3.7 Binomial Regression

Suppose the independent response variable $Y_1, \ldots, Y_n$ where $Y_i \sim \text{Binomial}(n_i, \pi_i)$, so that

$$\text{P}(Y_i = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

Suppose that for the ith response we also observe covariates $x_{i,1}, x_{i,2} \ldots$. Following the linear model approach, we construct a linear predictor:

$$\eta_i = \sum_{j=1}^{p} x_{ij} \beta_j.$$

We shall discuss the three common choices of link functions are used for binomial regression:

1) logit: $\eta = \log\left(\dfrac{\pi}{1-\pi}\right)$

2) probit: $\eta = \Phi^{-1}(\pi)$, where Φ^{-1} is the inverse cdf of a standard normal distribution.

3) complementary log log: $\eta = \log(-\log(1-\pi))$

Note

- A logistic binomial regression models is a GLM with binomial response and logit link function and a binary logistic binomial regression model is one with $n_i = 1$.
- The link above is defined through π and not $\mu \equiv n\pi$; this is annoying, but sometimes used in practice. You can use either π or μ to get the same results.

Worked Example

`challenger.R` `challenger-data.RData`

In January 1989, the space shuttle Challenger exploded shortly after launch. An investigation was conducted into the cause of the crash, paying particular attention to the O-ring seals in the rocket boosters. Could the failure of the O-rings have been predicted? Table 3.4 contains information about the damage of O-rings from 22 previous shuttle missions.

Temperature	No. Failure O-rings	Total No. O-rings
66	0	6
70	1	6
69	0	6
68	0	6
67	0	6
72	0	6
73	0	6
70	0	6
57	1	6
63	1	6
70	1	6
78	0	6
67	0	6
53	2	6
67	0	6
75	0	6
70	0	6
81	0	6
76	0	6
79	0	6
76	0	6
58	1	6

Table 3.4: Challenger Dataset

Let us proceed to fit a logistic regression model in R:

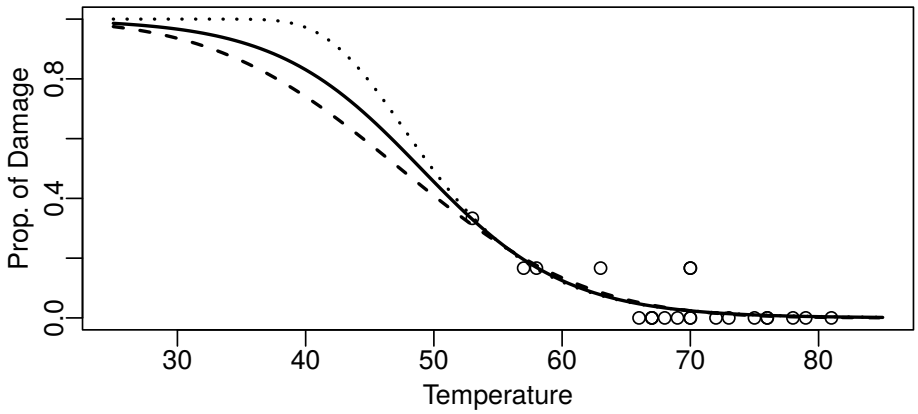
```
> mat <- cbind(dat$Fail,6-dat$Fail)
> logitmod <- glm(mat~Temp,family=binomial(link=logit),data=dat)
> probitmod <- glm(mat~Temp,family=binomial(link=probit),data=dat)
> cloglogmod <- glm(mat~Temp,family=binomial(link=cloglog),data=dat)
```

The estimated parameter values using the different link functions look very different:

```
                 logit     probit    cloglog
(Intercept)  8.6615667  4.1452794  7.9384946
Temp        -0.1768048 -0.0875188 -0.1665137
```

However, the actual fit given by each model are similar, especially around the observations:

```
> plot(x=dat$Temp,y=dat$Fail/6)
> x <- seq(25,85,by=0.5) # dummy x values
> logiteta <- 8.6615667-0.1768048*x
> probiteta <- 4.1452794-0.0875188*x
> cloglogeta <-  7.9384946-0.1665137*x
> lines(x,1/(1+exp(-logiteta)),lwd=2)
> lines(x,pnorm(probiteta),lwd=2,lty=2)
> lines(x,inv.clog(cloglogeta),lwd=2,lty=3)
```



We can predict the response given by each model at 31F:

```
> xstar <- 31
> 1/(1+exp(-(8.6615667-0.1768048*xstar)))
```

```
[1] 0.9600983
```

```
> pnorm(4.1452794-0.0875188*xstar)
```

```
[1] 0.9239562
```

```
> inv.clog(7.9384946-0.1665137*xstar)
```

```
[1] 0.9999999
```

The models suggest that there is a very high probability of damage at this temperature.

Let us check the deviance for the logistic model using the χ^2 test:

```
> pchisq(deviance(logitmod),df.residual(logitmod),lower=FALSE)
```

```
[1] 0.9776587
```

This p-value based on the χ^2 test of the deviance in well in excess of e.g. the 5% level. Thus we may conclude by saying model fits the data well. We may *not* say that the model is correct.

To construct a confidence interval for the prediction for the model using the logit link at 31F. Then

```
> ustar <- c(1,31)
> etastar <- ustar%*%logitmod$coefficients
> etastar
```

```
         [,1]
[1,] 3.180617
```

```
> 1/(1+exp(-etastar))
```

```
          [,1]
[1,] 0.9600983
```

```
> J <- vcov(logitmod)
> se <- sqrt(t(ustar)%*%J%*%ustar)
```

Then for an approximate 95% confidence interval (see section 3.4) on the probability scale:

```
> 1/(1+exp(-c(etastar - 1.96*se,etastar + 1.96*se)))
```

```
[1] 0.3957676 0.9988700
```

We can obtain the predictions directly using the `predict` command:

```
> predict(logitmod,newdata=list(Temp=31))
```

```
       1
3.180617
```

```
> predict(logitmod,newdata=list(Temp=31),type="response")
```

```
        1
0.9600983
```

The `predict` function can also be used to extract the fitted mean response values, $\widehat{\boldsymbol{\mu}}$:

```
> predict(logitmod,type="response")
```

The same procedure above can be applied for the other link functions.

Odds

Odds are sometimes a better scale than probability to represent chance.

- A 4-1 against bet would pay £4 for every £1.
- A 4-1 on bet would pay £1 for every £4.

Let p be the probability and o be the odds, where we represent e.g. 4-1 against as $1/4$ and 4 - 1 on as 4. In general, we have the relationship

$$o = \frac{p}{1-p} \quad \text{and} \quad p = \frac{o}{1+o}.$$

Odds Ratio

Suppose we have two groups where the probability of an event being in the first group is p_1 and the probability for the second group is p_2. Then the odds ratio is:

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

An odds ratio

- equal to 1 indicates the event is equally likely to occur in both groups.
- greater than 1 indicates the event is more likely to occur in the first group.
- less than 1 indicates the event is less likely to occur in the first group.

Odds Interpretation

Suppose we have a logistic Binomial regression model using two covariates x_1 and x_2. The linear predictor is given by

$$\eta = \log\left(\frac{p}{1-p}\right) = \log(o) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

We can interpret β_1 as follows: a unit increase in x_1 keeping x_2 held fixed increases the log-odds of success by β_1, or equivalently increases the odds of success by a factor of $\exp(\beta_1)$.

Note

This interpretation follows from using the logit link — no simple interpretation exists for other links.

3.7.1 Worked Example

`breastfeeding.R` `breastfeeding-data.RData`

Consider the data in Table 3.5 containing information from a study on infant respiratory disease by type of breast feeding and gender. The ratios presented are of the form "# with disease / total #".

	Bottle only	Some breast with supplement	Breast only
Boys	77/458	19/147	47/494
Girls	48/384	16/127	31/464

Table 3.5: Breast Feeding data

Can gender and feeding type be used to measure whether or not infants contract a respiratory disease?

Let us proceed to fit a logistic regression model in R:

```
> mat <- cbind(babyfood$disease, babyfood$nondisease)
> lrmod <- glm(mat~ sex + food, family=binomial, data=babyfood)
> summary(lrmod)
```

```
Call:
glm(formula = mat ~ sex + food, family = binomial, data = babyfood)

Deviance Residuals:
      1        2        3        4        5        6
 0.1096  -0.5052   0.1922  -0.1342   0.5896  -0.2284

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6127     0.1124 -14.347  < 2e-16 ***
sexGirl      -0.3126     0.1410  -2.216   0.0267 *
foodBreast   -0.6693     0.1530  -4.374 1.22e-05 ***
foodSuppl    -0.1725     0.2056  -0.839   0.4013
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.37529  on 5  degrees of freedom
Residual deviance:  0.72192  on 2  degrees of freedom
AIC: 40.24

Number of Fisher Scoring iterations: 4
```

We expect the χ^2 approximation for the deviance to be accurate for this data (why?).

Consider the interpretation of the coefficient for breast feeding. We have

```
> exp(-0.6693)
```

```
[1] 0.5120669
```

Following the log-odds interpretation: breast feeding reduces the odd of respiratory disease by 51% of that for bottle feeding. We could compute the confidence interval on the odds scale. However, to get better coverage properties we compute the interval on the log-odds scale and then transforming the endpoints as follows:

```
> z <- qnorm(0.025,lower.tail=FALSE) # critical value
> z
```

```
[1] 1.959964
```

```
> exp(c(-0.669-z*0.153,-0.669+z*0.153))
```

```
[1] 0.3795099 0.6913386
```

3.7.2 Prospective and Retrospective Sampling

Prospective Sampling : In prospective sampling, the covariates are fixed and then the outcome is observed.

Retrospective Sampling : In retrospective sampling, the outcome is fixed and then the covariates are observed.

Consider a study in which the contraction of a disease is of interest. Let

- ϖ_0 be the probability that an individual is included in the study if they do not have the disease,
- ϖ_1 be the probability that an individual is included in the study if they do have the disease.

For a prospective study, $\varpi_0 = \varpi_1$ as we have no knowledge of the outcome. For a retrospective study, typically ϖ_1 is much greater than ϖ_0.

For a given covariate x, let

- $p^*(x)$ denote the conditional probability that an individual has the disease given inclusion in the study.
- $p(x)$ denote the unconditional probability that an individual has the disease, as we would obtain from a prospective study.

Then by Bayes Theorem:

$$p^*(x) = \frac{\varpi_1 p(x)}{\varpi_1 p(x) + \varpi_0(1 - p(x))}.$$

Rearranging yields

$$\text{logit}(p^*(x)) = \log\left(\frac{\varpi_1}{\varpi_0}\right) + \text{logit}(p(x)).$$

So the only difference between the retrospective and prospective study is the intercept term $\log(\varpi_1/\varpi_0)$.

- Generally, ϖ_1/ϖ_0 is unknown — meaning we cannot estimate β_0.
- However, knowledge of the other β can be used to assess the relative error of the covariates.

Now return to the respiratory disease example:

Prospective Sampling : In the infant respiratory example, we would select a sample of newborns whose parents had chosen a particular method of feeding and then monitor them for their first year.

Retrospective Sampling : In the infant respiratory example, typically, we would find infants visiting a doctor with a respiratory disease in the first year and then record their gender and method of feeding. We would also obtain a sample of respiratory disease-free infants. How these samples are collected is important — we require that the probability of inclusion in the study is independent of the predictor.

Suppose that the respiratory disease example had been a prospective study. Then, focussing on the boys only,

- Given the infant is breast fed, the log-odds of having a respiratory disease are $\log(47/447) = -2.25$.
- Given the infant is bottle fed, the log-odds of having a respiratory disease are $\log(77/381) = -1.60$.

The difference between these two log-odds, $\Delta = -1.60 - (-2.25) = 0.65$, represents the increased risk of respiratory disease incurred by bottle feeding relative to breast feeding. This is the log-odds ratio.

Now suppose that the respiratory disease example had been a retrospective study — we could compute the log-odds of feeding type given respiratory disease status and then find the difference. The log-odds ratio would be exactly the same:

$$\Delta = \log(77/47) - \log(381/447) = 0.65$$

This shows that a retrospective design is as effective as a prospective design for estimating Δ.**

Notes

- Retrospective designs are cheaper, faster and more efficient, so it is convenient that the same results may be obtained from a prospective study.
- Retrospective studies are typically less reliable than prospective studies - relies on historical records that may be incomplete or inaccurate.

**See Supplementary handout for further details.

Summary for using the logit link

Canonical choice for binomial response GLMs is the logit link. Using the logit link

- leads to simplier mathematics,
- easy interpretation using odds;
- allows for easy analysis of retrospective data.

3.8 Poisson Regression

So far we have examined the following cases for the response:

- Response is real $\rightarrow$ normal linear model.
- Response is a probability $\rightarrow$ Binomial regression.
- Response is a bounded integer $\rightarrow$ Binomial regression.

What if the response is an unbounded integer?

One possible approach is to use a Poisson distribution as the response. Let the response Y follow a Poisson distribution with mean $\lambda > 0$:

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

Examples where the response can be modelled as Poisson are:

- counting the occurrence of a rare event (i.e. a small probability of success).
- counting the number of events in a time interval.

Recall that the Poisson distribution is a member of the exponential family as its pdf can be written as

$$\exp\left\{y\log(\lambda) - \lambda - \log(y!)\right\},$$

where $\theta = \log(\lambda)$, $b(\theta) = \lambda = \exp(\theta)$, $a(\phi) = \phi = 1$ and $c(y, \phi) = -\log(y_i!)$. Therefore the canonical link is

$$\eta = \theta = \log(\lambda)$$

Using the canonical link, the likelihood, for independent $Y_1, \ldots, Y_n$ where $Y_i \sim$ Poisson(λ_i), is:

$$L(\boldsymbol{\beta}; \boldsymbol{y}) =$$

and the log-likelihood is:

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) =$$

The discrepancy of the model can be measured using the deviance, which is

$$D = 2\sum_{i=1}^{n}\left(y_i \log(y_i/\widehat{\lambda}_i) - (y_i - \widehat{\lambda}_i)\right)$$

We can use the deviance to

- judge the goodness of fit of the model using the χ^2_{n-p} approximation.
- compare two nested models.

Alternatively, we can use Pearson's X^2 statistic as a goodness of fit measure, which for a Poisson response distribution, takes the form:

$$X^2 = \sum_{i=1}^{n}\frac{\left(y_i - \widehat{\lambda}_i\right)^2}{\widehat{\lambda}_i}$$

Note

A key property of the Poisson distribution is that its mean is equal to its variance.

3.8.1 Worked Example

In `R` fit the linear model using the following command:

```
> mylm1 <- lm(interlocks~assets+sector+nation,data=Ornstein)
```

The normal linear model fitted has response vector and design matrix:

$$\boldsymbol{Y} = \qquad\qquad , \quad X =$$

The `Ornstein` dataset contains 248 rows of data. The observations are the 248 largest Canadian firms specifying their `assets` in millions of dollars, the `sector` the firm belongs to, the `nation` that controls the firm and the number of `interlocking` director and executive positions shared with other firms.

Can we use a companys' `assets`, `sector` and `nation` values to measure the number of interlocking positions?

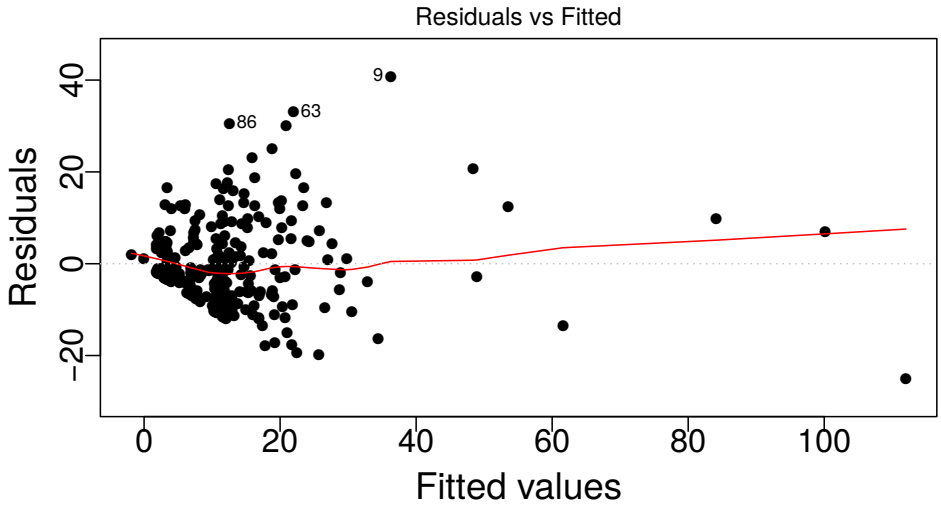
Let us take a look at the summary of linear model fitted in R:

```
> summary(mylm1)
```

```
Call:
lm(formula = interlocks ~ assets + sector + nation, data = Ornstein)

Residuals:
    Min      1Q  Median      3Q     Max
-25.001  -6.602  -1.629   4.780  40.728

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.027e+01  1.561e+00   6.575 3.14e-10 ***
assets       8.096e-04  6.119e-05  13.231  < 2e-16 ***
sectorBNK   -1.781e+01  5.906e+00  -3.016  0.00284 **
sectorCON   -4.709e+00  4.728e+00  -0.996  0.32034
sectorFIN    5.153e+00  2.646e+00   1.948  0.05266 .
sectorHLD    8.777e-01  4.004e+00   0.219  0.82669
sectorMAN    1.149e+00  2.065e+00   0.556  0.57849
sectorMER    1.491e+00  2.636e+00   0.566  0.57206
sectorMIN    4.880e+00  2.067e+00   2.361  0.01905 *
sectorTRN    6.171e+00  2.760e+00   2.236  0.02629 *
sectorWOD    8.228e+00  2.679e+00   3.072  0.00238 **
nationOTH   -1.241e+00  2.695e+00  -0.461  0.64555
nationUK    -5.775e+00  2.674e+00  -2.159  0.03184 *
nationUS    -8.618e+00  1.496e+00  -5.760 2.64e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.827 on 234 degrees of freedom
Multiple R-squared:  0.6463,        Adjusted R-squared:  0.6267
F-statistic: 32.89 on 13 and 234 DF,  p-value: < 2.2e-16
```

What can we say about the model fit from this summary?

Lets take a look at a residual plot:



This plot shows signs of heteroscedasity. Perhaps using a GLM with a Poisson response distribution could explain the data better. We now proceed to fit the model with independent $Y_1, \ldots, Y_n$ with $Y_i \sim \text{Poisson}(\lambda_i)$ and

$$\text{E}(Y_i) = \exp(\beta_1 + \beta_2 \text{assets}_i + \beta_3 \text{sector}_i + \beta_4 \text{nation}_i), \quad i = 1, \ldots, n.$$

Therefore, the link function is log link (canonical link).

```
> myglm <- glm(interlocks~assets+sector+nation,data=Ornstein,family=poisson)
> summary(myglm)
```

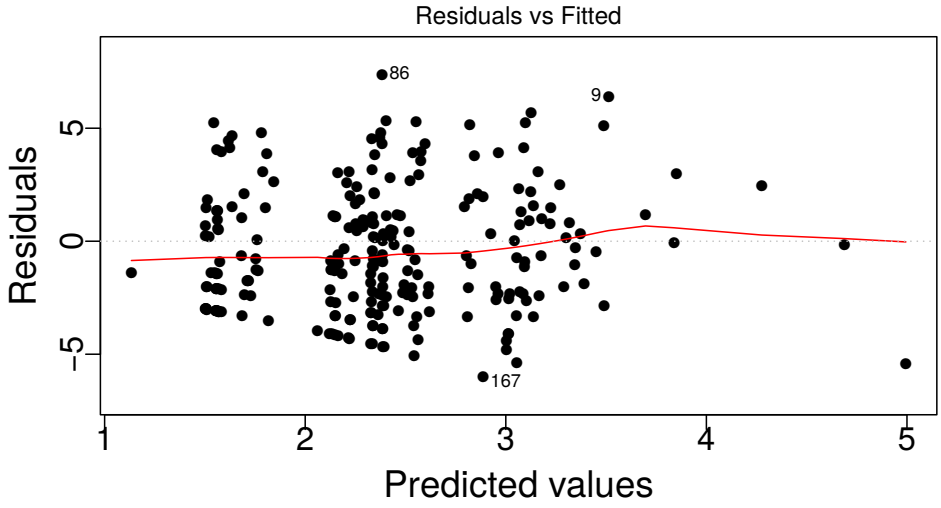
```
Call:
glm(formula = interlocks ~ assets + sector + nation, family = poisson,
    data = Ornstein)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.9908  -2.4767  -0.8582   1.3472   7.3610

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.325e+00  5.193e-02  44.762  < 2e-16 ***
assets       2.085e-05  1.202e-06  17.340  < 2e-16 ***
sectorBNK   -4.092e-01  1.560e-01  -2.623  0.00872 **
sectorCON   -6.196e-01  2.120e-01  -2.923  0.00347 **
sectorFIN    6.770e-01  6.879e-02   9.841  < 2e-16 ***
sectorHLD    2.085e-01  1.189e-01   1.754  0.07948 .
sectorMAN    5.260e-02  7.553e-02   0.696  0.48621
sectorMER    1.777e-01  8.654e-02   2.053  0.04006 *
sectorMIN    6.211e-01  6.690e-02   9.283  < 2e-16 ***
sectorTRN    6.778e-01  7.483e-02   9.059  < 2e-16 ***
sectorWOD    7.116e-01  7.532e-02   9.447  < 2e-16 ***
nationOTH   -1.632e-01  7.362e-02  -2.217  0.02663 *
nationUK    -5.771e-01  8.903e-02  -6.482 9.05e-11 ***
nationUS    -8.259e-01  4.897e-02 -16.867  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3737.0  on 247  degrees of freedom
Residual deviance: 1887.4  on 234  degrees of freedom
AIC: 2813.4

Number of Fisher Scoring iterations: 5
```

Based on the (residual) deviance, we can perform a χ^2_{n-p} significance test:

```
> pchisq(deviance(myglm), df.residual(myglm), lower=FALSE)
```

```
[1] 5.793237e-256
```

This is an extremely small p-value indicating an ill-fitting model if the Poisson response distribution is correct. Why is this a poor fit? Let us look at the diagnostic plots.



This residual plot looks better than that for the linear model — there is hardly any biased. However, the residuals are quite large — larger than the Poisson distribution suggest. This is likely caused by taking a too simple structure for the covariates. Note that the problem lies with the standard errors not the estimates — the model can be used to make predictions, but not to make inference.

The Poisson distribution is restrictive in the sense that it has only one parameter, forcing the mean to equal the variance of the observations,which is not very flexible for fitting purposes. This problem can be alleviated by estimating ϕ, which then leads to better standard errors. To this end, we compute Pearson's dispersion estimate $\widehat{\phi}_p$:

```
> phihat <- sum(residuals(myglm,type="pearson")^2)/df.residual(myglm)
> phihat
```

```
[1] 7.943697
```

We can then "plug-in" this estimate into the model:

```
> summary(myglm, dispersion=phihat)
```

```
Call:
glm(formula = interlocks ~ assets + sector + nation, family = poisson,
    data = Ornstein)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.9908  -2.4767  -0.8582   1.3472   7.3610

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.325e+00  1.464e-01  15.882  < 2e-16 ***
assets       2.085e-05  3.389e-06   6.152 7.64e-10 ***
sectorBNK   -4.092e-01  4.397e-01  -0.931 0.352038
sectorCON   -6.196e-01  5.974e-01  -1.037 0.299703
sectorFIN    6.770e-01  1.939e-01   3.492 0.000480 ***
sectorHLD    2.085e-01  3.350e-01   0.622 0.533800
sectorMAN    5.260e-02  2.129e-01   0.247 0.804857
sectorMER    1.777e-01  2.439e-01   0.728 0.466323
sectorMIN    6.211e-01  1.886e-01   3.294 0.000989 ***
sectorTRN    6.778e-01  2.109e-01   3.214 0.001309 **
sectorWOD    7.116e-01  2.123e-01   3.352 0.000803 ***
nationOTH   -1.632e-01  2.075e-01  -0.787 0.431534
nationUK    -5.771e-01  2.509e-01  -2.300 0.021456 *
nationUS    -8.259e-01  1.380e-01  -5.984 2.17e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 7.943697)

    Null deviance: 3737.0  on 247  degrees of freedom
Residual deviance: 1887.4  on 234  degrees of freedom
AIC: 2813.4

Number of Fisher Scoring iterations: 5
```

Note that the estimation of β is independent of the dispersion parameter, ϕ, therefore modifying ϕ does not change the estimated coefficients. Further, notice, in this example, that now some of the coefficients have become non-significant.

3.8.2 Overdispersion

The term overdispersion means that the observed variance of the response is larger than the variation implied by the distribution used to fit the model. Overdispersion can be caused by several different problems — we state some below:

- Observations for different individuals with the same covariates do not have exactly the same distribution; that is, there are unaccounted for individual differences not included in the model.
- Observations may be correlated or clustered, while the specified variance function wrongly assumes uncorrelated data

One approach to mitigate the problem of overdispersion is to estimate the dispersion parameter, ϕ rather than assume $\phi = 1$ for the binomial and Poisson distributions. The procedure is to "plug-in" the estimated dispersion parameter $\widehat{\phi}$ into the analysis, as done in the worked example above. As mentioned above, estimating the dispersion parameter has no effect on the estimate of β but it inflates all their standard errors.

3.8.3 Estimation Problems

It can be the case that the `glm` function in `R` fails to convergence. Problems may arise due to problems with the Fisher scoring method or a "bad" initial starting point, however sometimes it is a problem with the data themselves, which is exhibited in the following example.

The following data set contains the values of androgen and estrogen (types of hormones) from 26 healthy males with their sexual orientation.

	androgen	estrogen	orientation
1	3.9	1.8	s
2	4.0	2.3	s
3	3.8	2.3	s
4	3.9	2.5	s
5	2.9	1.3	s
6	3.2	1.7	s
7	4.6	3.4	s
8	4.3	3.1	s
9	3.1	1.8	s
10	2.7	1.5	s
11	2.3	1.4	s
12	2.5	2.1	g
13	1.6	1.1	g

	androgen	estrogen	orientation
14	3.9	3.9	g
15	3.4	3.6	g
16	2.3	2.5	g
17	1.6	1.7	g
18	2.5	2.9	g
19	3.4	4.0	g
20	1.6	1.9	g
21	4.3	5.3	g
22	2.0	2.7	g
23	1.8	3.6	g
24	2.2	4.1	g
25	3.1	5.2	g
26	1.3	4.0	g

Table 3.6: Hormone Dataset

Suppose we fit a binomial model to see if the orientation can be predicted from the hormone values.

```
> myglm <- glm(orientation~estrogen+androgen,data=hormone,family=binomial)
```

Executing the above command gives the following warnings

```
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Lets take a look at the summary

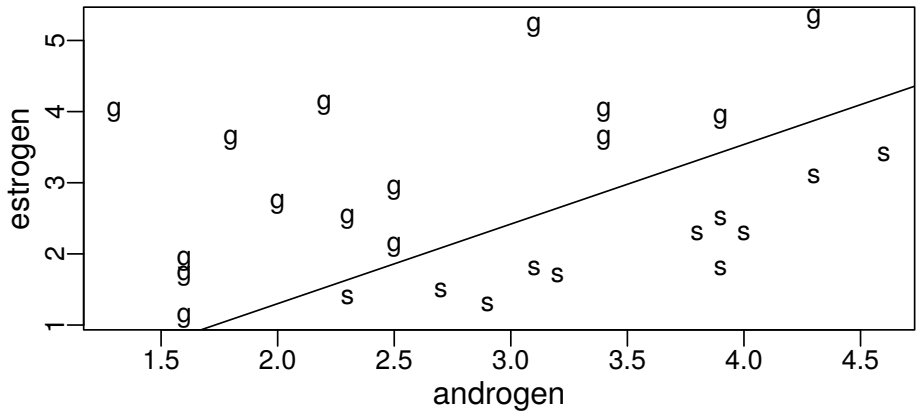
```
> summary(myglm)
```

```
Call:
glm(formula = orientation ~ estrogen + androgen, family = binomial,
    data = hormone)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.759e-05  -2.100e-08  -2.100e-08   2.100e-08   3.380e-05

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -84.49  136095.03  -0.001    1.000
estrogen       -90.22   75910.98  -0.001    0.999
androgen       100.91   92755.62   0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.5426e+01  on 25  degrees of freedom
Residual deviance: 2.3229e-09  on 23  degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 25
```

Notice

- the residual deviance is extraordinarily small indicating a good fit, yet none of the parameters are significant as they have very high standard errors.
- the default maximum number of iterations (25) has been reached.

The reason for these results come from the data.



The plot of the data reveals that the two classes of orientation are separable so that a perfect fit is possible.

This problem is called an embarrassment of riches — we can fit the data perfectly. The problem results in unstable estimates of the parameters and their standard errors, which suggest (probably incorrectly) perfect predictions.

Possible approaches to deal with these types of problems:

- Exact logistic regression.
- Bias reduction method — `R` package available `brlr`.

However, these methods are outside the scope of the course.

3.9 Quasi-Likelihood

Recall that a GLM is determined by the

-
-

We now discuss an approach that only requires specification of the link and variance functions of the model, but not the distribution of the response.

Motivation

Suppose we have the independent random variables $Y_1, \cdots, Y_n$. Let Y_i have mean μ_i and variance $\phi V(\mu_i)$. Define the score as

$$U_i = \frac{y_i - \mu_i}{\phi V(\mu_i)}.$$

It follows that

$$\begin{aligned} \mathrm{E}(U_i) &= 0 \\ \mathrm{var}(U_i) &= \frac{1}{\phi V(\mu_i)} \\ \text{and } -\mathrm{E}\left(\frac{\partial U_i}{\partial \mu_i}\right) &= \frac{1}{\phi V(\mu_i)}. \end{aligned}$$

These properties are shared by the score function of members of the exponential family. This suggests that we may use U_i as a score. Hence

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt,$$

should behave like a log-likelihood function for μ_i (if the integral exists). We shall refer to Q_i as the log quasi-likelihood (or confusingly as just the quasi-likelihood) for μ_i. As we assume the observations are independent, the log quasi-likelihood for the complete data is just the sum of the components: $Q = \sum_{i=1}^{n} Q_i$.

Example Take $V(\mu) = 1$ and $\phi = \sigma^2$. Then

$$U =$$

$$\text{and } Q =$$

which is the same as the log-likelihood of a normal distribution up to constants. ■

In general, using variance functions associated with members of the exponential family recovers the log-likelihood. Further, other choices of $V(\mu)$ may not correspond to any known distribution or may even lead to something that is not a distribution.

Estimation

The estimation of β in the model is obtained by maximising the log quasi-likelihood, Q. We can again use the IWLS algorithm (Algorithm 3.1) . The only exception is now the dispersion parameter ϕ is estimated by

$$\widehat{\phi}_P = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)},$$

which is based on Pearson's X^2 statistic. We do not use the deviance estimator for ϕ as it is based on the likelihood — not reliable here.

Inference

By analogy, the quasi-deviance function for a single observation is [††]

$$D_i = -2\phi Q_i = 2 \int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)} dt,$$

and the total quasi-deviance is the sum over the D_i. This total quasi-deviance can be used like the ordinary deviance to perform inference on the model.

[††] see problem sheet

Example in R

`quasi-seeds.R  seeds-data.RData`

We continue with the seeds example presented in section 3.6.1. We can fit a corresponding quasi-binomial model as follows:

```
> my.quasi.bin <- glm(prop ~ seed + extract + seed * extract,
+     family = quasibinomial(link = "logit"), data = dat)
> summary(my.quasi.bin)
```

```
Call:
glm(formula = prop ~ seed + extract + seed * extract, family = quasibinomial(link = "logit"),
    data = dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.88834  -0.18458   0.01555   0.13622   0.38167

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.5262     0.2764  -1.904  0.07398 .
seed          -0.2000     0.3969  -0.504  0.62079
extract        1.4479     0.3865   3.747  0.00161 **
seed:extract  -0.8478     0.5496  -1.543  0.14135
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.08915264)

    Null deviance: 3.9112  on 20  degrees of freedom
Residual deviance: 1.8151  on 17  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Notice that the estimated dispersion parameter is not 1 (as in the previous analysis), but $\phi = 0.0892$, far less than 1. Further, the models deviance has been significantly reduced.

Comparison of multiple quasi models can be done using the F-test e.g.

```
> my.quasi.bin.2 <- glm(prop ~ seed + extract, family = quasibinomial,
+     data = dat)
> anova(my.quasi.bin.2, my.quasi.bin, test = "F")
```

```
Analysis of Deviance Table

Model 1: prop ~ seed + extract
Model 2: prop ~ seed + extract + seed * extract
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        18     2.0280
2        17     1.8151  1  0.21282 2.3871 0.1407
```

This F-value is not significant, therefore there is insufficient evidence to reject the null hypothesis, i.e. use the smaller model.

The other options for use of quasi-likelihoods in R are

```
quasi(link = "identity", variance = "constant")

quasibinomial(link = "logit")

quasipoisson(link = "log")
```

■

Note

- The dispersion, ϕ, can be modelled as a free parameter which is useful in modelling overdispersion.
- Although using the quasi-likelihood approach is attractive as it uses fewer assumptions — the quasi based estimators are generally less efficient than corresponding regular likelihood-based estimator — so if information about the distribution is available, use it.