**Logistic regression for any type of categorical target variable**

Logistic regression is a fundamental statistical analysis for data science and analytics. It part of a class of modeling techniques known as classification models since they are trying to predict categorical target variables. This target variable can be binary, ordinal, or even nominal in its structure.

The primary focus will be binary logistic regression.

It is the most common type of logistic regression, and sets up the foundation for both ordinal and nominal logistic regression.

**Why is least squares bad?**

assumptions of ordinary least squares don’t really hold in this situation

probabilities do not tend to follow the properties of linearity in relation to their predictors. Also, the linear probability model possibly produces predictions outside of the bounds of 0 and 1

**Binary logistic regression**

The predicted probability from the above equation will always be between 0 and 1.

The parameter estimates do not enter the function linearly (this is a non-linear regression model), and

the rate of change of the probability varies as the predictor variables vary

**Logit link function**

**A red line with a red line

Description automatically generated**

To create a linear model, a **link function** is applied to the probabilities. The specific link function for logistic regression is called the **logit** function.

Relation btw parameters and logits are linear

Logits are unbounded

**Assumptions:**

Indepence of observations

Logit is linearly related to variables

**Interpretation on coefficient**

 A math equations on a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

Concordant

A **concordant** pair is a 0 and 1 pair where the bonus eligible home (the 1 in our model) has a higher predicted probability than the non-bonus eligible home (the 0 in our model) - our model successfully ordered these two observations by probability. It does not matter what the actual predicted probability values are as long as the bonus eligible home has a higher predicted probability than the non-bonus eligible home

**Concordance is only measuring how often we are able to predict 1’s with higher probability than 0’s - again, correctly ranking the observations.**

**There is no good or bad value as this can only be compared with another model to see which is better. Let’s compare this to our model with the categorical variables.**

Discordant

A **discordant** pair is a 0 and 1 pair where the bonus eligible home (the 1 in our model) has a lower predicted probability than the non-bonus eligible home (the 0 in our model) - our model unsuccessfully ordered the homes. It does not matter what the actual predicted probability values are as long as the bonus eligible home has a lower predicted probability than the non-bonus eligible home.

tied pairs

A **tied** pair is a 0 and 1 pair where the bonus eligible home has the same predicted probability as the non-bonus eligible home - the model is confused and sees these two different things as the same. In general, you want a high percentage of concordant pairs and low percentages of discordant and tied pairs

Variable selection

logistic regression uses the same approaches to doing variable selection.

For, back,Reg reg

9lasso)