Stat 343: Logistic Regression

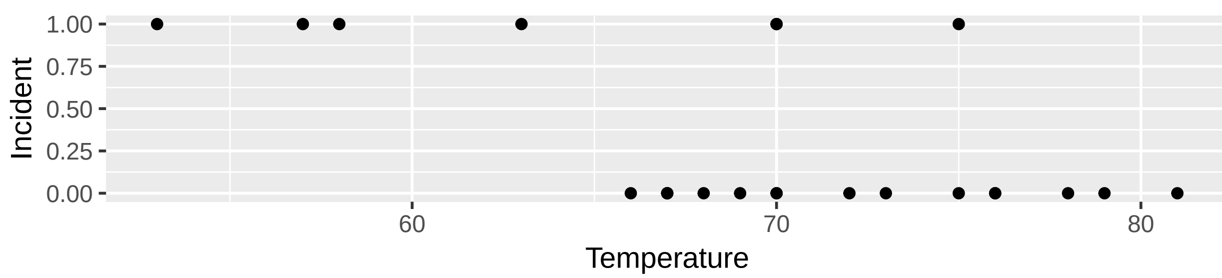
In the warm-up, we just described overall patterns:

- The proportion of Challenger launches with O-ring damage is $7/23 \approx 0.304$
- The proportion of Loblolly pine trees that were mature is $223/644 \approx 0.346$

Key Question: (How) can we say more if we have more information/covariates?

Data Set 1: Challenger Space Shuttle O-Rings

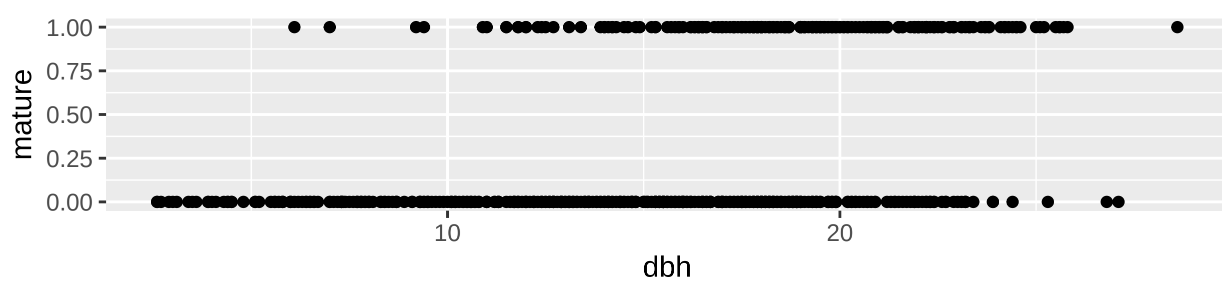
$$Y_i = \begin{cases} 1 & \text{if there was evidence of damage to on O-ring on launch number } i \\ 0 & \text{otherwise} \end{cases}$$
$$X_i = \text{ temperature at launch for launch number } i$$



Observation: O-ring damage is more likely if the temperature is low

Data Set 2: Loblolly Pines

$$Y_i = \begin{cases} 1 & \text{if pine tree number } i \text{ is mature} \\ 0 & \text{otherwise} \end{cases}$$
$$X_i = \text{ diameter at breast height (a measure of the tree's size) for pine tree number } i$$



Observation: Larger trees are more likely to be mature.

How to model $Y_i | X_i = x_i$?

Logistic Regression:

Model:

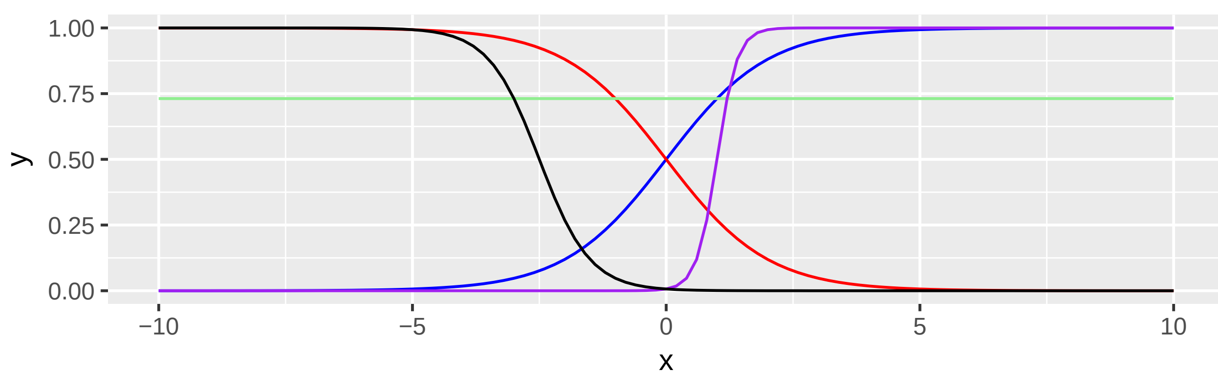
Y_i follows a Bernoulli distribution where the probability of success depends on x_i:

$$Y_i | X_i = x_i \sim \text{Bernoulli}(p(x_i | \beta_0, \beta_1))$$
$$p(x_i | \beta_0, \beta_1) = P(Y_i = 1 | X_i = x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

This function is called the **logistic function**.

```
logistic <- function(x, beta_0, beta_1) {
  return(plogis(beta_0 + beta_1 * x))
  # the above is equivalent to return(exp(beta_0 + beta_1 * x) / (1 + exp(beta_0 + beta_1 * x)))
}

ggplot(mapping = aes(x = c(-10, 10))) +
  stat_function(fun = logistic, args = list(beta_0 = 0, beta_1 = 1), color = "blue") +
  stat_function(fun = logistic, args = list(beta_0 = 0, beta_1 = -1), color = "red") +
  stat_function(fun = logistic, args = list(beta_0 = 1, beta_1 = 0), color = "lightgreen") +
  stat_function(fun = logistic, args = list(beta_0 = -5, beta_1 = 5), color = "purple") +
  stat_function(fun = logistic, args = list(beta_0 = -5, beta_1 = -2), color = "black") +
  xlab("x")
```

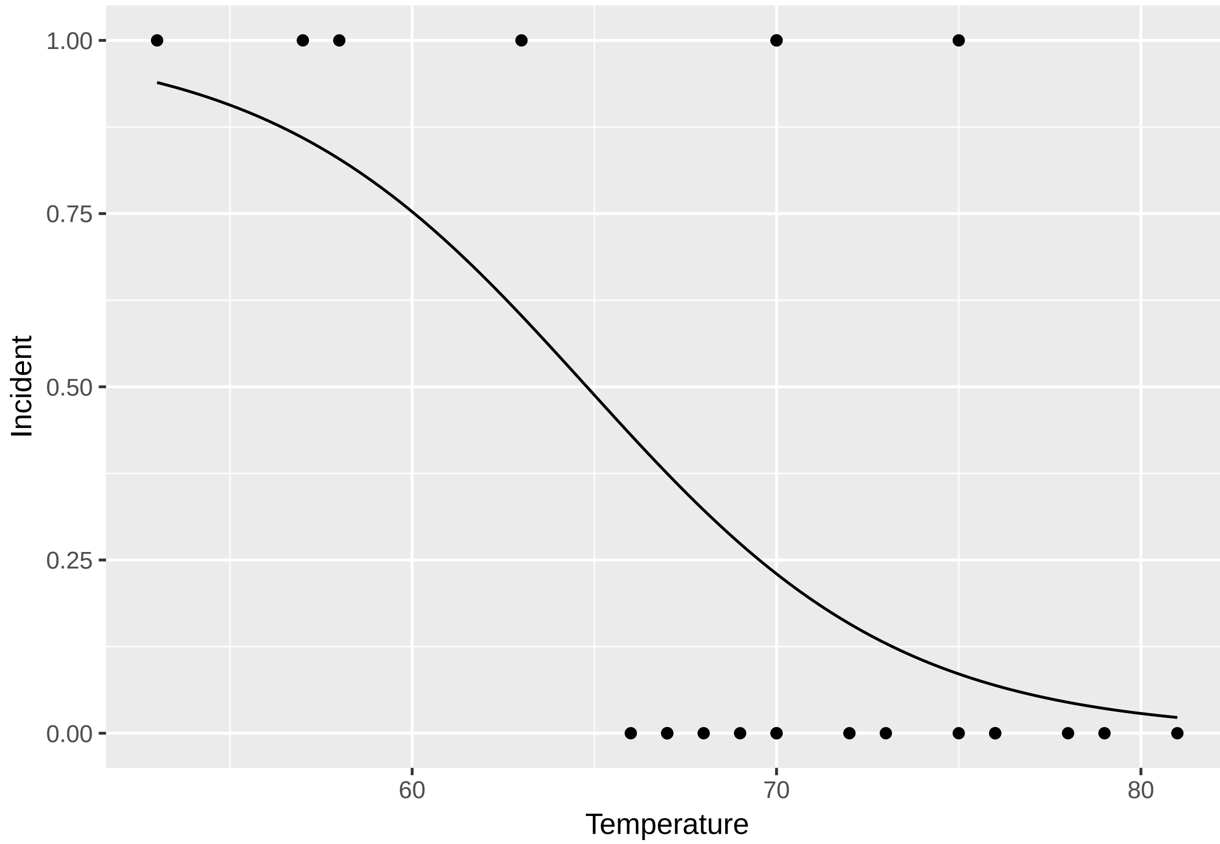


Observations:

- For all possible values of x_i, $P(Y_i = 1 | X_i = x_i) \in (0, 1)$
- β_1 controls direction of curve:
  - if $\beta_1 > 0$, then $p(x | \beta_0, \beta_1)$ is increasing in x
  - if $\beta_1 < 0$, then $p(x | \beta_0, \beta_1)$ is decreasing in x
  - if $\beta_1 = 0$, then $p(x | \beta_0, \beta_1)$ does not depend on the value of x.
- β_1 also controls "slope" of curve:
  - if $|\beta_1|$ is large, then $p(x | \beta_0, \beta_1)$ changes between 0 and 1 quickly
  - if $|\beta_1|$ is small, then $p(x | \beta_0, \beta_1)$ changes between 0 and 1 slowly
  - The maximum slope is $\beta_1/4$, and occurs at the value of x where $p(x | \beta_0, \beta_1) = 0.5$
- β_0 shifts the curve left and right

Applied to O-Rings Data:

Maximum likelihood estimates are $\hat{\beta}_0 = 15.043$, $\hat{\beta}_1 = -0.232$.



On the day of the Challenger explosion, the temperature was 33 degrees F.

The model's predicted probability of O-ring damage is

$p(33|\hat{\beta}_0, \hat{\beta}_1) = P(Y_i = 1|X_i = 33) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}} = \frac{e^{15.043 - 0.232 * 33}}{1+e^{15.043 - 0.232 * 33}} \approx 0.999$

(We may not trust an estimate that extrapolates 20 degrees below the observed data…)

Questions:

1. How could we obtain point estimates of the model parameters?
2. How could we obtain interval estimates of the model parameters?