Maximum likelihood and binomial regression

Generalised Linear Models (GLMs)?

Linear models suppose that we have observations

$$Y \sim N(X\beta, \sigma^2 I).$$

Can we use linear models for genetic case-control study?

Table: Data: Type 2 Diabetes example

Sample id	Y_i	X_i
1	type 2 diabetes	CC
2	no type 2 diabetes	CC
3	no type 2 diabetes	CT
4	no type 2 diabetes	TT
...	...	...
2000	type 2 diabetes	CC

$$Y_i \quad : \quad \begin{array}{ccc} \text{no type 2 diabetes} & \rightarrow & 0 \\ \text{type 2 diabetes} & \rightarrow & 1 \end{array}$$

$$X_i \quad : \quad \begin{array}{ccc} \text{CC} & \rightarrow & 0 \\ \text{CT} & \rightarrow & 1 \\ \text{TT} & \rightarrow & 2 \end{array}$$

GLMs allow for non-normal Y!

Challenger example

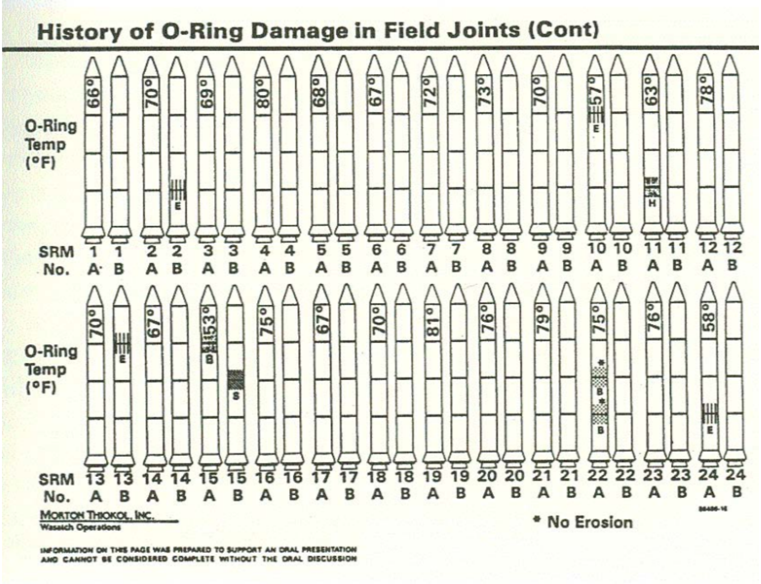
Challenger disaster

On the 28th of January 1986 the Space Shuttle Challenger broke apart after an O-ring seal failed at liftoff, leading to the deaths of its seven crew members.

Despite concerns about the O-rings failing due to the cold—the forecast temperature was $29 \pm 3\,^oF$—no one was able to provide an analysis that convinced NASA (who were under pressure to launch following several previous delays) not to go ahead.

The way the data was presented didn't help matters.

Challenger disaster: data



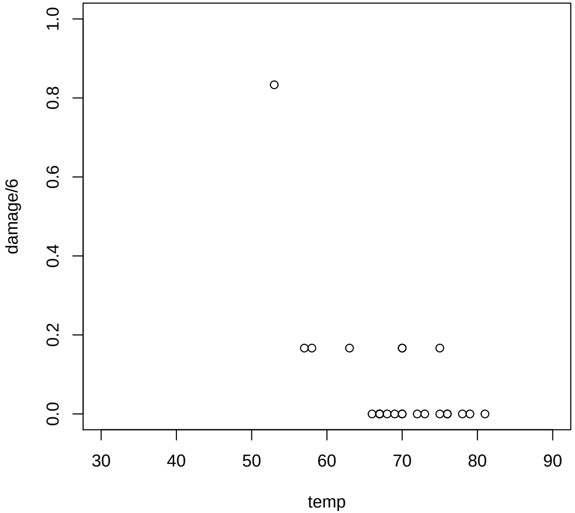
Challenger disaster: data

Summarise data using number of damaged O-rings (out of 6) per launch.

Data is in dataframe `orings`.
Response `damage` is number of damaged O-rings (out of 6).
Predictor `temp` is temperature (^oF)

```
> library(faraway)
> data(orings)
> str(orings)
data.frame: 23 obs. of  2 variables:
 $ temp  : num  53 57 58 63 66 67 67 67 68 69 ...
 $ damage: num  5 1 1 1 0 0 0 0 0 0 ...
> plot(damage/6 ~ temp, data = orings,
+      xlim = c(30, 90), ylim = c(0, 1))
```

Challenger disaster: visualize data



Challenger disaster: model

Make the assumption that Y_i, the number of damaged O-rings on the i-th launch, has distribution

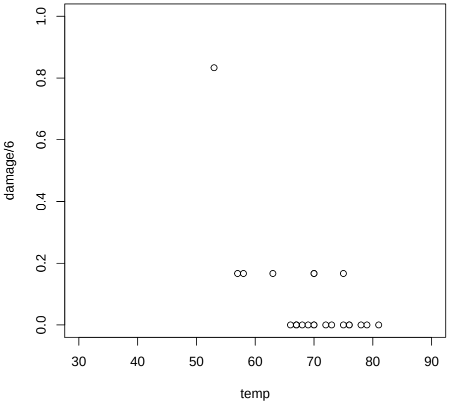
$$Y_i \sim \text{bin}(6, p_i)$$

where p_i depends on the temperature t_i. We also assume that the Y_i are independent.

How to relate p_i with t_i.

- Question: just use a linear relationship? $p_i = \beta_0 + \beta_1 t_i$?

No. We need parameter bounds so $0 \leq p_i \leq 1$. We need alternate function to relate p_i with t_i.



For a single observation, best estimate of p_i is just $y_i/6$. From plot of $y_i/6$ against t_i it is reasonable to assume that $p_i = p(t_i)$ where p is a smooth function of t, decreasing from 1 down to 0 as the temperature increases.

We choose logistic function: suppose that for some β_0 and β_1

$$p(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 t)}} = \frac{e^{\beta_0 + \beta_1 t}}{1 + e^{\beta_0 + \beta_1 t}}$$

- Restricts $0 \leq p(t) \leq 1$
- Monotonic increasing/decreasing function
- Can be linearized through the *logit* transformation

$$\log \frac{p(t)}{1 - p(t)} = \beta_0 + \beta_1 t$$

We choose logistic function: suppose that for some β_0 and β_1

$$p(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 t)}} = \frac{e^{\beta_0 + \beta_1 t}}{1 + e^{\beta_0 + \beta_1 t}}$$

- $p(t) = 1/2$ for $t = -\beta_0/\beta_1$, so $-\beta_0/\beta_1$ controls the location of the curve.
- $p'(-\beta_0/\beta_1) = \beta_1/4$, so β_1 controls the steepness of the curve.

Example: see R script and result in "logistic function with different values for beta0 and beta1" of Challenger.pdf

Challenger disaster: model fitting

Make the assumption that Y_i, the number of damaged O-rings on the i-th launch, has distribution

$$Y_i \sim \mathrm{bin}(6, p_i)$$

where

$$p_i = \frac{e^{\beta_0 + \beta_1 t_i}}{1 + e^{\beta_0 + \beta_1 t_i}}.$$

We also assume that the Y_i are independent.

Maximum likelihood estimators (MLE) $\hat{\beta}_0$, $\hat{\beta}_1$: values for β_0, β_1 which maximize the log-likelihood.

Challenger disaster: model fitting

The log-likelihood is

$$
\begin{aligned}
l(\beta_0, \beta_1) &= \log \mathcal{L}(\beta_0, \beta_1) \\
&= \log \mathbb{P}(\mathbf{Y} = \mathbf{y} \mid \beta_0, \beta_1) = \log \prod_i \mathbb{P}(Y_i = y_i \mid \beta_0, \beta_1) \\
&= \sum_i \log \left(\binom{6}{y_i} p_i^{y_i} (1 - p_i)^{6 - y_i} \right) \\
&= c + \sum_i \left(y_i \log p_i + (6 - y_i) \log(1 - p_i) \right) \\
&= c + \sum_i \left(y_i \log \frac{p_i}{1 - p_i} + 6 \log(1 - p_i) \right)
\end{aligned}
$$

Put $\eta_i = \beta_0 + \beta_1 t_i$, then $\log p_i/(1 - p_i) = \eta_i$ and $\log(1 - p_i) = -\log(1 + e^{\eta_i})$.

Challenger disaster: model fitting

There is no closed form solution for MLE $\hat{\beta}_0$, $\hat{\beta}_1$ in this model. Numerical search procedures are required to find MLE.

- the `glm` function uses the iterative weighted least squares (IWLS) algorithm - I will cover this later.
- For now, let's use the `optim` function.

Example:

- See R script and result in "maximum likelihood fitting" of Challenger.pdf
- $\hat{\beta}_0 = 11.667$ and $\hat{\beta}_1 =$ -0.216

Challenger disaster: questions

- Forecast probability of an O-ring being damaged when the launch temperature is $29\,^oF$.
- How good is our forecast? Can we provide a confidence interval?
- Is temperature useful to predict the O-ring failing?

Binomial regression

Binomial regression model

We suppose that we observe $Y_i \sim \text{bin}(m_i, p_i)$, $i = 1, \ldots, n$, independent.

The m_i are known and we suppose that for some **link function** g,

$$g(p_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$$

where $\mathbf{x}_i$ are known predictors and $\boldsymbol{\beta}$ are unknown parameters.

Binomial regression model: link function

Usual choices for g:

logit

$$\eta = \log \frac{p}{1-p}, \quad p = \frac{1}{1+\exp(-\eta)} = \frac{\exp(\eta)}{1+\exp(\eta)}$$

complementary log-log

$$\eta = \log(-\log(1-p)), \quad p = 1 - \exp(-e^{\eta})$$

probit

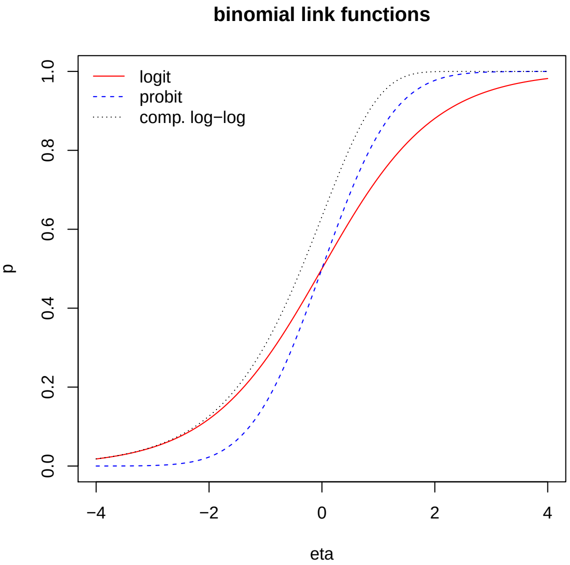
$$\eta = \Phi^{-1}(p), \quad p = \Phi(\eta),$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution.

Binomial regression model: link function

```
> curve(1/(1+exp(-x)), -4, 4, ylim=c(0,1),
+       xlab="eta", ylab="p", col="red",
+       main="binomial link functions")
> curve(pnorm(x), -4, 4, add=TRUE, col="blue", lty=2)
> curve(1-exp(-exp(x)), -4, 4, add=TRUE, col="black", lty=3)
> legend("topleft", c("logit", "probit", "comp. log-log"),
+        col=c("red", "blue", "black"), lty=c(1,2,3), bty="n")
```

Binomial regression model: link function



Binomial regression model: likelihood

Given observations y_i of $Y_i \sim \text{bin}(m_i, p_i = g^{-1}(\eta_i))$, where $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, the log-likelihood is

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log \mathbb{P}(Y_i = y_i) \\
&= \sum_{i=1}^{n} \log \left(\binom{m_i}{y_i} p_i^{y_i} (1 - p_i)^{m_i - y_i} \right) \\
&= c + \sum_{i=1}^{n} y_i \log(g^{-1}(\eta_i)) + (m_i - y_i) \log(1 - g^{-1}(\eta_i))
\end{aligned}
$$

We maximise this numerically.

MLEs have many desirable properties...

Reminder: questions from Challenger disaster

- Forecast probability of an O-ring being damaged when the launch temperature is 29 ^oF.
- How good is our forecast? Can we provide a confidence interval?
- Is temperature useful to predict the O-ring failing?

Forecast probability of an O-ring being damaged when the launch temperature is 29 ^oF.

$\hat{p} = g^{-1}(\hat{\eta})$ when $t = 29$.

$\hat{p} = \frac{\exp(\hat{\eta})}{1+\exp(\hat{\eta})}$, where $\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 29$

$\hat{p} = 0.995$: see R script and result in "prediction for temp of 29" of Challenger.pdf

Reminder: questions from Challenger disaster

- Forecast probability of an O-ring being damaged when the launch temperature is $29\,^{o}F$.
- How good is our forecast? Can we provide a confidence interval?
- Is temperature useful to predict the O-ring failing?

We need to know properties of MLE!!

Asymptotic properties MLE

Reminder: Maximum likelihood estimation (MLE)

Suppose that Y_i, $i = 1, \ldots, n$, are indepedendent, with densities/mass-functions $f_i(\cdot; \boldsymbol{\theta})$.

Given observations y_i of the Y_i, the log-likelihood is

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{y}) = \sum_i \log f_i(y_i; \boldsymbol{\theta}).$$

MLE $\hat{\boldsymbol{\theta}}$ is that value of $\boldsymbol{\theta}$ which maximises $l(\boldsymbol{\theta})$.

Note that allowing f_i to depend on i means that we can include the case where the distribution of Y_i depends on some covariate $\mathbf{x}_i$. That is, we can have $f_i(\cdot; \boldsymbol{\theta}) = f(\cdot; \mathbf{x}_i, \boldsymbol{\theta})$ for some common f.

Asymptotic properties of MLE

Under certain regularity conditions (to be introduced later), the MLE is

- asymptotically consistent,
- asymptotically normal,
- asymptotically efficient.

MLE: asymptotic consistency

Let $\boldsymbol{\theta}^*$ denote a true value for $\boldsymbol{\theta}$. As $n \to \infty$, $\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}^*$.

That is for any $\epsilon > 0$

$$\mathbb{P}(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*| > \epsilon) \to 0 \text{ as } n \to \infty.$$

MLE: asymptotic normality

The observed information is the matrix $\mathcal{J}(\boldsymbol{\theta}) = (\mathcal{J}_{ij}(\boldsymbol{\theta}))$ where $\mathcal{J}_{ij}(\boldsymbol{\theta}) = -\partial^2 l(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j$. In matrix notation

$$\mathcal{J}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}.$$

Clearly $\mathcal{J}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta};\mathbf{y})$ depends on $\mathbf{y}$ through $l(\boldsymbol{\theta}) = l(\boldsymbol{\theta};\mathbf{y})$.

For binomial regression with one predictor (e.g., $\boldsymbol{\theta} = \boldsymbol{\beta} = (\beta_0, \beta_1)$):

$$\mathcal{J}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\beta}) = \begin{pmatrix} -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_0^2} & -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_0} \\ -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_0\partial\beta_1} & -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_1^2} \end{pmatrix}$$

MLE: asymptotic normality

The Fisher information is

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\mathcal{J}(\boldsymbol{\theta}; \mathbf{Y}).$$

In practice $\mathcal{J}(\hat{\boldsymbol{\theta}})$ is often used as an approximation to $\mathcal{I}(\boldsymbol{\theta}^*)$.

Exercise: binomial regression with a logit link

We suppose that we observe $Y_i \sim \text{bin}(m_i, p_i)$, $i = 1, \ldots, n$, independent, where $p_i = \frac{1}{1+\exp(-\eta_i)}$ and $\eta_i = \beta_0 + \beta_1 x_i$. The log-likelihood is

$$\begin{aligned} l(\beta_0, \beta_1) &= c + \sum_i \left[y_i \log \frac{p_i}{1-p_i} + m_i \log(1 - p_i) \right] \\ &= c + \sum_i \left[y_i(\beta_0 + \beta_1 x_i) - m_i \log(1 + e^{\beta_0 + \beta_1 x_i}) \right]. \end{aligned}$$

Exercise: binomial regression with a logit link

Then,

$$\mathcal{J}(\beta) = \begin{pmatrix} \sum_i m_i p_i (1-p_i) & \sum_i m_i x_i p_i (1-p_i) \\ \sum_i m_i x_i p_i (1-p_i) & \sum_i m_i x_i^2 p_i (1-p_i) \end{pmatrix}.$$

So, since there are no y_i terms left,

$$\mathcal{I}(\beta) = \begin{pmatrix} \sum_i m_i p_i (1-p_i) & \sum_i m_i x_i p_i (1-p_i) \\ \sum_i m_i x_i p_i (1-p_i) & \sum_i m_i x_i^2 p_i (1-p_i) \end{pmatrix}.$$

MLE: asymptotic normality

As $n \to \infty$,

$$\mathcal{I}(\boldsymbol{\theta}^*)^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\mathrm{d}} N(\mathbf{0}, I).$$

That is,

$$\hat{\boldsymbol{\theta}} \approx^d N(\boldsymbol{\theta}^*, \mathcal{I}(\boldsymbol{\theta}^*)^{-1}).$$

MLE: asymptotic efficiency

- Asymptotic consistency: $\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}^*$
- Asymptotic normality: $\hat{\boldsymbol{\theta}} \approx^d N(\boldsymbol{\theta}^*, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$

MLE: asymptotically unbiased estimator with smallest variance $\mathcal{I}(\boldsymbol{\theta}^*)^{-1}$.

MLE: Wald CIs for $\mathbf{t}^T\boldsymbol{\theta}$ and θ_i

$$\hat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}^*, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$$

leads to

$$\mathbf{t}^T\hat{\boldsymbol{\theta}} \approx N(\mathbf{t}^T\boldsymbol{\theta}^*, \mathbf{t}^T\mathcal{I}(\boldsymbol{\theta}^*)^{-1}\mathbf{t}).$$

In particular, taking $\mathbf{t} = \mathbf{e}_i$,

$$\hat{\theta}_i \approx N(\theta_i^*, (\mathcal{I}(\boldsymbol{\theta}^*)^{-1})_{i,i}).$$

An approximate $100(1-\alpha)\%$ confidence interval for $\mathbf{t}^T\boldsymbol{\theta}$:

$$\mathbf{t}^T\hat{\boldsymbol{\theta}} \pm z_\alpha\sqrt{\mathbf{t}^T\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{t}},$$

where $\Phi(z_\alpha) = 1 - \alpha/2$. If $\mathcal{I}$ is unavailable then we can approximate it using the observed information $\mathcal{J}$.

An approximate $100(1-\alpha)\%$ CI for θ_i:

$$\hat{\theta}_i \pm z_\alpha\sqrt{(\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1})_{i,i}}.$$

Reminder: questions from Challenger disaster

- Forecast probability of an O-ring being damaged when the launch temperature is 29 ^oF.
- How good is our forecast? Can we provide a confidence interval?
- Is temperature useful to predict the O-ring failing?

How good is our forecast? Can we provide a confidence interval?

CI for $p = g^{-1}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$, where $\eta = \beta_0 + \beta_1 29$.

- Step 1: Compute a CI for η, (η_l, η_r).
- Step 2: CI for $p = g^{-1}(\eta)$ is $(g^{-1}(\eta_l), g^{-1}(\eta_r))$.

CI for p: (0.864307, 0.9998686). See R script and result in "Confidence Interval for p" of Challenger.pdf

MLE: log likelihood ratio CI

We have

$$2l(\hat{\boldsymbol{\theta}}) - 2l(\boldsymbol{\theta}^*) \approx \chi^2_k$$

where k is the dimension of $\boldsymbol{\theta}^*$.

This result can also, in principle, be used to construct a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$:

$$\{\boldsymbol{\theta} \,:\, 2l(\hat{\boldsymbol{\theta}}) - 2l(\boldsymbol{\theta}) \leq \chi^2_k(1-\alpha)\}$$

where $\chi^2_k(1-\alpha)$ is the $100(1-\alpha)\%$ point for a χ^2_k distribution.

One lab problem will be to plot CI using the log likelihood ratio.

This approximation is generally better than the normal approximation for $\hat{\boldsymbol{\theta}}$. That is, it holds for smaller sample sizes.

MLE: regularity conditions

For maximum likelihood theory to hold we require

- l smooth enough with respect to $\boldsymbol{\theta}$ (third derivatives exist and continuous)
- Third order derivatives of l have bounded expectations
- Support of Y_i does not depend on $\boldsymbol{\theta}$
- The domain Θ of $\boldsymbol{\theta}$ is finite dimensional and doesn't depend on Y_i
- $\boldsymbol{\theta}^*$ is not on the boundary of Θ.

References

- McCullagh & Nelder (1989), Appendix A.
- F.W. Scholz, Maximum likelihood estimation. *Encyclopedia of Statistical Sciences* Vol. 7, p.4629ff. Wiley, 2006.

Reminder: questions from Challenger disaster

- Forecast probability of an O-ring being damaged when the launch temperature is 29 ^oF.
- How good is our forecast? Can we provide a confidence interval?
- Is temperature useful to predict the O-ring failing?

Is temperature useful to predict the O-ring failing?

Y_i, the number of damaged O-rings on the i-th launch, has distribution

$$Y_i \sim \text{bin}(6, p_i)$$

where

$$\log p_i/(1-p_i) = \eta_i = \beta_0 + \beta_1 t_i.$$

Test for association between the number of damaged O-rings and temperature:

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$$

- Wald Test
- Likelihood Ratio Test

Wald Test

Reminder: asymptotic normality of MLE

$$\hat{\theta}_i \sim asy.\ N(\theta_i^*, (\mathcal{I}(\boldsymbol{\theta}^*)^{-1})_{i,i}).$$

Test for association between the number of damaged O-rings and temperature

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$$

Wald test statistic:

$$z^* = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim asy.\ N(0,1) \text{ under } H_0.$$

Challenger disaster

- See R script and result in “Wald Test” of Challenger.pdf
- $|z^*| = 4.07 > 1.96$ (critical value $N(0,1)$ at $\alpha = 0.05$) $\Rightarrow$ reject H_0.
- p-value $= 0.0000476$.

Likelihood Ratio Test (LRT)

Test for association between the number of damaged O-rings and temperature

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$$

Full model (F):

- $\eta_i = \beta_0 + \beta_1 t_i$
- Maximum log likelihood: $\log \mathcal{L}(\hat{\boldsymbol{\theta}}^F)$

Reduced model (R):

- $\eta_i = \beta_0$
- Maximum log likelihood: $\log \mathcal{L}(\hat{\boldsymbol{\theta}}^R)$

Compare two models.

- Likelihood ratio test statistic:

$$LR^* = -2\left[\log \mathcal{L}(\hat{\boldsymbol{\theta}}^R) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}^F)\right] \sim asy.\ \chi^2_1 \text{ under } H_0.$$

- $LR^* >$ critical value from χ^2_1 at $\alpha \Rightarrow$ reject H_0.

Likelihood Ratio Test (LRT)

- Wald test and likelihood ratio test are asymptotically equivalent.

$$[z^*]^2 = \left[\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right]^2 \sim asy.\ \chi^2_1$$

- However, the chi-squared approximation to the log likelihood ratio is generally better than the normal approximation to the MLE.
- In general, likelihood ratio test is used to select between two nested models (one model can be obtained by constraining parameters of another model).

Likelihood Ratio Test (LRT)

In general, likelihood ratio test is used to select between two nested models.

Full model (F):

- Maximum log likelihood: $\log \mathcal{L}(\hat{\boldsymbol{\theta}}^F)$

Reduced model (R):

- Maximum log likelihood: $\log \mathcal{L}(\hat{\boldsymbol{\theta}}^R)$

Let k indicate the difference in the number of parameters between two models.

Compare two models.

- Under the reduced model

$$LR^* = -2\left[\log \mathcal{L}(\hat{\boldsymbol{\theta}}^R) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}^F)\right] \sim asy.\ \chi^2_k.$$

- $LR^* >$ critical value from χ^2_k at $\alpha \Rightarrow$ select the full model.

Likelihood Ratio Test: Challenger disaster

- See R script and result in “Likelihood Ratio test” and “Wald Test vs Likelihood Ratio test” of Challenger.pdf
- $LR^* = 21.98 > 3.84$ (critical value from χ^2_1 at $\alpha = 0.05$) $\Rightarrow$ reject H_0.
- p-value $= 0.0000027$.

(Scaled) Deviance

The scaled deviance is used to judge model adequacy.

For the binomial regression model the deviance is the same as the *scaled deviance*, which is defined as the log likelihood ratio for the fitted model compared to the saturated model.

Full model (F):

- The saturated model has the same number of parameters and the observations.
- Maximum log likelihood: $\log \mathcal{L}(\hat{\boldsymbol{\theta}}^F)$

Reduced model (R):

- The fitted model.
- Maximum log likelihood: $\log \mathcal{L}(\hat{\boldsymbol{\theta}}^R)$

The scaled deviance:

$$D = -2\left[\log \mathcal{L}(\hat{\boldsymbol{\theta}}^R) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}^F)\right].$$

(Scaled) Deviance

Warning: the number of parameters in the saturated model is n, which is not fixed, so the theory of maximum likelihood does not apply, and D may not converge to a chi-squared distribution.

(Scaled) Deviance for binomial regression model

For binomial regression the saturated model allows a different p_i for each observation. Clearly, for the saturated model we estimate p_i by y_i/m_i. Let $\hat{p}_i = g^{-1}(\mathbf{x}_i^T \beta)$ be our (not saturated) model estimate of p_i, then the scaled deviance is

$$
\begin{aligned}
D &= -2 \sum_{i=1}^{n} \Bigg(y_i (\log \hat{p}_i - \log \frac{y_i}{m_i}) \\
&\qquad + (m_i - y_i)(\log(1 - \hat{p}_i) - \log(1 - \frac{y_i}{m_i})) \Bigg) \\
&= -2 \sum_{i=1}^{n} \left(y_i \log \frac{\hat{y}_i}{y_i} + (m_i - y_i) \log \frac{m_i - \hat{y}_i}{m_i - y_i} \right)
\end{aligned}
$$

where $\hat{y}_i = m_i \hat{p}_i$ is the i-th fitted value.

(Scaled) Deviance for binomial regression model

It just so happens: if $m_i p_i$ and $m_i(1 - p_i)$ are large enough (≥ 5 is a common rule of thumb), then for a binomial regression model, if the model is correct then $D \approx \chi^2_{n-k}$, where k is the number of parameters (including β_0).

In this case the (scaled) deviance can be used as a test for model adequacy. If D is too large (as compared to a χ^2_{n-k}), then the model is missing something.

Use the scaled deviance for LRT

For a binomial model with small m_i we can't use the (scaled) deviance directly to test model adequacy, but we can still use it to compare models.

If model A has (scaled) deviance D^A and model B has (scaled) deviance D^B, and A is nested within B, then

$$D^A - D^B = -2\left[\log \mathcal{L}(\hat{\boldsymbol{\theta}}^A) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}^B)\right].$$

That is, the log likelihood for the saturated model cancels, and we are left with the log likelihood ratio.

AIC

The Akaike Information Criterion is used for model selection:

$$\mathrm{AIC} = 2k - 2\log\mathcal{L}(\hat{\boldsymbol{\theta}})$$

where k is the number of parameters in the model. Given a choice, we prefer that model with the smaller AIC.

If model B has s more parameters than model A (not necessarily nested within B), then

$$\begin{aligned}\mathrm{AIC}^B - \mathrm{AIC}^A &= 2s - 2\log\mathcal{L}(\hat{\boldsymbol{\theta}}^B) + 2\log\mathcal{L}(\hat{\boldsymbol{\theta}}^A)\\ &= 2s - D^A + D^B.\end{aligned}$$

Challenger disaster: use the scaled deviance for LRT

- See R script and result in “Deviance” of Challenger.pdf

Reminder: questions from Challenger disaster

- Forecast probability of an O-ring being damaged when the launch temperature is $29\,^oF$.
- How good is our forecast? Can we provide a confidence interval?
- Is temperature useful to predict the O-ring failing?