Logistic Regression

The question is how should we model the relationship between

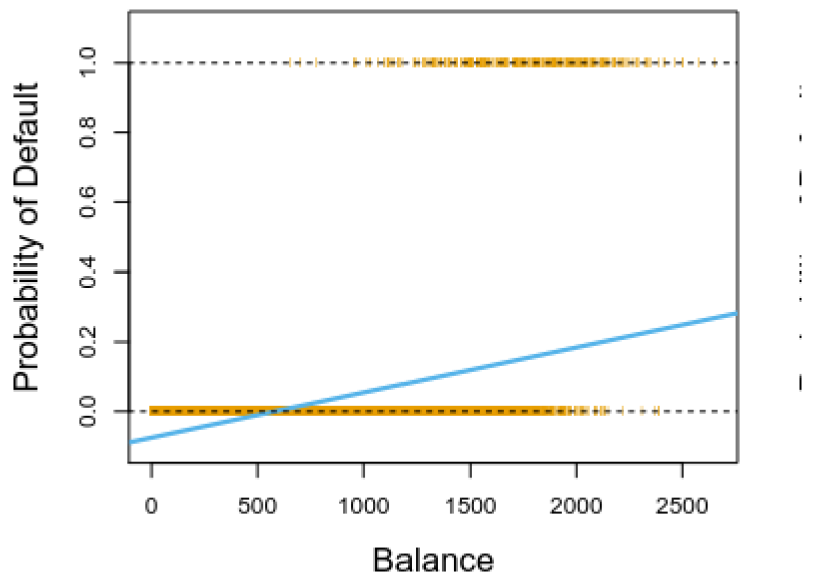
$$P(X) = \Pr(Y = 1|X)$$

- Relationship between the **Probability** of X and the **Classifying** Prediction for X
- Using 1 and 0 for the Response

Using Linear Regression model to represent these probabilities

$$p(X) = \beta_0 + \beta_1 X$$

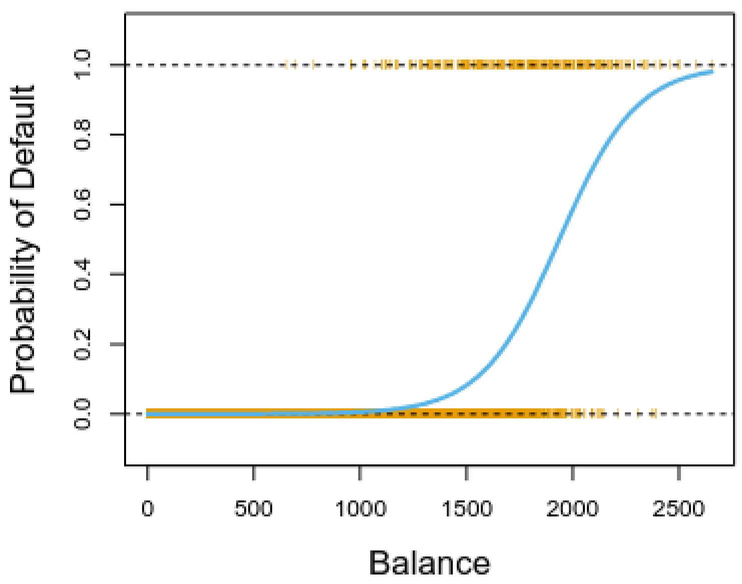
- If we fit the line to predict the **Probability**



- Notice that the `Balance` Lower than 500 our prediction for the probability is **negative**

To avoid this problem we model $p(X)$ to only fall between 1 and 0 for all values X **Logistic Function** which is a [Sigmoid Function](Sigmoid%20Function)

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$



- Any output of the **Logistic Function** Falls between 1 or 0

$$\frac{p(X)}{1-p(X)} = \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}} = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \times \frac{1}{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+e^{\beta_0+\beta_1 X}}} = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \times 1 + e^{\beta_0+\beta_1 X}$$

$$\text{odds} = \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

- the **quantity** $p(X)/[1-p(X)]$ is called the $odds$ can can only take values between 0 to ∞
- Values close to 0 indicates low probability
- Values close to ∞ indicate higher probability

- if $p(X) = 0.5$ the $\text{odds} = 1$ equal chance
- if $p(X) = 0.8$ the $\text{odds} = 4$ 4x success chances

Probabilities vs Odds

- The odds are the **ratio** of something happening *divided by* something not happening
- The probability is the **ratio** of something happening *divided by* to everything could happen

$$\text{Probability} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

$$\text{Odds} = \frac{\text{Probability of even occurs}}{\text{Probability event does not occur}} = \frac{p}{1-p}$$

Odds in logistic regression

Taking logarithm of both sides : log odds or logit

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- the $\log$ of odds gives us a **Linear equation** which is easy to model and **interpret**
  Why we use the odds?
- **Probabilities** lives in the interval $[0, 1]$
- Linear combinations like $\beta_0 + \beta_1 X$ lives on $(-\infty, +\infty)$
- The **odds** solves that by being in $(0, +\infty)$

In the **Linear Regression** β_1 gives the average change in Y associated with one unit increase in X
In the **Logistic Regression** β_1 does not correspond directly to the change in $p(X)$, if β_1 is positive increasing one unit in X will increase the **Probability** $p(X)$ but the increase depends on the current value of $p(X)$

$$\beta_1 > 0 \to \text{One unit increase X} \to \text{Increase } p(X)$$

$$\beta_1 < 0 \to \text{One unit increase X} \to \text{Decrease } p(X)$$

- The amount "Degree" of increase in the **Probability** $p(X)$ depends on the current value of X

Why the effect of the probability depends on current $p(X)$

- if $p(X) = 0.5$ the Sigmoid Function curve is steep, small changes in $X \to$ big changes in $p(X)$
- if $p(X) \approx 0.01 \text{ or } 0.09$ the **Sigmoid Function** curve is flat, changes in $X \to$ small changes in $p(X)$

Estimating The Regression Coefficients

In **Linear Regression** we use the least squares to estimate the coefficients, in the logistic regression the **Maximum Likelihood**.
Intuition :

- We seek estimates for $\hat{\beta}_0, \hat{\beta}_1$ such that we maximize the probability $\hat{p}(x_i)$
- More explanation in Maximum Likelihood Estimation

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i:y_{i=1}} p(x_i) \prod (1 - p(x_{i'}))$$

- This called the **Likelihood function**
- Our estimates $\hat{\beta}_0, \hat{\beta}_1$ are chosen to **maximize** this likelihood function
- **The least squares** is a special case of $\text{maximum Likelihood}$ Residual Sum of Squares
- Unlike Linear Regression the **Likelihood function** in logistic regression is nonlinear in the parameter β so there is no closed form solution for β Maximum Likelihood Estimator Derivation Logistic Regression

Same as **Linear Regression** we measure the accuracy of the estimated coefficients for the Logistic Regression by computing the **Standard Errors** , **Z-Test** , **Hypothesis Testing**

Once the **Coefficients** are estimated (Using iterative numerical optimization methods),we can calculate the predicted probability for the observation X_i

Multiple Logistic Regression

Generalizing the **Simple Logistic Regression** equation of the **log-odds**

$$\log\left(\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)$$

While $p(X)$ can be written as :

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- The **Maximization** of the coefficients is done via **MLE** [Maximum Likelihood Estimator Derivation Logistic Regression]
- The [Response] is till binary 1 or 0
- With multiple predictors **Variables** to make the prediction
- it uses **Linear decision boundary** in the log-odds space

Same case as in multiple linear regression it give different results than the simple linear regression for the same predictor cause often the predictors will be **correlated**

Multinomial Logistic Regression

Also know as **softmax regression**
When we want to classify a response Y that has more than two classes, The [Logistic Regression] and the Multiple Logistic Regression both follow a Bernoulli distribution which have only two outcomes, **Multinomial Logistic Regression** Follows a Binomial Distribution with multiple outcomes.

- Select a single class as a **baseline** (That's why its $K-1$)

$$\Pr(Y=k|X=x) = \frac{e^{\beta_{l,0}+\beta_{l,1}X_1+\cdots+\beta_{l,p}X_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l,0}+\beta_{l,1}X_1+\cdots+\beta_{l,p}X_p}}$$

$$\Pr(Y=K|X=x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l,0}+\beta_{l,1}X_1+\cdots+\beta_{l,p}X_p}}$$

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(\text{Y=K})|\text{X=x}}\right) = \beta_{k,0}+\beta_{k,1}X_1+\cdots+\beta_{k,p}X_p$$

- $e^{\beta_{l,0}+\beta_{l,1}X_1+\cdots+\beta_{l,p}X_p}$ Represent the probability of the k class divided by the probability of all other classes $K-1$ (same as the odds calculation)
- The log of odd is linear same as the logistic regression
- Here the probability of all the classes will add up to 1 which can be taken as a probability value
- Unlike when performing Bernoulli Logistic Regression on each class alone their results wont add up to one

Example

- Studying the **CRP** and it effect on infection types **(No infection ,Viral infection, Bacterial infection)**
  Here we have 3 **logistic regressions** : with $CRP = 25$
- $0 =$ (Viral ,Bacterial infections) and $1 =$ No infection, resulted in 0.009
- $0 =$ (No,Bacterial infections) and $1 =$ Viral infection, resulted in 0.360
- $0 =$(No,Viral infection) and $1 =$ Bacterial infection, resulted 0.001

These results don't sum up to one which means they we cannot be represented as **probabilities** to each class, Here using **softmax regression** is the option to go

Baseline

- To set a model as a baseline we subtract the **coefficients** of the **baseline** model from the coefficients of all model so its will be a reference for all models
- The decision of choosing the **baseline** will only effect the coefficients estimates but the predictions remains the same [Baseline Models Difference]