**Binary Logistic Regression**

1. Supervised Classification model
   1. When you have a target variable, and the model’s job is to predict that target variable
2. Unsupervised Classification Scoring
   1. Target variable is undefined clustering
   2. Where you know groups exist, and then the model makes those groups

Regression is modeling the expected response conditional to the predictors

* For binary, the expected value is just the probability of the event

Linear Probability Model

* Probabilities are bounded, but the linear function can take on any value
* Relationship between probabilities and X is usually non-linear
  + One unit change in X will have different effects when the probability is near 1 or 0.5
* Properties of OLS do not hold

Logistic Regression Model

* Predicted probability is always between 0 and 1
* The parameter estimates do not enter the model linearly
* The rate of change of the probability varies as the X’s vary

Logit Link Transformation

* To create a linear model, a link function (logit) is applied to the probabilities (aka the logistic regression model)
* The relationship between the parameters and the logits are linear
* Logits are unbounded

Odds ratio from logistic regression

* If the odds ratio is below 1 then flip numerator and denominator to make easier to interpret
* Binary variables
  + Ecoefficient = odds ratio for that variable
    - E3.952 = 52.03 🡪 odds ratio for yes/no central air in relation to bonus eligibility, holding all else constant
    - “Homes with central air are 52.03 times as likely to be bonus eligible than homes without central air, on average, holding all else constant
  + 100 \*( Ecoefficient -1) = higher expected odds
    - Odds are centered around 1, so we need to center around 0 for percents
      * 100 \*( Ecoefficient -1) = 5103%
      * Homes with central air have 5103% higher expected odds than homes without central air to be bonus eligible, holding all else constant
* Continuous variables
  + Ecoefficient = estimated odds ratio (for every additional unit added)
    - E.0041 = 1.0041 odds increase for every additional square foot of space added, holding all else constant
  + 100 \*( Ecoefficient -1) = higher expected odds (for every additional unit added)
    - 100\*(E.0041 -1) = 0.41%
    - Every additional square foot of space expects to have 0.41% higher odds of being bonus eligible
* Amount to double the odds (for continuous variables)
  + Double odds =
    - = 169.06
    - Every additional 169.06 square feet added doubled the odds to be bonus eligible, on average, and holding all else constant

Assumptions for OLS regression

* Random error term has a Normal distribution with a mean of 0
* The random error term has constant variance
* The error terms are independent
* Linearity of the mean
* No perfect collinearity
  + In logistic regression, the 1st and 2nd assumption fail miserably

Maximum Likelihood Estimation

* In logistic regression, estimates are obtained via maximum likelihood estimation (MLE)
* The likelihood function measures how probable a specific grid of beta values is to have produced your data 🡪 so we want to maximize that
  + “Which beta values most likely produced the data we have”
* Can be used for any model
* Based off the probability density function
* Will use a Log-Likelihood Function, as it is easier to mathematically work with

Likelihood Ratio Test

* If extra predictors do not add much information, then a model that includes them shouldn’t be substantially more likely than the model that doesn’t include them
* Do not do with categorical variables that have more than 2 levels
* Likelihood Ratio Test (LRT) compares the full and reduced models
* anova 🡪 compares any two models as long as they are nested
  + Full: Bigger of the two models you are comparing
  + Reduced: Smaller model that is nested within the full model
    - For intercept only model, use target ~ 1, data =…
  + Only look at 1 p-value for the whole output
  + Ho: full and reduced models are the same
    - Low p-value would mean the extra variables add value and should be used in the model
  + Can be used to see if any variables in a group adds to the model, so this way you do not have to test the variables one at a time
    - High p-value means the model with the categorical variables does provide additional information, and you can drop the variables
  + Is an overall comparison of the two models
  + Use anova(full.model, reduced.model, test = ‘LRT’)
    - DF in output is the difference in the # of coefficients between full and reduced models that are being tested (aka how many coefficients are being tested)
    - If you have a categorical variable with 20 levels, then the DF should be 19
* Car::Anova 🡪 compares the full model versus a model without one of the variables
  + Only need to input one model
  + Good for models with categorical variables
  + In R code generally use type = ‘III’
    - Type = ‘1’ would be for ordered terms (interactions and/or higher order terms), where model hierarchy is important, and it will drop the variables in the order they are listed in the model
  + Can be used to see if individual variables should be included in the model
  + One p-value per variable
  + Ho: the model with and without that variable are the same
    - Low p-value 🡪 that variable does provide additional information and should be kept in model
    - High p-value 🡪 the model with that categorical variable doesn’t provide any additional information and that variable should be dropped

Assumptions for logistic regression

* Continuous variables
  + Only assumption is linearity
  + Checked by seeing if that variable is a straight line
* Categorical variables
  + Only need to check independence
    - As binary variables do not have higher order terms
* We generally test linearity one continuous variable at a time
  + Use gam(), and do s(continuous variable)
    - gam() makes a spline model
    - s() is for smoothing when you suspect a non-linear term
    - it is alright if binary variables are included in the gam() code, as they will not affect the output
    - We do not care about the p-value from the gam() summary, as we care about the relationship
    - In the plot(fit.gam), the middle-dashed line is the variable of interest, and the dashed line is the C.I. for the variable of interest
    - Want the center dark line to be straight
    - Want edf = 1
      * Which would indicate that the model thinks (continuous variable)1 is the best representation to represent the relationship between the logit target and the continuous variable
  + To test numerically, compare logit model to gam model with anova()
    - anova(logit.model, fit.gam, test = ‘LRT’)
      * Logit.model is the glm() model with one continuous variable
      * Fit.gam() is the gam() model with one continuous variable
      * Ho: spline model is same as strait line model
        + Low p-value 🡪 spline model and straight-line model are different, which means term is not linear
        + High p-value 🡪 spline model and straight-line model are not different, which means term is linear
  + If assumptions fail for at least one continuous variable
    - Use GAM logistic model instead with more limited interpretation on variables that break assumption
    - Bin the continuous variables that break the assumption of linearity
      * This keeps interpretability

**Data Considerations**

Rare Event Modeling

* 5% or smaller in a category can lead to classification problems (target variable in unbalanced)
* Oversampling
  + Duplicate current event cases in training set to better balance with non-event cases
    - Increase number of rare event occurrence artificially
  + Keep test set are original population proportion
  + Better with small sample sizes
* Undersampling
  + Randomly sample current non-event cases to keep in the training set to balance with the event cases
    - Decrease number of common events to make rare event less rare
  + Keep test set as original population proportion
  + Better with large sample sizes
* For the notes, Undersampling was used
* When Undersampling/oversampling used, the resulting model will be biased, and so we need to alter the model to make it un-biased
* Nomenclature for below (1 is the rare event)
  + - Pi = unadjusted predictor (see R code to calculate)
    - Π = population proportion of 1’s and 0’s
    - p0/p1 = sample proportion
  + Adjusting the intercept (fix after breaking)
    - Numerator 🡪pi\* p0 π1
    - Denominator 🡪 (1-pi\*)p1 π0 + pi\*p0 π1
    - This will adjust the outputted prediction probabilities for each observation
  + Weighted observations (“take your vitamins before you get sick”)
    - Here, we need to overweight the 0’s, since their effect was reduced in the sampling
    - Numerator = p1 π0
    - Denominator = p0 π1
    - Aka🡪
    - Use ifelse to create a weight variable to be referenced in the glm() model
* When to use adjusted intercept vs weighted observations
  + Generally use weighted observations for everything
  + Only time to use adjusted intercept is when you are rebuilding a model you know is correct on new data
* Handling missing values
  + If over 50% of the category is missing, then consider deleting that category
  + If the variable with missing values is categorical, create a new category for missing
  + If variable with missing is continuous, then impute
    - And create a binary variable that indicates if that variable was imputed
    - Must include that binary variable in model if the original variable is also present
* Linear Separation (convergence problems)
  + Complete linear separation occurs when some combination of the predictors perfectly predicts every outcome
    - Every person in group A is a yes and every person in group B is a no
  + Quasi-Complete Separation occurs when the outcome can be perfectly predicted for only a subset of the data
    - Group A is 77 yes, 23 no, but group B is 0 yes, 50 no
    - Need a count of at least 1 in every box
  + R and Python will rarely report this as a problem, so you should look for it in your data beforehand
* Solution to Quasi-Complete Separation
  + Combining categories to eliminate the 0 count
  + For ordinal variables, make sure to collapse categories that are next to another
  + For Nominal variables, add the level with the 0 count to the level with the next closest distribution
    - Add to the level that has the lowest % of the predictor
  + Watch out for interactions
    - Tough to check all of the interactions for quasi-separation
    - Should not be any 0 counts within the interaction as well

**Diagnostic and Subset Selection**

Stepwise/backwards/forward selection can be used for variable selection for your model

* P-value for one category is the p-value for that category level compared to the reference level
* First do stepwise selection all main effects, and then do variable selection with interaction terms and significant main effects (like in the logistic team homework 2)

**Diagnostics**

Many types of residuals are used in binary response model setting

* Most common is deviance residuals

Deviance

* Model is a summary of the dataset
* The saturated model fits the data perfectly, but isn’t a useful summary
  + Fits the current data perfectly, but is unable to extrapolate
* Deviance is a measure of how far our fitted model is from the saturated model
  + Essentially our error
* Logistic regression minimizes the sum of squared deviance
* Deviance residuals tell us how much each observation reduces the deviance
* “How close is the generalized model to the un-usable but perfect saturated model”
* Outliers are the data points where the generalized model is not close to the saturated model

Influence Statistics

* DIFDEV 🡪 measures change in deviance with deletion of the observation
* DFBETAS 🡪 measure standardized change in each parameter estimate with deletion of observation
  + Does it one variable at a time, and can show if certain observations effect which coefficients
* Cooks D 🡪 measures the overall impact to the coefficients in the model
  + Gives one number to summarize how much coefficients change when an observation is removed

**Model Assessment**

Estimation 🡪 quantifying the expected change in response associated with predictors

Prediction 🡪 Use the model to predict new response

Logistic regression is a model for the probability of an event 🡪 not the occurrence of an event

* Logistic regression can be a classification model as well
* Logistic model can rank observations, and classification model only yes or no, and not the probability of churn

Deviance/Likelihood measures

* AIC🡪 Crude, large sample approximation of leave one out cross validation
  + Uses number of parameters in the model as one of its criteria
* BIC 🡪 favors smaller models/penalizes model complexity more
  + Use sample size as one of its criteria
* Lower the better
* Do not always both agree

Pseudo R2

* There are a bunch of different pseudo R2 quantities for logistic regression
* Higher values indicate “better” model
* Generalized / Nagelkerke R2
  + How much better than intercept only model
* Not interpretable

Discrimination 🡪 ability to separate events from the non-events

* Good in statistics
* How good is the model at distinguishing between the 1’s and 0’s

Calibration 🡪 how well the predicted probabilities agree with the actual frequency of the outcomes

* Are predicted probabilities systematically too high or too low

Coefficient of Discrimination

* Want the model to assign a higher probability to events and a lower probability to non-events
* Coefficient of discrimination (or discrimination slope) is the difference in the average predicted probability between the 1’s and 0’s
  + D = p1 – p0
  + Further apart the 1’s and 0’s are means the model does a better job at separating the 1’s and 0’s
  + Bigger the better

Rank order statistics

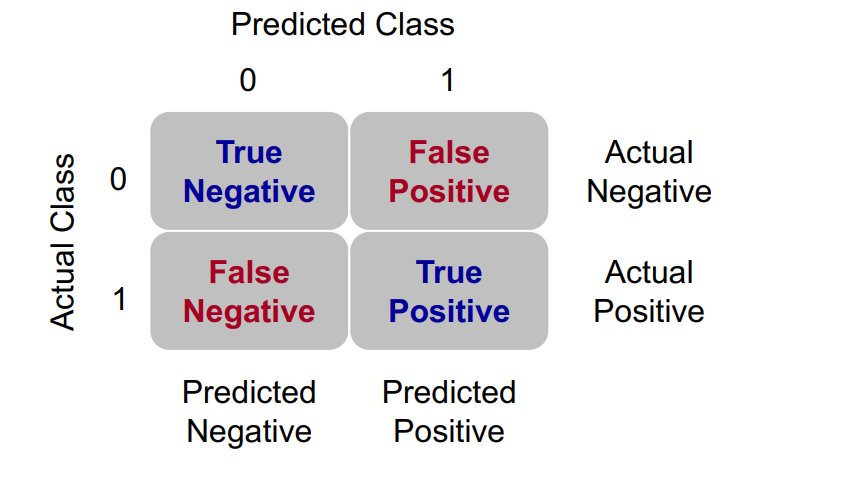
* How well does the model order predictions
* Compares all of the 1’s with all of the 0’s (every possible pair of 1’s and 0’s)
* Concordance🡪 for a pair of subjects with and without the event, the one with the event has the higher probability
* Discordance 🡪 for a pair of subjects with and without the event, the one with the event has the lower probability
  + Opposite of concordance
  + bad
* Tied 🡪 for a pair of subjects with and without the event, they both have the same probability

Concordance 🡪 for all possible (1,0) pairs, the model assigns the higher predicted probability to the observation with the event concordance% of the time

* C-statistic 🡪 c = concordance % + [(1/2)(tied%)]
* Somers D 🡪 Dxy = 2c – 1
* Kendall’s taua 🡪

Classification 🡪 forces the model to predict yi = 0 or yi = 1 based on whether the predicted probability exceeds some threshold

Classification table: Will depend on what the cutoff value is, and will change if the cutoff value is changed



Sensitivity 🡪 what percentage of the actual 1’s did you get correct

* TPR =

A diagram of negative and true positive

Description automatically generated

Specificity 🡪 what percent of the actual 0’s did you get right

A diagram of negative and false positive

Description automatically generated

Youden J statistic (or Youden Index) 🡪 J = sensitivity + specificity – 1

* If cost is not a factor and false positives and false negatives are weighted equally, select the cutoff with the highest Youden J statistic

ROC curve 🡪 plots True Positive Rate and False Positive Rate for a grid of thresholds

* AUC (area under the curve) summarizes the overall quality of the ROC curve
  + Equivalent to the C-statistic
  + Want high sensitivity and high specificity
  + AUC = %Concordance + [(1/2)(% tied)]
* We want the ROC curve to have as large of an AUC as possible

K-Statistic 🡪 The two sample K statistic can determine if there is a difference between two cumulative distribution functions

* Mathematically equivalent to Youden’s J statistic
* Has a corresponding hypothesis test with a D test statistic
  + Want a large D statistic
  + D = max(True Positive Rate – False Positive Rate)
  + = max(Sensitivity + Specificity – 1)
  + = max(Youden J)

Assessing Predictive Power

Recall (TPR) 🡪

Precision (PPV) 🡪

When not considering costs, the F1 score can be used (precision-recall version of Youden’s Index)

* F1 =2\*( )
* Want to maximize the F1 score

Lift 🡪 PPV / π1

* Common in marketing
* Great for interpretation around validity of model ranking / classifying observations correctly

Accuracy 🡪

Misclassification (Error) Rate 🡪

Lift 🡪 quantifies how much better a models predictions are compared to random chance or a baseline model/subset of the sample

* Great for interpretation around validity of model ranking / classifying observations correctly
* The top *Depth%* of your customers, based on predicted probability, you get *Lift* times as many responses compared to targeting a random sample of *Depth%* of your customers

Closing Thoughts on Classification

* Classification is a decision that is extraneous to statistical modeling
* Although logistic regression tends to work well in classification, it is a probability model and does not output 1's and 0’s
* Classification assumes the cost for each individual is the same
  + Useful for groups
  + Careful about single observation decisions

**Ordinal Logistic Regression**

Ordinal Logistic Regression 🡪 probability that observation i has at most event j, and j = 1,…, m ordered events

Methods for Modeling

1. Cumulative Logit Model
   1. Most common and easy to implement/interpret
   2. M-1 equations (where m is equal to the number of levels in the ordinary target variable)
2. Adjacent Categories Model
3. Continuous

Cumulative Logits

* Instead of modeling the typical logit, we will model the cumulative logits
* A math equation with black text

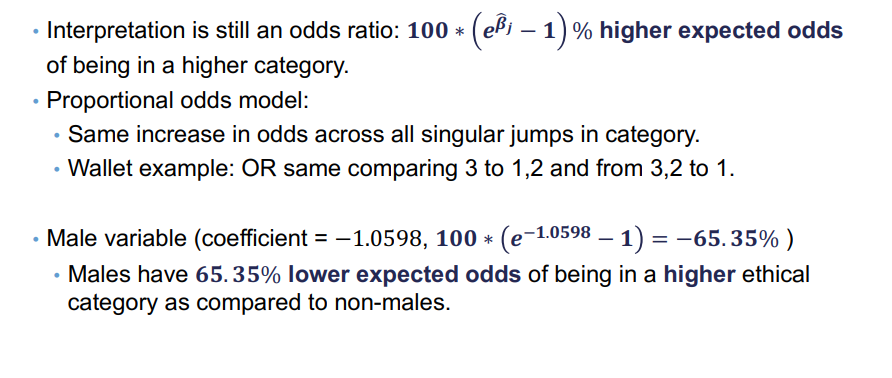
  Description automatically generatedIf an ordinal variable has M levels with probabilities (p1,p2,…,pm), then the cumulative logits are:

Proportional Odds Assumption

* Intercept changes, but slope parameters stay the same
* the relationship between each pair of outcome categories is the same across all thresholds or cutpoints of the ordinal dependent variable
* the odds ratios between the different levels of the predictor variables are constant, regardless of which specific threshold you are considering
  + Uniform Effect Across Categories: The assumption implies that the effect of predictor variables on the odds of being in a particular category or below it (versus being in a higher category) is the same across all possible cut points of the ordinal outcome. For example, if you're predicting levels of customer satisfaction (e.g., "Unsatisfied," "Neutral," "Satisfied"), the effect of a predictor like income on moving from "Unsatisfied" to "Neutral" is assumed to be the same as moving from "Neutral" to "Satisfied."

Brant Test (for proportional odds)

* need to see if the models are statistically different from each other in the proportional odds model
  + Ho: Proportional Odds correct (slopes equal across models)
  + Ha: Proportional Odds incorrect (slopes NOT equal across models)
    - Low p-value indicates on of the variables failed
* If the Brant Test fails…
  + Some variables fail the assumption 🡪Partial Proportional Odds Model
    - Will create different slopes for the variable(s) that failed
  + All variables fail the assumption 🡪 Multinomial Logistic Regression

Odds Ratio Interpretation

Similarities between Ordinal Logistic Regression and Binary Logistic Regression

* Ordinal Logistic Regression has a lot of the same aspects/issues as binary logistic regression
  + Multicollinearity still exists
  + Non-convergence problems still exist
  + Concordance, Discordance, and Tied pairs still exist – so the C statistic still exists
  + Generalized R2 remains the same
  + Can still make a confusion matrix

Differences between Ordinal Logistic Regression and Binary Logistic Regression

* Ordinal Logistic Regression has few aspects/issues that differ from a binary logistic regression
  + A lot of the diagnostics for binary logistic regression cannot be calculated easily since there are actually multiple models – ROC curve for each model?
  + Diagnostics/ Influence plots are not available – residuals for each model?
  + Predicted probabilities are for each category