1 Risk analysis of a space shuttle launch (the Challenger disaster)

<u>Tool</u>: Logistic regression

<u>The Analytics Edge</u>: Launching shuttles into space is an exercise requiring long periods of planning and testing. The risk involved in such space shuttle launches can be estimated using pre-launch tests and data collected from these tests. Using a logistic regression model, we estimate the risk of failure for one of the major disasters in space program launches. The value of the model lies in the fact that the risk is estimated using observable data from pre-launch tests and can be used to make predictions for new values of the input that might be unobservable in the original dataset. The model shows that it is important to perform careful risk assessment of systems and subsystems in space programs and has affected the way space organizations conduct risk evaluation.

2 Overview

On January 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 shuttle programs and the formation of the Rogers Commission to investigate the accident. The members of the commission included the first man to walk on the moon Neil Armstrong, Nobel prize winner theoretical physicist Richard Feynman, along with several other members of prominence. The Commission found that the accident was most probably caused by a failure in the O-ring sealing the aft field joint in one of the booster rockets, causing pressurized hot gases and flames to blow by the O-ring, making contact with the external tank. The Commission also concluded that O-rings did not seal well with low temperature. The weather forecast on January 28, 1986 was 31° Fahrenheit (or -0.55° Celsius). Engineers were concerned that the rubber O-rings were vulnerable to failure at low temperatures. The space shuttle still took off however.

2.1 Excerpt from Rogers Commission Report

The report of the fateful series of NASA and Thiokol meetings, telephone conferences, notes and facsimile transmissions on January 27, 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 occurrences 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° and 75° Fahrenheit.

2.2 Key question and available data

What are the chances of a catastrophic O-ring failure if the Space Shuttle is launched at 31° Fahrenheit?

There were 24 launches prior to the fateful Challenger disaster. For one of the flights, motors were lost at the sea. So motor failure data was available for each of the six O-rings for each of the 23 flights. The data on the temperature and pressure information for these launches is also available.
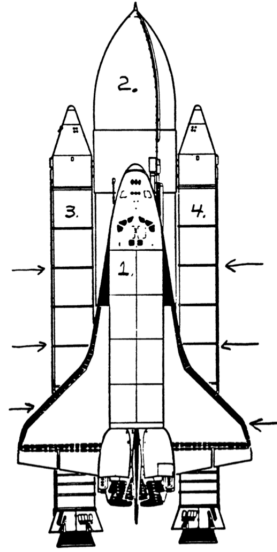
Figure 1: Schematic of Challenger shuttle. Subsystem 1 houses crew and system control, subsystem 2 is the external liquid-fuel tank, and subsystems 3 and 4 are the solid rocket motors (Morton Thiokol)

3   Summary

Data: The data is a small dataset that provides information on test launches.

Model: A logistic regression model helps characterize the probability of failure of an event with variables that are identified as potential predictors. The output variable here is binary and the model has to account for this explicitly.

Decision: The model provides an estimate of the probability of failure. If this number is even reasonably high, one would expect to not launch such a space shuttle. It is of course unlikely that in any practical complex system, this number will be exactly zero.

Value: The value of the model is to provide estimates for new scenarios that one might not see in the original dataset (very low temperatures as in this example). Lives might be saved in such applications.
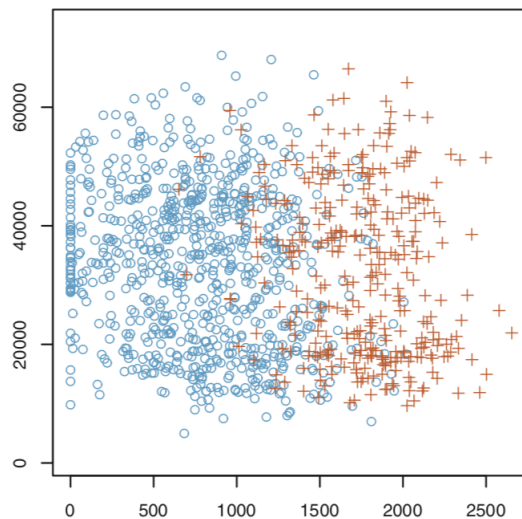
4   Logistic Regression

Figure 2: Classification

Problem setup:

1. Number of observations $= n$.

2. Number of predictor variables (excluding the constant 1) $= p$.

3. Dependent variable (binary outcome, yes or no, qualitative): $y \in \{0, 1\}$.

4. Independent variables (predictors): $x_1, \ldots, x_p$.

Our data set consists of $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$ where $y_i \in \{0, 1\}$ and $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})$.

This problem can be viewed as a classification problem where the problem is to predict which class (0 or 1) an observation belongs to or alternatively as a regression problem where the probability that the output is of a particular class (a number in the range $[0, 1]$) is predicted in terms of the independent variables.

We are interested in estimating the linear model:

$$y^* \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon.$$

However in this case, we do not observe y^* (a latent variable), but rather we observe a variable y which is defined as follows:

$$y = 1 \text{ if } y^* \geq 0, \text{ and } y = 0 \text{ otherwise.}$$

4.1   Key ideas

1. The problem is to estimate $\Pr(Y = 1)$ and $\Pr(Y = 0)$ in terms of the predictor variables $\{x_1, \ldots, x_p\}$. Using a linear regression model is not suitable since the probability must lie between 0 and 1. We think of Y as the random variable representing the response.
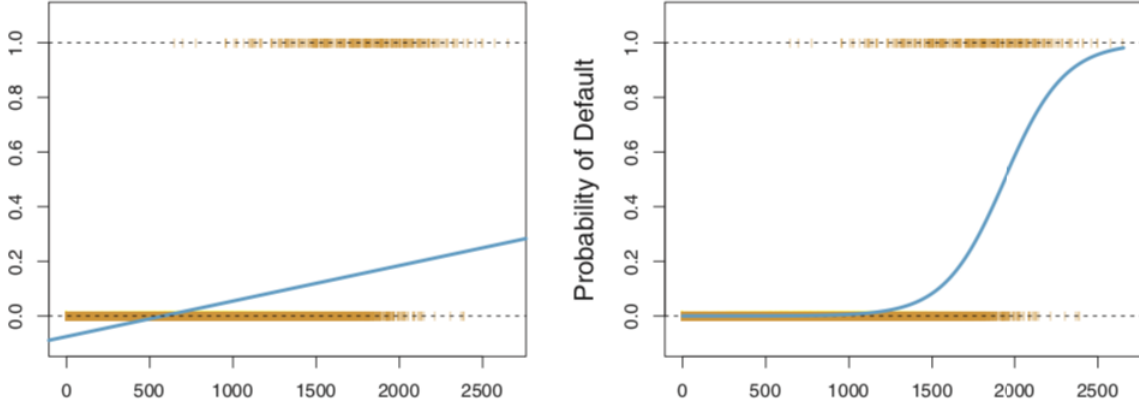

Figure 3: Linear regression and Logistic Regression

Define the vector of parameters:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

The logistic function (S-shaped) provides a nice formulation to capture this:

$$p(\boldsymbol{x}, \boldsymbol{\beta}) := \Pr(Y = 1 | \boldsymbol{x}, \boldsymbol{\beta}) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}} = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}}.$$

This number is always between 0 and 1, irrespective of the value of the x variables.

2. Odds is defined as:

$$\text{Odds} \quad = \quad \frac{\Pr(Y = 1 | \boldsymbol{x}, \boldsymbol{\beta})}{\Pr(Y = 0 | \boldsymbol{x}, \boldsymbol{\beta})} = \frac{p(\boldsymbol{x}, \boldsymbol{\beta})}{1 - p(\boldsymbol{x}, \boldsymbol{\beta})}$$

$$= \quad e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p} = e^{\boldsymbol{\beta}^T \boldsymbol{x}}.$$

Odds > 1 if $Y = 1$ is more likely and Odds < 1 if $Y = 0$ is more likely, given a particular x.

The logit or log-odds is defined as:

$$\text{Log(Odds)} \quad = \quad \boldsymbol{\beta}^T \boldsymbol{x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

and is linear in the beta coefficients. Note that a positive β_j coefficient indicates that $\Pr(Y = 1)$ increases, if x_j increase. Similarly, a negative β_j coefficient indicates that $\Pr(Y = 1)$ decreases, if x_j increases. However in contrast to linear regression, increasing the x_j variable by 1 unit (keeping all of the other x_k values fixed), changes the log-odds by β_j or multiplies the odds by e^{β_j}.

To estimate the coefficients, we maximize the likelihood function which is defined as:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Pr(Y = 1 | \boldsymbol{x} = \boldsymbol{x}_i; \boldsymbol{\beta})^{y_i} \Pr(Y = 0 | \boldsymbol{x} = \boldsymbol{x}_i; \boldsymbol{\beta})^{1-y_i}.$$

The maximum likelihood problem is defined as:

$$\max_{\beta} \prod_{i=1}^{n} \Pr(Y = 1 | \boldsymbol{x} = \boldsymbol{x}_i; \boldsymbol{\beta})^{y_i} \Pr(Y = 0 | \boldsymbol{x} = \boldsymbol{x}_i; \boldsymbol{\beta})^{1-y_i},$$

where the objective function is the likelihood of the observations, assuming that each observation is independent of the other. The beta coefficients are estimated to maximize the probability of observing the actual responses under the model for each observation $i = 1, \ldots, n$. This problem is equivalently solved by the maximum log-likelihood problem:

$$\max_{\beta} \log L(\boldsymbol{\beta})$$

where the log-likelihood is defined as:

$$\log L(\boldsymbol{\beta}) = LL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log p(\boldsymbol{x}_i, \boldsymbol{\beta}) + (1 - y_i) \log(1 - p(\boldsymbol{x}_i, \boldsymbol{\beta})) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \boldsymbol{\beta}^T \boldsymbol{x}_i - \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}) \right].$$

This objective function is concave in the beta variables and we are solving a maximization problem over these variables. In this case, the local optimum is the global optimum and the problem is efficiently solvable. However, we need to use an iterative algorithm to find the optimal solution in this case.

Algorithm: The optimality condition is given by the system of equations:

$$\nabla LL(\boldsymbol{\beta}) := \sum_{i=1}^{n} (y_i - p(\boldsymbol{x}_i, \boldsymbol{\beta})) \, \boldsymbol{x}_i = 0.$$

Let $\boldsymbol{y}$ denote a column vector of size n with entries y_i and $\boldsymbol{X}$ as the input data matrix of size $n \times (p+1)$ with rows denoting the observations and columns denoting the intercept and the predictors. Applying Newton's method to solve this problem gives the iterates:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + (\nabla^2 LL(\boldsymbol{\beta}^{(k)}))^{-1} \nabla LL(\boldsymbol{\beta}^{(k)}).$$

The gradients and the Hessian matrix are computed as:

$$\nabla LL(\boldsymbol{\beta}) = \boldsymbol{X}^T (y - p(\boldsymbol{X}, \boldsymbol{\beta}))$$

$$\nabla^2 LL(\boldsymbol{\beta}) = -\boldsymbol{X}^T W(X, \boldsymbol{\beta}) \boldsymbol{X}$$

where $p(\boldsymbol{X}, \boldsymbol{\beta})$ is a vector with entries as $p(\boldsymbol{x}_i, \boldsymbol{\beta})$ and $W(X, \boldsymbol{\beta})$ is a diagonal matrix with entries $W_{ii} = p(\boldsymbol{x}_i, \boldsymbol{\beta})(1 - p(\boldsymbol{x}_i, \boldsymbol{\beta}))$.

4.2 Quality of fit

1. Deviance is a measure of the fit of the generalized linear model (of which logistic regression is an example) where higher numbers indicates a worse fit.

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

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

   A significant decrease in the value from the null to residual deviance indicates that the predictor variables are useful in explaining the response variable. For logistic regression problems, suppose we estimate these coefficients as $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$, then:

   $$\text{Null deviance } = -2LL(\hat{\beta}_0),$$

   where $\hat{\beta}_0$ is the estimated coefficient when only the intercept is fit and

   $$\text{Residual deviance } = -2LL(\hat{\boldsymbol{\beta}}).$$

   We can verify that when only $\hat{\beta}_0$ is fit in the maximum likelihood estimation, it is chosen that the estimated fraction of 1's from the model is exactly equal to the observed fraction of 1's in the training set.

   The Akaike information criterion (AIC) is based on deviance but penalizes for making the model more complicated (similar to adjusted R-squared). However the AIC does not have a benchmark range like R-squared which lies in the inteval $[0, 1]$. Smaller the AIC, the better is the model. AIC is defined as:

   $$AIC = -2LL(\hat{\boldsymbol{\beta}}) + 2(p+1),$$

   where the p coefficients and the intercept coefficient are estimated.

2. Confusion matrix: The confusion matrix is a table that allows to visualize the performance of a classification algorithm as follows.

   Suppose we use the following rule to classify or predict an output:

   (a) Choose a threshold t

   (b) For any observation i with predictor variables x_i and estimated coefficients $\hat{\boldsymbol{\beta}}$:
   If $\Pr(Y = 1|\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}) \geq t$, then predict 1, else if $\Pr(Y = 1|\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}) < t$, predict 0

   This results in the confusion matrix as follows:

	Actual = 0	Actual = 1
Predict = 0	True Negative (TN)	False Negative (FN)
Predict = 1	False Positive (FP)	True Positive (TP)

   Table 1: Confusion matrix

We can define the following quantities:

$$\text{True negative rate:} \quad TNR = \frac{TN}{FP + TN} \qquad \text{(Specificity)},$$

$$\text{False positive rate:} \quad FPR = \frac{FP}{FP + TN} \qquad \text{(Type I error)},$$

$$\text{True positive rate:} \quad TPR = \frac{TP}{TP + FN} \qquad \text{(Sensitivity)},$$

$$\text{False negative rate:} \quad FNR = \frac{FN}{TP + FN} \qquad \text{(Type II error)},$$

$$\text{Overall Accuracy:} \quad Accuracy = \frac{TP + TN}{FP + TN + TP + FN}.$$

Note that by definition, $FPR + TNR = 1$ and $TPR + FNR = 1$. By varying the threshold, we change the entries in the confusion matrix and the corresponding FPR, TPR, FNR, TNR.

3. Receiver operating characteristic (ROC) curve: Rather than computing the TPR and FPR for just a fixed threshold t, the ROC curve plots $TPR(t)$ and $FPR(t)$ as a function of t
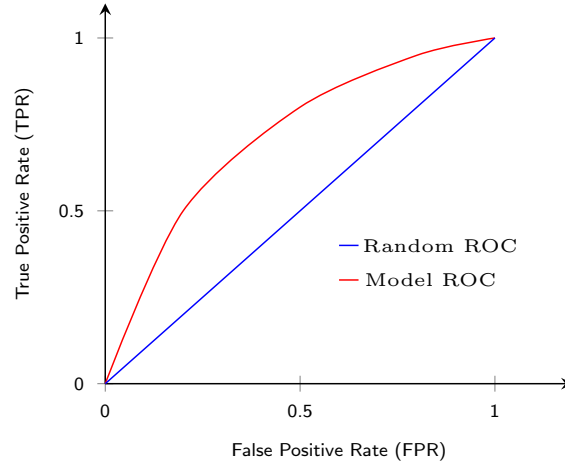

Figure 4: ROC Curve

(a) Setting $t = 0$: All predictions are $y = 1$. Then $FPR = 1$ and $TPR = 1$.

(b) Setting $t = 1$: All predictions are $y = 0$. Then $FPR = 0$ and $TPR = 0$.

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

(d) A system that perfectly separates $y = 1$ from $y = 0$ has a ROC curve with $FPR = 0$ and $TPR = 1$.

4. Area under the curve: The overall performance of a classifier over all thresholds is given by the area under the ROC curve (AUC). A good predictive model has AUC closer to 1.

The AUC of a classifier is the probability that a classifier will rank a randomly chosen instance with $y = 1$ higher than a randomly chosen $y = 0$ instance. A classifier that randomly guesses the class $y = 1$ half the time is expected to get half the labels with $y = 1$ and 0 correctly (the point $(0.5,0.5)$ on the ROC curve). A classifier that randomly guesses the class $y = 1$, 90% of the time is expected to get 90% of the positives correct but FPR will also increase to 90% (the point $(0.9,0.9)$ on the ROC curve). The AUC of a random classifier is hence 0.5.