Challenger case study

STA303/1002 Winter 2021

Your name here

This template runs through the same analysis as the Challenger shuttle disaster case study.

Shuttle data

On January 28, 1986, the Space Shuttle Challenger broke apart 73 seconds into its flight, killing all seven crew members. The spacecraft **disintegrated** over the Atlantic Ocean. The disintegration of the vehicle began after a joint in its right rocket booster failed at liftoff. The failure was caused by the **failure of O-ring seals** used in the joint that were not designed to handle the unusually cold conditions that existed at this launch.

We will look at a data set about the number of rubber O-rings showing thermal distress for 23 flights of the space shuttle, with the ambient temperature and pressure at which tests on the putty next to the rings were performed.

```
library(tidyverse)
```

```
# You may need to install this package to get the data
# install.packages("SMPracticals")
data('shuttle', package='SMPracticals')
rownames(shuttle) <- as.character(rownames(shuttle))
shuttle[1:4,]
```

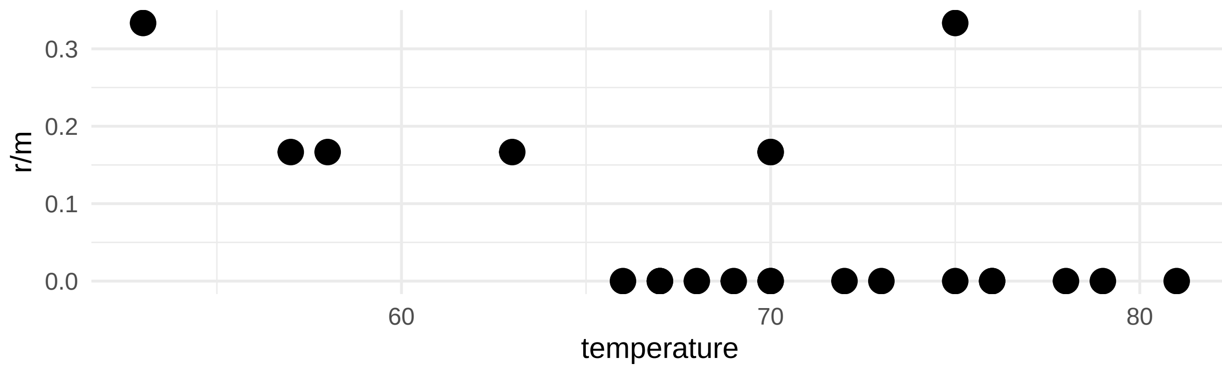
```
##   m r temperature pressure
## 1 6 0          66       50
## 2 6 1          70       50
## 3 6 0          69       50
## 4 6 0          68       50
```

- `m`: number of rings
- `r`: number of damaged rings

Thus we have a situation where we are interested in the number of successes out of a fixed number of trials. Hopefully your memories of the Binomial distibution are being triggered by that language.

```
# Base R plot
# plot(shuttle$temperature, shuttle$r/shuttle$m)

# ggplot
shuttle %>%
  ggplot(aes(x = temperature, y = r/m)) +
  geom_point(size = 4) +
  theme_minimal()
```



Are shuttle rings more likely to get damaged in cold weather?

We can think of **m** as the number of trials, and **r** as the number of "successes". (It feels weird to call damage a success, but it is our outcome of interest, so we treat it as such.)

$$
\begin{aligned}
Y_i \sim & \text{Binomial}(N_i, \mu_i) \\
\log\left(\frac{\mu_i}{1-\mu_i}\right) = & X_i\beta
\end{aligned}
$$

- m: number of rings, N_i
- r: number of damaged rings Y_i
- pressure, temperature: covariates X_i
- μ_i: probability of a ring becoming damaged given X_i
- $\beta_{\text{temperature}}$: parameter of interest

Inference: parameter estimation

$$
\begin{aligned}
Y_i \sim & G(\mu_i, \theta) \\
h(\mu_i) = & X_i\beta \\
\pi(Y_1 ... Y_N; \beta, \theta) = & \prod_{i=1}^{N} f_G(Y_i; \mu_i, \theta) \\
\log L(\beta, \theta; y_1 ... y_N) = & \sum_{i=1}^{N} \log f_G(y_i; \mu_i, \theta)
\end{aligned}
$$

- The Y_i are *independently distributed*
- **Joint density** π of random variables $(Y_1 ... Y_N)$ is the product of the marginal densities f_G.
- **Likelihood function** L given observed data $y_1 ... y_N$ is a function of the parameters.
- **Maximum Likelihood Estimation**:

$$\hat{\beta}, \hat{\theta} = \text{argmax}_{\beta,\theta} L(\beta, \theta; y_1 \dots y_N)$$

- The best parameters are those which are most likely to produce the observed data

Shuttle example in R

- `glm` works like `lm` with a `family` argument.
- Binomial models can take two types of inputs:
  - If, as in this case, we have groups of trials, we need our response to be a matrix with two columns: `y` and `N-y`.
  - If our `y` is a single 0/1 (or otherwise binary categorical variable) then we can set it up as usual, just a single column.

```
shuttle$notDamaged <- shuttle$m - shuttle$r
shuttle$y <- as.matrix(shuttle[,c('r','notDamaged')])
shuttleFit <- glm(y ~ temperature + pressure,
  family=binomial(link='logit'), data=shuttle)
shuttleFit$coef
```

```
##  (Intercept)  temperature     pressure
##  2.520194641 -0.098296750  0.008484021
```

Summary of fit

```
summary(shuttleFit)
```

```
##
## Call:
## glm(formula = y ~ temperature + pressure, family = binomial(link = "logit"),
##     data = shuttle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0361  -0.6434  -0.5308  -0.1625   2.3418
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.520195   3.486784   0.723   0.4698
## temperature -0.098297   0.044890  -2.190   0.0285 *
## pressure     0.008484   0.007677   1.105   0.2691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.546  on 20  degrees of freedom
## AIC: 36.106
##
## Number of Fisher Scoring iterations: 5
```

```
confint(shuttleFit)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %      97.5 %
## (Intercept) -4.322926283  9.77264497
## temperature -0.194071699 -0.01356289
## pressure    -0.004346403  0.02885221
```

There is no evidence that pressure is significantly associated with failure of O-rings... but how do we interpret these values?

Interpreting logistic models

$$
\begin{aligned}
Y_i \sim & \text{Binomial}(N_i, \mu_i) \\
\log\left(\frac{\mu_i}{1-\mu_i}\right) = & \sum_{p=1}^{P} X_{ip}\beta_p \\
\left(\frac{\mu_i}{1-\mu_i}\right) = & \prod_{p=1}^{P} \exp(\beta_p)^{X_{ip}}
\end{aligned}
$$

- μ_i is a probability
- $\log[\mu_i/(1-\mu_i)]$ is a log-odds
- $\mu_i/(1-\mu_i)$ is an odds
- If $\mu_i \approx 0$, then $\mu_i \approx \mu_i/(1-\mu_i)$

$$
\begin{aligned}
\beta_q = & \log\left(\frac{\mu_2}{1-\mu_2}\right) - \log\left(\frac{\mu_1}{1-\mu_1}\right) \\
\exp(\beta_q) = & \left(\frac{\mu_2}{1-\mu_2}\right) \bigg/ \left(\frac{\mu_1}{1-\mu_1}\right)
\end{aligned}
$$

- β_q is the log-odds ratio
- $\exp(\beta_q)$ is the odds ratio
- exp(intercept) is the baseline odds, when $X_1 \dots X_n = 0$.

Centring parameters

```r
quantile(shuttle$temperature)
```

```
##   0%  25%  50%  75% 100% 
##   53   67   70   75   81
```

```r
quantile(shuttle$pressure)
```

```
##   0%  25%  50%  75% 100% 
##   50   75  200  200  200
```

- Currently the intercept is log-odds when temperature = 0 and pressure = 0

- centre the covariates so the intercept refers to:
    - temperature = 70 (degrees Farenheit)
    - pressure = 200 (pounds per square inch)

```
shuttle$temperatureC <- shuttle$temperature  - 70
shuttle$pressureC <-  shuttle$pressure - 200
shuttleFit2 <-  glm(y ~ temperatureC + pressureC, family='binomial', data=shuttle)
```

Shuttle odds parameters

```
par_table = cbind(est = summary(
  shuttleFit2)$coef[,1],
  confint(shuttleFit2))
```

```
## Waiting for profiling to be done...
```

```
rownames(par_table)[1]= "Baseline"
```

```
round(exp(par_table),3)
```

```
##                est 2.5 % 97.5 %
## Baseline     0.070 0.023  0.155
## temperatureC 0.906 0.824  0.987
## pressureC    1.009 0.996  1.029
```

Table 1: MLEs of baseline odds and odds ratios, with 95% confidence intervals.

Interpreting shuttle parameters

- The odds of a ring being damaged when temperature = 70 and pressure = 200 is 0.0697, which corresponds to a probability of

```
round(exp(par_table[1,'est']) / (1+exp(par_table[1,'est'])), 3)
```

```
## [1] 0.065
```

- Each degree increase in temperature (in Farenheit) decreases the odds of damage by (in percent)

```
round(100*(1-exp(par_table[2,'est']) ), 3)
```

```
## [1] 9.362
```