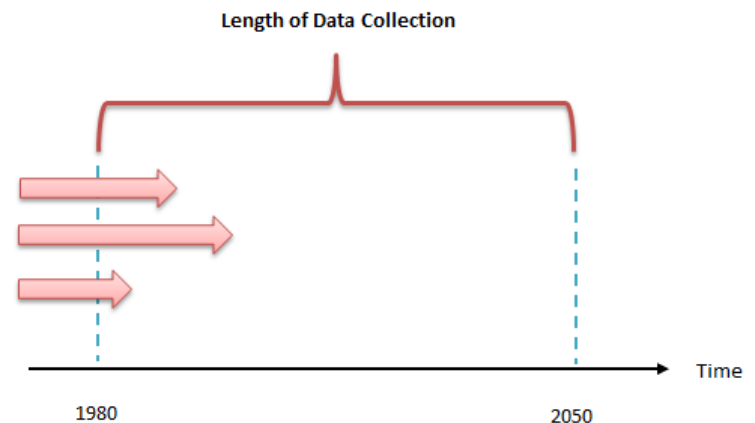
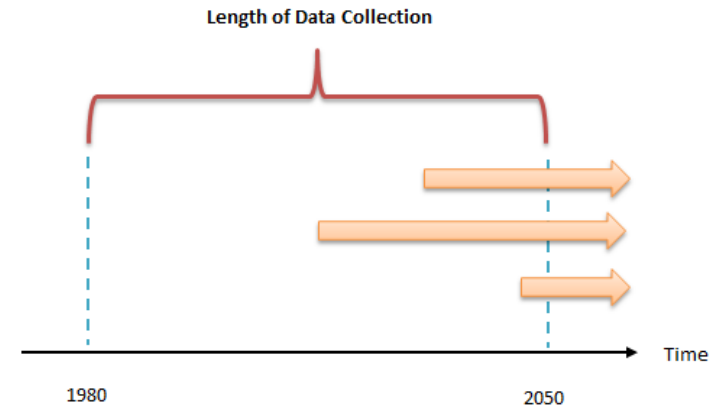
Censoring



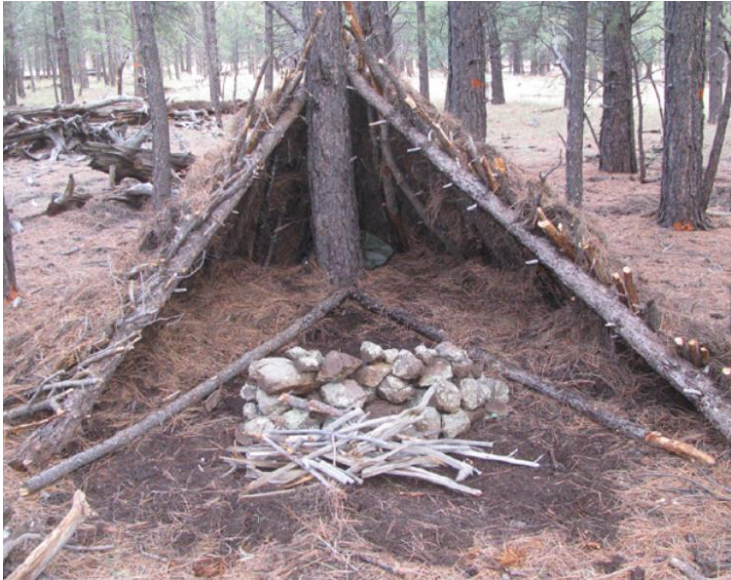
- ❖ If we are collecting random samples of people to include within our study, then it is entirely possible to believe that we will find individuals that do not fit neatly into the walls of 1980 – 2050.
- ❖ Let us showcase this idea by representing the lives of individual subjects with an arrow.
- ❖ Individuals who fit within these walls are shown with a green arrow and those which do not (censored) will be shown by either a red or orange arrow.

Censoring

- ❖ Therefore, we can think about censoring as a form of a missing data problem.
- ❖ Additionally there are 2 types of censoring that we should be aware of: **Right Censoring** & **Left Censoring**.
- ❖ **Right Censoring** will occur for those subjects whose birth date is known but who are still alive when the study ends (Orange Arrows).
- ❖ If a subject's lifetime is known to be less than a certain duration, the lifetime is said to be **Left Censored** (Red Arrows).
- ❖ The data scientist will need to formulate a plan on how to treat censored observations during the EDA.



Survival Function



- ❖ The Survival function is the probability that the time of death (event) is greater than some specified time.

$$S(t) = \Pr(T > t)$$

The Survival Function is composed of:

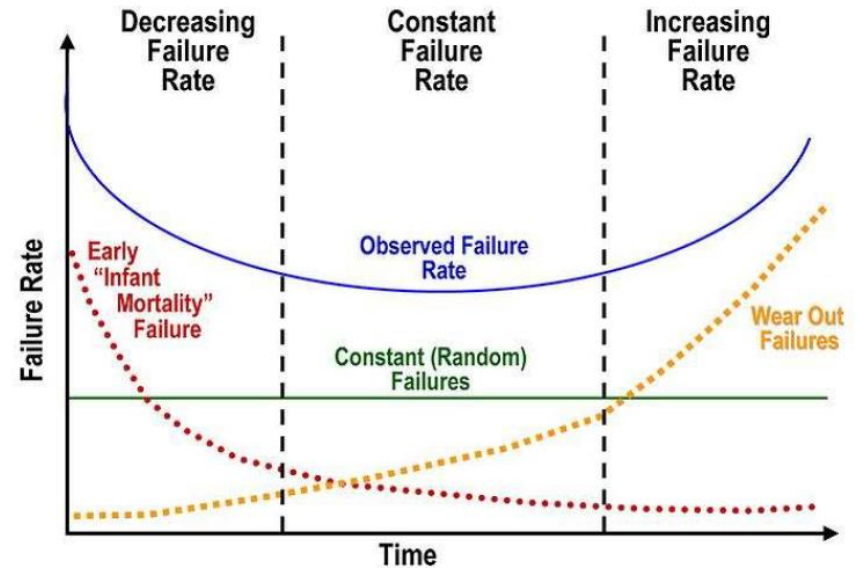
- ❖ The underlying Hazard function (How the risk of death per unit time changes over time at baseline covariates).
 - ❖ The effect parameters (How the hazard varies in response to the covariates) .
-
- ❖ Usually one assumes $S(0) = 1$, although it could be less than 1 if there is the possibility of immediate death or failure.

Hazard Function

- ❖ The hazard function, conventionally denoted λ , is defined as the event rate at time t conditional on survival until time t or later (that is, $T \geq t$):

$$\lambda(t) = \lim_{dt \rightarrow 0} \frac{\Pr(t \leq T < t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}$$

- ❖ The hazard function must be non-negative, $\lambda(t) \geq 0$, and its integral over $[0, \infty]$ must be infinite, but is not otherwise constrained.
- ❖ It may be increasing or decreasing, non-monotonic, or discontinuous (as depicted on the right).
- ❖ Hazard and Survival functions are mathematically linked - by modelling the Hazard we obtain the Survival Function.

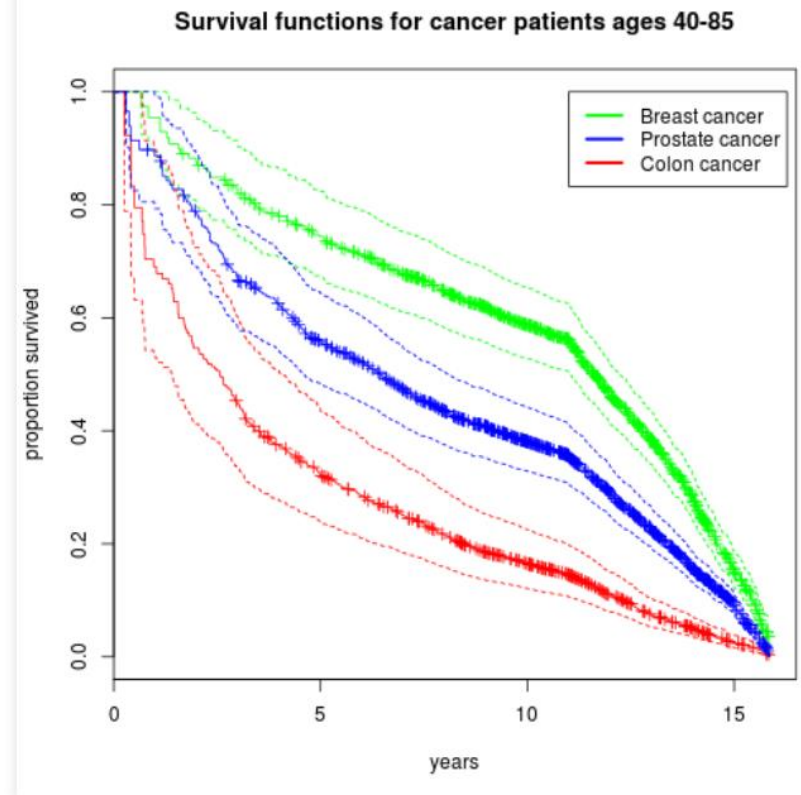


Survival Analysis

Here is an example of the survival functions for individuals with differing type of cancers.

- ❖ The X axis consists of a length of time and the Y axis is the Survival Probability %.
- ❖ In this example, all of the subjects were alive at the start of the study. $Y = 1$ or $S(0) = 100\%$.
- ❖ At year 10, approximate 20% of individuals with Colon Cancer survived, 40% survived with Prostate Cancer, and 65% survived with Breast Cancer.

Key Thought: When evaluating these charts for analytical insights, look for the separation between the various curves.



Cox Proportional- Hazards Model

- ❖ Sir David Cox observed that if the proportional hazards assumption holds (or, is assumed to hold) then it is possible to estimate the effect parameter(s) without any consideration of the hazard function.
- ❖ Let Y_i denote the observed time (either censoring time or event time) for subject i , and let C_i be the indicator that the time corresponds to an event:

$$\lambda(t|X) = \lambda_0(t) \exp(\beta_1 X_1 + \cdots + \beta_p X_p) = \lambda_0(t) \exp(X\beta')$$

- ❖ This expression gives the hazard at time t for an individual with covariate vector (explanatory variables) X . Based on this hazard function, a partial likelihood can be constructed from the datasets as:

$$L(\beta) = \prod_{i:C_i=1} \frac{\theta_i}{\sum_{j:Y_j \geq Y_i} \theta_j}$$

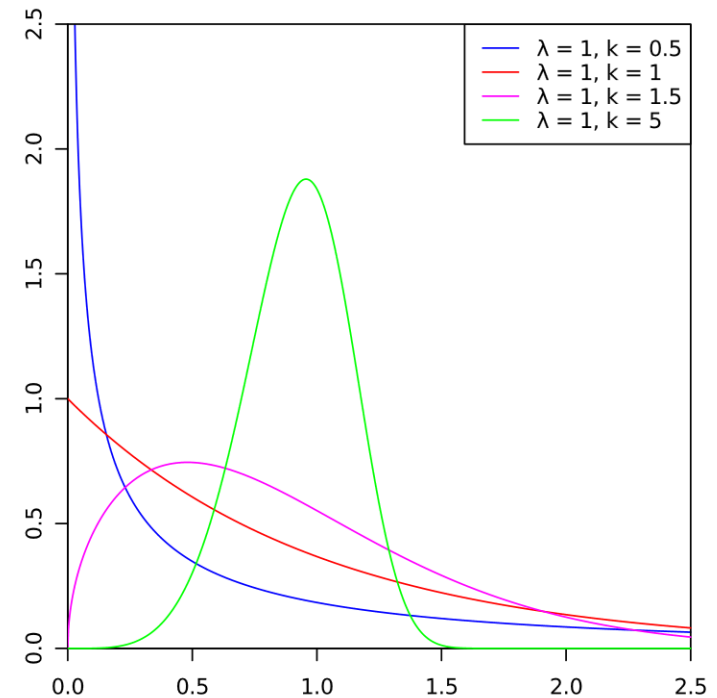
- ❖ where $\theta_j = \exp(X_j \beta')$ and X_1, \dots, X_n are the covariate vectors for the n independently sampled individuals in the dataset (treated here as column vectors).

Cox Proportional- Hazards Model

- ❖ The corresponding log partial likelihood is:

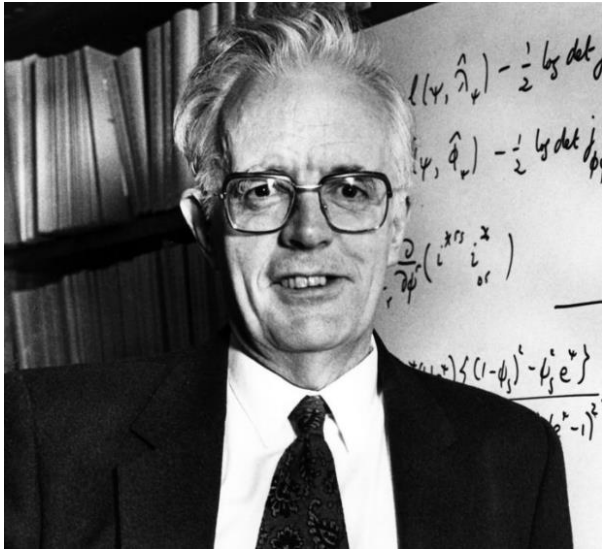
$$\ell(\beta) = \sum_{i:C_i=1} \left(X_i \beta' - \log \sum_{j:Y_j \geq Y_i} \theta_j \right)$$

- ❖ This function can be maximized over β to produce maximum partial likelihood estimates of the model parameters.
- ❖ The Cox Proportional Hazard's model may be specialized (changing the underlying baseline function) if a reason exists to assume that the baseline hazard follows a particular form.
- ❖ An example would include the Weibull Distribution (Weibull Hazard function).



Weibull Probability Distributions
for various λ values

Cox Proportional- Hazards Model



Sir David Cox

- ❖ The most common model used to determine the effects of covariates on survival.

It is a semi-parametric model:

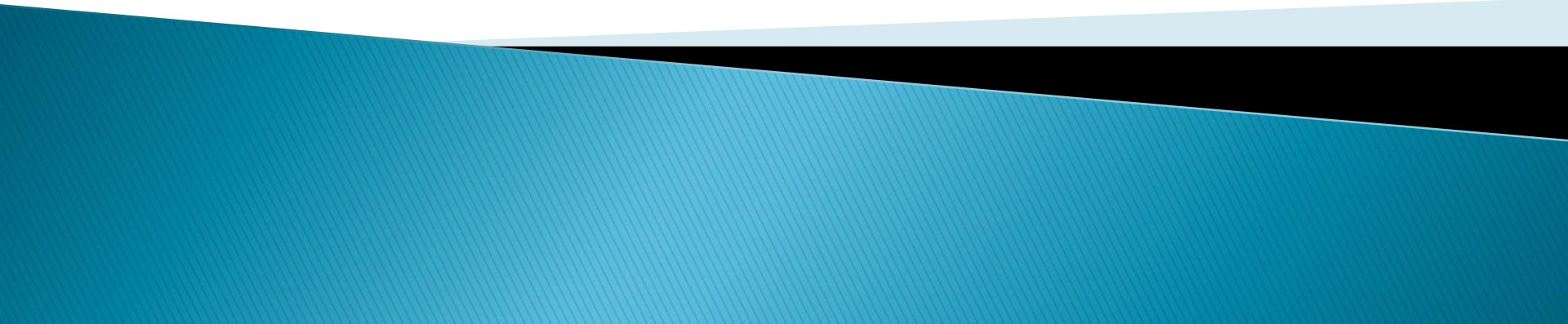
- ❖ The baseline hazard function is unspecified.
- ❖ The effects of the covariates are multiplicative.
- ❖ Doesn't make arbitrary assumptions about the shape/form of the baseline hazard function.

Key Assumptions:

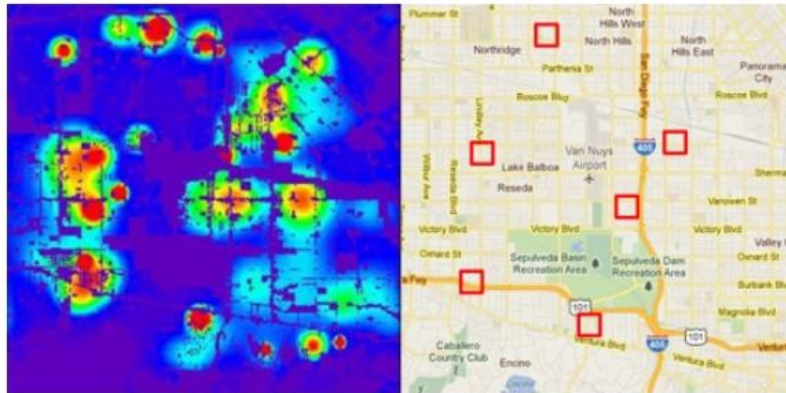
- ❖ Covariates multiply the hazard by some constant.
- ❖ Ex. A drug may halve a subjects risk of death at any time.
- ❖ The effect is the same at any time point.

Violating the PH assumption can seriously invalidate your model!!!

Practical Example – Predicting Crime in the US



Crime Prediction



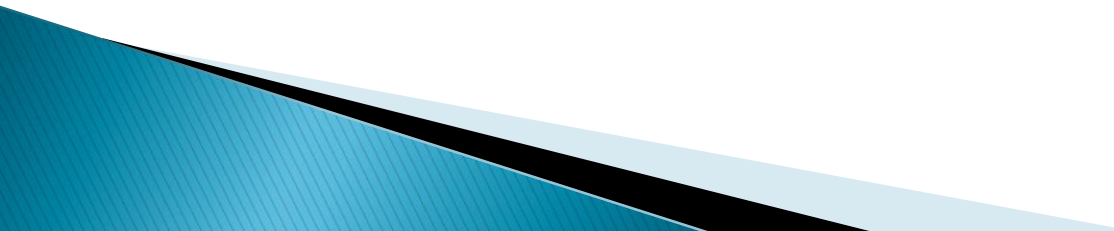
- ❖ Being able to reasonably predict the individuals who are most likely to commit crimes is of significant value to the criminal justice system.
- ❖ Specifically if the individual is within custody and a judge needs to decide whether they are at risk for recidivism.
- ❖ Data Science can help to shed light on the underlying factors, and when used appropriately, can aid the judicial system in decision making, particularly in pre-trial.
- ❖ This improvement in decision making reduces the costs of inmate housing and has a societal benefit through keeping the low risk offenders out of the prison and the high risk offenders behind bars.
- ❖ This case study is one from my consulting career which I have personally prepared from the data munging stage to the final predictive model.

Crime Prediction

- ❖ In addition to writing up my findings in an official report, I have presented the techniques to a major US cities leadership committee where they have used the results in the following ways:
 - ❖ To build a risk assessment system that is leveraged in pre-trial decision making for recidivism.
 - ❖ Refinement of the build plan for a multi-million dollar prison facility based upon the predictive model results.
- ❖ Due to the sensitive nature, I will be randomizing the data/results to fit the tutorial.
- ❖ I will however go through some of the “real-world” considerations that I had throughout the process and showcase the preliminary results of our logistic regression model.



Why use Logistic Regression?

- ❖ The odds are high that if you are reviewing this material, you have a strong interest in data science and know all of the latest and greatest Machine Learning techniques.
 - ❖ However, I would suspect that a substantial number of us data scientists could not explain most of these algorithms to business decision makers in an intuitive manner. (Ex. SVM higher dimensions, Neural Networks).
 - ❖ When I first started on this project, I was working under the former mayor of multiple major US cities, a gifted attorney, and a Harvard professor.
 - ❖ He had understood the value of advanced analytics and launched a series of Big Data initiatives during his tenure as deputy mayor in New York City.
 - ❖ While preparing my plan with him, I was fairly shocked to realize the difference in understanding between the public and private sectors in regards to statistical modeling.
- 

Why use Logistic Regression?

- ❖ The reality is that the government sector is a little slower to adopt the technological advancements in ML.
- ❖ He expressed that if we are to be successful in establishing the value of these techniques within the judicial system, that we would need to emphasize interpretability of the algorithm over the predictive performance.
- ❖ This initially was somewhat counterintuitive to me but I now understand the rationale behind this.
 - ❖ If you cant explain what the algorithm is doing in laymen's terms, the users (judges) will become confused and then begin to doubt the value of the technique.
 - ❖ Once they are comfortable with the nuances of machine learning algorithms and its integrated into the judicial process, then we expand on the predictive framework emphasizing predictive accuracy.
- ❖ This is ultimately why I settled on using a **logistic regression** to demonstrate how the tool can be built (and using a random forest model for higher predictive performance at the expense of interpretability).

Understanding the data

- ❖ The final cohort was constructed from the various databases in the prison system, judicial system, and bail bond system. (n=4,600) The dataset was also pared down to ensure that there were no instances of right or left censoring.

CASE-ID	Gender	Age	Non White	Felony	Month at Address	Employment	FTAEver	Drug Charge	Property Charge	Personal Charge	Num of Priors	Charge Class	Prior Felonies	Prior Misdmr	Case Charge	FTA
A12345	1	61	0	1	1	0	0	0	1	0	0	C	0	0	1	0
A12346	1	62	0	0	744	1	0	0	0	1	0	A	0	0	1	1
A12347	1	61	1	1	0	0	0	0	0	0	4	A	0	4	1	0
A12348	1	65	1	0	6	0	0	0	1	0	0	D	0	0	1	0
A12349	1	61	1	0	4	0	0	0	0	0	0	D	0	0	1	0
A12350	1	59	1	1	4	0	1	1	0	1	14	B	2	12	1	0
A12351	1	68	1	0	3	0	0	0	0	1	2	D	1	1	1	0
A12352	1	63	0	1	1	0	0	0	0	0	1	B	0	1	1	0
A12353	1	59	1	1	12	0	1	1	0	0	8	B	2	6	1	0
A12354	1	59	1	1	12	0	1	1	1	1	10	A	3	7	1	1
A12355	1	59	1	1	12	0	1	1	0	0	18	A	6	12	1	0
A12356	1	59	1	1	12	0	1	1	0	0	22	A	6	16	1	0
A12357	1	62	1	1	744	1	1	1	1	1	7	A	5	2	1	0
A12358	1	59	1	0	5	0	1	1	0	0	3	A	1	2	1	0
A12359	1	59	1	1	4	0	0	0	1	0	0	D	0	0	1	0
A12360	1	72	1	0	12	1	0	0	0	0	4	D	2	2	1	1
A12361	1	58	0	1	696	0	0	0	0	1	6	C	6	0	1	0
A12362	1	59	1	0	708	1	1	1	0	0	0	D	0	0	1	0

- ❖ This was important to ensure that our sample was representative of the population at large. Coincidentally, the dataset now allows for us to readily perform a survival analysis.

Understanding the data

- ❖ Some variables were constructed around the expert opinions of criminologists and extensive literature review.
- ❖ I spent a large portion of my time just trying to understand how to connect the pieces of data when working through the various disparate information systems.
- ❖ I like to believe that I am fairly proficient in SQL through my time in the business world, however, getting the data into modeling ready form posed significant hurdles that business users rarely encounter.
- ❖ In total, the data munging process took about 3 months of intensive on site work.

Validated Risk Factors	Study
Prior FTA	Virginia, 2 or more (VanNostrand, 2003) New York City (Siddiqi, 2006) Harris County, TX (Austin and Murray, 2008) Hennepin County, MN (Podkopacz, 2006) Allegheny County, PA, 2 or more (Pretrial Justice Institute, 2007) Ohio (Lowenkamp, Lemke and Latesasa, 2008) Federal System (VanNostrand and Keebler, 2009)
Prior convictions	Virginia, 1 or more (VanNostrand, 2003) New York City, Prior misdemeanor convictions (Siddiqi, 2006) Harris County, TX (Austin and Murray, 2008) Hennepin County, MN, Having a higher number (Podkopacz, 2006) Allegheny County, PA, 2 or more (Pretrial Justice Institute, 2007) Ohio, 3 or more prior jail incarcerations (Lowenkamp, Lemke and Latesasa, 2008) Federal System (VanNostrand and Keebler, 2009)
Present charge a felony	Virginia (VanNostrand, 2003) Federal System (VanNostrand and Keebler, 2009)
Being unemployed	Virginia (VanNostrand, 2003) New York City (Siddiqi, 2006) Harris County, TX (Austin and Murray, 2008) Hennepin County, MN (Podkopacz, 2006) Ohio (Lowenkamp, Lemke and Latesasa, 2008) Federal System (VanNostrand and Keebler, 2009)
History of drug abuse	Virginia (VanNostrand, 2003) Federal System (VanNostrand and Keebler, 2009) Ohio (Lowenkamp, Lemke and Latesasa, 2008)
Having a pending case	Virginia (VanNostrand, 2003) New York City (Siddiqi, 2006) Federal System (VanNostrand and Keebler, 2009)

19 Austin, J. and T. Murray (2009) *Re-Validation of the Actuarial Risk Assessment Instrument for Harris County Pretrial Services*. Washington, D.C.: The JFA Institute. Clark, J. and D. Levin (2007) *The Transformation of Pretrial Services in Allegheny County, Pennsylvania: Development of Best Practices and Validation of Risk Assessment*. Washington, D.C.: Pretrial Justice Institute. Lowenkamp, C., R. Lemke and E. Latesasa (2008) *The Development and Validation of a Pretrial Screening Tool*. *Federal Probation*. Vol. 72 (3). Podkopacz, M. (2006) *Fourth Judicial District of Minnesota Pretrial Evaluation: Scale Validation Study*. Power Point Presentation. Siddiqi, Q. (2006) *Predicting the likelihood of pretrial re-arrest for violent felony offenses and examining the risk of pretrial failure among New York City defendants: An analysis of the 2001 dataset*. New York, NY: New York City Criminal Justice Agency, Inc. VanNostrand, M. (2003) *Assessing risk among pretrial defendants in Virginia: The Virginia pretrial risk assessment*. Richmond, VA: Virginia Department of Criminal Justice Services. VanNostrand, M., and G. Keebler (2009) *Pretrial Risk Assessment in Federal Court*. *Federal Probation*. Vol. 72 (2).

Understanding the data

- ❖ There were a total of 15 variables which were included as potential predictor variables and with response variable being “FTA” or failure to appear, a proxy for recidivism.
- ❖ Here is the listing of the final variables used for the analysis and a brief description:

Gender: Categorical Variable with a 0 = Female, 1 = Male.

Age: This is the age of the defendant at the time of bail interview.

NonWhite: Categorical Variable with 0 for Caucasian and 1 for other races

FelonyIndicator: Represents whether or the most serious offense stemming from the arrest for the current case ID was a Felony (1) or a Misdemeanor (0).

Month at Address: # of Months at current address.

Employment: Categorical Variable with 0 for Not Employed and 1 for Employed.

FTAEver: Categorical Variable with 0 representing the defendant never having failed to appear in court and 1 representing the defendant indicating at least one failure to appear; self-reported from bail interview.

DrugCharge: Categorical Variable with a value of 0 for no drug charge stemming from the arrest for a case ID and 1 with a drug charge.

PropertyCharge: Categorical Variable with a value of 0 for no property charge stemming from the arrest for a case ID and 1 with a property charge.

PersonalCharge: Categorical Variable with a value of 0 for no personal charge stemming from the arrest for a case ID and 1 with a personal charge.

NumofPriors: The total count of the prior charges for a specific individual's Gallery ID.

ChargeClass: Class variable; Type A, B, C, or D crime.

NumofMisdmr: The total count of the prior misdemeanor charges for a specific individual's Gallery ID.

NumofFelonies: The total count of the prior felonies charges for a specific individual's Gallery ID. Calculated

NumofCaseCharge: The total count of charges related to a specific Case ID.

Understanding the data

- ❖ The “FTA” definition we used is depicted as a dichotomous categorical variable and formally defined as follows:

0 - An individual who has not had a failure to appear to court between the time the defendant was released from either the jail and their initial court trial in the county.

1 - An individual who has had a failure to appear to court between the time the defendant was released from either the jail and their initial court trial in the county.

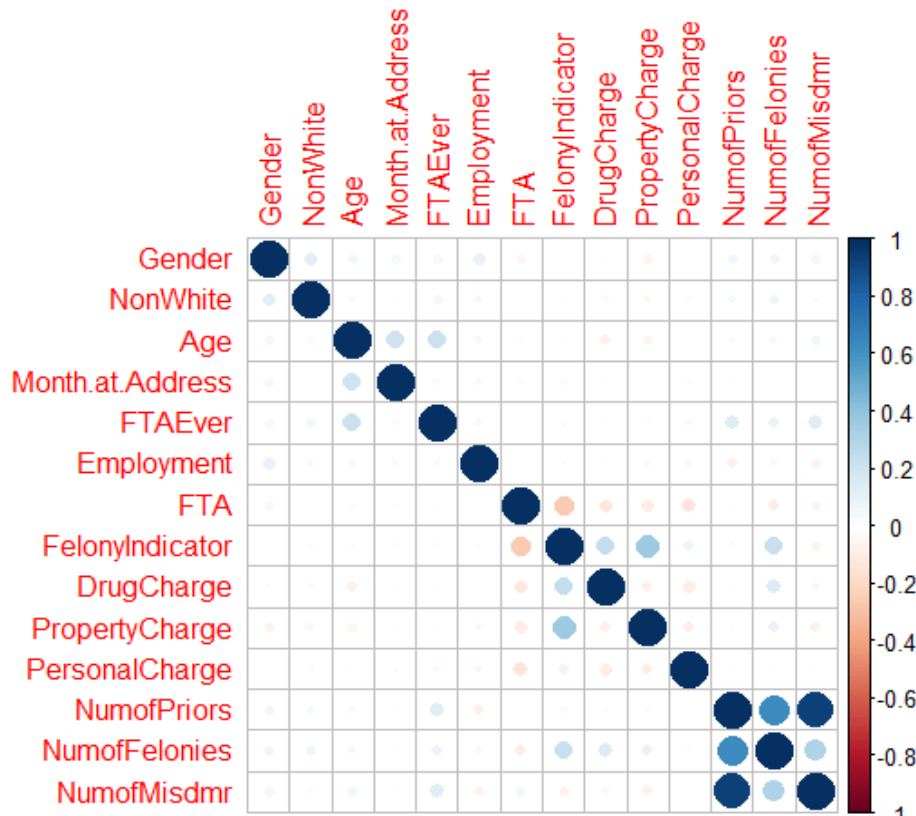
- ❖ The breakdown of the various considerations for the Personal, Property, and Drug Arrest types is as follows:

Personal – Assault, Battery, Kidnapping, Homicide, & Sexual

Property – Larceny, Robbery, Burglary, Arson, Embezzlement, Forgery, Receipt of Stolen Goods

Drug – Dealing, Possession, Prescription, and various drug types

Exploratory Data Analysis



- ❖ The correlation matrix provides a visual indication of the relative strength of the correlation between each variable.
- ❖ A large red bubble depicts a more significant negative correlation whereas a large blue bubble signifies a stronger positive correlation.
- ❖ By referencing the 'FTA' row, this correlation matrix is showing that the variables, FelonyIndicator, DrugCharge, and PersonalCharge, all of which have red bubbles associated with them, may be indicators of potential predictor variables within our model.
- ❖ In general, there does not appear to be extremely strong correlations when we visually inspect the data.

Model Selection Procedure

- ❖ The next step in the overall approach involved looking at some automated variable selection procedures, such as forward selection, backward selection, and stepwise selection:
- ❖ The forward selection procedure had produced the following output in R:

```
Step:  AIC=5741.43  
FTA ~ FelonyIndicator + PersonalCharge + ChargeClass + DrugCharge +  
      Gender + PropertyCharge + NumofMisdmr + NumofPriors + Age +  
      NumofCaseCharge
```

	Df	Deviance	AIC
<none>		933.05	5741.4
+ NonWhite	1	932.81	5742.2
+ FTAEver	1	932.92	5742.8
+ Month.at.Address	1	932.98	5743.1
+ Employment	1	933.05	5743.4

- ❖ Only variables which contributed to the reduction of the AIC statistic (lower is better) were incorporated into the final model. The variables "Month at Address", "Employment", "NonWhite", and "FTAEver" were found to be statistically insignificant and did not improve the models performance, thus, were removed from the model.

Key Point: The final model we presented included the "Employment" and "FTA Ever" variable despite the fact that they did not meet the $p < 0.05$ threshold here. The thresholds were loosened slightly to accommodate subject matter expertise in our model.

Model Selection Procedure

- ❖ The next step was to produce a binomial logistic regression model from the results. We are utilizing the glm model within the base R.

```
Call:
glm(formula = FTA ~ FelonyIndicator + PersonalCharge + chargeclass +
    DrugCharge + Gender + PropertyCharge + NumofMisdmr + Age +
    NumofCaseCharge, family = "binomial", data = trainData)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9805  -1.1592   0.7242   0.8231   2.0861
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.201115	8.970	< 2e-16	***
FelonyIndicator	0.171012	-10.730	< 2e-16	***
PersonalCharge	0.090600	-9.087	< 2e-16	***
chargeclassB	0.090362	-0.367	0.713309	
chargeclassC	0.130897	2.809	0.004966	**
chargeclassD	0.187740	5.493	3.94e-08	***
DrugCharge	0.088679	-5.313	1.08e-07	***
Gender	0.079869	-3.529	0.000417	***
PropertyCharge	0.102558	-3.166	0.001547	**
NumofMisdmr	0.008632	2.364	0.018084	*
Age	0.002999	-2.172	0.029888	*
NumofCaseCharge	0.147476	-1.994	0.046140	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

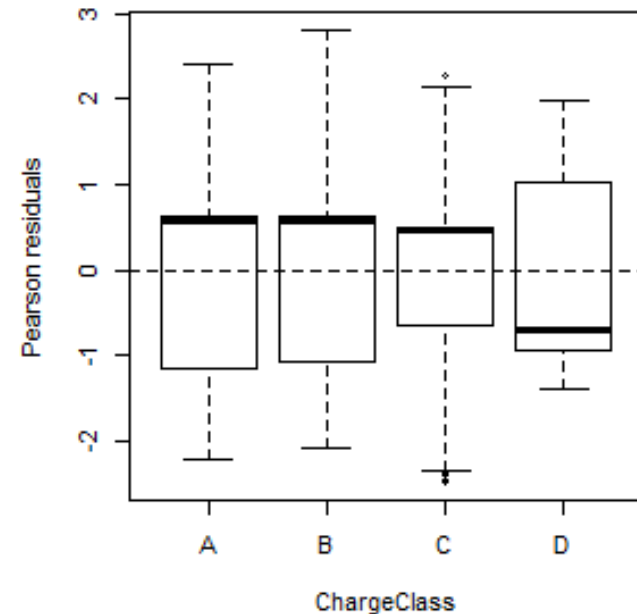
Null deviance: 5932.6 on 4590 degrees of freedom
Residual deviance: 5463.3 on 4579 degrees of freedom
AIC: 5487.3

Number of Fisher Scoring iterations: 4

- ❖ The diagnostics of the model produced indicate that each of the variables included are statistically significant at a confidence level $P < 0.05$.

Model Selection Procedure

- ❖ The ChargeClass has been retained in the model due to the overall contributions to the model even though some of the classes (Type= B) are not statistically significant.
- ❖ This boxplot shows that the Pearson residuals for the charge class variable are consistent.
- ❖ This indicates that the variable does not suffer from heteroscedasticity (variability of the FTA is unequal across the range of values for the different ChargeClass).
- ❖ This is an indication that the ChargeClass variable can be retained in the model.



Presentation of the Model Algorithm

$P(FTA)$

$$= \frac{1}{e^{-(1.72 - 1.631(FI) - 0.923(PC) - 0.003(CC1) + 0.567(CC2) - 1.041(CC3) - 0.677(DC) - 0.481(G) - 0.222(PrC) + 0.09(\#M) + 0.03(Age) - 0.3(\#C))}}$$

Where:

FI = FelonyIndicator	CC3 = ChargeClassD	#M = NumofMisdmr
PC = PersonalCharge	DC = DrugCharge	A = Age
CC1 = ChargeClassB	G = Gender	#C = NumofCaseCharge
CC2 = ChargeClassC	PrC = PropertyCharge	

$$p = P(E) = \frac{1}{1 + e^{-b_0 - \sum_{j=1}^k b_j x_j}}$$

- ❖ The P(FTA) represents the probability that an offender will have a failure to appear based off of the models parameters.
- ❖ The default threshold is $p = 0.50$, where a value of probability less than 0.50 will produce a 0 (no FTA) and probability greater than or equal to 0.50 will produce a 1 (will commit an FTA). A discussion around calibrating the cutoff threshold will be addressed later.
- ❖ The probability calculated can also be evaluated a potential risk score for a failure to appear outcome. If the probability generated for a particular case is 0.05 this implies that there is a 5% risk for a FTA outcome based upon the specified model.

Odds Ratios

- ❖ The interpretations of the coefficients are not intuitive within a logistic regression model and require further explanation.
- ❖ There is an alternative representation of the variables which can be used to help drive decision making.
- ❖ This approach involves transforming the variables into Odds Ratio's which can be then interpreted in a more intuitive fashion.
- ❖ Odds Ratio's are used to compare the relative odds of the occurrence of FTA, given exposure to the variable of interest.

95% Confidence Interval			
Variable	Odds Ratio	Lower	Upper
(Intercept)	6.074	4.154	9.100
FelonyIndicator	0.160	0.114	0.222
PersonalCharge	0.439	0.368	0.524
ChargeClassB	0.967	0.811	1.156
ChargeClassC	1.444	1.122	1.875
ChargeClassD	2.805	1.949	4.071
DrugCharge	0.624	0.525	0.743
Gender	0.754	0.644	0.881
PropertyCharge	0.723	0.591	0.884
NumofMisdmr	1.021	1.004	1.039
Age	0.994	0.988	0.999
NumofCaseCharge	0.745	0.550	0.968

Odds Ratios

95% Confidence Interval			
Variable	Odds Ratio	Lower	Upper
(Intercept)	6.074	4.154	9.100
FelonyIndicator	0.160	0.114	0.222
PersonalCharge	0.439	0.368	0.524
ChargeClassB	0.967	0.811	1.156
ChargeClassC	1.444	1.122	1.875
ChargeClassD	2.805	1.949	4.071
DrugCharge	0.624	0.525	0.743
Gender	0.754	0.644	0.881
PropertyCharge	0.723	0.591	0.884
NumofMisdmr	1.021	1.004	1.039
Age	0.994	0.988	0.999
NumofCaseCharge	0.745	0.550	0.968

- ❖ The Odds Ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome.
- ❖ OR=1 Variable does not affect odds of FTA.
- ❖ OR>1 Variable associated with higher odds of FTA.
- ❖ OR<1 Variable associated with lower odds of FTA.
- ❖ The Odds Ratio can compare the magnitude of various risk factors for that outcome.

Odds Ratios

These Odds Ratio's can be interpreted in the following manner:

- ❖ **Gender** = An individual who is Male has an odds that is 0.754 times less likely to have a failure to appear than a female, controlling for the other variables (holding the values constant).
- ❖ **Age** = For each year an individual ages, there is a 0.6% decrease in the odds that they will have a failure to appear, controlling for the other variables. The odds value being so close to 1 implies that this should not have a considerable impact in determining risk factors.
- ❖ **ChargeClassD** = An individual with a case ID where the most serious offense is a charge class of D has an odds of FTA that is 2.8 times higher than those who do not, controlling for other variables.
- ❖ **DrugCharge** = An individual with a drug charge has an odds of FTA that is 0.624 times lower than those who do not, controlling for other variables.

Odds Ratios

- ❖ The Odds Ratio's can be seen as indicators of an underlying risk for a failure to appear.
- ❖ With this understanding, the variables which indicate the greatest risk of a FTA include whether the crime is Misdemeanor with a stronger odds that a Class C (odds 1.4 or 2.8) and also individuals who commit class D Felonies.

For example, let's consider the Class D type (Felony specific).

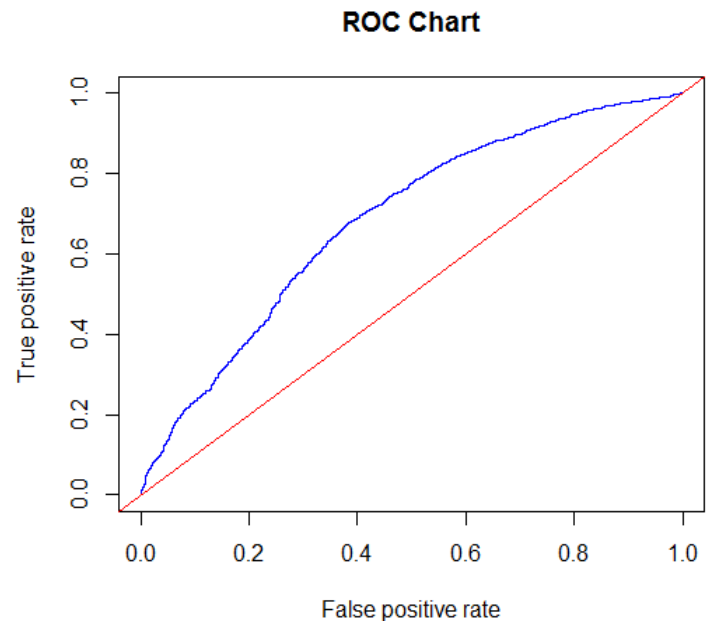
- ❖ There is a substantial increase in the odds (2.8 times) for having a FTA when the arrestee has this lower classification crime.
- ❖ This insight can be used by a judge when considering release of an offender and the costs associated with a failure to appear.

Additional Insight: For each case ID in which the most serious offense is a misdemeanor, the odds for a failure to appear increases by 2.1%.

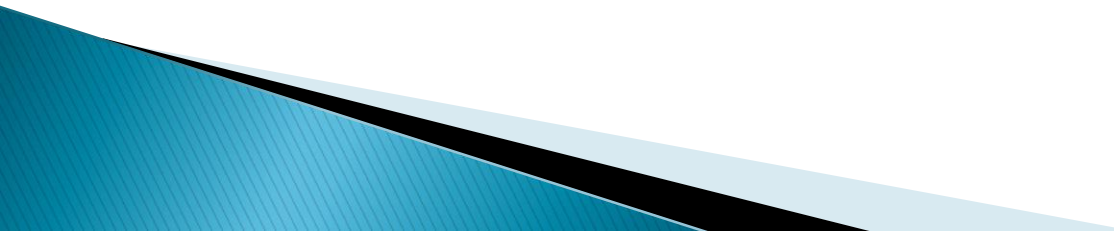
Predictive Accuracy

- ❖ The model with a cutoff of $p=0.5$ has correctly classified 69.07% of the instances ($n=3,171$) and incorrectly classified 30.93% ($n=1,420$).
- ❖ The confusion matrix shows the various classification errors.
- ❖ This initial model has a specificity of 0.295542 and a sensitivity of 0.9014696. These values describe the Type 1 and Type 2 errors.
- ❖ The predictive accuracy of the model can be visually represented through the ROC chart. The AUC for the model is equal to 0.685.
- ❖ This indicates that the model is 18.5% better at predicting FTA then through randomly guessing the outcome.

		<i>Predicted</i>	
		Yes (1)	No (0)
<i>Observed</i>	Yes (1)	472	295
	No (0)	1125	2699



The Cost of an Error

- ❖ There is an underlying concern when producing predictive analytics models related to the cost of a misclassification, particularly with crime.
 - ❖ If the model predicts with 69% accuracy then it must be incorrect 31% of the time.
 - ❖ These classification errors represent real costs to the municipality and could be the difference between releasing an individual back into society and then having an FTA that costs the courts time and money. Or even worse... the released person commits a serious crime like Murder!!!!
 - ❖ The other type of error would be that the county would place an individual into the jail system when they would not have had a FTA in the first place. This also costs the county time and resources and needs to be considered as well.
- 

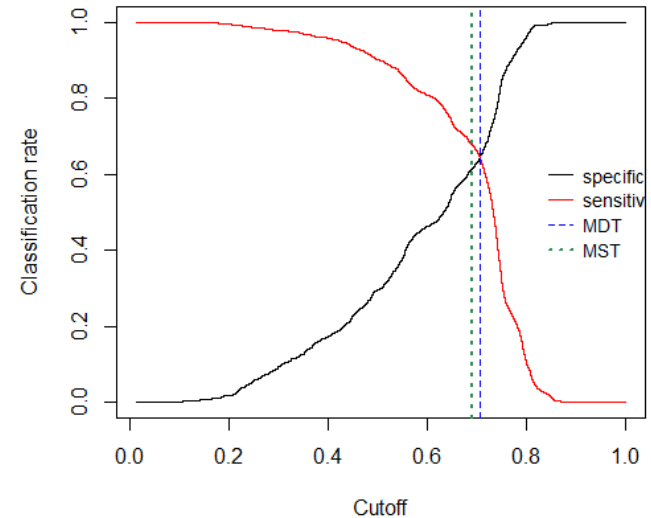
The Cost of an Error

- ❖ If the intention would be to more uniformly balance the classification performance, the following approach can be utilized.
- ❖ We adjust the cutoff threshold from 0.50 (<0.5 is 0, ≥ 0.5 is 1) to a different probability threshold in order to mitigate the classification performance and Type I and Type II errors.
- ❖ Additional calibration work from subject matter specialists was necessary before this technique was to be used in practice.
- ❖ What is the cost of these classification errors and how do we find a more cost effective threshold?
- ❖ These questions became the catalyst for additional follow-up research.

Probability Threshold	Cutoff = 0.5	FTA Risk Level
1.00	FTA (1)	Very High
0.90	FTA (1)	Very High
0.80	FTA (1)	High
0.70	FTA (1)	High
0.60	FTA (1)	Medium
0.50	FTA (1)	Medium
0.40	No FTA (0)	Medium
0.30	No FTA (0)	Low
0.20	No FTA (0)	Low
0.10	No FTA (0)	Very Low
0.00	No FTA (0)	Very Low

The Cost of an Error

- ❖ We explored 2 different techniques to further balance the performance based upon the cost matrix being identical for each classification.
- ❖ The chart identifies the ideal balancing point for calibration of the probability threshold.
- ❖ The Minimum Difference Threshold (MDT) approach indicated to adjust the probability threshold from $p=0.50$ to $p=0.71$.
- ❖ This results in a decrease of correctly classified of the instances (64,34% $n=2,954$) from the initial model and an increase of incorrectly classified instances. (35.66% $n=1,637$).
- ❖ The MDT balanced model has a specificity of 0.6443 and a sensitivity of 0.6429. The confusion matrix is shown here.



		<i>Predicted</i>	
		Yes (1)	No (0)
<i>Observed</i>	Yes (1)	1029	1069
	No (0)	568	1925

The Cost of an Error

		<i>Predicted</i>	
		Yes (1)	No (0)
<i>Observed</i>	Yes (1)	984	966
	No (0)	613	2028

- ❖ The Maximized Sum Threshold (MST) approach indicated to adjust the probability threshold from $p=0.50$ to $p=0.691$.
- ❖ This results in a decrease of correctly classified of the instances (65.6% $n=3,012$) from the initial model and an increase of incorrectly classified instances. (34.5% $n=1,579$).
- ❖ The MST balanced model has a specificity of 0.6161 and a sensitivity of 0.6773.
- ❖ The MST approach indicates a reduction in predictive performance from the initial model but higher than the MDT based approach.
- ❖ The change in the probability threshold balanced out the specificity and sensitivity ratios more evenly for both the MST and MDT approach than the initial model.

Final Results

- ❖ However, if the intent would be to use this model and the opportunity costs for a False Positive was equal to the cost of a False Negative classification, the MST approach provides a stronger mechanism to draw from the analytics for decision making.
- ❖ The MST calibrated model contains a 65.6% predictive accuracy (34.4% erroneous predictions) while more evenly distributing the classification errors. This model would indicate that an additional 966 individuals would be predicted to not have an FTA and 613 would be classified as having an FTA based upon the algorithm and the classification errors.
- ❖ This calibrated model could be leveraged to potentially reduce the influx of defendants awaiting trial within the jail population by approximately 7.5%.
- ❖ This model can then be used to extrapolate the total amount of inmates within the prison complex, provided the logistic model is being used exclusively. The predictions were then applied to the existing prison expansion plan and used to drive down the cost of the new facility by reducing the number of beds by 130 and saving the county \$1.4 - \$2.8 million annually for the county.

Final Results

- ❖ The next phase of the predictive model deployment plan would be to introduce a computer based risk assessment system.

The basic concept is as follows:

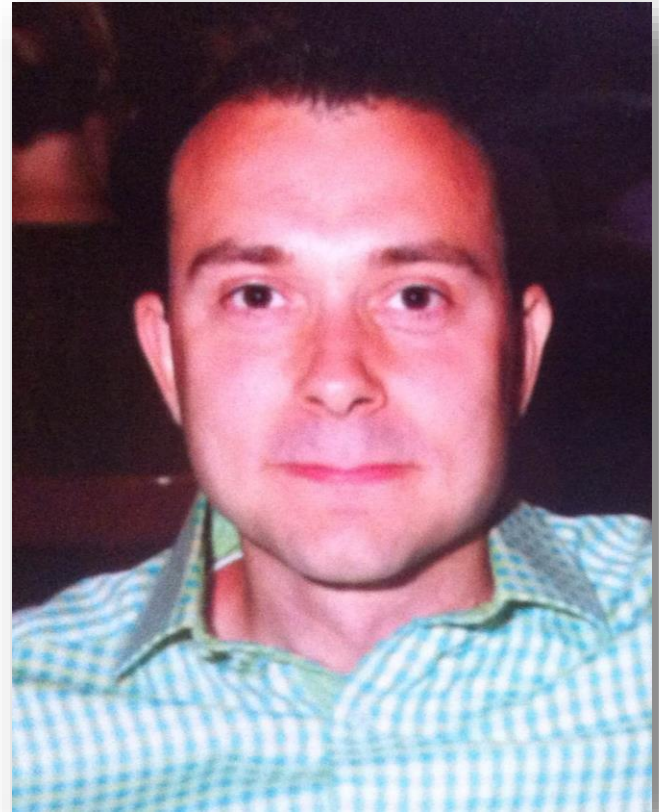
- ❖ A judge enters the defendant's ID into the system when deciding whether or not to keep the person in jail.
- ❖ The system will take into consideration demographic variables (age, sex, education, etc...) as well as behavioral variables (# of prior felonies, time at address, etc...), runs the algorithm, and produces a probability level.
- ❖ The software assigns the probability threshold to a specific risk category (Ex. Low, Medium, High Risk). This risk category is then shown to the judge.
- ❖ The judge then utilizes this information and their professional judgment when deciding whether or not to release the individual into the population at large.



Probability Threshold	Cutoff = 0.5	FTA Risk Level
1.00	FTA (1)	Very High
0.90	FTA (1)	Very High
0.80	FTA (1)	High
0.70	FTA (1)	High
0.60	FTA (1)	Medium
0.50	FTA (1)	Medium
0.40	No FTA (0)	Medium
0.30	No FTA (0)	Low
0.20	No FTA (0)	Low
0.10	No FTA (0)	Very Low
0.00	No FTA (0)	Very Low

About Me

- ❖ Reside in Wayne, Illinois
- ❖ Active Semi-Professional Classical Musician (Bassoon).
- ❖ Married my wife on 10/10/10 and been together for 10 years.
- ❖ Pet Yorkshire Terrier / Toy Poodle named Brunzie.
- ❖ Pet Maine Coons' named Maximus Power and Nemesis Gul du Cat.
- ❖ Enjoy Cooking, Hiking, Cycling, Kayaking, and Astronomy.
- ❖ Self proclaimed Data Nerd and Technology Lover.



Acknowledgements

- ❖ <http://www.salford-systems.com/products/mars>
- ❖ https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_splines
- ❖ <http://www.r-bloggers.com/fit-and-visualize-a-mars-model/>
- ❖ <https://rpubs.com/daspringate/survival>
- ❖ <http://www.math.ntnu.no/~bo/TMA4275/Download/R.tutorialDiez.pdf>
- ❖ http://www.statistics4u.com/fundstat_eng/cc_linvsnonlin.html
- ❖ https://en.wikipedia.org/wiki/Survival_analysis
- ❖ <https://rpubs.com/daspringate/survival>
- ❖ https://en.wikipedia.org/wiki/Proportional_hazards_model

Fine