

Exercises in Generalized Linear Models

Week 39

Vinnie Ko, Jonas Moss and Per August Moen

Fall 2024

Exercises for the course [STK3100/STK4100: Introduction to Generalized Linear Models](#) at the University of Oslo, fall 2023. The exercises are from the textbook Alan Agresti: *Foundations of Linear and Generalized Linear Models*. Wiley, 2015. ISBN: 978-1-118-73003-4, and from the additional exercises posted on the course web page. The first version of these exercise solutions were written by Vinnie Ko and Jonas Moss. The solutions have been rewritten and are continuously extended by Per August Jarval Moen.

Exercise 4.14

Consider the GLM

$$Y_i \stackrel{\text{ind}}{\sim} f(y; \theta_i, \varphi_i),$$

where $g(\mathbb{E}(Y_i)) = g(\mu_i) = x_i^\top \beta$ for $i \in [n]$, for some unknown coefficient vector $\beta \in \mathbb{R}^p$ and for some known link function g . We assume that f is in the exponential dispersion family with natural parameter θ and dispersion parameter φ , i.e. that $f(y; \theta, \varphi) = \exp(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y; \varphi))$ for some functions a, b, c .

Let $\hat{\beta}$ be the MLE of β , and let $\hat{\mu}_i$ be the fitted value of observation i , i.e., $\hat{\mu}_i = g^{-1}(x_i^\top \hat{\beta})$. We begin by showing that if g is the canonical link (i.e. $g(\cdot) = (b')^{-1}(\cdot)$), the model contains an intercept, and $\varphi_i = \varphi$ for all i , then

$$\sum_{i=1}^n \hat{\mu}_i = \sum_{i=1}^n Y_i.$$

So suppose g is the canonical link, $\varphi_i = \varphi$ for all i and that $x_{i,1} = 1 \forall i \in [n]$, so that β_1 is the intercept coefficient¹. The likelihood equations (p124) imply that the MLE $\hat{\beta}$ of β satisfies

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_1}(\hat{\beta}) &= \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)x_{i,1}}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}(\hat{\beta}) \\ &= \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}(\hat{\beta}) \\ &= 0. \end{aligned}$$

Since g is the canonical link and $\mu_i = g^{-1}(\eta_i)$, we get that $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial}{\partial \eta_i} g^{-1}(\eta_i) = b''(\eta_i) = b''(\theta_i)$. The last equality follows from $\eta_i = \theta_i$ whenever g is the canonical link (i.e. $g(\cdot) = (b')^{-1}(\cdot)$). To see this, note that

¹Although we usually denote the intercept by β_0 , we use β_1 and $x_{i,1} = 1$ here to align the notation with the one used in the book when presenting the likelihood equations.

$\eta_i = g(\mu_i) = g(b'(\theta_i)) = (b')^{-1}(b'(\theta_i)) = \theta_i$. Since we also have that $\text{Var}(Y_i) = b''(\theta_i)a(\varphi)$, we obtain

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_1}(\hat{\beta}) &= \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}(\hat{\beta}) \\ &= \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)}{b''(\hat{\theta}_i)a(\varphi)} b''(\hat{\theta}_i) \\ &= \frac{1}{a(\varphi)} \sum_{i=1}^n (Y_i - \hat{\mu}_i) \\ &= 0.\end{aligned}$$

Hence, $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{\mu}_i$, and so the residuals have sample mean zero.

In the above, we needed

- (1) inclusion of an intercept,
- (2) canonical link, and
- (3) a constant value of φ_i for all i ².

If one or some of (1), (2) or (3) are not satisfied, we can cook up examples where the likelihood equations imply that $\sum_{i=1}^n Y_i \neq \sum_{i=1}^n \hat{\mu}_i$ by choosing x_i , $b(\theta)$, $g(\cdot)$ and/or φ_i suitably.

Exercise 5.10

We consider the logistic regression model

$$Y_i \stackrel{\text{ind}}{\sim} \frac{1}{n_i} \text{Binom}(n_i, \pi_i),$$

for $i \in [n]$, where $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$ for some $\beta_0, \beta_1 \in \mathbb{R}$ and covariates $x_i \in \mathbb{R}$.

Let x_0 be the value of x such that $\pi = \pi(x) = \text{logit}^{-1}(\beta_0 + \beta_1 x) = \pi_0$ for some given value of π_0 . Then x_0 depends on the unknown parameters β_0 and β_1 .

We seek an (approximate) confidence interval for x_0 by inverting hypothesis tests.

Remark. Why does inverting a test give a valid confidence interval? Let $T_\alpha = T_\alpha(\tilde{\theta})$ be any testing procedure for testing $H_0 : \theta = \tilde{\theta}$ versus $H_1 : \theta \neq \tilde{\theta}$. Here θ is some parameter of interest, like x_0 in the exercise. We let $T_\alpha = 1$ correspond to rejection and $T_\alpha = 0$ correspond to no rejection of the null hypothesis. Suppose T_α is a valid testing procedure, in the sense that $\mathbb{P}_{\tilde{\theta}}(T_\alpha = 1) = \mathbb{P}_{\tilde{\theta}}(\text{Reject } H_0 \text{ at level } \alpha) \leq \alpha$. Here $\mathbb{P}_{\tilde{\theta}}$ denotes the probability measure of the model under the assumption that $H_0 : \theta = \tilde{\theta}$ is true. Say we construct a confidence interval for θ to be the values of $\tilde{\theta}$ for which T_α does not reject. That is, define

$$\begin{aligned}\text{CI}_\alpha &= \left\{ \tilde{\theta} : \tilde{\theta} \text{ is not rejected by } T_\alpha(\tilde{\theta}) \right\} \\ &= \left\{ \tilde{\theta} : T_\alpha(\tilde{\theta}) = 0 \right\}.\end{aligned}$$

Then CI_α is a valid confidence interval for θ . Indeed, we have that if θ_{true} is the true value of θ , then

$$\begin{aligned}\mathbb{P}(\theta_{\text{true}} \in \text{CI}_\alpha) &= \mathbb{P}(\theta_{\text{true}} \text{ is not rejected by } T_\alpha(\theta_{\text{true}})) \\ &\geq 1 - \alpha,\end{aligned}$$

²The book usually lets φ be the same for all i , and instead allows $a(\varphi)$ differ for each i . This is not transparent notation, so I prefer to let a be the same across observations and rather let φ_i depend on i .

by the assumption on the testing procedure being valid. Hence, inverting a test gives a valid confidence interval.

Back to the exercise. We will first construct a confidence interval for x_0 by inverting a Wald test.

i)

We consider a Wald test for testing $H_0 : x_0 = \tilde{x}$ versus $H_1 : x_0 \neq \tilde{x}$. Note that $\beta_0 + \beta_1 x_0 = \text{logit}(\pi_0)$. By maximum likelihood theory (see remark on this later), we have that

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \stackrel{d}{\approx} N \left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} \text{Var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{Var}(\hat{\beta}_1) \end{bmatrix} \right).$$

Consequently³,

$$\hat{\beta}_0 + \tilde{x} \hat{\beta}_1 \stackrel{d}{\approx} N(\beta_0 + \tilde{x} \beta_1, V),$$

where $V = \text{Var}(\hat{\beta}_0) + \tilde{x}^2 \text{Var}(\hat{\beta}_1) + 2\tilde{x} \text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$. It then follows that, under $H_0 : x_0 = \tilde{x}$,

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 \tilde{x} - \text{logit}(\pi_0)}{\sqrt{V}} \stackrel{d}{\approx} N(0, 1).$$

By replacing the true variance V by the estimated version $\hat{V} = \widehat{\text{Var}}(\hat{\beta}_0) + \tilde{x}^2 \widehat{\text{Var}}(\hat{\beta}_1) + 2\tilde{x} \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1)$, we obtain a Wald statistic

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 \tilde{x} - \text{logit}(\pi_0)}{\sqrt{\hat{V}}} \stackrel{d}{\approx} N(0, 1).$$

For testing $H_0 : x_0 = \tilde{x}$ versus $H_1 : x_0 \neq \tilde{x}$, we can reject H_0 if

$$\left| \frac{\hat{\beta}_0 + \hat{\beta}_1 \tilde{x} - \text{logit}(\pi_0)}{\sqrt{\hat{V}}} \right| > z_{1-\alpha/2},$$

where $z_{1-\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution (the book uses the notation $z_{\alpha/2}$ for the same quantity). This is an approximately valid test. Hence, by the above remark, we obtain that an approximately valid confidence interval for x_0 is given by all values of $\tilde{x}$ for which

$$\left| \frac{\hat{\beta}_0 + \hat{\beta}_1 \tilde{x} - \text{logit}(\pi_0)}{\sqrt{\hat{V}}} \right| \leq z_{1-\alpha/2},$$

which is the confidence interval described in the exercise text.

ii)

How could we invert a likelihood ratio test to form another confidence interval for x_0 ? Note that $H_0 : x_0 = \tilde{x}$ and $H_1 : x_0 \neq \tilde{x}$ are equivalent to $H_0 : \beta_0 + \beta_1 \tilde{x} = \text{logit}(\pi_0)$ and $H_1 : \beta_0 + \beta_1 \tilde{x} \neq \text{logit}(\pi_0)$. A likelihood ratio testing procedure for testing H_0 versus H_1 is constructed as follows.

Let

$$\begin{aligned} \text{LR}(\tilde{x}) &:= 2 \left\{ \sup_{H_1} \ell(\beta_0, \beta_1) - \sup_{H_0} \ell(\beta_0, \beta_1) \right\} \\ &= 2 \left\{ \sup_{\beta_0, \beta_1 \in \mathbb{R}} \ell(\beta_0, \beta_1) - \sup_{(\beta_0, \beta_1) : \beta_0 + \beta_1 \tilde{x} = \text{logit}(\pi_0)} \ell(\beta_0, \beta_1) \right\} \\ &= 2 \left\{ \ell(\hat{\beta}_0, \hat{\beta}_1) - \sup_{(\beta_0, \beta_1) : \beta_0 + \beta_1 \tilde{x} = \text{logit}(\pi_0)} \ell(\beta_0, \beta_1) \right\}, \end{aligned}$$

³If $Z \sim N_p(\xi, \Sigma)$, then for any $v \in \mathbb{R}^p$ we have that $v^T Z \sim N(v^T \xi, v^T \Sigma v)$.

where $\ell(\beta_0, \beta_1)$ is the log likelihood. By Wilks' Theorem (see the solution for Addition Exercise 16), we have that

$$\text{LR}(\tilde{x}) \stackrel{d}{\approx} \chi_1^2$$

under H_0 .

Hence, an approximate level α test for testing H_0 versus H_1 is to reject H_0 if $\text{LR}(\tilde{x}) > \chi_1^2(1 - \alpha)$, where $\chi_1^2(1 - \alpha)$ is the upper α quantile of the χ_1^2 distribution. A $100(1 - \alpha)\%$ confidence interval for x_0 is then given by

$$\{\tilde{x} : \text{LR}(\tilde{x}) \leq \chi_1^2(1 - \alpha)\},$$

since we above showed that inverting valid hypothesis tests gives valid confidence intervals.

Exercise 5.16

In this exercise we will see that grouping binomial data is equivalent to not grouping the data in terms of the likelihood, but that grouped and un-grouped data will have different deviances.

Suppose we have ungrouped binomial data

$$Y_{i,j} \stackrel{\text{ind}}{\sim} \text{Bin}(1, \pi_i) \text{ for } j \in [n_i], i \in [N].$$

We assume that $\text{logit}(\pi_i) = x_i^\top \beta$ for $i \in [N]$ and some unknown $\beta \in \mathbb{R}^p$.

a)

We shall show that the log-likelihood of the un-grouped data is equal to the log-likelihood of the grouped data up to a constant term.

The log-likelihood of the ungrouped data $(Y_{i,j})$ is given by

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^N \sum_{j=1}^{n_i} Y_{i,j} \log(\pi_i) + (1 - Y_{i,j}) \log(1 - \pi_i) \\ &= \sum_{i=1}^N n_i \tilde{Y}_i \log(\pi_i) + n_i(1 - \tilde{Y}_i) \log(1 - \pi_i), \end{aligned}$$

where $\tilde{Y}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}$. Note that $\tilde{Y}_i \stackrel{\text{ind}}{\sim} \frac{1}{n_i} \text{Bin}(n_i, \pi_i)$, and represents the data in a grouped form. Now, the likelihood of the grouped data is given by

$$\tilde{\ell}(\beta) = \sum_{i=1}^N n_i \tilde{Y}_i \log(\pi_i) + n_i(1 - \tilde{Y}_i) \log(1 - \pi_i) + \log \binom{n_i}{n_i \tilde{Y}_i}.$$

Notice that $\ell(\beta)$ and $\tilde{\ell}(\beta)$ only differ by the term $\log \binom{n_i}{n_i \tilde{Y}_i}$, which is constant in β .

b)

In a saturated GLM, each observation is given its own freely varying mean. When we consider un-grouped data, this means that each response $Y_{i,j}$ is given a free parameter $\pi_{i,j} \in (0, 1)$. When we consider grouped data, this means that each response $\tilde{Y}_i$ is given its own free parameter π_i .

Hence, the saturated model will differ between grouped and un-grouped data. This implies that the log-likelihoods of the saturated models will not necessarily be the same. Hence the deviance, which depends on the log-likelihood of the saturated model, will in general be different for grouped and un-grouped binomial data.

c)

We shall explain why the difference between deviances for two unsaturated models does not depend on the form of data entry (i.e. is the same for grouped and un-grouped data).

So suppose we wish to compare two competing nested models M_0 and M_1 with the same form of data entry (either both use grouped data or both use un-grouped data). To compare them, we can consider the difference between their deviances $D(Y; \hat{\mu}_0) - D(Y; \hat{\mu}_1)$, where $\hat{\mu}_0$ are the fitted values under model M_0 and $\hat{\mu}_1$ are the fitted values under model M_1 . Let $\ell(\cdot; Y)$ denote the log-likelihood of the data, parameterized by the mean values upon observing $Y = (Y_1, \dots, Y_N)^T$. Then, the difference between the deviances between the two competing models becomes ⁴

$$\begin{aligned} D(Y; \hat{\mu}_0) - D(Y; \hat{\mu}_1) &= -2 \{ \ell(\hat{\mu}_0; Y) - \ell(Y; Y) \} + 2 \{ \ell(\hat{\mu}_1; Y) - \ell(Y; Y) \} \\ &= -2 \{ \ell(\hat{\mu}_0; Y) - \ell(\hat{\mu}_1; Y) \}. \end{aligned}$$

Thus, the log-likelihood of the saturated model cancels, and the difference in deviance only depends on the log-likelihoods of the unsaturated models. In (a) we showed that these do not depend on the form of data entry. Hence, the difference in deviance does not depend on the form of data entry either.

Remark. We recognize that the difference between the deviances in this case is the likelihood ratio statistic for testing model M_0 versus model M_1 .

Additional exercise 15

Recall that we considered the logistic regression model for the Beetle data:

$$Y_i \stackrel{\text{ind}}{\sim} \frac{1}{n_i} \text{Bin}(n_i, \pi_i), \quad \text{for } i \in [n],$$

where we use the canonical logit link $\text{logit}(\pi_i) = \eta_i$ and linear predictor $\eta_i = \beta_0 + \beta_1 x_i$. The response Y_i is the proportion of deaths in group i and the covariate x_i is the dose given to group i .

The LD50 is the value of the covariate x for which $\pi = \pi(x)$ is equal to $\frac{1}{2}$.

i)

We estimate the value of LD50. Note that LD50 is a function of the unknown coefficients β_1 and β_2 , and so LD50 is also unknown.

We know that $\text{logit}(\pi) = \beta_0 + \beta_1 x$. Rearranging the equation, we get that $x = \frac{\text{logit}(\pi) - \beta_0}{\beta_1}$ and thus $LD50 = \frac{\text{logit}(\frac{1}{2}) - \beta_0}{\beta_1} = -\frac{\beta_0}{\beta_1}$. An natural estimate of LD50 is therefore $\widehat{LD50} = -\frac{\hat{\beta}_0}{\hat{\beta}_1} = 1.7716$. In fact,

⁴In the definition of the deviance, we multiply by φ , which is the common dispersion parameter. If each observation has a separate dispersion parameter, say $\varphi_i = a_i \varphi$, as in the case of Binomial data, then the a_i factor is captured by the likelihood term, and we only multiply by φ in the deviance. In the case of Binomial data, each observation i has dispersion parameter $\varphi_i = \frac{1}{n_i}$, and so we may take $\varphi = 1$.

by the invariance⁵ of Maximum Likelihood, $\widehat{LD50}$ is the MLE of LD50.

Let us estimate LD50 from the data:

```
> # Read data.
> Beetle = read.table("http://www.stat.ufl.edu/~aa/glm/data/Beetles2.dat", header = T)
> head(Beetle)
  logdose  n dead
1   1.691 59    6
2   1.724 60   13
3   1.755 62   18
4   1.784 56   28
5   1.811 63   52
6   1.837 59   53
>
> # Fit logistic regression
> Beetle.model.1 = glm(cbind(dead,n-dead) ~ logdose, family = binomial, data = Beetle)
> summary(Beetle.model.1)
```

Call:
 glm(formula = cbind(dead, n - dead) ~ logdose, family = binomial,
 data = Beetle)

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-1.5878	-0.4085	0.8442	1.2455	1.5860

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-60.740	5.182	-11.72	<2e-16 ***
logdose	34.286	2.913	11.77	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.202 on 7 degrees of freedom
 Residual deviance: 11.116 on 6 degrees of freedom
 AIC: 41.314

Number of Fisher Scoring iterations: 4

```
>
> # Estimation of LD50
> LD50 = -as.numeric(Beetle.model.1$coef[1]) / as.numeric(Beetle.model.1$coef[2])
> show(LD50)
[1] 1.771576
```

ii)

⁵The invariance of maximum likelihood is the following. Say $\hat{\theta}$ is the MLE of some parameter vector $\theta \in \mathbb{R}^p$. Then for any function $g : \mathbb{R}^p \rightarrow \mathbb{R}$, the MLE of $g(\theta)$ is given by $g(\hat{\theta})$.

From exercise 5.10 of the book, we have that a 95% confidence interval for LD50 is given by all x that satisfy the inequality

$$\left| \frac{\widehat{\beta}_0 + \widehat{\beta}_1 x - \text{logit}(0.50)}{\sqrt{\widehat{\text{Var}}(\widehat{\beta}_0) + x^2 \widehat{\text{Var}}(\widehat{\beta}_1) + 2x \widehat{\text{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1)}} \right| < 1.96$$

or equivalently

$$\begin{aligned} (\widehat{\beta}_0 + \widehat{\beta}_1 x)^2 &< 1.96^2 \cdot (\widehat{\text{Var}}(\widehat{\beta}_0) + x^2 \widehat{\text{Var}}(\widehat{\beta}_1) + 2x \widehat{\text{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1)) \\ (\widehat{\beta}_1^2 - 1.96^2 \cdot \widehat{\text{Var}}(\widehat{\beta}_1)) x^2 + 2(\widehat{\beta}_0 \widehat{\beta}_1 - 1.96^2 \widehat{\text{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1)) x + \widehat{\beta}_0^2 - 1.96^2 \cdot \widehat{\text{Var}}(\widehat{\beta}_0) &< 0 \end{aligned}$$

Thus, the confidence interval is all x that satisfy the second degree inequality

$$ax^2 + bx + c < 0$$

where

$$\begin{aligned} a &= \widehat{\beta}_1^2 - 1.96^2 \cdot \widehat{\text{Var}}(\widehat{\beta}_1) \\ b &= 2\widehat{\beta}_0 \widehat{\beta}_1 - 2 \cdot 1.96^2 \widehat{\text{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1) \\ c &= \widehat{\beta}_0^2 - 1.96^2 \cdot \widehat{\text{Var}}(\widehat{\beta}_0). \end{aligned}$$

By solving the inequality, we obtain the confidence interval

$$\left(\frac{-b - \sqrt{b^2 - 4ac}}{2a}, \frac{-b + \sqrt{b^2 - 4ac}}{2a} \right)$$

```
> # 95% confidence interval of LD50
> alpha = 0.05
> z.value = qnorm(1 - alpha/2)
>
> beta.hat = as.numeric(Beetle.model.1$coeff)
> beta.hat.cov.mat = vcov(Beetle.model.1)
>
> a.val = beta.hat[2]^2 - (z.value^2)*beta.hat.cov.mat[2,2]
> b.val = 2*beta.hat[1]*beta.hat[2] - 2*(z.value^2)*beta.hat.cov.mat[1,2]
> c.val = beta.hat[1]^2 - (z.value^2)*beta.hat.cov.mat[1,1]
>
> LD50.CI95 = c(
+   (-b.val -sqrt(b.val^2 -4*a.val*c.val))/(2*a.val),
+   (-b.val +sqrt(b.val^2 -4*a.val*c.val))/(2*a.val)
+ )
> show(LD50.CI95)
[1] 1.763722 1.779054
```

Additional Exercise 16

Assume that $V \sim \text{Binomial}(n, \pi)$, so that V has pmf

$$f(v; \pi, n) = \binom{n}{v} \pi^v (1 - \pi)^{n-v}, \quad v \in \{0, 1, \dots, n\}.$$

(a)

Given the observation $V = v$, the log-likelihood is given by

$$\ell(\pi; v) = \log \binom{v}{n} + v \log \pi + (n - v) \log(1 - \pi).$$

The score function equals the derivative of the log-likelihood with respect to π , namely

$$S(\pi; v) = \frac{v}{\pi} - \frac{n - v}{1 - \pi} = \frac{v(1 - \pi) - \pi(n - v)}{\pi(1 - \pi)} = \frac{v - \pi n}{\pi(1 - \pi)}.$$

Now, since

$$\frac{d}{d\pi} \left[\frac{v}{\pi} - \frac{n - v}{1 - \pi} \right] = -\frac{v}{\pi^2} - \frac{n - v}{(1 - \pi)^2},$$

the *observed* information matrix is given by

$$I(\pi; v) = \frac{v}{\pi^2} + \frac{n - v}{(1 - \pi)^2},$$

and the *expected* information matrix is given by

$$\begin{aligned} J(\pi) &= \mathbb{E}(I(\pi; V)), \\ &= \frac{n}{\pi(1 - \pi)}, \end{aligned}$$

where we used that $\mathbb{E}(V) = n\pi$.

We remark that the MLE of π is $\hat{\pi} = V/n$.

(b)

When n is large, maximum likelihood theory states that the MLE $\hat{\pi}$ of π satisfies

$$\hat{\pi} \stackrel{d}{\approx} N(\pi, J^{-1}).$$

Remark (for those who are curious). More formally, maximum likelihood theory⁶ states that if $Y_i \stackrel{i.i.d.}{\sim} f(y; \theta)$ for some $\theta \in \mathbb{R}^d$, and f is sufficiently smooth (and some other regularity conditions), then the MLE $\hat{\theta}$ of θ satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \stackrel{d}{\rightarrow} N(0, J_0^{-1}), \quad \text{as } n \rightarrow \infty, \quad (1)$$

where

$$J_0 = -\mathbb{E} \left(\frac{\partial^2 (\log(f(Y; \theta)))}{\partial \theta \partial \theta^\top} \right).$$

Notice that $J_0 = \frac{1}{n}J$, where J is defined as in (a). The notation $\stackrel{d}{\rightarrow}$ means "converges in distribution".⁷ Equation (1) is the formal statement of $\hat{\theta} \stackrel{d}{\approx} N(\theta, J^{-1})$. Indeed, from (1) we get that $\sqrt{n}(\hat{\theta} - \theta) \stackrel{d}{\approx} N(0, J_0^{-1})$,

⁶See, for instance, *A course in Large Sample Theory* by Thomas Ferguson (1995), or *Confidence, Likelihood and Probability* by Tore Schweder and Nils Lid Hjort (2016).

⁷We say that a sequence of real-valued random variables $(X_n)_{n \in \mathbb{N}}$ converges in distribution to some r.v. X if $F_n(x) \rightarrow F(x)$ as $n \rightarrow \infty$ for all values of x at which F is continuous, where F_n is the CDF of X_n and F is the CDF of X . Defined more generally, we say that a sequence r.v.s $(X_n)_{n \in \mathbb{N}}$ with values in $\mathbb{R}^d$ converges in distribution to some r.v. X if, for all bounded and continuous $g : \mathbb{R}^d \rightarrow \mathbb{R}$, it holds that $\mathbb{E}(g(X_n)) \rightarrow \mathbb{E}(g(X))$ as $n \rightarrow \infty$. Intuitively, if $X_n \stackrel{d}{\rightarrow} X$, then X_n has approximately the same distribution as X whenever n is large.

and so $\hat{\theta} \stackrel{d}{\approx} N(\theta, \frac{1}{n}J_0^{-1}) = N(\theta, J^{-1})$. The reason the LHS of (1) is scaled up by $\sqrt{n}$ is that $\hat{\theta} - \theta$ shrinks with n . You can see this by noting that $J^{-1} \rightarrow 0$ as $n \rightarrow \infty$. Hence it does not strictly make sense to write $(\hat{\theta} - \theta) \stackrel{d}{\rightarrow} N(0, J^{-1})$ as $n \rightarrow \infty$ because J^{-1} depends on n and also vanishes as $n \rightarrow \infty$.

Importantly, if θ is one-dimensional, the convergence in (1) implies that $(\hat{\theta} - \sqrt{J_0^{-1}z_{1-\alpha/2}/\sqrt{n}}, \hat{\theta} + \sqrt{J_0^{-1}z_{1-\alpha/2}/\sqrt{n}})$ is an *asymptotically* valid confidence interval, in the sense that

$$\begin{aligned} & \mathbb{P} \left(\theta \in (\hat{\theta} - \sqrt{J_0^{-1}z_{1-\alpha/2}/\sqrt{n}}, \hat{\theta} + \sqrt{J_0^{-1}z_{1-\alpha/2}/\sqrt{n}}) \right) \\ &= \mathbb{P} \left(\theta \in (\hat{\theta} - \sqrt{J^{-1}z_{1-\alpha/2}}, \hat{\theta} + \sqrt{J^{-1}z_{1-\alpha/2}}) \right) \\ &\rightarrow 1 - \alpha \text{ as } n \rightarrow \infty. \end{aligned}$$

That is, the constructed confidence interval will with probability *tending to* $1 - \alpha$ contain the true value of θ . This fact is not entirely obvious, but it can be shown using the properties of the notion of convergence in distribution (in one dimension).

Now, back to the exercise. A Wald statistic for testing $H_0 : \pi = \pi_0$ versus $H_1 : \pi \neq \pi_0$ is given by

$$Z_W(\pi_0) := \frac{\hat{\pi} - \pi_0}{\sqrt{J^{-1}(\hat{\pi})}} = \sqrt{n} \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1 - \hat{\pi})}},$$

and is approximately distributed as $N(0, 1)$ under H_0 , due to the asymptotic normality of the MLE and the fact that, for large n , $\hat{\pi} \approx \pi_0$ under H_0 (this can also be formalized; look up "convergence in probability" or "consistency" and "Slutsky's Lemma").

(c)

The score statistic for testing $H_0 : \pi = \pi_0$ versus $H_1 : \pi \neq \pi_0$ is given by

$$Z_S(\pi_0) := \frac{S(\pi_0)}{\sqrt{J(\pi_0)}} = \frac{\frac{v - \pi_0 n}{\pi_0(1 - \pi_0)}}{\sqrt{\frac{n}{\pi_0(1 - \pi_0)}}} = \sqrt{n} \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)}},$$

which is approximately distributed as $N(0, 1)$, due to the asymptotic normality of the MLE. Notice that the score statistic only differs from the Wald statistic in that the score statistic divides by the true asymptotic variance under H_0 ($\pi_0(1 - \pi_0)$), while the Wald statistic divides by the estimated asymptotic variance ($\hat{\pi}(1 - \hat{\pi})$).

(d)

The likelihood ratio statistic for testing $H_0 : \pi = \pi_0$ versus $H_1 : \pi \neq \pi_0$ is given by

$$\begin{aligned} Z_{LR}^2(\pi_0) &:= 2 \left\{ \sup_{\pi \in (0,1)} \ell(\pi) - \sup_{\pi \in \{\pi_0\}} \ell(\pi) \right\} \\ &= -2(l(\pi_0; v) - l(\hat{\pi}; v)) \\ &= 2v \log \left(\frac{\hat{\pi}}{\pi_0} \right) + 2(n - v) \log \left(\frac{1 - \hat{\pi}}{1 - \pi_0} \right), \end{aligned}$$

where $\hat{\pi}$ is the MLE of π . When n is large, $Z_{LR}^2(\pi_0)$ is approximately χ_1^2 -distributed under the null hypothesis that $\pi = \pi_0$.

Remark (for those who are curious). It is not at all obvious why $Z_{LR}^2(\pi_0)$ has a chi square distribution. This follows from a general theorem called Wilks' Theorem (note that there are several theorems called Wilks' theorem). We state it (somewhat unprecisely) in general form here. Suppose $Y_i \stackrel{i.i.d.}{\sim} f(y; \theta)$ for some $\theta \in \mathbb{R}^d$, and f is sufficiently smooth (and some other regularity conditions we do not state here). Let $\ell(\theta) = \log(\prod_{i=1}^n f(Y_i; \theta))$ be the log-likelihood. Suppose we want to test the null hypothesis $H_0 : \theta \in \Theta_0$ versus the alternative hypothesis $H_1 : \theta \in \Theta_1$, where Θ_0 is a set of dimension p_0 and Θ_1 is a set of dimension p_1 . Assume that the *true* θ generating the data is contained in Θ_0 ; that is, the null hypothesis is true. Then we have that the likelihood ratio statistic satisfies

$$Z_{LR}^2 := 2 \left\{ \sup_{\theta \in \Theta_1} \ell(\theta) - \sup_{\theta \in \Theta_0} \ell(\theta) \right\} \xrightarrow{d} \chi_{p_1 - p_0}^2 \text{ as } n \rightarrow \infty.$$

Informally, this means that $Z_{LR}^2 \stackrel{d}{\approx} \chi_{p_1 - p_0}^2$ for large n , when the null hypothesis is true. When the null hypothesis is not true, the statistic becomes large as n becomes large.

In the above, $\Theta_1 = (0, 1)$ with dimension $p_1 = 1$ and $\Theta_0 = \{\pi_0\}$ with dimension $p_0 = 0$.

(e)

The solutions are as follows.

Wald:

$$\sqrt{n} \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1 - \hat{\pi})}} \approx -4.4$$

Score:

$$\sqrt{n} \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)}} \approx -4$$

Likelihood:

$$Z_{LR}^2(\pi_0) \approx 16.5$$

All tests agree to roughly the same degree that H_0 is false. Notice that $\sqrt{16.5} \approx 4.05$, and since the root of a χ_1^2 is the absolute value of the normal, the likelihood ratio test agrees with the others as well.

(f)

The solutions are as follows.

Wald:

$$\sqrt{n} \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1 - \hat{\pi})}} \approx -11.5$$

Score:

$$\sqrt{n} \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)}} \approx -6.7$$

Likelihood:

$$Z_{LR}^2(\pi_0) \approx 34$$

Since $\sqrt{34} \approx 5.85$, the likelihood test is more conservative.

(g)

This is the case since

$$\sqrt{n} \frac{\hat{\pi} - \pi}{\sqrt{\hat{\pi}(1 - \hat{\pi})}} \approx N(0, 1)$$

and ± 1.96 are the two-sided confidence limits for a standard normal variable.

(h)

The score is asymptotically standard normal, which implies the square is asymptotically χ_1^2 . The $(1 - \alpha)$ -percentile is $-\Phi^{-1}(\alpha/2) = c_\alpha$. We find the confidence interval by inverting the hypothesis test that rejects when the square of the score is larger than c_α . This is done by solving the equation

$$n \frac{(\hat{\pi} - \pi_0)^2}{\pi_0(1 - \pi_0)} = c_\alpha^2.$$

The equation is equivalent to

$$n(\hat{\pi}^2 - 2\hat{\pi}\pi_0 + \pi_0^2) = c_\alpha^2\pi_0 - c_\alpha^2\pi_0^2.$$

Rearrange to get

$$\pi_0^2(n + c_\alpha^2) - \pi_0(2v + c_\alpha^2) + n\hat{\pi}^2 = 0.$$

The solutions are

$$\frac{2v + c_\alpha^2 \pm \sqrt{(2v + c_\alpha^2)^2 - 4(n + c_\alpha^2)n\hat{\pi}^2}}{2(n + c_\alpha^2)}.$$

Simplify to get

$$\frac{v + \frac{1}{2}c_\alpha^2}{n + c_\alpha^2} + \frac{\sqrt{(2v + c_\alpha^2)/4 - (n + c_\alpha^2)n\hat{\pi}^2}}{n + c_\alpha^2}.$$

(i)

wald

[1] 0.2101832 0.3898168

score

[1] 0.2189489 0.3958485

lrt

[1] 0.2160263 0.39409

(j)

```
# wald  
0.007283575 0.092716425
```

```
# score  
0.02154368 0.11175047
```

```
# lrt  
0.01823454 0.10438783
```

Additional Exercise 17

(a)

We consider the normal linear model

$$Y_i = x_i^\top \beta + \varepsilon_i \quad (2)$$

where $\varepsilon_i \stackrel{\text{ind}}{\sim} N\left(0, \frac{\sigma^2}{w_i}\right)$ for $i \in [n]$ and some unknown $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$.

By multiplying both sides of equation (2) with $\sqrt{w_i}$, we get

$$\begin{aligned} \sqrt{w_i} Y_i &= \sqrt{w_i} x_i^\top \beta + \sqrt{w_i} \varepsilon_i \\ Y_i^* &= (x_i^*)^\top \beta + \varepsilon_i^*, \end{aligned} \quad (3)$$

for all i , where $\varepsilon_i^* = \sqrt{w_i} \varepsilon_i \stackrel{\text{ind}}{\sim} \sqrt{w_i} N\left(0, \frac{\sigma^2}{w_i}\right) = N(0, \sigma^2)$. In particular, (3) is a linear model with constant variance.

(b)

The OLS estimate $\hat{\beta}$ of β for model (3) is given by

$$\hat{\beta} = ((X^*)^\top X^*)^{-1} (X^*)^\top Y^*,$$

where $Y^* = (Y_1^*, \dots, Y_n^*)^\top$ and $X^* = \begin{bmatrix} (x_1^*)^\top \\ \vdots \\ (x_n^*)^\top \end{bmatrix}$.

Now,

$$\begin{aligned} (X^*)^\top X^* &= [x_1^*, \dots, x_n^*] \begin{bmatrix} (x_1^*)^\top \\ \vdots \\ (x_n^*)^\top \end{bmatrix} \\ &= \sum_{i=1}^n x_i^* (x_i^*)^\top \\ &= \sum_{i=1}^n w_i x_i x_i^\top \\ &= X^\top W X, \end{aligned}$$

where $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ and $W = \text{diag}(w_1, \dots, w_n)$. Similarly, one finds that

$$(X^*)^\top Y^* = X^\top W Y,$$

where $Y = (Y_1, \dots, Y_n)^\top$. We therefore get that

$$\hat{\beta} = (X^\top W X)^{-1} X^\top W Y,$$

as desired.

Additional Exercise 18

(a)

We consider the Inverse Gaussian distribution, which has density given by

$$f(y) = \frac{1}{\sigma \sqrt{2\pi y^3}} \exp \left[-\frac{1}{2y} \left(\frac{y - \mu}{\mu \sigma} \right)^2 \right],$$

for $y > 0, \mu > 0$ and $\sigma^2 > 0$.

Remark. The Inverse Gaussian distribution is *not* the distribution of $\frac{1}{Z}$ when $Z \sim N(0, 1)$. The name is somewhat misleading.

To show that the Inverse Gaussian distribution is contained in the Exponential Dispersion Family, we see that

$$\begin{aligned} f(y) &= \frac{1}{\sigma \sqrt{2\pi y^3}} \exp \left[-\frac{1}{2y} \left(\frac{y - \mu}{\mu \sigma} \right)^2 \right] \\ &= \exp \left[-\frac{1}{2} \left(\frac{y}{\mu^2 \sigma^2} - \frac{2}{\mu \sigma^2} + \frac{1}{y \sigma^2} \right) - \log \left(\sigma \sqrt{2\pi y^3} \right) \right] \\ &= \exp \left[-\frac{1}{2} \left(\frac{y}{\mu^2 \sigma^2} - \frac{2}{\mu \sigma^2} \right) - \frac{1}{2y \sigma^2} - \log \left(\sigma \sqrt{2\pi y^3} \right) \right] \\ &= \exp \left[\frac{-\frac{y}{\mu^2} + \frac{2}{\mu}}{2\sigma^2} - \frac{1}{2y \sigma^2} - \log \left(\sigma \sqrt{2\pi y^3} \right) \right] \\ &= \exp \left[\frac{\theta y - b(\theta)}{a(\varphi)} + c(y, \varphi) \right], \end{aligned}$$

where $\theta = -\frac{1}{\mu^2}$, $b(\theta) = -2\sqrt{-\theta}$, $\varphi = \sigma^2$, $a(\varphi) = 2\varphi = 2\sigma^2$, $c(y, \varphi) = -\frac{1}{2y\varphi} - \log \left(\sqrt{2\pi y^3 \varphi} \right) = -\frac{1}{2y\sigma^2} - \log \left(\sigma \sqrt{2\pi y^3} \right)$.

Hence, $f(y)$ is within the exponential dispersion family.

(b)

Recall the general formulas for the expectation and variance of a random variable Y following a distribution in the exponential dispersion family:

$$\begin{aligned}\mathbb{E}(Y) &= b'(\theta), \text{ and,} \\ \text{Var}(Y) &= b''(\theta)a(\varphi).\end{aligned}$$

In the case of the Inverse Gaussian distribution, we have that

$$\begin{aligned}b'(\theta) &= \frac{1}{\sqrt{-\theta}}, \\ b''(\theta) &= \frac{1}{2}(-\theta)^{-3/2}, \text{ and,} \\ a(\varphi) &= 2\varphi.\end{aligned}$$

We therefore get that

$$\begin{aligned}\mathbb{E}[Y] &= b'(\theta) = \frac{1}{\sqrt{-\theta}} = \mu, \text{ and,} \\ \text{Var}(Y) &= b''(\theta)a(\varphi) = \frac{1}{2}(-\theta)^{-3/2}2\varphi = \mu^3\sigma^2.\end{aligned}$$

(c)

Recall that, given a distribution in the exponential dispersion family, the canonical link function is the function g satisfying $g(\mathbb{E}[Y]) = \theta$. Since $\mathbb{E}(Y) = b'(\theta)$, the canonical link is therefore given by $g(\cdot) = (b')^{-1}(\cdot)$.

Now, we already know that $b'(\theta) = \frac{1}{\sqrt{-\theta}}$. Solving for θ in $y = \frac{1}{\sqrt{-\theta}}$, we find that the inverse function of b' is given by $(b')^{-1}(\theta) = -\frac{1}{\theta^2}$. Hence the canonical link g is given by

$$g(x) = -\frac{1}{x^2}.$$

(d)

Consider the GLM

$$Y_i \stackrel{\text{ind}}{\sim} \text{InverseGaussian}(\mu_i, \sigma^2),$$

where $g(\mu_i) = x_i^\top \beta$ for some $\beta \in \mathbb{R}^p$, some link function g , for $i \in [n]$. Suppose further that σ is known.

The log-likelihood function parameterized by the vector of means $\mu^\top = (\mu_1, \dots, \mu_n)$ is then given by

$$\ell(\mu) = -\sum_{i=1}^n \log(\sqrt{2\pi Y_i^3 \sigma}) - \sum_{i=1}^n \frac{1}{2Y_i} \left(\frac{Y_i - \mu_i}{\mu_i \sigma} \right)^2.$$

Let $\hat{\beta}$ be the MLE of β , and, as usual, let the fitted values be given by $\hat{\mu}_i = g^{-1}(x_i^\top \hat{\beta})$. Then, the deviance of the fitted model is given by

$$D(Y; \hat{\mu}) = -2\varphi \{ \ell(\hat{\mu}) - \ell(Y) \},$$

where $\widehat{\mu}^\top = (\widehat{\mu}_1, \dots, \widehat{\mu}_n)$ and $Y^\top = (Y_1, \dots, Y_n)$. Remember that the log-likelihood is parameterized by the vector of means. Inserting the log-likelihood we found above, and recalling that $\varphi = \sigma^2$, we obtain

$$\begin{aligned} D(Y; \widehat{\mu}) &= 2\sigma^2 \left(-\sum_{i=1}^n \log(\sqrt{2\pi Y_i^3 \sigma}) - \sum_{i=1}^n \frac{1}{2Y_i} \left(\frac{Y_i - \mu_i}{\mu_i \sigma} \right)^2 \right) + \sum_{i=1}^n \log(\sqrt{2\pi Y_i^3 \sigma}) + \sum_{i=1}^n \frac{1}{2Y_i} \left(\frac{Y_i - Y_i}{\mu_i \sigma} \right)^2 \\ &= \sum_{i=1}^n \frac{1}{Y_i} \left(\frac{Y_i - \widehat{\mu}_i}{\widehat{\mu}_i} \right)^2. \end{aligned}$$

Additional Exercise 20

We consider a model where

$$\begin{aligned} A &\sim \text{Bin}(n_1, \pi(1)), \\ B &\sim \text{Bin}(n_0, \pi(0)), \end{aligned}$$

where n_1 and n_0 are natural numbers, and $\pi(0), \pi(1) \in (0, 1)$ are unknown, and $A \perp\!\!\!\perp B$.

Remark. The indexing in this exercise is slightly confusing. In the exercise text, n_0 and n_1 seem to have switched places.

Remark. The above model is equivalent to a GLM with $n = 2$, $Y_1 = A/n_1$, $Y_2 = B/n_0$, $\text{logit}(\pi(1)) = \beta_0$, and $\text{logit}(\pi(0)) = \beta_0 + \beta_1$ for some $\beta_0, \beta_1 \in \mathbb{R}$.

(a)

The Likelihood of the model is given by

$$L(\pi(0), \pi(1)) = \binom{n_1}{A} \pi(1)^A (1 - \pi(1))^{n_1 - A} \binom{n_0}{B} \pi(0)^B (1 - \pi(0))^{n_0 - B}.$$

Hence, the log-likelihood is given by

$$\ell(\pi(0)) = A \log(\pi(1)) + (n_1 - A) \log(1 - \pi(1)) + B \log(\pi(0)) + (n_0 - B) \log(1 - \pi(0)).$$

To find the MLEs of $\pi(1)$ and $\pi(0)$, we differentiate, set equal to zero and solve. It is not difficult to show that ℓ is a concave function, which implies that we indeed find the maximizers of ℓ . We have that

$$\begin{aligned} \frac{\partial \ell}{\partial \pi(1)} &= \frac{A}{\pi(1)} - \frac{n_1 - A}{1 - \pi(1)}, \\ \frac{\partial \ell}{\partial \pi(0)} &= \frac{B}{\pi(0)} - \frac{n_0 - B}{1 - \pi(0)}. \end{aligned}$$

Solving for $\pi(1)$ and $\pi(0)$, we obtain

$$\begin{aligned} \widehat{\pi(1)} &= \frac{A}{n_1}, \\ \widehat{\pi(0)} &= \frac{B}{n_0}. \end{aligned}$$

A natural estimator for the odds ratio

$$\text{OR} = \frac{\frac{\pi(1)}{1 - \pi(1)}}{\frac{\pi(0)}{1 - \pi(0)}}$$

is to plug in our estimates of $\pi(0)$ and $\pi(1)$:

$$\widehat{\text{OR}} = \frac{\frac{\widehat{\pi(1)}}{1 - \widehat{\pi(1)}}}{\frac{\widehat{\pi(0)}}{1 - \widehat{\pi(0)}}}.$$

In fact, $\widehat{\text{OR}}$ is the MLE for the Odds ratio of illness between the two groups.

(b)

The canonical parameters $\theta(0), \theta(1)$ are simply functions of $\pi(0)$ and $\pi(1)$, respectively. By the invariance principle of MLEs, it follows that $\widehat{\theta}(j) = \log(\widehat{\pi}(j)/(1 - \widehat{\pi}(j)))$ is the MLE of $\theta(j)$. Formally, the invariance principle for MLEs states that $\widehat{\varepsilon}$ is the MLE for some unknown parameter ε , then $h(\widehat{\varepsilon})$ is the MLE of $h(\varepsilon)$ for any function h .

Now, recall from Maximum Likelihood Theory (see solutions from week 38) that

$$\sqrt{n_j}(\widehat{\pi}(j) - \pi(j)) \xrightarrow{d} N(0, J_j^{-1}),$$

where $J_j = -\mathbb{E}\left(\frac{1}{n_j} \frac{\partial^2 \ell}{\partial \pi(j)^2}\right)$. Perhaps easier to read, this means that $\widehat{\pi}(j) \overset{d}{\approx} N(\pi(j), \frac{1}{n_j} J_j^{-1})$.

In order to compute J_j for $j = 0, 1$, we first note that

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \pi(1)^2} &= -\frac{A}{\pi(1)^2} - \frac{n_1 - A}{(1 - \pi(1))^2}, \\ \frac{\partial^2 \ell}{\partial \pi(0)^2} &= -\frac{B}{\pi(0)^2} - \frac{n_0 - B}{(1 - \pi(0))^2}. \end{aligned}$$

Therefore,

$$\begin{aligned} J_1 &= -\mathbb{E}\left(\frac{1}{n_1} \frac{\partial^2 \ell}{\partial \pi(1)^2}\right) \\ &= \frac{1}{n_1} \mathbb{E}\left(\frac{A}{\pi(1)^2} - \frac{n_1 - A}{(1 - \pi(1))^2}\right) \\ &= \frac{1}{\pi(1)} + \frac{1}{1 - \pi(1)}. \end{aligned}$$

Hence, the inverse of J_1 is given by

$$\begin{aligned} J_1^{-1} &= \left(\frac{1}{\pi(1)} + \frac{1}{1 - \pi(1)}\right)^{-1} \\ &= \pi(1)(1 - \pi(1)). \end{aligned}$$

Hence,

$$\sqrt{n_1}(\widehat{\pi}(1) - \pi(1)) \xrightarrow{d} N(0, \pi(1)(1 - \pi(1))),$$

as $n_1 \rightarrow \infty$. Informally, this means that $\widehat{\pi}(1) \overset{d}{\approx} N\left(\pi(1), \frac{\pi(1)(1 - \pi(1))}{n_1}\right)$. Arguing in the same fashion, we get that

$$\sqrt{n_0}(\widehat{\pi}(0) - \pi(0)) \xrightarrow{d} N(0, \pi(0)(1 - \pi(0))),$$

as $n_0 \rightarrow \infty$.⁸

Above we found that $\hat{\pi}(j)$ is approximately normally distributed for $j = 0, 1$. However, what we really want is the approximate distributions of $\hat{\theta}(j)$. To this, we will employ the Delta Method.

Remark. The Delta method is actually a Theorem. We state it in general form:

Theorem 0.1. Suppose $(X_i)_{i=1}^n$ is a sequence of random variables such that

$$\sqrt{n}(X_n - \xi) \xrightarrow{d} N(0, \tau^2), \quad \text{as } n \rightarrow \infty.$$

Let $h : \mathbb{R} \rightarrow \mathbb{R}$ be a function with a continuous derivative. Then,

$$\sqrt{n}(g(X_n) - g(\xi)) \xrightarrow{d} N(0, g'(\xi)^2 \tau^2), \quad \text{as } n \rightarrow \infty.$$

Informally, the Delta method states that if $X_n \stackrel{d}{\approx} N(\xi, \tau^2/n)$ for large n , then $g(X_n) \stackrel{d}{\approx} N(g(\xi), g'(\xi)^2 \tau^2/n)$. Note that a multivariate version of the Delta method also exists.

We are now ready to find the approximate distribution of $\hat{\theta}(j)$. Let $h(x) = \log\left(\frac{x}{1-x}\right)$. Then $\theta(j) = h(\pi(j))$ and $\hat{\theta}(j) = g(\hat{\pi}(j))$. Since $g'(x) = \frac{1}{x(1-x)}$, and since $\hat{\pi}(j) \stackrel{d}{\approx} N\left(\pi(j), \frac{\pi(j)(1-\pi(j))}{n_1}\right)$, we get from the delta method that

$$\hat{\theta}(j) \stackrel{d}{\approx} N\left(\theta(j), \frac{1}{n_j \pi(j)(1-\pi(j))}\right),$$

for large n_j , or more formally:

$$\sqrt{n_j}(\hat{\theta}(j) - \theta(j)) \xrightarrow{d} N\left(0, \frac{1}{\pi(j)(1-\pi(j))}\right),$$

as $n_j \rightarrow \infty$.

Hence, the variance of the approximate distribution of $\hat{\theta}(j)$ is $\frac{1}{n_j \pi(j)(1-\pi(j))}$ for $j = 0, 1$. A natural estimator of this quantity is to insert the estimate of $\pi(j)$. For $j = 1$, this becomes

$$\begin{aligned} \frac{1}{n_1 \hat{\pi}(1)(1-\hat{\pi}(1))} &= \frac{1}{n_1 \frac{A}{n_1} \frac{n_1-A}{n_1}} \\ &= \frac{1}{(A+C) \frac{A}{A+C} \frac{C}{A+C}} \\ &= \frac{1}{A} + \frac{1}{C} \end{aligned}$$

recalling that $C = n_1 - A$. For $j = 0$ we proceed in a similar fashion.

(c)

We shall show that a natural estimator for the variance of $\log(\widehat{\text{OR}})$, denoted SE^2 , is given by

$$\widehat{\text{Var}}(\log(\widehat{\text{OR}})) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}.$$

⁸We could also have argued by using the Central Limit Theorem. Consider $\hat{\pi}(1) = \frac{A}{n_1}$, for instance. As A is a Binomial variable, it can be written as the sum of n_1 i.i.d. Bernoulli($\pi(1)$) variables, denoted by Z_j for $j = 1, \dots, n_1$. The Central Limit Theorem says that $\sqrt{n_1}(\frac{1}{n_1} \sum_{j=1}^{n_1} Z_j - \mathbb{E}(Z_1)) \xrightarrow{d} N(0, \text{Var}(Z_1))$ as $n_1 \rightarrow \infty$. Now $\hat{\pi}(1) = \frac{1}{n_1} \sum_{j=1}^{n_1} Z_j$, $\mathbb{E}(Z_1) = \pi(1)$ and $\text{Var}(Z_1) = \pi(1)(1-\pi(1))$, and so the Central Limit Theorem agrees with the above.

To motivate this estimator, we first notice that $\log(\widehat{\text{OR}}) = \widehat{\theta}(1) - \widehat{\theta}(0)$. Hence, the true variance satisfies

$$\begin{aligned}\text{Var}\left(\log(\widehat{\text{OR}})\right) &= \text{Var}\left(\widehat{\theta}(1) - \widehat{\theta}(0)\right) \\ &= \text{Cov}\left(\widehat{\theta}(1) - \widehat{\theta}(0), \widehat{\theta}(1) - \widehat{\theta}(0)\right) \\ &= \text{Var}(\widehat{\theta}(1)) + \text{Var}(\widehat{\theta}(0)) - 2 \text{Cov}(\widehat{\theta}(1), \widehat{\theta}(0)).\end{aligned}$$

Since $\widehat{\theta}(1)$ is only based on A , and $\widehat{\theta}(0)$ is only based on B , and A and B are independent, we must have that $\text{Cov}(\widehat{\theta}(1), \widehat{\theta}(0)) = 0$. Hence,

$$\text{Var}\left(\log(\widehat{\text{OR}})\right) = \text{Var}(\widehat{\theta}(1)) + \text{Var}(\widehat{\theta}(0)). \quad (4)$$

But we have previously shown that $\widehat{\theta}(j) \stackrel{d}{\approx} N\left(\theta(j), \frac{1}{n_j \pi(j)(1-\pi(j))}\right)$. We can estimate the variance with $\frac{1}{A} + \frac{1}{C}$ for $j = 1$ and with $\frac{1}{B} + \frac{1}{D}$ for $j = 0$. Hence, motivated by (4), we can estimate $\text{Var}\left(\log(\widehat{\text{OR}})\right)$ by

$$\widehat{\text{Var}}\left(\log(\widehat{\text{OR}})\right) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}.$$

Furthermore, since $\widehat{\theta}(j)$ is approximately normal and independent for $j = 0, 1$, it must also be that $\widehat{\theta}(1) - \widehat{\theta}(0)$ is approximately normal. Hence,

$$\log(\widehat{\text{OR}}) = \widehat{\theta}(1) - \widehat{\theta}(0) \stackrel{d}{\approx} N\left(\log(\text{OR}), \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}\right).$$

Hence $\log(\widehat{\text{OR}}) \pm 1.96SE$ is an approximate 95% confidence interval for $\log(\text{OR})$. Since confidence intervals are invariant under monotonic transformations, an approximate 95% confidence interval for $\text{OR} = \exp(\log(\text{OR}))$ is given by

$$\begin{aligned}\exp\left(\log(\widehat{\text{OR}}) \pm 1.96SE\right) \\ = \widehat{\text{OR}} \cdot \exp(\pm 1.96SE).\end{aligned}$$

(d)

The (Pearson) Chi-Squared test in this case refers to the (old and famous) test for the null hypothesis that the probability of success is the same in two or more groups. See [this Wikipedia article](#). The test statistic (which we do not state here) has an asymptotic Chi-Square distribution, hence the name of the test. We load the data and perform the test by using a built-in function:

```
> # Enter the data
> diabetes = as.data.frame(matrix(c(377, 17864, 336, 20099), 2, 2))
> rownames(diabetes) = c("diseased", "healthy")
> colnames(diabetes) = c("male", "female")
> show(diabetes)
      male female
diseased   377   336
healthy 17864 20099
>
```

```
> # Perform chi-square test
> chisq.test(diabetes)
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: diabetes
X-squared = 9.277, df = 1, p-value = 0.00232
```

The p-value is $0.00232 < 0.05$. We conclude that there is significance difference between the occurrence of diabetes between men and women.

(e)

By using the result from c), we obtain the 95% confidence interval of odds ratio for diabetes between men and women.

```
> # Estimated odds ratio
> OR.hat = diabetes[1,1]*diabetes[2,2]/(diabetes[1,2]*diabetes[2,1])
> show(OR.hat)
[1] 1.262402
> # Standard error of log of odds ratio
> std.error = sqrt(sum(1/diabetes))
>
> # 95% confidence interval of odds ratio
> low.CI = OR.hat*exp(-qnorm(0.975)*std.error)
> upp.CI = OR.hat*exp(qnorm(0.975)*std.error)
> c(low.CI, upp.CI)
[1] 1.088277 1.464387
```

Thus, $\widehat{\text{OR}} = 1.2624$ and 95% confidence interval for OR is given by $[1.0883, 1.4644]$.

(f)

(i)

Recall that

$$\text{OR} = \frac{\pi(1)/(1 - \pi(1))}{\pi(0)/(1 - \pi(0))} = \frac{\pi(1) \, 1 - \pi(0)}{\pi(0) \, 1 - \pi(1)}.$$

From the above we see that if $\pi(0) = \pi(1)$, then $\text{OR} = 1$. On the other hand, if $\pi(1) \neq \pi(0)$, say $\pi(0) = c\pi(1)$ for some $c > 0$, then we have that

$$\begin{aligned} \text{OR} &= \frac{\pi(1) \, 1 - \pi(0)}{\pi(0) \, 1 - \pi(1)} \\ &= \frac{\pi(1) \, 1 - c\pi(1)}{c\pi(1) \, 1 - \pi(1)} \\ &= \frac{1 - c\pi(1)}{c - c\pi(1)} \\ &= \frac{1 - c\pi(1)}{(c - 1) + 1 - c\pi(1)} \\ &\neq 0 \text{ if } c \neq 1. \end{aligned}$$

Hence, we obtain that $OR = 1 \iff \pi(0) = \pi(1)$. Hence we can test if $\pi(0) = \pi(1)$ by testing $OR = 1$.

(ii)

We want to test $H_0 : \pi(0) = \pi(1)$. Under H_0 , it must be that $OR = 1$, and it is not the case under H_1 . We can thus equivalently test $H_0 : OR = 1$ versus $H_1 : OR \neq 1$. A valid 5% level test is to reject H_0 if 1 is not contained in the 95% confidence interval for OR . But we computed this in (e), and since

$$1 \notin [1.0883, 1.4644],$$

we reject the null hypothesis at 5% level. We conclude that there is significance difference between the occurrence of diabetes between men and women.

(g)

If we fit logistic regression with **sex** as covariate, we have

$$\log \widehat{OR} = \log \left(\frac{\frac{\hat{\pi}(1)}{1-\hat{\pi}(1)}}{\frac{\hat{\pi}(0)}{1-\hat{\pi}(0)}} \right) = \log \left(\frac{\hat{\pi}(1)}{1-\hat{\pi}(1)} \right) - \log \left(\frac{\hat{\pi}(0)}{1-\hat{\pi}(0)} \right) = \hat{\beta}_1.$$

We can use this relationship to convert the confidence interval of $\hat{\beta}_1$ into the confidence interval of $\widehat{OR}$.

```
> # Modify data such that it's suitable for logistic regression
> diabetes.glm.form = data.frame(
+   y = as.numeric(diabetes["diseased",]),
+   n = as.numeric(colSums(diabetes)),
+   sex = c(1,0))
> diabetes.glm.form[, "sex"] = as.factor(diabetes.glm.form[, "sex"])
> head(diabetes.glm.form)
      y      n sex
1 377 18241    1
2 336 20435    0
>
> # Fit logistic regression
> diabetes.model.1 = glm(cbind(y, n - y) ~ sex,
+   family = binomial(link = "logit"),
+   data = diabetes.glm.form)
> summary(diabetes.model.1)

Call:
glm(formula = cbind(y, n - y) ~ sex, family = binomial(link = "logit"),
    data = diabetes.glm.form)
```

Deviance Residuals:

```
[1] 0 0
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.09131	0.05501	-74.376	< 2e-16 ***
sex1	0.23302	0.07573	3.077	0.00209 **

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 9.4891e+00 on 1 degrees of freedom
Residual deviance: -2.7651e-12 on 0 degrees of freedom
AIC: 19.389

Number of Fisher Scoring iterations: 2

```
>
>
> # Confidence interval based on logistic regression
> beta.1.hat = summary(diabetes.model.1)$coefficients["sex1","Estimate"]
> std.error.beta.1 = summary(diabetes.model.1)$coefficients["sex1","Std. Error"]
> low.CI = exp(beta.1.hat - qnorm(0.975)*std.error.beta.1)
> upp.CI = exp(beta.1.hat + qnorm(0.975)*std.error.beta.1)
> c(low.CI, upp.CI)
[1] 1.088278 1.464387
```

As expected, this matches the confidence interval we obtained from e).