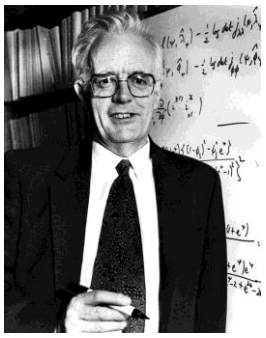
Logistic regression

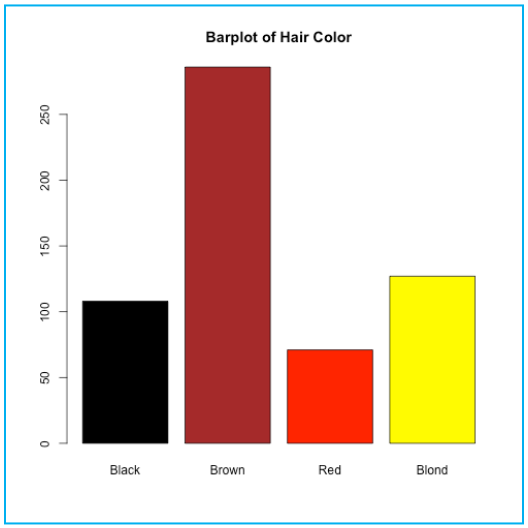
Remember :
Logistic regression is a linear method, but the predictions are transformed using the logistic function.

David Cox



Numeric (X) = Numeric (Y) Linear regression
We don't apply it on every situation

You can't use any algorithm in any condition



What If you have **Y as categorical variable**?

Logistic Regression, Decision Trees,
SVM, Random Forest

Categorical dependent variable

What is Logistic Regression ?

It is a **Classification algorithm** to predict a **binary outcome** given a set of independent variables (can be numeric or categorical)

It predicts the probability of occurrence of an event by fitting data to a logit function

Important Points:

1. GLM does not assume a linear relationship between dependent and independent variables.
2. Dependent variable need not to be normally distributed
3. It does not uses Ordinary Least Square for parameter estimation. Instead, it uses **maximum likelihood estimation**.
4. Errors need to be independent but not normally distributed

Example:

Predict the probability →

?

Sample of 1000 customers

A customer will buy or not

$g(y) = \beta o + \beta(Age)$ → Linear function →

Concerned about probabilities

Function is established using two things:

- Probability of Success(p)
- Probability of Failure(1-p)

And p should meet following criteria

1. It must always be positive (since p >= 0)
2. It must always be less than equals to 1 (since p <= 1)

People's sex as male or female from their height ?

P(sex = male | height)

We can write in this way

P(X) = P(Y = 1 | X) → The probability that an input (X) belongs to the default class (Y = 1)

$$p(X) = \frac{e^{B0+B1\times X}}{1+e^{B0+B1\times X}}$$

Remove the e from one side by adding a ln( ) to the other

$$ln(\frac{p(X)}{1-p(X)}) = B0 + B1 \times X$$

The input on the left is a natural logarithm of the probability of the default class.

Ratio on the left is called the odds of the default class

The output on the right is linear again same like linear regression

Odds are calculated as a ratio of the probability of the event divided by the probability of not the event

odds(success) = p/(1-p) or p/q = .8/.2 = 4
odds(failure) = q/p = .2/.8 = .25

ln (odds) = B0 + B1 X

move the exponent back to the right

$$odds = e^{B0+B1\times X}$$

The model is still a linear combination of the inputs, but that this linear combination relates to the log-odds of the default class

Learning the Logistic Regression Model

Coefficients estimated from your training data → Maximum-likelihood estimation

The best coefficients would result in a model that would predict a value very close to 1 (e.g. male) for the default class and a value very close to 0 (e.g. female) for the other class.

Probability of 1 if the data is the primary class

Making Predictions with Logistic Regression

Given a height of 150 cm is the person male or female?

coefficients of $B0 = -100$ and $B1 = 0.6$

$$y = \frac{e^{B0+B1\times X}}{1+e^{B0+B1\times X}}$$

$$y = \frac{EXP(-100+0.6\times 150)}{1+EXP(-100+0.6\times X)}$$

$$y = 0.0000453978687$$

$$prediction = 0 \text{ IF } p(male) < 0.5$$

$$prediction = 1 \text{ IF } p(male) \geq 0.5$$

Performance of Logistic Regression Model

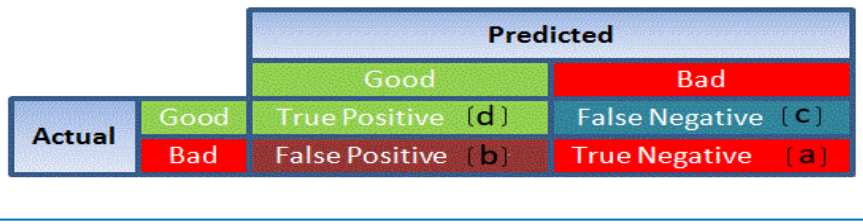
AIC (Akaike Information Criteria):

Just like adjusted R square in linear regression we take a AIC for logistic regression and AIC is measure of fit which penalizes model for the number of model coefficients

Prefer model with minimum AIC value

Confusion Matrix:

A tabular representation of Actual vs. Predicted values, mainly to find the accuracy of the model and avoid over fitting



To calculate accuracy:

$$\frac{\text{True Positive + True Negatives}}{\text{True Positive + True Negatives + False Positives + False Negatives}}$$

Performance of Logistic Regression Model (cont.)

ROC Curve: Receiver Operating Characteristic

A tabular representation of Actual vs. Predicted values, mainly to find the accuracy of the model and avoid over fitting

What it does:

Summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity).

$p > 0.5$

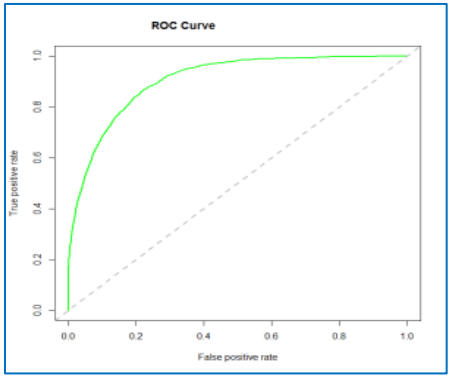
we are more concerned about success rate

$$\text{True Negative Rate (TNR), specificity} = \frac{A}{A+B}$$
$$\text{False Positve Rate (FPR), } 1 - \text{specificity} = \frac{B}{A+B}$$

} sum to 1

$$\text{True Positive Rate (TPR), sensitivity} = \frac{D}{C+D}$$
$$\text{False Negative Rate (FNR)} = \frac{C}{C+D}$$

} sum to 1



The ROC of a perfect predictive model has TP equals 1 and FP equals 0

Prepare Data for Logistic Regression

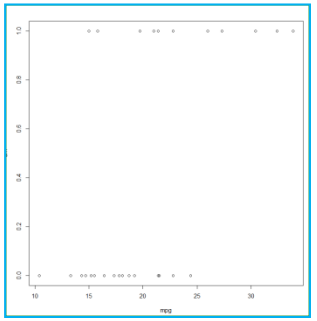
1. **Binary Output Variable**
2. **Remove Noise**
3. **Gaussian Distribution**: log, root and Box-Cox
4. **Remove Correlated Inputs**: model can over fit if you have multiple highly-correlated inputs
5. **Fail to Converge**: remove highly sparse data (e.g. lots of zeros in your input data).

Working with R

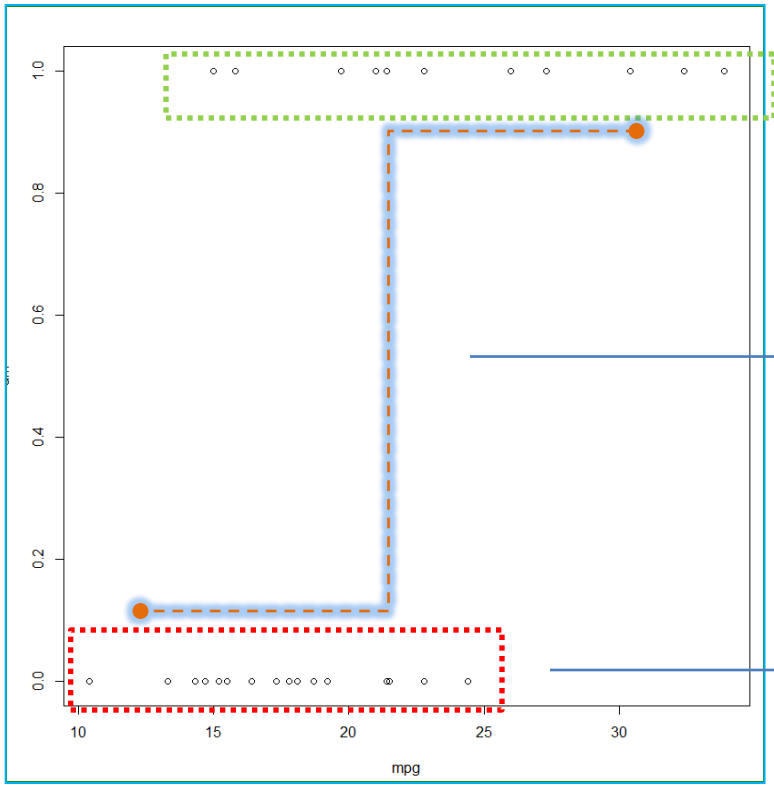
Predict:

Automatic or manual transmission depending upon miles per gallon

Actual plot



When plotted:



MT has high mileage

You can see backward z shape
With no intermediate points on Y axis

Defines odds of 0 or 1 classification these two given X value

AM has less mileage

Summary

Deviance residuals (5 number summary or distribution of the residuals):
Measures model fit it provides

Coefficients:
change of log odds of event occurring based on single unit of increase in X variable.

Intercept:
Interpreted as Log probability of zero value if all the predicted values are equal to zero
But since odds are having automatic transmission with a car that has mpg of zero doesn't make nay sense at all, for this example you can remove intersect.

P value indicates the significance of the model
$P < 0.05$ for mpg indicates that mpg is a significant predictor of cars transmission type.

For every increase in one unit of mpg the log odd of a manual transmission car represented as 1 increases by 0.3070.

Since it is hard to have intuitive understanding of log odds transform the coefficients to regular units.
Exp

For each one unit increase in mpg the likelihood or odds of a car having manual transmission increases by approximately 1.35 or approx. a 36% in the odds

To visualize:
Curve will be added to the exiting plot

"If you give people a good enough 'why', they will always figure out the 'how'."

Jordan Belfort