Predicting the failures of space shuttles (Challenger)

Tool : Logistic regression

The Analytics Edge :

The risk of launch of space shuttles can be estimated using prelaunch tests and data collected from these tests. Using a simple logistic regression model, it is possible to estimate a risk of failure for one of the major disasters in space program launches - The Challenger launch of 1986. This is important to perform probabilistic risk assessments of systems and subsystems in space systems and effects the way NASA conducts evaluation.

Overview :

On Jan 28, 1986, the NASA space shuttle orbiter Challenger broke apart 73 seconds into its flight leading to the death of 7 crew members. The disaster resulted in a 32 month hiatus in the shuttle program and the formation of the Rogers Commission to investigate the accident. This included Neil Armstrong and Richard Feynman (famous physicist) along with several other members of prominence. The commission found the accident was caused by a failure in the O-ring sealing the aft field joint in one of the booster rockets, causing pressurized hot gases and flame to blow by the O ring and make contact with external tank. The commission also concluded that O-rings did not seal well at low temp

Weather forecast on 28 Jan 1986 : Temperature 31°F (= −0.55°C)

Engineers were concerned that the rubber O-rings were vulnerable to failure at low temperature

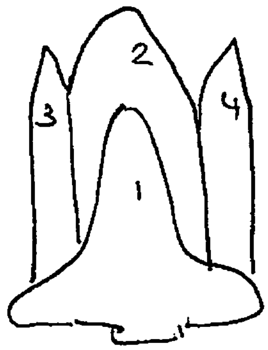
"Excerpt from Rogers Report"

The record of the fateful series of NASA and Thiokol meetings, telephone conferences, notes, and facsimile transmissions on January 27th, the night before the launch of 51-L, shows that only limited consideration was given to the past history of O-ring damage in terms of temperature. The managers compared as a function of temperature the flights for which thermal distress of O-rings had been observed – not the frequency of occurrence based on all flights. In such a comparison, there is nothing irregular in the distribution of O-ring distress over the spectrum of joint temperatures at launch between 53°F and 75°F.

Key question:

What is the chances of Catastrophic O-ring failure if the space shuttle is launched at 31° F ?

Shuttle system



4 subsystems

1 = House crew & control
2 = External fuel tank
3, 4 = Solid rocket motors manufactured by Morton Thiokol

24 launches prior to Challenger.
For one flight, motors were lost at sea,
So motor data was available for 23 flights.

No. of O-ring failures out of 6
Pressure, Temperature

Analytics on Pre-launch Challenger data : R

Data

Orings <- read.csv("Orings.csv")
str(orings)
Summary(orings)

144 observations of 5 variables
Flight (name of flight)
Date (date of flight)
Field (1 if an oring fail
0 otherwise)
Temp (temperature) °F
Pres (Leak check Pressure) PSI
Each flight had 6 orings

tapply(orings$Field, orings$Flight, sum)

Provides the number of Orings that failed out of 6 in each of the flights launched

table(tapply(orings$Field, orings$Flight, sum))

0	1	2	3	(No. of failure)
16	5	1	1	(No. of flights)

plot(orings$Temp[orings$Field > 0], orings$Field[orings$Field > 0])

plot(orings$Temp[orings$Field > 0], jitter(orings$Field[orings$Field > 0]))

plot(jitter(orings$Temp[orings$Field > 0]), orings$Field[orings$Field > 0])

The jitter command helps jitter the data to be able to better visualise points one on top of another

```
plot(jitter(orings$Temp), orings$Field)
```

The plots of temperature with failures only and with failures and non failures provides different information. In the former there are failures across a range with some more at the extremes. In the second case, it is clear that there are lesser failures at higher temperatures. It is believed that analysis of plots such as the first one led the managers to conclude that there was not significant effect of low temperatures.

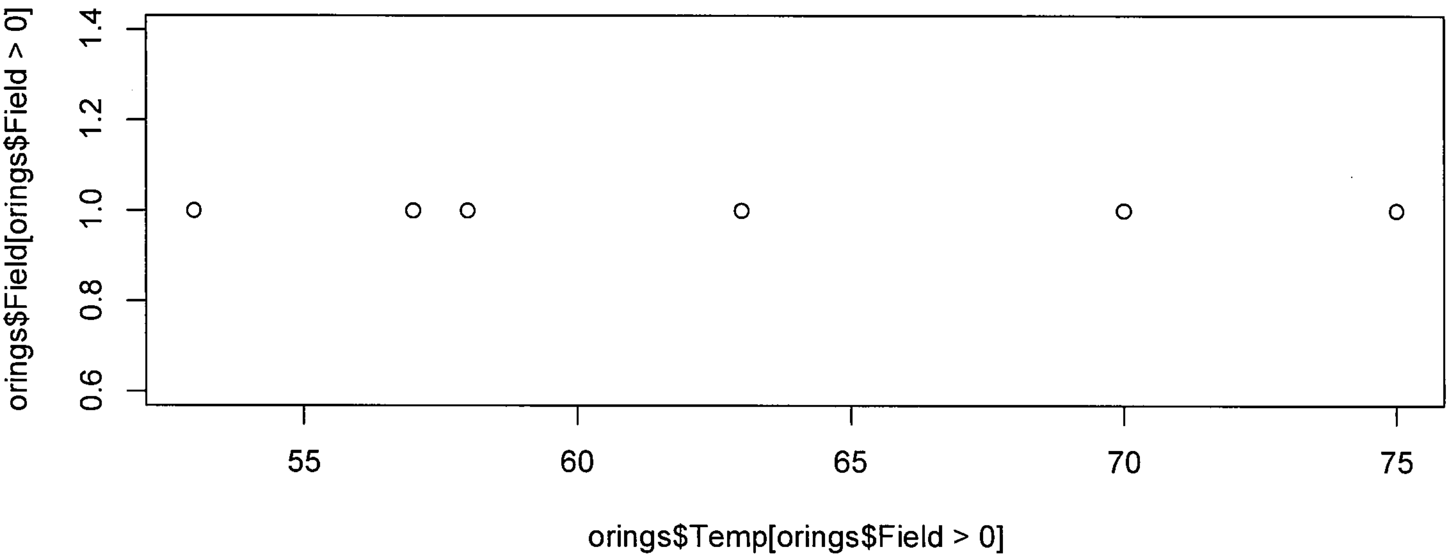
Fitting a model

```
model1 <- lm(Field ~ Temp + Pres, data = orings)
summary(model1)
model2 <- lm(Field ~ Temp, data = orings)
summary(model2)
```

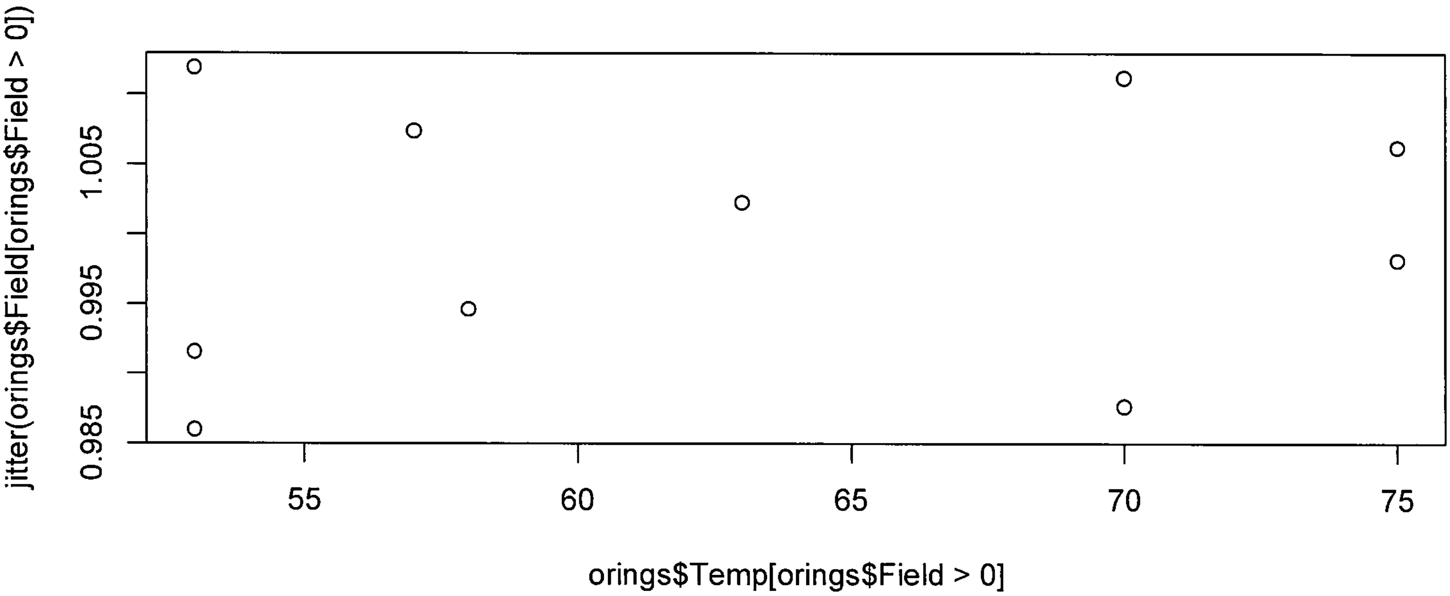
Model 2 gives $R^2 = 0.077$ and a linear fit which is not particularly convincing though it does identify the significance of the temperature and the fact that it has a negative impact.

```
plot(orings$Temp, orings$Field)
abline(model2)
```

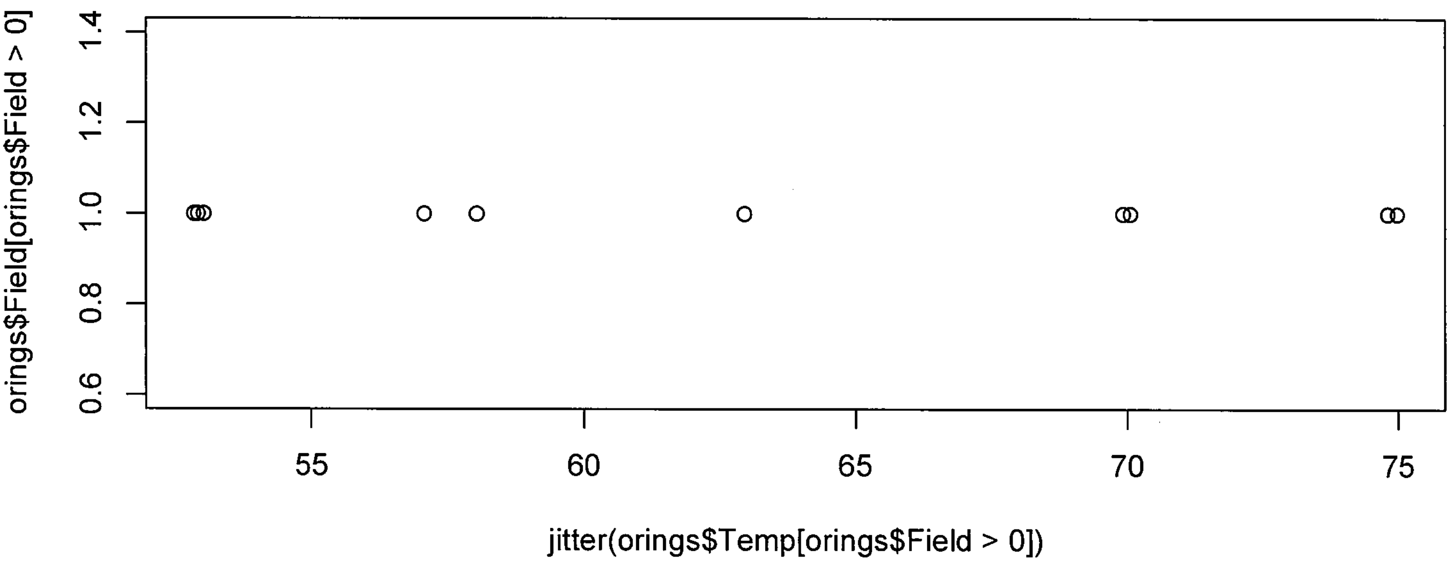
Note that with a linear fit we can predict below 0 and above 1.

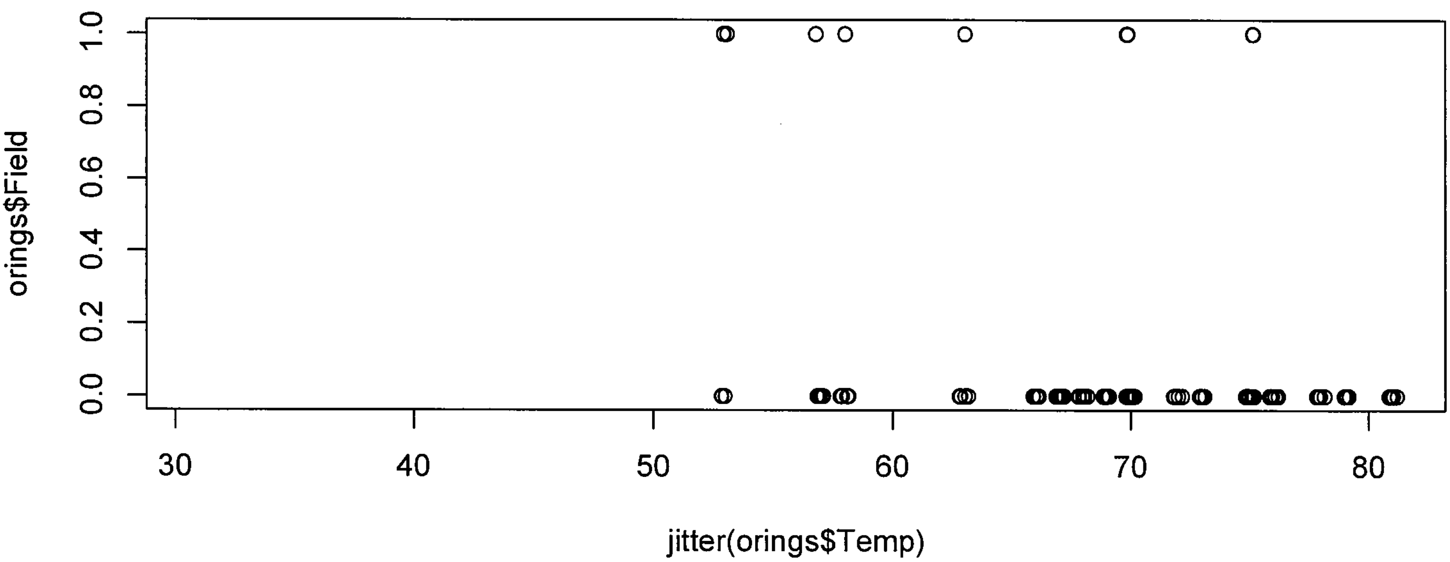


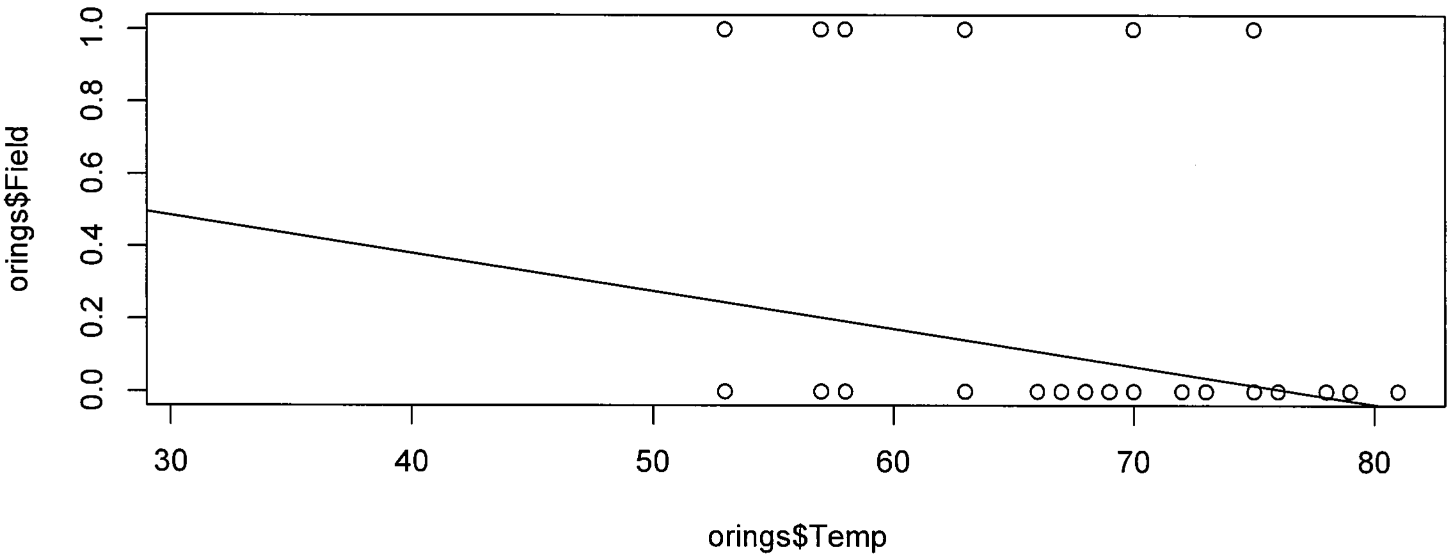
orings$Field[orings$Field > 0]
0.6
0.8
1.0
1.2
1.4
55
60
65
70
75
orings$Temp[orings$Field > 0]



jitter(orings$Field[orings$Field > 0])
0.985
0.995
1.005
55
60
65
70
75
orings$Temp[orings$Field > 0]


orings$Field[orings$Field > 0]
jitter(orings$Temp[orings$Field > 0])
0.6
0.8
1.0
1.2
1.4
55
60
65
70
75


orings$Field
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80
jitter(orings$Temp)


orings$Field
1.0
0.8
0.6
0.4
0.2
0.0
30
40
50
60
70
80
orings$Temp

Fitting a more suitable model

model 3 ← glm(Field ~ Temp + Pres, data = orings, family = binomial)

glm() is a generalized linear model that can be used to fit a logistic regression model by choosing family = binomial

Summary(model 3)

AIC = 66.47

$$\log\left(\frac{P(\text{Fail}=1)}{1-P(\text{Fail}=1)}\right) = \underbrace{\hat{\beta}_0}_{= 3.96} + \underbrace{\hat{\beta}_1}_{= -0.12}\text{Temp} + \underbrace{\hat{\beta}_2}_{0.008}\text{Pres}$$

$$P(\text{Fail}=1) = \frac{e^{3.96-0.12\text{Temp}+0.008\text{Pres}}}{1+e^{3.96-0.12\text{Temp}+0.008\text{Pres}}}$$

Significance level indicates that Temp is significant at 5 % level.

model 4 ← glm(Fail ~ Temp, data = orings , family = binomial)

Summary (model 4)

AIC = 66.083 (Balances log likelihood & number of parameters)

$$P(\text{Fail}=1) = \frac{e^{6.75-0.1397\text{Temp}}}{1+e^{6.75-0.1397\text{Temp}}}$$

Both the intercept & ~~te~~ temperature are significant at 5% level.

We drop the pressure variable here and use model 4.

```
predict(model4, newdata = orings[144,])
```

Prediction gives 2.42 (link value) which is $\hat{\beta}_0 + \hat{\beta}_1(31)$

$$\frac{e^{\hat{\beta}_0 + \hat{\beta}_1(31)}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1(31)}} \quad \text{where } \hat{\beta}_0 = 6.75,\ \hat{\beta}_1 = -0.139$$

```
predict(model4, newdata = orings[144,],
        type = "response")
```

gives predicted probability of failure = 0.918

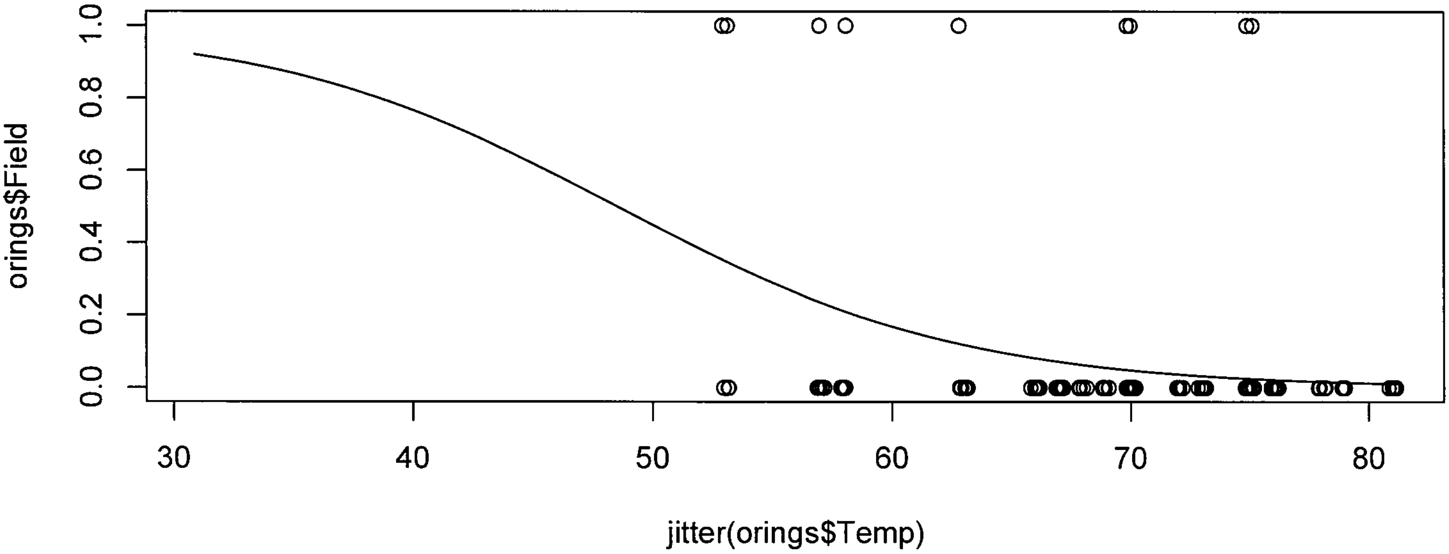
```
plot(jitter(orings$Temp), orings$Field)
~~[illegible] )~~

curve(exp(6.75 - 0.139 * x)/(1+exp(6.75-0.139*x)),
      add = T)
```

Plots data and fitted curve.

Alternatively

```
curve(predict(model4, newdata = data.frame(Temp = x),
      type = "response"), add = T)
```



orings$Field
1.0
0.8
0.6
0.4
0.2
0.0
30
40
50
60
70
80
jitter(orings$Temp)

Developing a predictive rule (classifier)

```
install.packages("ROCR")
library(ROCR)
```

Installs and loads a package that is useful for visualizing the performance of scoring classifiers

```
Pred <- predict(model4, newdata= orings,
                type = "response")
```

Provides the probability of failure for each obs.

(Note that here we are still using the training data. Typically you want test data to verify the results.)

```
Q <- as.numeric(Pred > 0.25)
```

Set 1 if predicted prob > 0.25 and 0 otherwise

```
table(Q[1:138], orings$Field[1:138])
```

Pred \ Actual	0	1
0	125	7
1	3	3

```
Q <- as.numeric(Pred > 0.2)
table(Q[1:138], orings$Field[1:138])
```

Pred \ Actual	0	1
0	115	5
1	13	5

```
Q <- as.numeric(Pred > 0.5)
table(Q[1:138], orings$Field[1:138])
```

Pred \ Actual	0	1
0	128	10

ROCRpred ← prediction (Pred[1:138], oring$Field[1:138])

ROCRperf ← performance (ROCRpred, measure = "tpr",
x.measure = "fpr")

The prediction function transforms data to standardized format and performance function does all kinds of prediction evaluations

plot (ROCRperf)

Uses commands from ROCR package to plot the ROC curve

performance (ROCRpred, measure = "auc")

AUC value for this example is 0.725

		Actual 0	Actual 1
Pred	0	125	7
	1	3	3

$$FPR = \frac{3}{3+125} = 0.023$$

$$TPR = \frac{3}{3+7} = 0.3$$

$$FNR = \frac{7}{7+3} = 0.7$$

$$TNR = \frac{125}{3+125} = 0.976$$

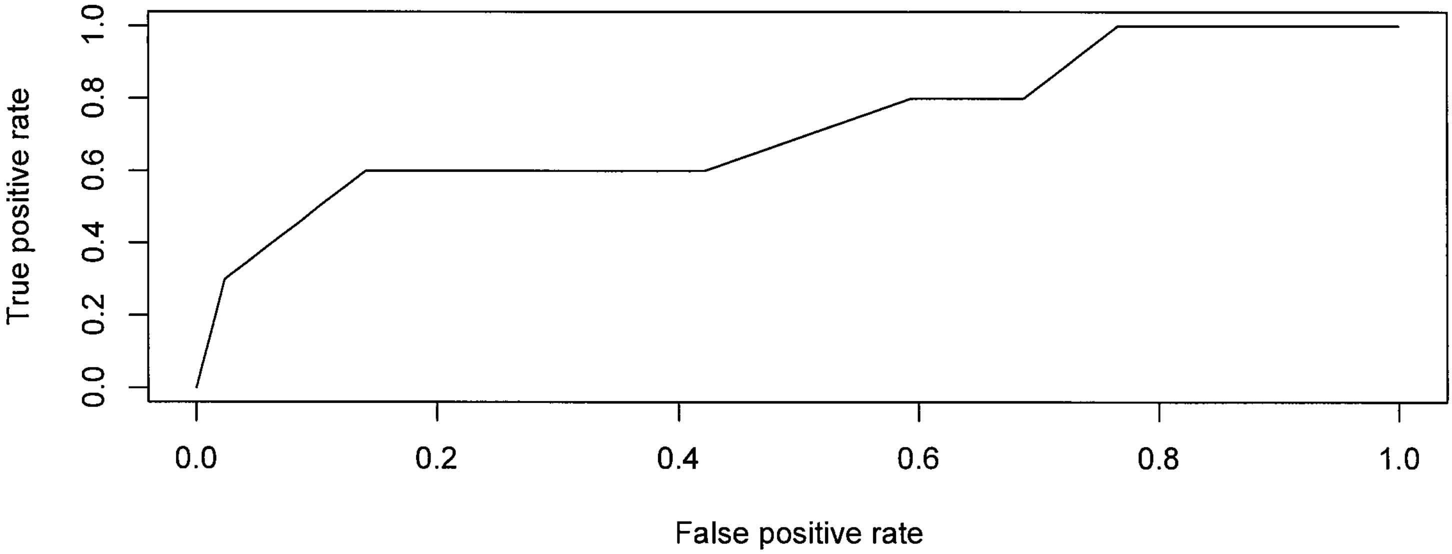
		Actual 0	Actual 1
Pred	0	115	5
	1	13	5

FPR = 0.101

TPR = 0.5

FNR = 0.5

TNR = 0.898

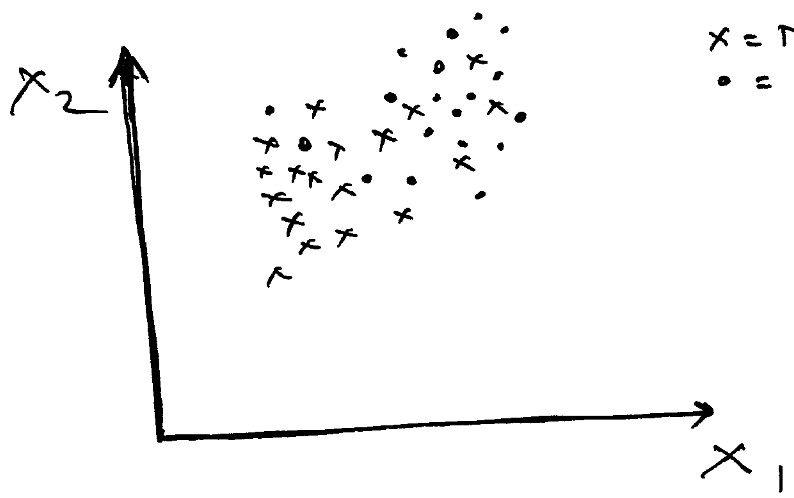

True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

Logistic Regression

Response variable : Qualitative (categorical)
Yes or No

Classification problem - Predicting a qualitative response given the predictor variables (quantitative)

The problem can also be interpreted as a regression problem where the probability that a response is Yes or No is predicted in terms of the predictor variables



$\times$ = No
$\bullet$ = Yes

Use X_1 and X_2 variables to predict if response is Yes or No.

$$\underbrace{Y \in \{0,1\}}_{\text{Output (response)}} \qquad \underbrace{(X_1, \ldots, X_p)}_{\text{Input (predictor)}}$$

Data : $Y_i \in \{0,1\}$ for $i = 1, \ldots, n$

$\bar{X}_i = (X_{i1}, \ldots, X_{ip})^T$ for $i = 1, \ldots, n$

n = Number of observations

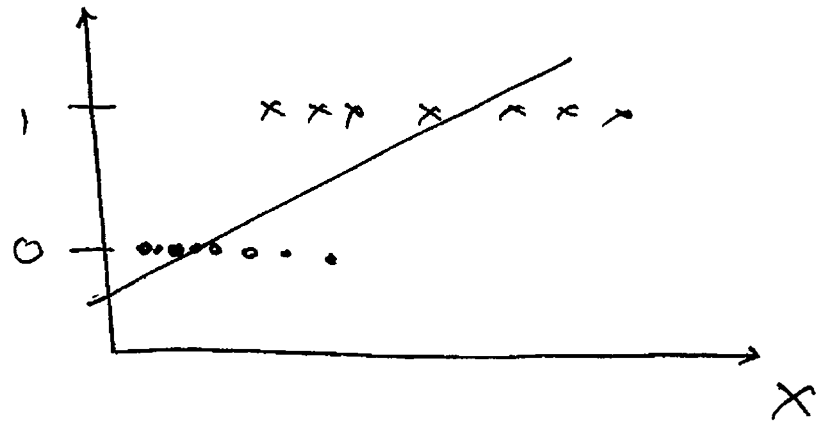
p = Number of predictor variables

Main Points: Logistic Regression

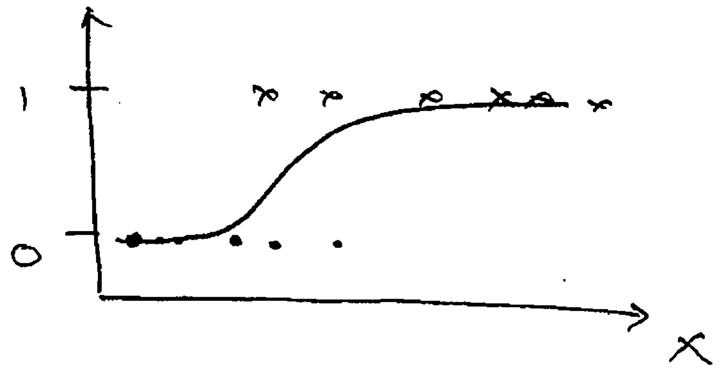
① Estimating $P(Y=1)$ and $P(Y=0)$ given the predictors $X_1, X_2, \ldots, X_p$.

Using linear regression is not suitable since the probability must lie between 0 and 1.

$$P(Y=1) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$



The logistic function provides a nice way to capture this.



S shaped Curve

$$P(Y=1) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}.$$

This number is always between 0 and 1 irrespective of the value of the coefficients and predictors.

$$\text{Odds} = \frac{P(Y=1)}{P(Y=0)} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\dots+\beta_p x_p}$$

Odds > 1 if $Y=1$ is more likely and
odds < 1 if $Y=0$ is more likely.

$$\underbrace{\log(\text{Odds})}_{\substack{\text{Logit or} \\ \text{log-odds.}}} = \beta_0+\beta_1 x_1+\dots+\beta_p x_p$$

is linear in the regression Coefficients

Positive β_j coefficient increases the $P(Y=1)$ if X_j increases. Negative β_j coefficient decreases the $P(Y=1)$ if X_j increases.

However in contrast to linear regression increases the X_j by 1 unit (keeping all other X values the same), changes the log odds by β_j or multiplies the odds by e^{β_j}.

② Maximizing the likelihood function

$$\max_{\beta_0, \beta_1, \dots, \beta_p} \prod_{i: y_i=1} P(Y=1 \mid X = x_i) \prod_{i: y_i=0} P(Y=0 \mid X = x_i)$$

Here x_i is the vector of i th observation predictor variables and y_i is the observed response (0 or 1)

The estimates $\hat{\beta}_0, \dots, \hat{\beta}_p$ are chosen so as to maximize the probability of the actual observed response for each $i = 1, \dots, n$. This is referred to as the likelihood of the observations (assuming each observation is independent of the other).

This problem is solved by taking a logarithm and solving the maximum log-likelihood problem:

$$\max_{\beta_0, \dots, \beta_p} \sum_{i: y_i=1} \log\left(\frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \dots + \beta_p x_{ip}}}\right) + \sum_{i: y_i=0} \log\left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}\right)$$

This objective function is concave and since we are maximizing over β variables, the problem is efficiently solvable and the global optimum can be found efficiently.

Concavity of objective function

Let $\hat{x}_i = \begin{pmatrix} 1 \\ \bar{x}_i \end{pmatrix} \Big\} p+1$ and $\hat{\beta} = \begin{pmatrix} \beta_0 \\ \bar{\beta} \end{pmatrix} \Big\} p+1$

Then
$$Z(\beta) = \sum_i y_i \log\left(\frac{e^{\hat{\beta}'\hat{x}_i}}{1+e^{\hat{\beta}'\hat{x}_i}}\right) + \sum_i (1-y_i) \log\left(\frac{1}{1+e^{\hat{\beta}'\hat{x}}}\right)$$

$$= \sum_i y_i(\hat{\beta}'\hat{x}_i) - \sum_i y_i \log(1+e^{\hat{\beta}'\hat{x}_i})$$
$$+ \sum_i (1-y_i)(0) - \sum_i (1-y_i) \log(1+e^{\hat{\beta}'\hat{x}_i})$$

$$= \underbrace{\sum_i y_i \hat{\beta}'\hat{x}_i}_{\text{Linear in } \hat{\beta}} - \underbrace{\sum_i \log(1+e^{\hat{\beta}'\hat{x}_i})}_{\text{we show this is concave in } \beta.}$$

1) Function $f(t) = -\log(1+e^t)$

$$\frac{df}{dt} = \frac{-e^t}{1+e^t}$$

$$\frac{d^2f}{dt^2} = \frac{-(1+e^t)e^t + e^t(e^t)}{(1+e^t)^2} = \frac{-e^t}{(1+e^t)^2} \leq 0 \ \forall$$

Hence $f(t)$ is concave in t

2) $g(\hat{\beta}) = f(\hat{\beta}'x)$ is concave in $\hat{\beta}$ if $f(t)$ is concave in t.

$\nabla g(\hat{\beta}) = f'(\hat{\beta}'x)x$, $\nabla^2 g(\hat{\beta}) = f''(\hat{\beta}'x)xx' \leq 0$

Hence $Z(\beta)$ is concave in β

Suppose we solve a logistic regression problem only with the intercept

$$\max_{\beta_0} \sum_i y_i \log\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) + \sum_i (1-y_i) \log\left(\frac{1}{1+e^{\beta_0}}\right)$$

Differentiating wrt β_0, we get by setting it to 0:

$$\sum_i y_i - \sum_i \frac{d}{d\beta_0} \log\left(1+e^{\beta_0}\right) = 0$$

$$\therefore \quad \sum_i y_i - \sum_i \frac{e^{\beta_0}}{1+e^{\beta_0}} = 0$$

$$\therefore \quad \sum_i \frac{e^{\beta_0}}{1+e^{\beta_0}} = \sum_i y_i$$

$$\therefore \quad \frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{\sum_i y_i}{n}$$

Choose β_0 such that the estimated fraction of 1's is equal to observed fraction of 1's

(3) Quality of fit

Deviance is a measure of fit of the generalized linear model. (Higher numbers indicate worst fit)

Null deviance measures how well the response variable is predicted by a model that includes just the intercept.

Residual deviance measures how well the response variable is predicted by the intercept and the additional predictor variables (p).

A significant decrease in the value from null to residual deviance indicates that the predictor variables are useful in making good predictions

For logistic regression problems,

$$\text{Null deviance} = -2\,LL(\text{Only intercept})$$

$$\text{Residual deviance} = -2\,LL(\underbrace{\hat{\beta}}_{\text{Intercept} + p \text{ variables}})$$

Akaike information criterion (AIC) is based on deviance but penalizes for making the model more complicated (similar to adjusted R^2). However the AIC does not have a range to benchmark unlike R^2 in $[0,1]$.

Smaller the AIC, better the fit.

$$AIC = -2\,\underbrace{LL(\hat{\beta})}_{\text{Log likelihood at } \hat{\beta}} + 2(\underbrace{p}_{\text{Parameters}} + \underbrace{1}_{\text{intercept}})$$

Confusion matrix, Sensitivity, specificity, ROC curves

Confusion matrix

Actual (Truth)

	Actual = 0	Actual = 1
Predict = 0	True negative (TN)	False negative (FN)
Predict = 1	False positive (FP)	True positive (TP)

$$P(Y=1) \geq t \Rightarrow \text{Predict } Y=1$$
$$P(Y=1) < t \Rightarrow \text{Predict } Y=0$$

Rule to classify or predict based on a number t (threshold value)

For example $t = 0.5$

Varying the threshold, changes the entries in the confusion matrix and affects the false positive rate, true positive rate, true negative rate, false negative rate.

False positive rate (Type I error) $\quad FPR = \dfrac{FP}{FP+TN}$

Specificity (True negative rate) $\quad TNR = \dfrac{TN}{FP+TN}$

True positive rate (sensitivity) $\quad TPR = \dfrac{TP}{TP+FN}$

False negative rate (Type II error) $\quad FNR = \dfrac{FN}{TP+FN}$

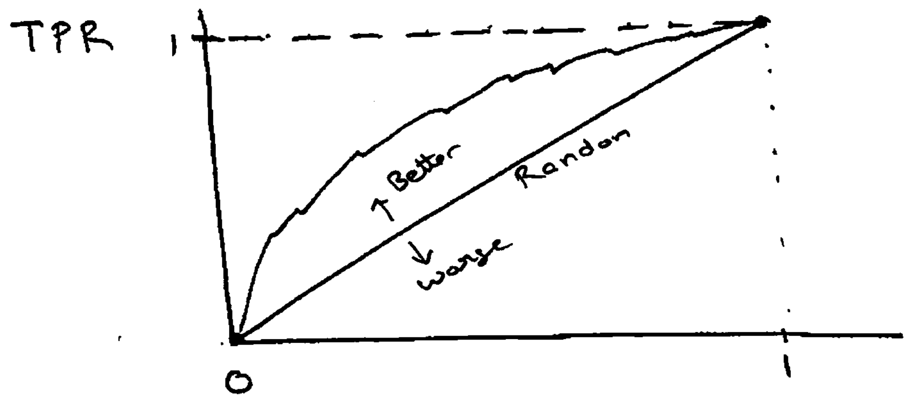
$$FPR + TNR = 1$$

$$TPR + FNR = 1$$

$$\text{Accuracy} = \frac{TN+TP}{TN+FN+TP+FP}$$

ROC curve (Receiver operating characteristic curve)

Rather than computing TPR and FPR for a fixed threshold t, the ROC curve plots TPR vs FPR as an implicit function of t



ROC curve

Setting $t = 0$ $\Rightarrow$ All predictions are $Y = 1$ (positive)
Then $FPR = TPR = 1$

Setting $t = 1$ $\Rightarrow$ All predictions are $Y = 0$ (negative)
Then $FPR = TPR = 0$

If a model is performing at the level of chance then we can achieve a point along the diagonal $FPR = TPR$. (random guessing by flipping a coin)

A system that perfectly separates $Y = 1$ from $Y = 0$ (positive and negative labels) has a ROC curve that hugs the left axis and then top axis $\begin{pmatrix} FPR = 0 \\ TPR = 1 \end{pmatrix}$

Overall performance of classifier over all possible thresholds = Area under the ROC curve (AUC)

A good model has AUC closer to 1. (Good predictive power)

AUC of a Classifier is the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

Random performance

A classifier that randomly guesses the positive class (1) half the time is expected to get half the positives and half the negatives correct (0.5, 0.5) on ROC curve.

A classifier that guesses the positive class randomly 90% of the time is expected to get 90% of positives correct but FPR will also increase to 90%. (0.9, 0.9) on ROC curve.

AUC of random guess = 0.5