MAST90104 - Lecture 7

Weichang Yu

Room 108, Old Geology South Bldg
School of Mathematics and Statistics, University of Melbourne

Extending the linear model

Under assumptions (V), our linear model setup is

$$y_i \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$$

where the responses y_i's are pairwise independent.

What if the responses y_i's are drawn from a discrete distribution? A non-negative distribution?
Example: Response variable in *Challenger* dataset is discrete and takes values in $\{0, 1, \ldots, 6\}$.

Solution: Generalised linear models allow for non-normal $\mathbf{y}$, in particular for count data - this is what we do here for the next few weeks.

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 $29 \pm 3\,^{o}F$—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.

Data is in dataframe `orings`.

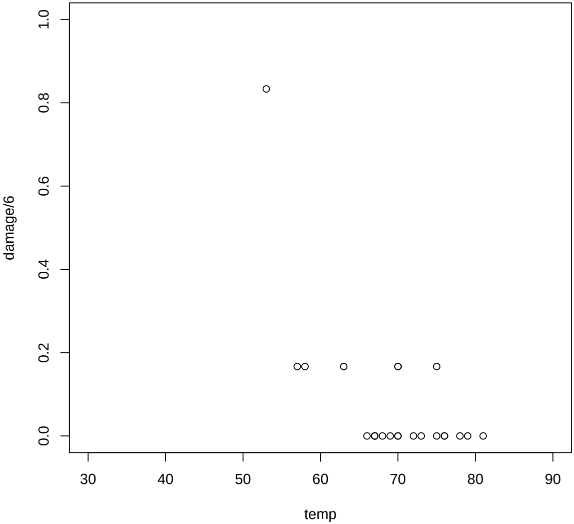
Response `damage` is number of damaged O-rings (out of 6).

Predictor `temp` is temperature (^oF)

```
orings <- data.frame(temp = c(53, 57, 58, 63, 66, 67, 67,
67, 68, 69, 70, 70, 70, 70, 72,
73, 75, 75, 76, 76, 78, 79, 81),
damage = c(5, 1, 1, 1, 0, 0, 0,
0, 0, 0, 1, 0, 1, 0, 0,
0, 0, 1, 0, 0, 0, 0, 0))
```

```
plot(damage/6 ~ temp, data = orings,
xlim = c(30, 90), ylim = c(0, 1))
```

Figure below shows the failure proportions for O-rings versus temperature at launch.



Challenger disaster: model

A natural assumption is 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.

For a single observation, the best estimate of p_i is just $y_i/6$.

From Figure 1 and engineering considerations, it seems 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 (what about temperatures greater than $85\,^oF$?).

Choice of p

We choose the function p from a family of sigmoid functions: suppose that for some α and β

$$p(t) = \frac{1}{1 + e^{-(\alpha+\beta t)}} = \frac{e^{\alpha+\beta t}}{1 + e^{\alpha+\beta t}}$$

Note that $p(t) = 1/2$ for $t = -\alpha/\beta$, so $-\alpha/\beta$ controls the location of the curve.

Also $p'(-\alpha/\beta) = \beta/4$, so β controls the steepness of the curve.

Note also that

$$\text{log odds for p } = \alpha + \beta t$$

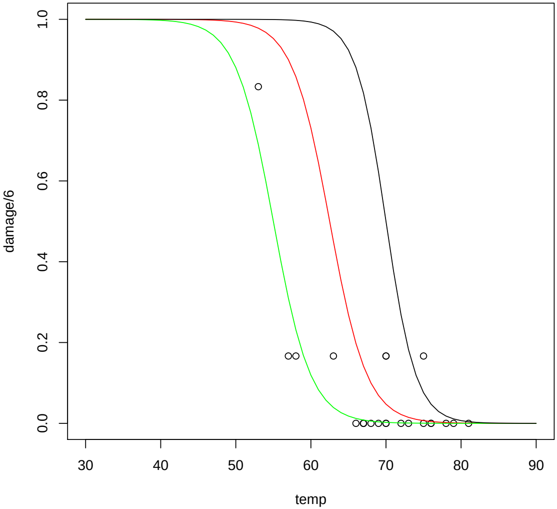
$f(p) = \log(p) - \log(1 - p)$ is called a *logit* function.

Adding some possible logistic curves to the data

```r
try <- function(a, b, col) {
t <- seq(30, 90, 1)
p <- 1/(1 + exp(-a - b*t))
lines(t, p, col = col)
}
```

```r
try(25,-0.4,"red")
try(22,-0.4,"green")
try(35,-0.5,"black")
```

Figure: O-ring failures with three logistic curves



Challenger disaster: model fitting

How to choose $\boldsymbol{\theta} = (\alpha, \beta)^T$?

General approach to curve fitting: minimise some loss function $L(\boldsymbol{\theta})$ which measures how close the model is to the data. For example $L(\boldsymbol{\theta}) = d(\hat{\mathbf{y}}, \mathbf{y})$, where $\hat{\mathbf{y}}$ are the fitted values (determined by $\boldsymbol{\theta}$) and d is some distance measure.

Our goodness of fit measure (loss function) will be minus the log-likelihood. Good theoretical basis for this and the most commonly used technique in Frequentist statistics.

Challenger disaster: log-likelihood

Recall, in general, that the log-likelihood is

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) \\
&= \log \mathbb{P}(\mathbf{Y} = \mathbf{y}) \\
&= \log \prod_i \mathbb{P}(Y_i = y_i) \\
&= \sum_i \log \mathbb{P}(Y_i = y_i) \\
&= \sum_i \log \left({}^6C_{y_i} p_i^{y_i} (1 - p_i)^{6-y_i} \right), \quad \text{where } {}^rC_s = \tfrac{r!}{s!(r-s)!} \\
&= c + \sum_i \left(y_i \log p_i + (6 - y_i) \log(1 - p_i) \right) \\
&= c + \sum_i \left(y_i \log \frac{p_i}{1 - p_i} + 6 \log(1 - p_i) \right)
\end{aligned}$$

Maximising log-likelihood numerically

Put $\eta_i = \alpha + \beta t_i$, then $\log\left(p_i/(1-p_i)\right) = \eta_i$ and $\log(1-p_i) = -\log(1+e^{\eta_i})$.

So

$$\ell(\boldsymbol{\theta}) = c + \sum_i \left(y_i \eta_i - 6\log(1+e^{\eta_i})\right) \tag{1}$$

The task is to minimise the negative of the log-likelihood or, equivalently, maximise the log-likelihood. That is to find $\boldsymbol{\theta}$ which maximises $\ell(\boldsymbol{\theta})$ numerically...

Maximising log-likelihood numerically in R

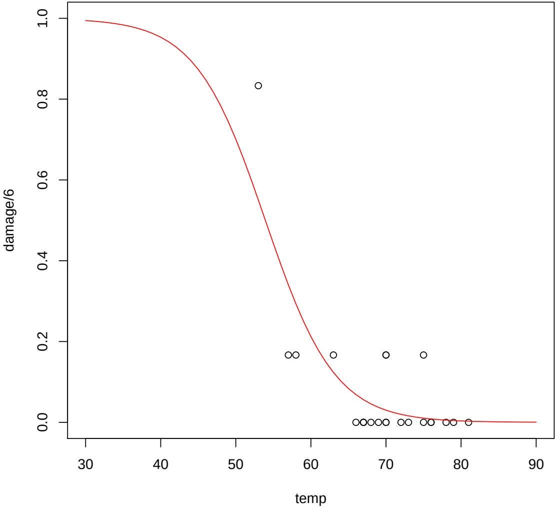
```r
# function to evaluate log-likelihood
l <- function(tha, y, t) {
eta <- tha[1] + tha[2]*t
return(sum(y*eta - 6*log(1 + exp(eta))))
}
# optim is general purpose optimiser:
# fnscale= -1 spec. max.
(betahat <- optim(c(10, -0.1), l,
y = orings$damage, t = orings$temp,
control = list(fnscale = -1,reltol=1e-16))$par)

## [1] 11.6629893 -0.2162337
```

And plot the results

```r
plot(damage/6 ~ temp, data = orings,
xlim = c(30, 90), ylim = c(0, 1))
t <- seq(30, 90, 1)
p <- 1/(1 + exp(-betahat[1] - betahat[2]*t))
lines(t, p, col = "red")
```

Figure: O-ring data with fitted logistic curve



What advice should be given before launch?

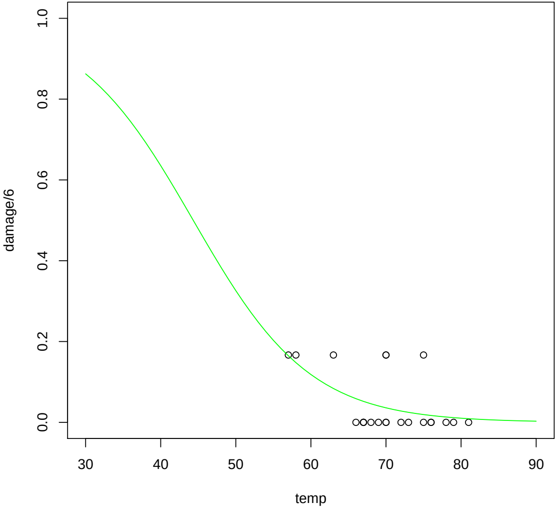
Forecast probability an O-ring is damaged when the launch temperature is 29 ^oF?

All indications are that the probability of O-ring failure at 29 ^oF is very close to 1

How good is our forecast? Can we be sure?

Leverage of the data point when 5 out of 6 failures occurred at 53 ^oF is high, so perhaps wise to repeat the analysis omitting this data point to see how much it changes the result.

Figure: O-ring data with fitted logistic curve - without 5 out of 6 at 53 ^oF



Even without the near disaster at $53\,^oF$, there was very strong evidence that launching at a forecast temperature of $29\,^oF$.

No data at lower temperatures should be omitted when safety is concerned - given the forecast - , so another relevant approach is to obtain a confidence or prediction interval.

To do this theory is needed.

Binomial regression model

We suppose that we observe $Y_i \sim \mathrm{Bin}(m_i, p_i)$, $i = 1, \dots, n$, 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.

Link function possibilities

Usual choices for g:

logit or logistic or log-odds

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

probit or normal quantiles

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

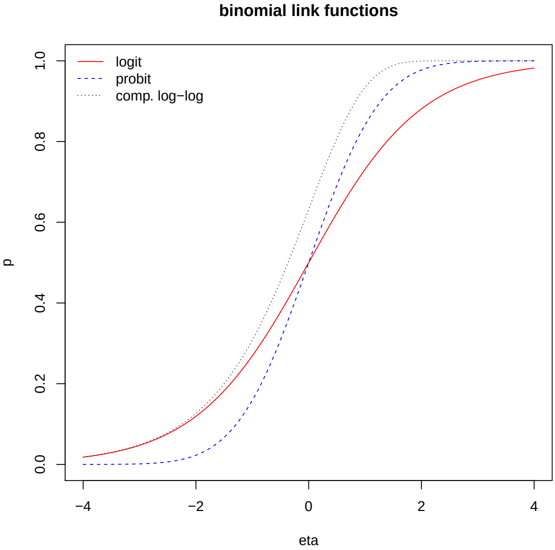
complementary log-log

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

Illustration of inverse link function possibilities

```r
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")
```

Figure: Illustration of logit, probit and comp. log log inverse link



Binomial regression model: likelihood

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

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

We maximise this numerically.

Maximum likelihood estimators have many desirable properties...

Maximum likelihood estimation

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

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

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

The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is that value of $\boldsymbol{\theta}$ which maximises $\ell(\boldsymbol{\theta})$.

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.

Under certain regularity conditions, the MLE is *consistent*, *asymptotically normal*, and *asymptotically efficient*.

MLE: consistency

As $n \to \infty$, if the true parameter value is $\boldsymbol{\theta}^*$, then $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$. 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 $\mathcal{J}(\boldsymbol{\theta}) = (\mathcal{J}_{jk}(\boldsymbol{\theta}))$ where $\mathcal{J}_{jk}(\boldsymbol{\theta}) = -\partial^2 \ell(\boldsymbol{\theta}) / \partial\theta_j \partial\theta_k$.

In matrix notation

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

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

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 $\mathcal{I}(\boldsymbol{\theta}^*)$.

When $|\mathcal{I}(\boldsymbol{\theta})|$ is large (that is, for large n), under regularity conditions,

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

That is,

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

Note: higher curvature of ℓ at $\boldsymbol{\theta}^*$ makes $\hat{\boldsymbol{\theta}}$ easier to find. From above it also means $\mathcal{I}(\boldsymbol{\theta}^*)$ is larger, and thus the variance of $\hat{\boldsymbol{\theta}}$ is smaller.

Lemma

$$\begin{aligned}
\mathbb{E}\frac{\partial \ell(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_j} &= 0 \\
-\mathbb{E}\frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_j \partial \theta_k} &= \mathbb{E}\left(\frac{\partial \ell(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_j}\frac{\partial \ell(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_k}\right)
\end{aligned}$$

Proof Reverse order of differentiation and integration - extension of MAST90105

Example: binomial regression

For binomial regression with a logit link, the log-likelihood (see equation 1 and the derivation before it) is

$$\ell(\boldsymbol{\beta}) = c + \sum_{i=1}^{n} (y_i \eta_i - m_i \log(1 + e^{\eta_i})) \tag{2}$$

where $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$ is the linear predictor of the probability p_i. Now

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

and

$$f(x) = \log(1 + e^x) \implies f'(x) = \frac{e^x}{1 + e^x}.$$

So

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left(y_i x_{ij} - m_i \frac{e^{\eta_i}}{1 + e^{\eta_i}} x_{ij} \right)$$

Example: binomial regression

Since

$$f''(x) = \frac{e^x}{(1+e^x)^2}.$$

differentiating the log-likelihood a second time gives

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k} &= -\sum_{i=1}^n m_i \frac{e^{\eta_i}}{(1+e^{\eta_i})^2} x_{ij} x_{ik} \\ &= -\sum_{i=1}^n m_i x_{ij} x_{ik} p_i (1-p_i),\end{aligned}$$

where $p_i = 1/(1+\exp(-\eta_i)) = 1/(1+\exp(-\mathbf{x}_i^T \boldsymbol{\beta}))$.

Example: binomial regression

So, since there are no y_i terms left, the expectation of each second derivative is the same as the value and the Fisher Information matrix is

$$\mathcal{I}(\boldsymbol{\beta}) = \left(\sum_{i=1}^{n} m_i x_{ij} x_{ik} p_i (1 - p_i) \right)_{j,k=0}^{q} . \tag{3}$$

MLE: efficiency

Cramér-Rao lower bound Under mild regularity conditions, if $\hat{\boldsymbol{\theta}}$ is any unbiased estimator of $\boldsymbol{\theta}^*$, then

$$\mathrm{V}\mathit{ar}\left(\hat{\boldsymbol{\theta}}\right) - \mathcal{I}(\boldsymbol{\theta}^*)^{-1}$$

is positive semidefinite.

An estimator that achieves the Cramér-Rao lower bound is said to be efficient.

The MLE is *asymptotically* efficient, as $n \to \infty$.

MLE: Wald CI

We can use large-sample theory to approximate the distribution of $\mathbf{t}^T\hat{\boldsymbol{\theta}}$ as $N(\mathbf{t}^T\boldsymbol{\theta}^*, \mathbf{t}^T\mathcal{I}(\boldsymbol{\theta}^*)^{-1}\mathbf{t})$. We then obtain confidence intervals for linear combinations of $\mathbf{t}^T\boldsymbol{\theta}$.

In particular, taking $\mathbf{t} = \boldsymbol{v}_i$, where $\boldsymbol{v}_i$ is a column vector with 1 as the i-th entry and 0 otherwise. Then, we get $\hat{\theta}_i \approx N(\theta_i^*, (\mathcal{I}(\boldsymbol{\theta})^{-1})_{i,i})$, thus an approximate $100(1-\alpha)\%$ CI for θ_i^* is

$$\hat{\theta}_i \pm z_{\alpha/2}\sqrt{(\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1})_{i,i}}$$

where $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

MLE: Wald CI

As well, this gives an approximate confidence interval for the linear predictor of observation i, namely a confidence interval for η_i is

$$\mathbf{x}_i^T\hat{\boldsymbol{\theta}} \pm z_{\alpha/2}\sqrt{\mathbf{x}_i^T(\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1})\mathbf{x}_i}$$

If $\mathcal{I}$ is unavailable then we can approximate it using the observed information $\mathcal{J}$.

Challenger disaster: Large sample variance matrix

```
# inverse logit function
ilogit <- function(x) 1/(1+exp(-x))

#calculate estimated probabilities from parameters
phat = ilogit(betahat[1]+betahat[2]*orings$temp)

# calculate large sample variance matrix from equation (3)
# substitute estimates for probabilities
I11 <- sum(6*phat*(1 - phat))
I12 <- sum(6*orings$temp*phat*(1 - phat))
I22 <- sum(6*orings$temp^2*phat*(1 - phat))
(Iinv <- solve(matrix(c(I11, I12, I12, I22), 2, 2)))

##            [,1]         [,2]
## [1,] 10.865351 -0.174240983
## [2,] -0.174241  0.002827797
```

Challenger disaster: Estimates of parameters and sd's

```
# parameter estimates
(betahat)

## [1] 11.6629893 -0.2162337

# estimates of their sd's
(sdp <- c(sqrt(Iinv[1,1]),sqrt(Iinv[2,2])))

## [1] 3.29626323 0.05317703
```

Confidence intervals for probabilities

The confidence intervals for linear combinations of the parameters can be turned into confidence intervals for the probabilities.

This is because the link function and its inverse are both increasing.

Hence the event that the confidence interval (L, U) (L, U are random variables) contains the linear combination $\mathbf{t}^T\boldsymbol{\theta}^*$ is the same as the event that $(g^{-1}(L), g^{-1}(U))$ contains $g^{-1}(\mathbf{t}^T\boldsymbol{\theta}^*)$.

So their probabilities are the same (for example, 95% for a 95% confidence interval).

And for the binomial regregssion example, $g^{-1}(\mathbf{t}^T\boldsymbol{\theta}^*)$ is the true Binomial probability when the input variables are $\mathbf{t}$.

Challenger disaster: CI for forecast when temp. $= 29$

```
q95 <- qnorm(0.975) # normal quantile for 95% ci
si2 <- matrix(c(1, 29), 1, 2) %*%
Iinv %*% matrix(c(1, 29), 2, 1) # estimated variance of linear
# predictor estimate at 29
ilogit(betahat[1] + betahat[2]*29) #estimate of probability

## [1] 0.9954687

ilogit(betahat[1] + betahat[2]*29 - q95*sqrt(si2)) # 95% ci lower

##            [,1]
## [1,] 0.8721945

ilogit(betahat[1] + betahat[2]*29 + q95*sqrt(si2)) # 95% ci upper

##            [,1]
## [1,] 0.9998586
```

R facilities for Generalised Linear Models - `glm`

There is an R command `glm` which is just like the R command `lm` but extends to Generalised Linear Models, including the Binomial Regression case.

It is necessary now to specify

1. the formula for the linear predictor in terms of variables in the data frame
2. the family of distributions - here Binomial
3. the link function - here logistic (which is the default)

Challenger disaster: Summary from glm

```
# using the glm command
logitmod <- glm(cbind(damage,6-damage) ~ temp,
family=binomial, orings)
summary(logitmod)
```

Challenger disaster: Summary from glm

```
##
## Call:
## glm(formula = cbind(damage, 6 - damage) ˜ temp, family = binomial,
##     data = orings)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9529  -0.7345  -0.4393  -0.2079   1.9565
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.66299    3.29626   3.538 0.000403 ***
## temp        -0.21623    0.05318  -4.066 4.78e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.898  on 22  degrees of freedom
## Residual deviance: 16.912  on 21  degrees of freedom
## AIC: 33.675
##
## Number of Fisher Scoring iterations: 6
```

Challenger disaster: CI `glm`

```r
# get the estimated probability - type = "response" indicates probability
predict(logitmod, newdata=data.frame(temp=29), type="response")

##         1
## 0.9954687

#  get the confidence interval
fitlp <- predict(logitmod, newdata=data.frame(temp=29), type="link",se.fit=T)
(lower <- ilogit(fitlp$fit - q95*fitlp$se.fit))

##         1
## 0.8721945

(upper <- ilogit(fitlp$fit + q95*fitlp$se.fit))

##         1
## 0.9998586
```

MLE: likelihood ratio

For large n

$$2\ell(\hat{\boldsymbol{\theta}}) - 2\ell(\boldsymbol{\theta}^*) \sim \chi^2_k$$

where

$$k = \begin{cases} q+1 & \text{if } \mathbf{X} \text{ is full rank,} \\ r(\mathbf{X}) & \text{if } \mathbf{X} \text{ is not full rank.} \end{cases}$$

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

$$\{\boldsymbol{\theta} \,:\, 2\ell(\hat{\boldsymbol{\theta}}) - 2\ell(\boldsymbol{\theta}) \leq \chi^2_{k;1-\alpha}\}$$

where $\chi^2_{k;1-\alpha}$ is the $100(1-\alpha)\%$ point for a **central** χ^2_k distribution.

MLE: regularity conditions (non-examinable)

The following conditions are enough to ensure that the asymptotic results hold in maximum likelihood theory:

- 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 $\boldsymbol{\Theta}$ of $\boldsymbol{\theta}$ is finite dimensional and doesn't depend on Y_i
- The true value $\boldsymbol{\theta}^*$ is not on the boundary of $\boldsymbol{\Theta}$.

References

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

Nested models

A *model* is a set of distribution functions. An example is the family of zero-centered Gaussian:

$$\mathcal{M} = \{\Phi_{0,\sigma} \ : \ \sigma \in \mathbb{R}^+\},$$

where $\Phi_{\mu,\sigma}$ is the CDF of a normal distribution mean μ and SD σ.

Consider two models $\mathcal{M}_A$ and $\mathcal{M}_B$. We say that $\mathcal{M}_A$ is *nested* within $\mathcal{M}_B$ if $\mathcal{M}_A \subseteq \mathcal{M}_B$.

Examples of nested models:

- $\mathcal{M}_A = \{\Phi_{0,\sigma} \ : \ \sigma \in \mathbb{R}^+\}$ and $\mathcal{M}_B = \{\Phi_{\beta_0,\sigma} \ : \ \beta_0 \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$
- $\mathcal{M}_A = \{\text{Binomial-CDF}_{\beta_1} \ : \ p = \text{logit}^{-1}(\beta_1 x), \ \ \beta_1 \in \mathbb{R}, n \in \mathbb{Z}^+\}$ and $\mathcal{M}_B = \{\text{Binomial-CDF}_{\beta_0,\beta_1} \ : \ p = \text{logit}^{-1}(\beta_0 + \beta_1 x), \ \ \beta_0, \beta_1 \in \mathbb{R}, n \in \mathbb{Z}^+\}$, where x is constant.

Note: If $\mathcal{M}_A \subset \mathcal{M}_B$, then $\mathcal{M}_B$ has more parameters than $\mathcal{M}_A$.

Nested models

Consider the data $\mathbf{y} = (y_1, \ldots, y_n)^T$. Denote their joint distribution by F.

Consider models A and B such that $\mathcal{M}_A \subset \mathcal{M}_B$. Suppose model B has s more parameters than model A. Then, we can write

$$\mathcal{M}_A = \{G_{(\boldsymbol{\theta}^A, \mathbf{0}_s)} : (\boldsymbol{\theta}^A, \mathbf{0}_s) \in \boldsymbol{\Theta}\}$$

and

$$\mathcal{M}_B = \{G_{\boldsymbol{\theta}^B} : \boldsymbol{\theta}^B \in \boldsymbol{\Theta}\}$$

We say that model A is *correct* if $F \in \mathcal{M}_A$.

Likelihood ratio test

Suppose model A is correct. Then, there exists $\boldsymbol{\theta}^{*A}$ such that $(\boldsymbol{\theta}^{*A}, \mathbf{0}_s) \in \boldsymbol{\Theta}$ and $F = G_{(\boldsymbol{\theta}^{*A}, \mathbf{0}_s)}$.

By large-sample theory, we have

$$\Lambda = -2 \log \frac{\mathcal{L}(\hat{\boldsymbol{\theta}}^A)}{\mathcal{L}(\hat{\boldsymbol{\theta}}^B)} \;=\; -2(\ell(\hat{\boldsymbol{\theta}}^A) - \ell(\hat{\boldsymbol{\theta}}^B)) \overset{\text{asymp.}}{\sim} \chi_s^2. \qquad (4)$$

Moreover, if model A is correct then the likelihood ratio will be large, so that $-2\{\ell(\hat{\boldsymbol{\theta}}^A) - \ell(\hat{\boldsymbol{\theta}}^B)\}$ is small. Hence we are unlikely to reject H_0 for a one-tailed test:

$$H_0 : \text{Model A is correct} \quad H_1 : \text{Model A is not correct}$$

Likelihood ratio test

Suppose model A is incorrect and model B is correct, i.e., $F \notin \mathcal{M}_A$ and $F \in \mathcal{M}_B$. Then, there exists $\boldsymbol{\theta}^{*B}$ such that $\boldsymbol{\theta}^{*B} \in \boldsymbol{\Theta}$, $F = G_{(\boldsymbol{\theta}^{*B})}$, and the last s entries are not all zeroes.

Then, the likelihood ratio $\mathcal{L}(\hat{\boldsymbol{\theta}}^A)/\mathcal{L}(\hat{\boldsymbol{\theta}}^B)$ will be small, so that Λ is large. Hence we reject H_0 in the test:

$$H_0 : \text{Model A is correct} \quad H_1 : \text{Model A is incorrect}$$

We reject H_0 for large values of the statistic Λ.

Inference

Likelihood ratios are used for inference. That is, they are used to choose between nested models, or equivalently decide if parameters are non-zero.

The Wald CI for a single parameter θ_i^* can also be used to decide if $\theta_i^* = 0$ (does the CI contain 0 or not?). However, the chi-squared approximation to the log likelihood ratio is generally better than the normal approximation to the MLE, so we prefer to use the likelihood ratio for model selection. Note that Wald's statistic is not equivalent to likelihood ratio statistic in Binomial (many other glms).

Compare this to linear models where the Wald test statistic and the likelihood ratio statistic are equivalent to the F statistic.

Saturated model

The saturated model refers to the model with as many parameters as sample size n.

In binomial regression model, the saturated model loglikelihood is

$$\ell(p_1, \ldots, p_n) = \sum_{i=1}^{n} \log \left\{ {}^{m_i}C_{y_i} p_i^{y_1} (1 - p_i)^{m_i - y_i} \right\}$$

The MLEs are

$$\widetilde{p}_i = y_i / m_i, \quad i = 1, \ldots, n.$$

Deviance

The *deviance* is used to judge model adequacy.

The binomial regression model the deviance is defined as -2 times the difference in the log likelihoods betwen the fitted model and the *saturated model*.

Under the fitted model, let $\widehat{p}_i = g^{-1}(\mathbf{x}_i^T \widehat{\beta})$ be our (not full) model estimate of p_i, then the deviance is

$$\begin{aligned} D &= -2 \sum_{i=1}^{n} \Bigg(y_i (\log \widehat{p}_i - \log \frac{y_i}{m_i}) \\ &\quad + (m_i - y_i)(\log(1 - \widehat{p}_i) - \log(1 - \frac{y_i}{m_i})) \Bigg) \end{aligned}$$

That is

$$D = -2 \sum_{i=1}^{n} \left(y_i \log \frac{\widehat{y}_i}{y_i} + (m_i - y_i) \log \frac{m_i - \widehat{y}_i}{m_i - y_i} \right) \tag{5}$$

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

The following statements are true for binomial regression:

- 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).
- Deviances 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.

Deviance difference as analogue of difference in SS_{res}

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

If model A has deviance D^A and model B has deviance D^B, and model A is nested within model B, then it can be shown that:

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

The difference in deviances is the generalised linear models' analogue for the difference in residual sums of squares in nested linear models.

Challenger disaster: significance of temperature

```
# deviance calculated from equation (6)
y <- orings$damage
n <- rep(6, length(y))
ylogxy <- function(x, y) ifelse(y == 0, 0, y*log(x/y))
(D <- -2*sum(ylogxy(n*phat, y)
+ ylogxy(n*(1-phat), n - y)))
```

```
## [1] 16.91228
```

```
(df <- length(y) - length(betahat))
```

```
## [1] 21
```

Challenger disaster: significance of temperature

```
# deviance calculated in glm
deviance(logitmod)

## [1] 16.91228

df.residual(logitmod)

## [1] 21

# significance of deviance of full model
pchisq(D, df,lower=FALSE)

## [1] 0.7164099
```

Challenger disaster: significance of temperature

```
# null model in which there is no temperature effect
# direct calculation of deviance difference and significance
(phatN <- sum(y)/sum(n))

## [1] 0.07971014

(DN <- -2*sum(ylogxy(n*phatN, y) + ylogxy(n*(1-phatN), n - y)))

## [1] 38.89766

(DfN <- length(y) - 1)

## [1] 22

pchisq(DN - D, DfN - df, lower=FALSE)

## [1] 2.747351e-06
```

```r
# using glm
logitnull <- glm(cbind(y, n - y) ~ 1, family=binomial)
summary(logitnull)

##
## Call:
## glm(formula = cbind(y, n - y) ~ 1, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9984  -0.9984  -0.9984   0.6947   4.4781
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.4463     0.3143  -7.783 7.06e-15 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.898  on 22  degrees of freedom
## Residual deviance: 38.898  on 22  degrees of freedom
## AIC: 53.66
##
## Number of Fisher Scoring iterations: 4
```

Challenger disaster: significance of temperature

```
# probability of failure using intercept
# ie no temperature effect
ilogit(-2.4463)

## [1] 0.07970954

# significance of temperature using normal theory
# Wald test (less powerful)
2*pnorm(abs(betahat[2]), 0, sqrt(Iinv[2,2]), lower=FALSE)

## [1] 4.776586e-05
```

AIC

The Akaike Information Criterion is used for model selection:

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

where k is the number of parameters in the model.

Given a choice, we prefer the model with the smaller AIC.

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

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