Introduction to Statistical Learning

Notes by Tim Brown and Owen Jones

Module 7: Maximum Likelihood and Binomial Regression

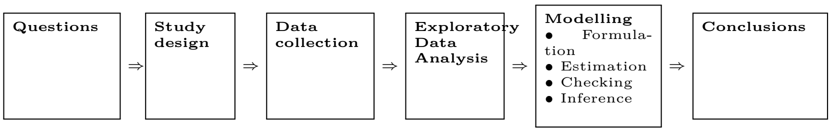
Contents

1 Introduction

1.1 Perspective on where we are

Statistics

Statistics is a collection of tools for quantitative research, the main aspects of which are:



Looking Back

We needed all the linear algebra and multivariate normal theory (including χ^2- and $F-$ distributions) to understand linear models Understanding linear algebra and statistical theory is what sets a data scientist apart from a computer coder or project manager You have the tools now to create linear model estimates or hypothesis tests that are not standard for problems that are not standard - and most problems are not standard! See quotes on the website in my blog - or just look on the net - for the importance of theory As we developed the tools, our focus has been more on practical application

Looking forward

We have enough basic theory now from last semester and this, to be able develop tools and applications more at the same time There will be a lot more practical applications from now on this semester References in this module, and others in Generalised Linear Models, are to Faraway - details on p.3 of Module 1

Extending the linear model

Linear models suppose that we have observations

$$\mathbf{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$$

There are many ways we can try to generalise them:

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. Linear model analogue?

Mixed effects models allow for more general correlation structures, such as for hierarchical data or longitudinal data. This will be pursued in MAST90084 Statistical Modelling.

Non-parametric regression models allow for a non-linear relationship between X and $\mathbb{E}\mathbf{y}$. This will also be studied in MAST90084 Statistical Modelling and includes many machine learning techniques.

2 Challenger example

2.1 Background - F Ch2 beginning

Challenger disaster

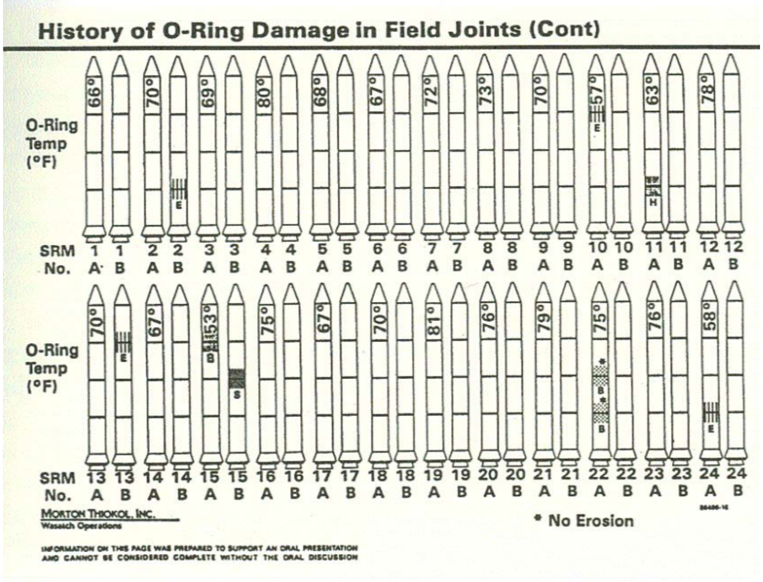
On the 28th of January 1986 the Space Shuttle Challenger broke apart after an O-ring seal failed at liftoff, leading to the deaths of its seven crew members.

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.

See p.280 of Hogg and Tanis for the story on how the engineers omitted data on the incidence of successful launches at higher temperatures.

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

Challenger disaster: data


History of O-Ring Damage in Field Joints (Cont)
O-Ring Temp (°F)
66° 70° 69° 80° 68° 67° 72° 73° 70° 57° 63° 78°
SRM No.
1 A 1 B 2 A 2 B 3 A 3 B 4 A 4 B 5 A 5 B 6 A 6 B 7 A 7 B 8 A 8 B 9 A 9 B 10 A 10 B 11 A 11 B 12 A 12 B
O-Ring Temp (°F)
70° 67° 53° 75° 67° 70° 81° 76° 79° 75° 76° 58°
SRM No.
13 A 13 B 14 A 14 B 15 A 15 B 16 A 16 B 17 A 17 B 18 A 18 B 19 A 19 B 20 A 20 B 21 A 21 B 22 A 22 B 23 A 23 B 24 A 24 B
E
E
H
E
B
S
B
B
E
MORTON THIOKOL, INC.
Wasatch Operations
* No Erosion
INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION

2.2 Data summary- F Ch2 beginning

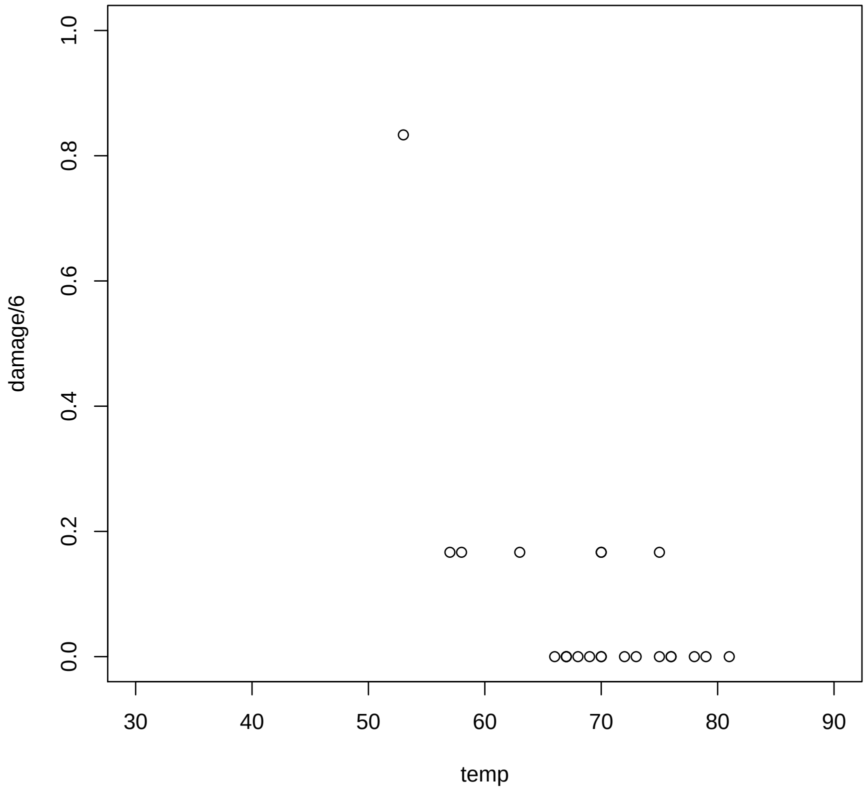
Data is in dataframe `orings`. Response `damage` is number of damaged O-rings (out of 6). Predictor `temp` is temperature (^oF)

```
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 1 shows the failure proportions for O-rings versus temperature at launch.

Figure 1: Plot of proportion of failed o-rings versus temperature (^oF)



2.3 Model? - F Ch2

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

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

- called *logistic* 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")
```

2.4 Fitting a model - F Ch2

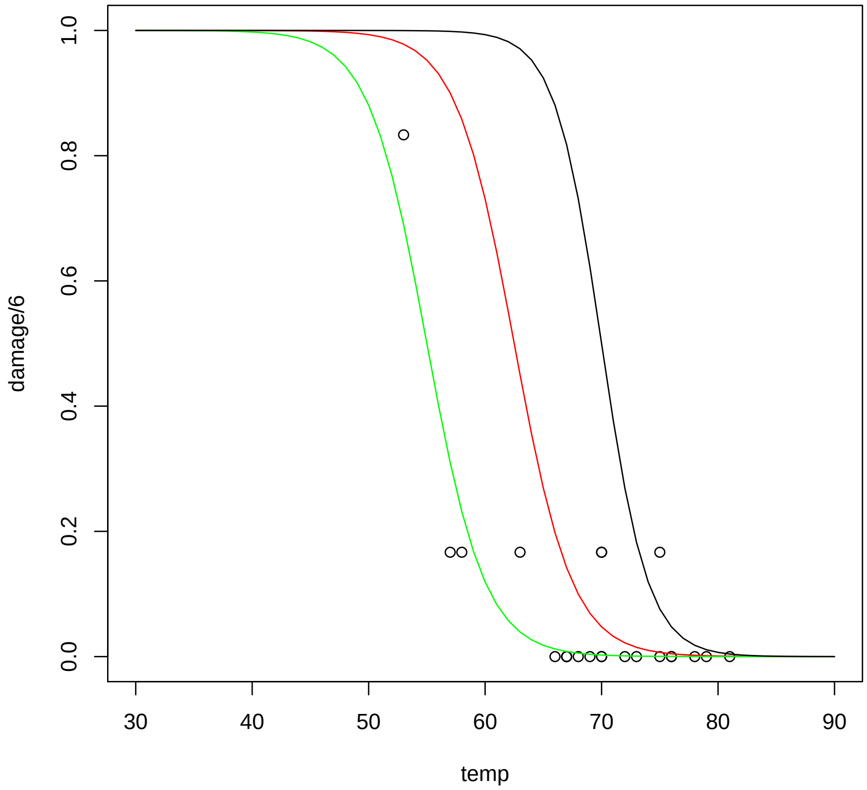
Challenger disaster: model fitting

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.

Figure 2: O-ring failures with three logistic curves



Challenger disaster: log-likelihood

Recall, in general, that the log-likelihood is

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

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

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

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 $l(\boldsymbol{\theta})$ numerically...

2.5 Using R - F Ch2

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 3: O-ring data with fitted logistic curve

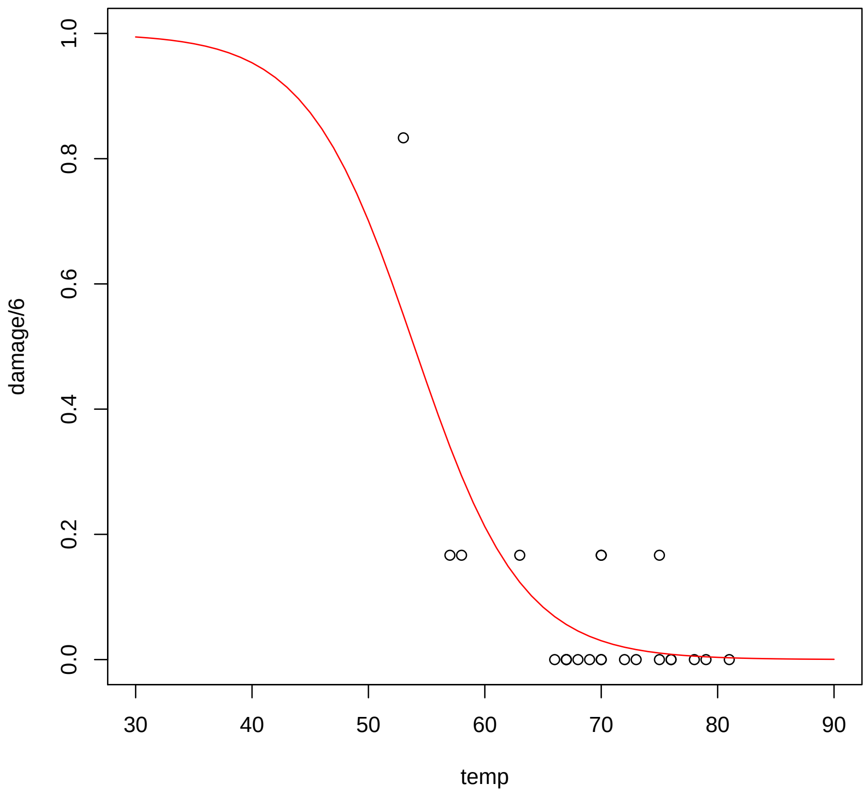
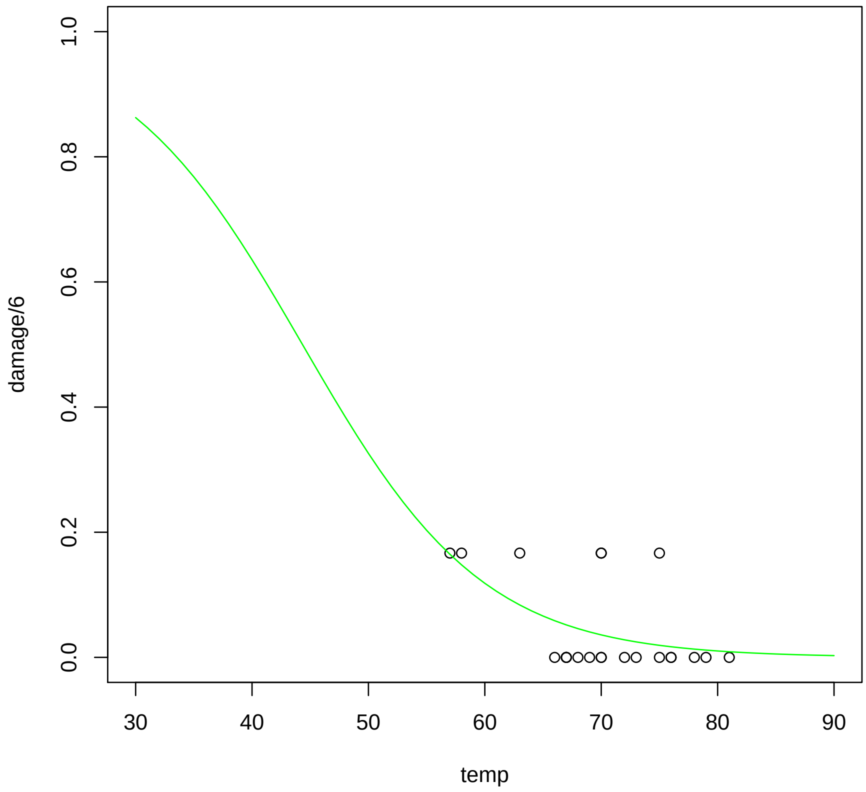


Figure 4: O-ring data with fitted logistic curve - without 5 out of 6 at 53 ^oF



2.6 Consequences and Leverage- F Ch2

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 failues 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.

Even without the near disaster at 53 ^oF, there was very strong eivdence 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.

3 Binomial regression

3.1 Setup - F Ch2

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.

3.2 Link functions - F Ch2

Link function possibilites

Usual choices for g:

logit or logistic or log-odds

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

probit or normal quantiles

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

complementary log-log

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

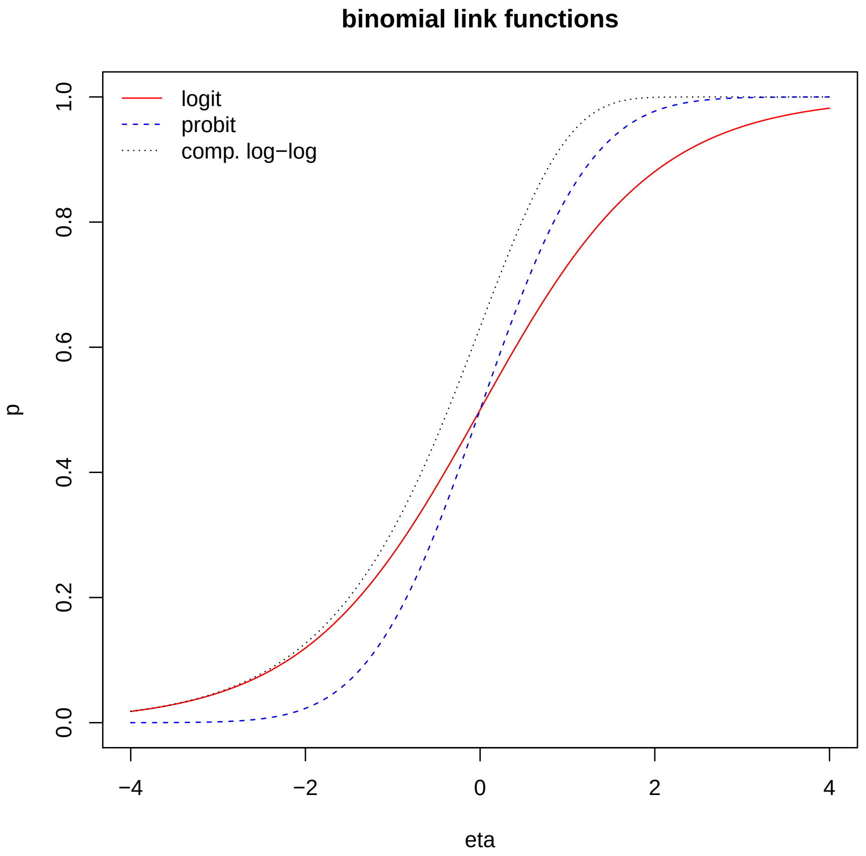
3.3 Illustration of link inverse - F Ch2

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")
```

Binomial regression model: likelihood

Figure 5: Illustration of logit, probit and comp. log log inverse link



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

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

We maximise this numerically.

Maximum likelihood estimators have many desirable properties...

4 Maximum likelihood

4.1 Likelihood Theory - F Ch2

Maximum likelihood estimation

Suppose that Y_i, $i = 1, \dots, n$, 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

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

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

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.

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$\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}^*$. 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 $l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{y})$.

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,

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

That is,

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

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

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

Proof Need to reverse order of differentiation and integration - extension of MAST90105

4.2 Example: binomial - F Ch2

Example: binomial regression

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

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

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

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

Example: binomial regression

Since
$$f''(x) = \frac{e^x}{(1+e^x)^2}.$$
differentiating the log-likelihood a second time gives
$$\begin{aligned}\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k} &= -\sum_{i=1}^n m_i \frac{e^{\eta_i}}{(1+e^{\eta_i})^2} x_{ij}x_{ik} \\ &= -\sum_{i=1}^n m_i x_{ij}x_{ik}p_i(1-p_i),\end{aligned}$$

where $p_i = 1/(1+\exp(-\eta_i)) = 1/(1+\exp(-\mathbf{x}_i^T\boldsymbol{\beta}))$. 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
$$\mathcal{I}(\boldsymbol{\beta}) = \left(\sum_{i=1}^n m_i x_{ij}x_{ik}p_i(1-p_i)\right)_{j,k=0}^q. \qquad (3)$$

4.3 Asymptotic Variance, Efficiency and CIs - F Ch2

MLE: efficiency

Cramér-Rao lower bound Under mild regularity conditions, if $\hat{\boldsymbol{\theta}}$ is any unbiased estimator of $\boldsymbol{\theta}^*$, then
$$\operatorname{Var}\hat{\boldsymbol{\theta}} \geq \mathcal{I}(\boldsymbol{\theta}^*)^{-1}$$
where by $A \geq B$ we mean $A - B$ is positive definite.

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

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

MLE: Wald CI

We can use the large sample approximation $\mathbf{t}^T\hat{\boldsymbol{\theta}} \approx N(\mathbf{t}^T\boldsymbol{\theta}^*, \mathbf{t}^T\mathcal{I}(\boldsymbol{\theta}^*)^{-1}\mathbf{t})$ to get confidence intervals for linear combinations of $\mathbf{t}^T\boldsymbol{\theta}$.

In particular, taking $\mathbf{t} = \mathbf{e}_i$ 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
$$\hat{\theta}_i \pm z(1-\alpha/2)\sqrt{(\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1})_{i,i}}$$
where $\Phi(z(1-\alpha/2)) = 1-\alpha/2$.

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
$$\mathbf{x}_i^T\hat{\boldsymbol{\theta}} \pm z(1-\alpha/2)\sqrt{\mathbf{x}_i^T(\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1})\mathbf{x}_i}$$
If $\mathcal{I}$ is unavailable then we can approximate it using the observed information $\mathcal{J}$.

4.4 Example Challenger - F Ch2

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
```

4.5 CIs for probabilities - F Ch2

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 $\mathbf{t}$.

4.6 Challenger example - F Ch2

Challenger disaster: CI for forecast when $t = 29$

```
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)

##
## 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`

```
# 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
```

4.7 Likelihood ratio- F Ch2

MLE: likelihood ratio

For large n

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

where k is the dimension of $\boldsymbol{\theta}^*$ ie $q+1$ provided X is of full rank, and the dimension of X if it is not of full rank.

This result can also be used, in principle, 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 $100(1-\alpha\text{ł})\%$ point for a χ^2_k distribution.

MLE: likelihood ratio for normal regression model

Suppose $\mathbf{y} \sim MVN(X\boldsymbol{\theta}^*, \sigma^2 I)$. Then, if the MLE for $\boldsymbol{\theta}^*$ is $\hat{\boldsymbol{\theta}}$,

$$2l(\hat{\boldsymbol{\theta}}) - 2l(\boldsymbol{\theta}^*) = SS_{Reg}/\sigma^2$$

and this has an exact χ^2_k distribution.

This result is in Theorem 5.1 in Module 5 and follows from the Analysis of Variance identity (4.13).

The reason is that the constant and σ^2 terms cancel out when the difference of log-likelihoods is computed.

The confidence regions in Module 4, Section 9.3 were constructed using this fact and the independence of the regression and residual sum of squares.

4.8 MLE theory conditions - F Ch2

MLE: regularity conditions

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 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.

5 Inference

5.1 Likelihood ratios - F Ch2

Likelihood ratios

The result on the limiting distribution of the log likelihood ratio extends to *nested* models.

Suppose we have observations $\mathbf{y}$ from $\mathbf{Y}$. Model A is nested within model B if the set of possible distributions for $\mathbf{Y}$ under A are a subset of those under B. We will assume that model A is parameterised by a subspace of the parameters used to describe model B, and that the parameter spaces differ in dimension by s. That is, we suppose that model B has s more parameters than model A. Let $\boldsymbol{\theta}^{*A}$ be the true parameter value if model A is correct, and $\boldsymbol{\theta}^{*B}$ the true parameter value if model B is correct. If model A is correct then w.l.o.g. we have $\boldsymbol{\theta}^{*B\,T} = (\boldsymbol{\theta}^{*A\,T}, \mathbf{0}^T)$.

If model A is correct, then as the sample size $n \to \infty$,

$$\begin{aligned} -2\log\frac{\mathcal{L}(\hat{\boldsymbol{\theta}}^A)}{\mathcal{L}(\hat{\boldsymbol{\theta}}^B)} &= -2(l(\hat{\boldsymbol{\theta}}^A) - l(\hat{\boldsymbol{\theta}}^B)) \qquad (4) \\ &\xrightarrow{\text{d}} \chi^2_s. \qquad (5) \end{aligned}$$

Moreover, if model B is correct then the likelihood ratio will be smaller, so that the statistic on the left of 4 is larger. Hence a one-tailed test is appropriate.

Note that (as for all ML results) this result assumes that the parameter spaces for models A and B are fixed, and the sample size grows to infinity.

We also require that $\boldsymbol{\theta}^{*A}$ and $\boldsymbol{\theta}^{*B}$ are in the *interior* of their parameter spaces (not on the boundary) when model A is correct.

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 likehihood ratio is generally better than the normal approximation to the MLE, so we prefer to use the likelihood ratio for model selection.

Compare this to linear models where the Wald test is equivalent to the F test. (We will see later that the F test is in fact a likelihood ratio test.)

5.2 Deviance - F Ch2

Deviance

The *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 *full model*.

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

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

5.3 Example binomial - F Ch2

That is

$$D = -2\sum_{i=1}^{n}\left(y_i\log\frac{\hat{y}_i}{y_i} + (m_i - y_i)\log\frac{m_i - \hat{y}_i}{m_i - y_i}\right) \tag{6}$$

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

Warning: the number of parameters in the full 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.

It just so happens: if m_ip_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.

5.4 Deviance difference - F Ch2

Deviance difference like difference of RSS

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

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

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

This is the analogue for generalised linear models of what hapens with the residual sums of squares in nested models for linear models.

5.5 Challenger example - F Ch2

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
```

```
# 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
```

5.6 AIC- F Ch2

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

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