- simple linear regression is one quantitative variable predicting another
- multiple regression is simple linear regression with more independent variables
- nonlinear regression is still two quantiative variables, but data is curvilinear
- binary data does not have a normal distribution - a condition needed for most other types of regression
- predicted variables of the dependent variable can be beyond 0/1 in other types of regressions - logistical regression deals with probabilities
- + probabilities have to be between 0 and 1
- probabilities are often not as linear such as U shapes where probability is very low/high at extreme x-values
- + e.g. probability (may be, as an example) higher for contracting the flu for oldies/young'uns, but lower for middle aged people
- odds of event X - P(X)/(1-P(X))
- the odds ratio for a variable in logistic regression represents how the odds change with a 1 unit increase in that variable holding all other variables constant
- odds vs probability
- + underlying probability may be low, but odds could increase in magnitude very quickly
- + e.g. probabilities of being struck by lightning, and being struck by a meteor
- * probability of being hit by lightning are *much much* higher than being struck than a meteor
- * however, being hit by lightning is a very low prob to begin with
- in logistic regresion, we are estimating an unknown p for any given lienar combination of the independent variables
- link together independent variables to essentially the Bernoulli distribution; that link is called the logit
- in logistical regression, we don't know p like we do in Binomial (Bernoulli) distribution problems - goal of logistical regression is to estimate p for a linear combination of the independent variables
- the natural log of the odds ratio, the logit, is that link function
- $ln(odds) ==> ln(\frac{p}{(1-p)})$ is the $logit(p)$ OR $ln(p) - ln(1-p) = logit(p)$
- reminder: $log_e x = lnx$
- boundaries - if p = 1 or 0, our function is undefined, when p = 0.5, our function = 0 (our function being the logit)
- logit function graphs to a sigmoid function
- 0-1 ran along x-axis but we want probabilities to be on the y-axis - achieve this by taking inverse of the logit function
- $logit^{-1}(\alpha) = \frac{1}{1+e^{-\alpha}} = \frac{e^{\alpha}}{1+e^{\alpha}}$ where α = some number
- + α is the linar combination of variables and their coefficients - inverse-logit will return the probabilyt of being a "1" or in the "event occurs" group
- inverse logit A.K.A mean function
- + $\mu_{y|x}$
- note about coefficients
- + coeffs calculated using maximum likelihood estimation
- estimated regression equation
  - natural log of odds ratio is equivalent to linear function of the independent variables - antilog of logit function allows ut to find the estimated regression equation, i.e. solve for p, the estimated probabibility
  - $logit(p) = ln() = \beta_0 + \beta_1 x_1$
  - isolate/solve p! - going to do this using algebra

- $\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1}$ // on the right here is Euler's constant raised to the power of a linear combinatino of all the independent variables
  - $p = e^{\beta_0+\beta_1 x_1}(1-p)$
  - $p = e^{\beta_0+\beta_1 x_1} - e^{\beta_0+\beta_1 x_1} p$
  - $p + e^{\beta_0+\beta_1 x_1} p = e^{\beta_0+\beta_1 x_1}$
  - $p(1 + e^{\beta_0+\beta_1 x_1}) = e^{\beta_0+\beta_1 x_1}$
  - estimated regression equation! $\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$
- watched up to Brandon Foltz' Statistics 101: Logistic Regression, Odds Ratio for Any Interval

1 Random Forests

- combination of learning models increases classification accuracy (bagging)
- bagging - average noisy and unbiased models to create a new model with low variance
- random forest algorithm works as a large collection of decorrelated decision trees [1]
- create random subsets of the original full set of values - e.g. out of 100, one particular subset might contain all features of rows 1, 5, 9, 24, 38, 49, 52, 68, 72, 82 and their classifications, others follow
- create decision tree from each subset
- 'forest' - lots of decision trees
- count votes for each class when an unknown class feature set is passed through all the decision trees - take majority for classification, average for regression
- random forests as a supervised learning algorithm came only second to boosted decision trees in a comparison study [2] (third was bagged decision trees, SVM 4th - essentially, aggregated tree methods were at the top)
- at each node in tree, choose random subset of (m - constant in trees) features - consider only splits on those features
- each individual tree has high variance - but by averaging over an ensemble of trees, we reduce the variance in the final estimator, and hence lowering error
- introduced by Breiman

2 Multivariate Gaussian

- univariate Gaussian - X
  $N|\mu, \sigma^2$
- means X has density $f(x) =< ..insertfhere... >$
- degenerate univariate Gaussian - σ^2 can be 0, i.e. $X == \mu$
- multivariate Gaussian - A random variable $X \in R^n$ is multivariate Gaussian (or normal or multivariate normal) if any linear combination of its components is univariate Gaussian i.e. $a^T X = \Sigma a_i X_i$ is Gaussian for every $a \in R^n$

[1] Gupta, Ashish. Learning Apache Mahout Classification. Packt Publishing Ltd, 2015.

[2] Caruana, Rich, and Alexandru Niculescu-Mizil. "An empirical comparison of supervised learning algorithms." Proceedings of the 23rd international conference on Machine learning. ACM, 2006. APA

- defn: $X\ N(\mu, C)$ (C is a pos semidef matrix) means X is Gaussian with $EX_i = \mu_i$ and $covariance(X_i, X_j) = C_{ij}$
- $\mu and C$ uniquely determine the distribution $N(\mu, C)$
- positive semidef matrix - all eigenvalues are ¿= 0

3 Gaussian Processes

- most often used in geostatistics - modeling spatial things - geography, meteorology, etc.
- things evolving over time, where its going, e.g. an airplane
- finance, physics, diffusion processes, all the things!
- Gaussian Processes - for any set S, a Gaussian process (GP) on S is a set of random variables $(Z_t : t \in S)$
- for any N, and for any $t_1 ... t_n \in S$, $(Z_{t_1}, ..., Z_{t_n})$ is multivariate Gaussian
- finite dimentional distributions - distribution of all the t in Z vectors
- Example: $S = 1, ..., d$ $(Z_t) = (Z_1, ..., Z_d) \in R^d$ - satisfies definition, trivial
- random lines: $S = R$, $Z_t = tW, W\ N(0, 1)$
-