L22: Logistic regression

Department of Statistics
University of South Carolina

Elementary Statistics for the Biological and Life Sciences

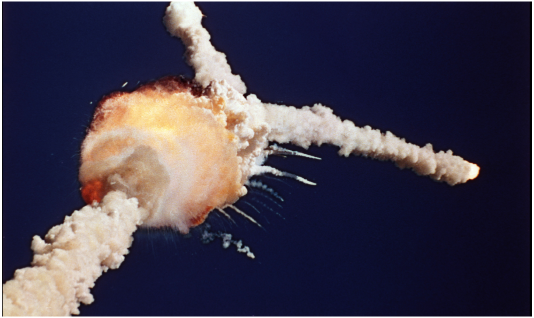
(STAT 205)

Logistic regression

- So far, we have learnt how to find a good simple linear regression model to fit the data, whose reponse variable Y is quantitative, e.g. continuous numbers.
- Sometimes we wish to predict a categorical response Y using a quantitive variable X.
- Consider Y to be binary ($0 =$ failure, $1 =$ success)
- Logistic regression is used to model how the probability of success (p) depends on X.
- Rather than normally distributed data we now have binomially distributed data.

Example: O-Ring Failure

The **Space Shuttle Challenger disaster** occurred on January 28, 1986, when the NASA space shuttle orbiter *Challenger* broke apart 73 seconds into its flight, leading to the deaths of its seven crew members. Disintegration of the vehicle began after an **O-Ring** seal in its right solid rocket booster failed at liftoff.



Example: O-Ring Failure

O-Ring seal failed becuase the launch temperature is lower than expected. Therefore, it is critical to carefully test the reliability of O-Ring under different circumstance. Here we have 24 data points, including the lauching temperature and whether at least one O-Ring failure has occured.

Table: My caption

O-Ring Failure	Temperature
1	52
1	56
1	57
0	63
0	66
...	...
0	81

Logit Function

- We want to model and predict the probability to success

$$\Pr\{\text{success}\} = \Pr\{Y = 1\} = p$$

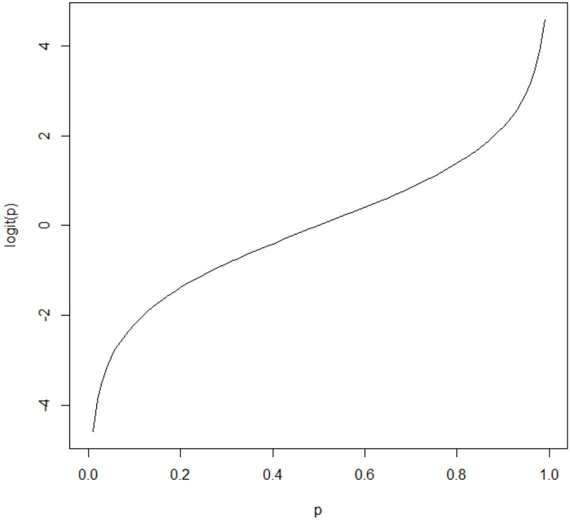
- If we still use the simple linar regression method to regress Y, the results might not be interpretable when the predicted value of Y is greater than 1 or less than 0.
- **Logit function** is frequently used in mathematics and statistics. It is defined as

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right), \qquad 0 < p < 1$$

- The reason it is popular is that it can transfer a random variable from $(0, 1)$ to the entire real line.

- Note that when p is close to 0, $\text{logit}(p)$ is close to $-\infty$, and when p is close to 1, $\text{logit}(p)$ is close to ∞.

Shape of the Logit Function



Logistic Regression Model

- Using the property of the logit function, the logistic regression model is that
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$
where p_i is the success probability for ith unit, x_i is the ith predictor, β_0 and β_1 are parameters.
- Solving for p_i, this gives
$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$$
- Remark: logistic regression is NOT a model simply transfer Y_i with the logit function. The logit transformation is conducted with respect to p_i, the probability for i to success.
- Remark: $\frac{p}{1-p}$ is **odds**, so that the logit function, $\log\left(\frac{p}{1-p}\right)$, is the **log odds**.

R code for O-Ring Failure

```
# input data
oring <- c(1,1,1,0,0,0,0,0,0,0,0,1,1,1,0,0,0,1,0,0,0,0,0,0)
temperature <- c(53, 56, 57, 63, 66, 67, 67, 67, 68, 69, 70,
  70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 80, 81)

# fit the logistic regression model
fit <- glm(oring ~ temperature, family=binomial)
summary(fit)

# R output
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.87535    5.70291   1.907   0.0565 .
temperature -0.17132    0.08344  -2.053   0.0400 *
```

Example: O-Ring Failure

- The fitted logistic regression model is

$$\log\left(\frac{p_i}{1-p_i}\right) = 10.875 - 0.171x_i$$

- $e^{\hat{\beta}_1} = e^{-0.171} = 0.843$, so every 1 degree of increase in temperature reduces the odds of failure by 0.843.
- It is equivalent to

$$p_i = \frac{1}{1 + e^{-(10.875 - 0.171x_i)}}$$

- Remark: in testing $H_0 : \beta_1 = 0$ v.s. $H_a : \beta_1 \neq 0$, the p-value is 0.04, indicating that the linear relationship between $\log\left(\frac{p_i}{1-p_i}\right)$ and x_i is significant.

Example: O-Ring Failure

- The actual temperature at the *Challenger* launch was 31 F.

$$p_i = \frac{1}{1 + e^{-(10.875 - 0.171(31))}} = 0.996$$

- The probability that at least one O-Ring failure is 99.6%! It is almost certainly going to happen!
- It is interesting to note that all of these data were available **prior** to launch. However, engineers were unable to effectively analyze the data and use them to provide a convincing argument against launching *Challenger* to NASA managers.