Logistic Regression

The qualitative data with which are dealing, the binary response variable, can always be coded as having two values, 0 or 1. Rather than predicting these two values we try to model the probabilities that the response takes one of these two values

It’s a classification algorithm, that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and probability of particular outcome.

Questions:

- Two critical questions:

➢Can we run a usual linear regression and interpret the outcome?

➢Since the response variable is qualitative in nature, what do you predict in this case?

We run a linear regression and answer both questions together.

A Business Problem:

- We consider 1340 bodily injury liability claims from a single state , USA using a 2002 survey conducted by the Insurance Research Council (IRC)
- The survey is conducted in order to understand the characteristics of the claimants who choose to be represented by an Attorney when settling a claim
- The profits of an Insurance firm is often found to be dependent on whether the claimant appoints an Attorney or not- appointment of an Attorney can often increase the amount the claimant can claim from the firm
- An insurance firm may be interested in finding the probability of a claimant appointing an Attorney
- Depending on the demographic characteristics of the claimants who appoint an Attorney, the firm aims to design different policy instruments for different target groups

Linear Regression

- Consider the "claimant" dataset
- Dummy Variable ATTORNEY: Attorney=0, if yes
  =1, if not
- Predict the outcome whether claimant is represented by an attorney or not on the following-
  - Claimant's age –CLMAGE (D1)(in years)
  - Claimant's gender- CLMSEX(D2)(0 if Male, 1 if Female)
  - Whether the claimant was wearing seatbelt –SEATBELT (D4) (0 if yes, 1 if no)
  - Whether the driver of the claimant's vehicle was uninsured-CLMINSUR (D5) (0 if yes, 1 if no)
  - The claimant's total economic loss (in thousands) -LOSS (X)

- Specify the regression:

$$Y = \beta_0 + \beta_1 D1 + \beta_2 D2 + \beta_3 D3 + \beta_4 D4 + \beta_5 D5 + \beta_6 X + \varepsilon$$

Linear Regression Output of proposed model:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3293872  0.0555616   5.928 4.11e-09
***
CLMSEX       0.0860886  0.0296149   2.907  0.00372
**
MARITAL     -0.0020250  0.0235048  -0.086  0.93136
CLMINSUR     0.1438686  0.0499335   2.881  0.00404
**
SEATBELT    -0.1194189  0.1102862  -1.083  0.27913
CLMAGE       0.0003384  0.0007514   0.450  0.65257
LOSS        -0.0105399  0.0014182  -7.432 2.17e-13
***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 0.4838 on 1084 degrees of freedom
  (249 observations deleted due to
missingness)
Multiple R-squared:  0.06706,   Adjusted R-squared:  0.0619
F-statistic: 12.99 on 6 and 1084 DF,  p-value:
3.202e-14
```

Observation no	Given y	Predicted value E(y|X)
1325	0	0.481165988
1326	0	0.457571633
1327	0	-0.158865176
1328	0	0.465798861
1329	0	0.494928856
1330	1	0.56770376
1331	1	0.527868666

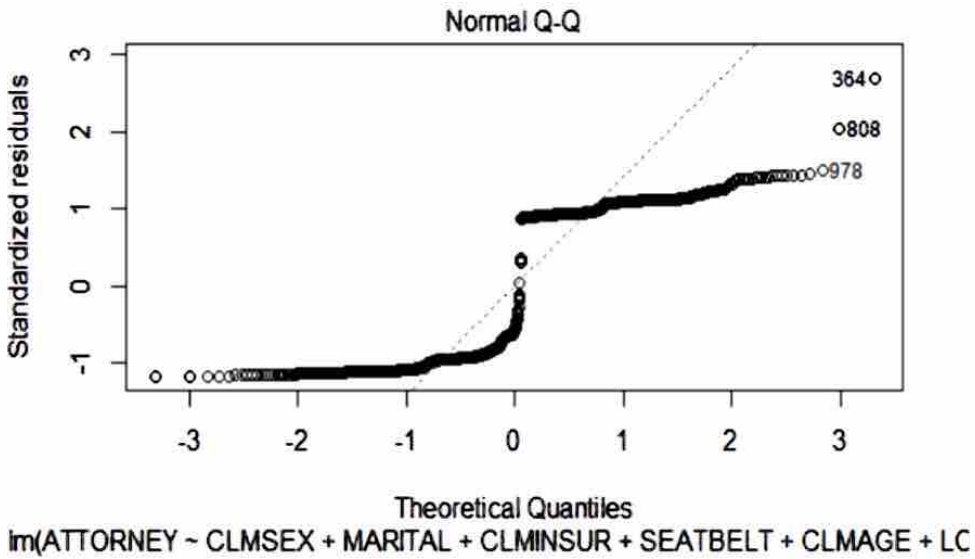
A snapshot output

- Not all predicted values lie between 0 and 1 !
- Some predicted values (predicted probability in LPM) is negative!

Problems with Linear Probability Model:

1. No guarantee that E(Y|X) will lie between 0 and 1

But then probability interpretation doesn't make sense!



A way...Logistic Function

Logistic regression is named for the function used at the core of the method, the logistic function.

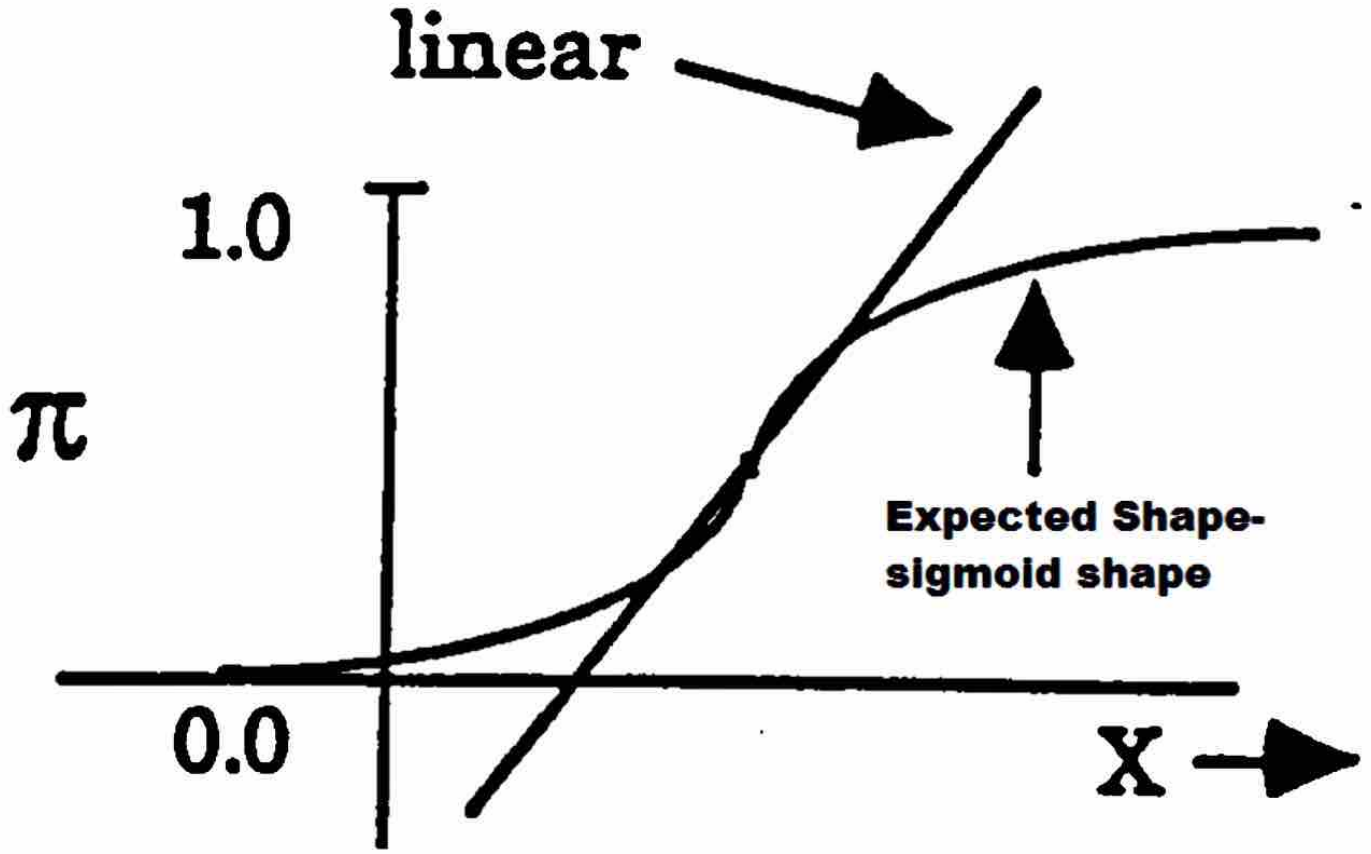
The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth of bacteria in ecology, rising quickly and maxing out at the carrying capacity of the environment.

It's an S-shaped ke any real-valued number and map it into a val 1, but never exactly at those limits.

$$\frac{1}{1 + e^{-value}}$$

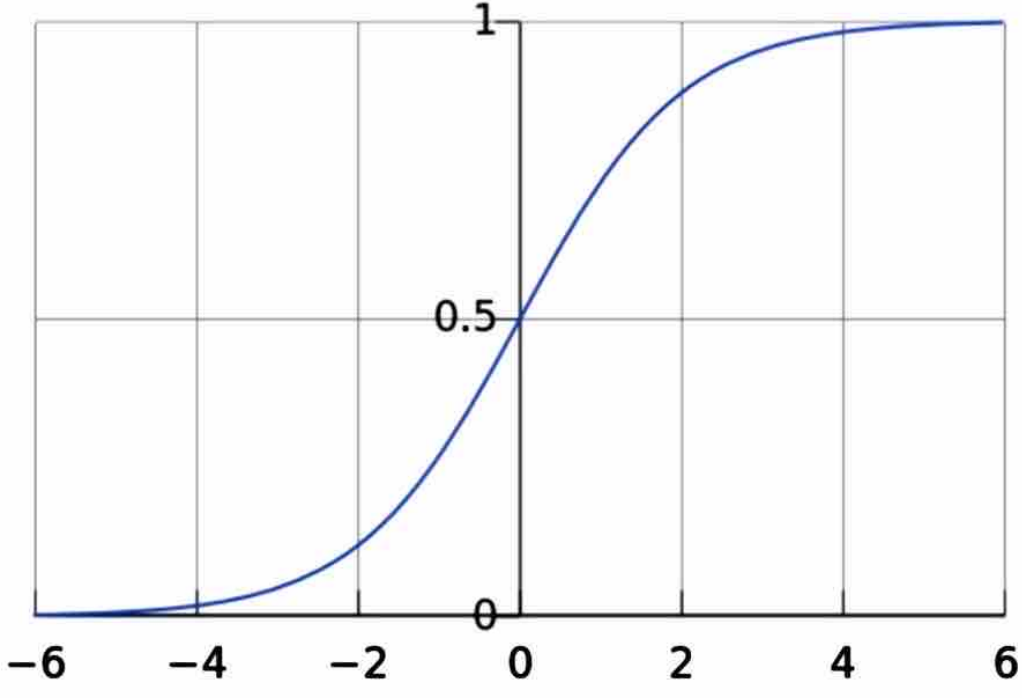
Where e is the base of the natural logarithms and value is the actual numerical value that you want to transform.

Sigmoid Shape Versus Linear Shape



Logistic Curve

- Standard Logistic Sigmoid Function



Logistic Function

Logistic Curve

Logistic regression uses an equation as the representation, very much like linear regression.

Input values (x) are combined linearly using weights or coefficient values to predict an output value (y).

A key deference from linear regression is that the output value being modelled is a binary values (0 or 1) rather than a numeric value.

Classification tables (Confusion matrix)

- Sensitivity
- Specificity
- ROC

Confusion matrix & Goodness of fit

	Observed 1's	Observed 0's
Predicted 1's	a	b
Predicted 0's	c	d

Many counts in a and d boxes and few in b and c boxes indicate good fit.
Sensitivity = a/(a+c)
Specificity = d/(b+d)

High sensitivity and specificity indicate good fit.

Sensitivity or Recall (True Positive Rate): Sensitivity is the conditional probability that the predicted class is positive given that the actual class is positive. Mathematically, sensitivity is given by

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity (True Negative Rate) : Specificity is the conditional probability that the predicted class is negative given that the actual class is negative. Mathematically, specificity is given by

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Precision : Precision is the conditional probability that the actual value is positive given that the prediction by the model is positive. Mathematically,

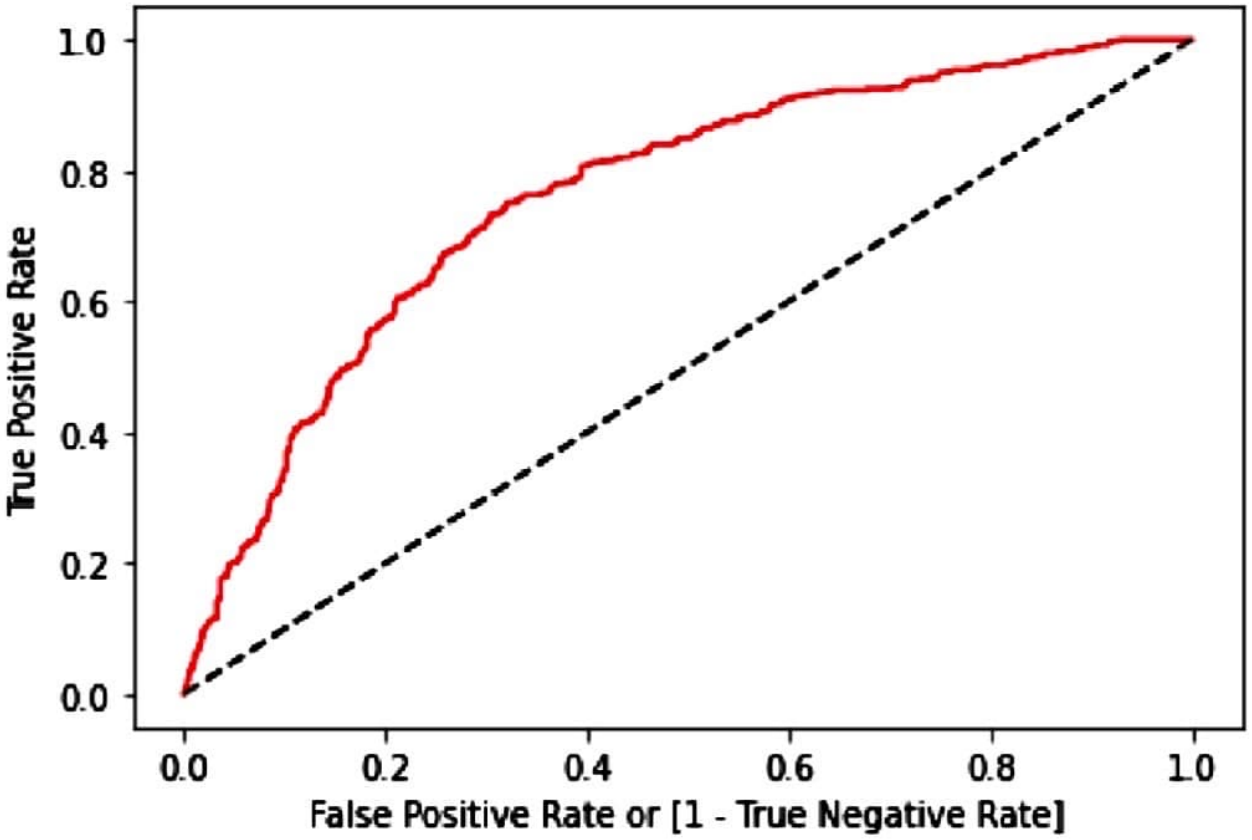
$$\text{Precision} = \frac{TP}{TP + FP}$$

F-Score : F-Score is a measure that combines precision and recall (harmonic mean between precision and recall).
Mathematically, F-Score is given by

$$\text{F-Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

ROC curves

- Receiver operating characteristic (ROC) curve can be used to understand the overall performance (worth) of a logistic regression model (and, in general, of classification models) and used for model
- Extending the above two-by-two table idea, rather than selecting a single cutoff, we can examine the full range of cutoff values from 0 to 1. For each possible cutoff value, we can form a two-by-two table.
- Plotting the pairs of sensitivity versus one minus specificity (True positive vs false positive) on a scatter plot provides an ROC (Receiver Operating Characteristic) curve.
- The area under this curve (AUC of the ROC) provides an overall measure of fit of the model.
- In particular, the AUC provides the probability that a randomly selected pair of subjects, one truly positive, and one truly negative, will be correctly ordered by the test. By "correctly ordered", we mean that the positive subject will have a higher fitted value (i.e., higher predicted


True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0