

Exercises in Generalized Linear Models

Week 46

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 9.29

GEE estimation is an analogue to Quasi-Likelihood estimation that allows for correlated data. The reason we are interested in GEE estimation is it is a tool to fit GLMMs. Say we have a GLMM in which the responses Y_i conditional on the U_i 's are *not* Gaussian. Then the marginal distribution of the Y_i 's (when we integrate out the U_i 's) can be very messy. In addition, the elements of Y_i will be correlated in some complicated way (due to random effects U_i), which can make maximum likelihood estimation difficult.

In GEE estimation we assume

1. $\mathbb{E}(Y_{i,j}) = \mu_{i,j} = g^{-1}(x_{i,j}^\top \beta)$ for some g ;
2. $\text{Var}(Y_{i,j}) = \nu(\mu_{i,j})$ for some variance function ν ;
3. Some (preferably simple) correlation matrix $R(\alpha)$ for $(Y_{i,1}, \dots, Y_{i,d})^\top$.

Then we estimate β using the generalized estimating equations given in (9.15), giving the estimator $\hat{\beta}$. Note that specifying $\text{Var}(Y_{i,j})$ and the correlation matrix $\text{Corr}(Y_i) = R(\alpha)$ determines $\text{Cov}(Y_i) = V$. If $\mathbb{E}(Y_{i,j}) = \mu_{i,j}g^{-1}(x_{i,j}^\top \beta)$ is a *correct* assumption, the GEE estimate $\hat{\beta}$ of β will be consistent¹. In fact, the estimate of β only requires the expectation to be correctly specified, but not the correlation.

We can estimate the asymptotic covariance matrix of the GEE estimate $\hat{\beta}$ using the sandwich estimator in (9.16). Then we will get an approximately correct estimate of the asymptotic covariance of $\hat{\beta}$ even when the assumption that $\text{Corr}(Y_i) = R(\alpha)$ is wrong.

Unfortunately, exercise 9.29 is imprecise and difficult to interpret. Therefore we need to include our own assumptions for it to make sense.

Assume that we have some GLMM for $Y_i = (Y_{i,1}, \dots, Y_{i,d})^\top$ with covariates $x_{i,j} \in \mathbb{R}^p$. By integrating out the random effects (the U_i 's), we then get a marginal distribution for each $Y_{i,j}$. Assume that the distribution of $Y_{i,j}$ from the marginalization is in the exponential dispersion family and specified by $\varphi, a(\varphi), \theta, b(\theta)$

¹This means that the estimate $\hat{\beta}$ will tend to the true value in some sense. Formally, consistency means that the estimate converges in probability to the correct value β . Check Wiki for the definition of convergence in probability.

and $c(y; \varphi)$. Assume also that $\mu_{i,j} := \mathbb{E}(Y_{i,j}) = g^{-1}(x_i \beta)$ for some $\beta \in \mathbb{R}^p$. That is, we assume that the GLMM implies a marginal GLM. Then we have specified 1. and 2. in the above list, so we only need to make some assumption on the correlation matrix $R(\alpha)$ of each Y_i . Take $R(\alpha) = I_{d \times d}$. Then, the (j, l) -th element of covariance matrix V_i of Y_i is given by

$$\begin{aligned} V_{i,j,l} &:= \text{Cov}(Y_{i,j}, Y_{i,l}) \\ &= \text{Corr}(Y_{i,j}, Y_{i,l}) \sqrt{\text{Var}(Y_{i,j}) \text{Var}(Y_{i,l})} \\ &= R(\alpha)_{j,l} \sqrt{\text{Var}(Y_{i,j}) \text{Var}(Y_{i,l})}. \end{aligned}$$

Since we have assumed that $R(\alpha) = I_{d \times d}$, we then get that V_i is a diagonal matrix with the j -th diagonal element being

$$\begin{aligned} V_{i,j,j} &= \text{Var}(Y_{i,j}) \\ &= b''(\theta_{i,j}) a(\varphi) \\ &= \frac{\partial \mu_{i,j}}{\partial \theta_{i,j}} a(\varphi), \end{aligned}$$

where $\theta_{i,j}$ is the natural parameter of the exponential dispersion distribution of $Y_{i,j}$. Hence, we have that

$$V_i = a(\varphi) \text{diag} \left(\frac{\partial \mu_{i,1}}{\partial \theta_{i,1}}, \dots, \frac{\partial \mu_{i,d}}{\partial \theta_{i,d}} \right).$$

Therefore, by the inverse function theorem,

$$V_i^{-1} = \frac{1}{a(\varphi)} \text{diag} \left(\frac{\partial \theta_{i,1}}{\partial \mu_{i,1}}, \dots, \frac{\partial \theta_{i,d}}{\partial \mu_{i,d}} \right).$$

As for D_i , we have that

$$D_i := \frac{\partial(\mu_{i,1}, \dots, \mu_{i,d})^\top}{\partial \beta},$$

so that the (l, k) -th element of D_i is

$$\begin{aligned} D_{i,j,k} &= \frac{\partial \mu_{i,j}}{\partial \beta_k} \\ &= \frac{\partial \mu_{i,j}}{\partial \eta_{i,j}} \frac{\partial \eta_{i,j}}{\partial \beta_k} \\ &= \frac{\partial \mu_{i,j}}{\partial \eta_{i,j}} x_{i,j,k}, \end{aligned}$$

where $\eta_{i,j} := x_{i,j}^\top \beta$, as usual. We then have that

$$D_i = \text{diag} \left(\frac{\partial \mu_{i,1}}{\partial \eta_{i,1}}, \dots, \frac{\partial \mu_{i,d}}{\partial \eta_{i,d}} \right) X_i.$$

Let $\mu_i = \mu_i(\beta) = (\mu_{i,1}(\beta), \dots, \mu_{i,d}(\beta))^\top = (g^{-1}(x_{i,j}^\top \beta), \dots, g^{-1}(x_{i,j}^\top \beta))^\top$. The GEE equation (9.15) becomes

$$\begin{aligned} \sum_{i=1}^n D_i^\top V_i^{-1} (Y_i - \mu_i(\beta)) &= \sum_{i=1}^n X_i^\top \text{diag} \left(\frac{\partial \mu_{i,1}}{\partial \eta_{i,1}}, \dots, \frac{\partial \mu_{i,d}}{\partial \eta_{i,d}} \right) \frac{1}{a(\varphi)} \text{diag} \left(\frac{\partial \theta_{i,1}}{\partial \mu_{i,1}}, \dots, \frac{\partial \theta_{i,d}}{\partial \mu_{i,d}} \right) (Y_i - \mu_i(\beta)) \\ &= \frac{1}{a(\varphi)} \sum_{i=1}^n X_i^\top \text{diag} \left(\frac{\partial \theta_{i,1}}{\partial \mu_{i,1}} \frac{\partial \mu_{i,1}}{\partial \eta_{i,1}}, \dots, \frac{\partial \theta_{i,d}}{\partial \mu_{i,d}} \frac{\partial \mu_{i,d}}{\partial \eta_{i,d}} \right) (Y_i - \mu_i(\beta)) \\ &= \frac{1}{a(\varphi)} \sum_{i=1}^n X_i^\top \text{diag} \left(\frac{\partial \theta_{i,1}}{\partial \eta_{i,1}}, \dots, \frac{\partial \theta_{i,d}}{\partial \eta_{i,d}} \right) (Y_i - \mu_i(\beta)) \\ &= \frac{1}{a(\varphi)} \sum_{i=1}^n X_i^\top \Delta_i (Y_i - \mu_i(\beta)), \end{aligned}$$

where Δ_i is as specified in the exercise.

Additional Exercise 27

a)

Suppose $Y \sim \text{Poisson}(\mu)$ for some $\mu > 0$. Then the PMF of Y is given by

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}, \quad \text{for } y \in \mathbb{N} \cup \{0\}.$$

Now,

$$\begin{aligned} f(y) &= \frac{\mu^y e^{-\mu}}{y!} \\ &= \exp(y \log \mu - \mu - \log(y!)) \\ &= \exp \left(\frac{y \log \mu - \mu}{1} - \log(y!) \right) \\ &= \exp \left(\frac{y\theta - b(\theta)}{a(\varphi)} - c(y; \varphi) \right), \end{aligned}$$

where $\theta = \log \mu$, $b(\theta) = e^\theta$, $\varphi = 1$, $a(\varphi) = 1$ and $c(y; \varphi) = -\log(y!)$. Thus the Poisson distribution is in the exponential dispersion family.

From the general formula for distributions in the Exponential Dispersion Family,

$$\mathbb{E}(Y) = b'(\theta),$$

we get that

$$\mathbb{E}(Y) = \exp(\theta),$$

which gives the relationship between the expectation $\mu = \mathbb{E}(Y)$ and the natural parameter θ .

b)

We now consider the model where

$$Y_i \stackrel{\text{ind}}{\sim} \text{Poisson}(\mu_i),$$

where $\log(\mu_i) = \beta_0 + \beta_1 x_i$ and x_i are known univariate covariates and $\beta_0, \beta_1 \in \mathbb{R}$ are unknown coefficients.

This is a particular case of a GLM because (i) the Poisson distribution is in the Exponential Dispersion Family, and (ii) the expression $\log(\mu_i) = \beta_0 + \beta_1 x_i$ relates the mean values to a linear predictor through a link function (the natural logarithm).

c)

The log-likelihood of the Poisson GLM in (c) is given by

$$\begin{aligned}\ell(\beta_0, \beta_1; Y) &= \log \left(\prod_{i=1}^n \frac{\mu_i^{Y_i} e^{-\mu_i}}{Y_i!} \right) \\ &= \sum_{i=1}^n Y_i \log(\mu_i) - \mu_i - \log(Y_i!),\end{aligned}$$

where $\mu_i = \mu_i(\beta_0, \beta_1) = \exp(\beta_0 + \beta_1 x_i)$. Hence, the log-likelihood is given by

$$\ell(\beta_0, \beta_1; Y) = \sum_{i=1}^n Y_i(\beta_0 + \beta_1 x_i) - \exp(\beta_0 + \beta_1 x_i) - \log(Y_i!).$$

The score function $u(\beta_0, \beta_1)$ is the gradient of the log-likelihood, so

$$\begin{aligned}u(\beta_0, \beta_1) &:= \frac{\partial \ell}{\partial(\beta_0, \beta_1)^\top} \\ &= \sum_{i=1}^n \begin{bmatrix} 1 \\ x_i \end{bmatrix} (Y_i - \mu_i).\end{aligned}$$

Remark. When we set the score function to zero we obtain the likelihood equations, which determine the MLE $\hat{\beta}$ of β .

The *observed* information matrix $I(\beta_0, \beta_1)$ is the negative hessian of the log-likelihood function. Thus,

$$\begin{aligned}I(\beta_0, \beta_1) &:= -\frac{\partial^2 \ell}{\partial(\beta_0, \beta_1)^\top \partial(\beta_0, \beta_1)}(\beta_0, \beta_1) \\ &= \sum_{i=1}^n \begin{bmatrix} 1 \\ x_i \end{bmatrix} \begin{bmatrix} 1 & x_i \end{bmatrix} \mu_i(\beta_0, \beta_1) \\ &= \sum_{i=1}^n \exp(\beta_0 + \beta_1 x_i) \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix}.\end{aligned}\tag{1}$$

The *expected* information matrix $J(\beta_0, \beta_1)$ is the expected value of $I(\beta_0, \beta_1)$, where the expectation is taken over the Y_i 's with parameters β_0, β_1 . The observed information matrix $I(\beta_0, \beta_1)$ does not depend on the Y_i 's, so the expected information matrix is also given by (1).

d)

The saturated model is the model in which we drop the assumption that $\mu_i = \exp(\beta_0 + \beta_1 x_i)$ and instead allow each observation i to have its own freely varying mean $\mu_i \in \mathbb{R}^+$. That is, the saturated model assumes that $Y_i \sim \text{Poisson}(\mu_i)$ with $\mu_i \in \mathbb{R}^+$ for all $i \in [n]$.

The log-likelihood of the saturated model is

$$\ell_{\text{saturated}}(\mu_1, \dots, \mu_n) = \sum_{i=1}^n Y_i \log(\mu_i) - \mu_i - \log(Y_i!).$$

The MLE $\tilde{\mu}_i$ for μ_i must solve $\frac{\partial \ell_{\text{saturated}}}{\partial \mu_i}(\mu_1, \dots, \mu_n) = \frac{Y_i}{\mu_i} - 1 = 0$, which implies that $\tilde{\mu}_i = Y_i$ for all $i \in [n]$.

e)

The deviance $D(Y; \hat{\mu})$ of a GLM is defined as

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

where $\ell(\mu)$ is the log-likelihood of the data parameterized by a vector of means $\mu = (\mu_1, \dots, \mu_n)^\top$, the vector $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_n)^\top$ contains the fitted values from the GLM, and $Y = (Y_1, \dots, Y_n)^\top$ is the vector of observed values, and φ is the dispersion parameter of the model distribution.

For the Poisson distribution, we have that $\varphi = 1$, and by the previous exercises, an explicit expression for the deviance is given by

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

Since the fitted values $\hat{\mu}$ are defined through the MLE $\hat{\beta}$ for β , the term $\ell(\hat{\mu}; Y)$ is the maximum log-likelihood under the GLM. By the previous exercise, $\tilde{\mu} = Y$ is the MLE for the vector of means in the saturated model, so the term $\ell(Y; Y)$ is the maximum log-likelihood under the saturated model. Thus, the deviance is $\varphi = 1$ times the Likelihood Ratio test statistic for testing H_0 : the GLM is true versus the alternative hypothesis H_1 : the saturated model is true². Now, the saturated model has $p_1 = n$ parameters, while the GLM has $p_0 = 2$ parameters. By Wilk's Theorem, we thus have that

$$D(Y; \hat{\mu}) \stackrel{d}{\approx} \chi^2_{p_1 - p_0 = n - 2},$$

under H_0 . So we can test H_0 versus H_1 (i.e. the model fit) at level α by comparing the observed value of $D(Y; \hat{\mu})$ to the $(1 - \alpha)$ quantile of the χ^2_{n-2} distribution.

f)

We have done this exercise before.

g)

Although the model fitting is performed using the binary data, exercise f) shows that we can still interpret the coefficients in terms of the Poisson GLM. The requested rate ratios are given below.

- (i) The estimated rate ratio between age category 2 and age category 1 (keeping all other covariates fixed) is given by $\exp(\hat{\beta}_{\text{agecat}=2}) = \exp(-0.217)$.
- (ii) The estimated rate ratio between age category 5 and age category 2 (keeping all other covariates fixed) is given by $\exp(\hat{\beta}_{\text{agecat}=5} - \hat{\beta}_{\text{agecat}=2}) = \exp(-0.51762 - (-0.217))$.
- (iii) The estimated rate ratio between two cars worth 25000 dollars and 5000 dollars, respectively (keeping all other covariates fixed), is given by $\exp(2.5\hat{\beta}_{\text{veh-value}} - 0.5\hat{\beta}_{\text{veh-value}}) = \exp(2 \cdot 0.11656)$.
- (iv) The estimated rate ratio between two drivers where one is in age category 2 and owns a car worth 20 000 dollars, while the other driver belongs to age category 1 and owns a car worth 10 000 dollars, is given by $\exp(\hat{\beta}_{\text{agecat}=2} + 2\hat{\beta}_{\text{veh-val}} - \hat{\beta}_{\text{veh-val}}) = \exp(-0.217 + 0.116)$.

²The GLM is nested in the saturated model, so we can use the Likelihood Ratio test.

h)

(i) We shall find a 95% CI for the rate ratio $\exp(\beta_{\text{agecat}=2})$. Since $\hat{\beta}_{\text{agecat}=2} \stackrel{d}{\approx} N(\beta_{\text{agecat}=2}, \widehat{\text{se}}(\hat{\beta}_{\text{agecat}=2})^2)$, an approximate 95% CI for $\beta_{\text{agecat}=2}$ is given by $\hat{\beta}_{\text{agecat}=2} \pm 1.96 \cdot \widehat{\text{se}}(\hat{\beta}_{\text{agecat}=2})$. Therefore, an approximate 95% CI for the rate ratio $\exp(\beta_{\text{agecat}=2})$ is given by $\exp(\hat{\beta}_{\text{agecat}=2} \pm 1.96 \cdot \widehat{\text{se}}(\hat{\beta}_{\text{agecat}=2}))$.

(ii) We shall find a 95% CI for the rate ratio $\exp(\beta_{\text{agecat}=2} + \beta_{\text{veh-val}})$. Now,

$$\begin{bmatrix} \hat{\beta}_{\text{agecat}=2} \\ \hat{\beta}_{\text{veh-value}} \end{bmatrix} \stackrel{d}{\approx} N \left(\begin{bmatrix} \beta_{\text{agecat}=2} \\ \beta_{\text{veh-val}} \end{bmatrix}, \begin{bmatrix} \widehat{\text{se}}(\hat{\beta}_{\text{agecat}=2})^2 & \widehat{\text{Cov}}(\hat{\beta}_{\text{agecat}=2}, \hat{\beta}_{\text{veh-val}}) \\ \widehat{\text{Cov}}(\hat{\beta}_{\text{agecat}=2}, \hat{\beta}_{\text{veh-val}}) & \widehat{\text{se}}(\hat{\beta}_{\text{veh-val}})^2 \end{bmatrix} \right).$$

Using the properties of the Normal distribution³ we get that

$$\hat{\beta}_{\text{agecat}=2} + \hat{\beta}_{\text{veh-value}} \stackrel{d}{\approx} N \left(\beta_{\text{agecat}=2} + \beta_{\text{veh-val}}, \widehat{\text{se}}(\hat{\beta}_{\text{agecat}=2})^2 + \widehat{\text{se}}(\hat{\beta}_{\text{veh-val}})^2 + 2\widehat{\text{Cov}}(\hat{\beta}_{\text{agecat}=2}, \hat{\beta}_{\text{veh-val}}) \right).$$

Using that $\text{Cov}(A, B) = \text{Corr}(A, B)\sqrt{\text{Var}(A)\text{Var}(B)}$ for any pair (A, B) of random variables, the estimated (asymptotic) variance of $\hat{\beta}_{\text{agecat}=2} + \hat{\beta}_{\text{veh-value}}$ is $\widehat{V} := 0.05846^2 + 0.01847^2 + 2(-0.0259)(0.05846)(0.01847) = 0.0037$. An approximate 95% CI for $\beta_{\text{agecat}=2} + \beta_{\text{veh-value}}$ is thus given by $\hat{\beta}_{\text{agecat}=2} + \hat{\beta}_{\text{veh-value}} \pm 1.96 \cdot \sqrt{\widehat{V}}$. An approximate 95% CI for the rate ratio $\exp(\beta_{\text{agecat}=2} + \beta_{\text{veh-value}})$ is therefore given by

$$\exp \left(\hat{\beta}_{\text{agecat}=2} + \hat{\beta}_{\text{veh-value}} \pm 1.96 \cdot \sqrt{\widehat{V}} \right).$$

³if $Z \sim N(\mu, \Sigma)$, then for any vector x we have that $x^T Z \sim N(x^T \mu, x^T \Sigma x)$.