Exercise 6C: Genalized Linear Models and Extensions
2018 Spring

---

1 Background Knowledge

1.1 Quasi-Likelihood

- The maximum likelihood approach is based on the assumption of a specific model or data generation process (like Poisson process).
- When we have insufficient information about the data for use to specify a model for the data.
- but we can specify some features:
  - it is continuous or discrete,
  - how the mean/median is affected by extern variables,
  - how the variability of the response changes with the average,
  - whether the observations are independent,
  - whether the response distribution is skewed
- With these features we can develop analyses based on approximations to the likelihood.

Suppose

- We have a vector of responses, $\mathbf{Y}$, which are independent with mean μ and covariance matrix $\sigma^2 V(\mu)$
- μ is a function of covariates, x, and some regression parameters β. We can write this into $\mu(\beta)$.
- Typically, σ^2 is unknown, and has to be estimated.
- $V(\mu)$ is made up of known functions:

$$V(\mu) = \text{diag}(V_1(\mu), \ldots, V_n(\mu))$$

- We also assume that $V_i(\mu)$ only depends on μ_i.
- For a single component Y of $\mathbf{Y}$:

$$U = u(\mu|Y) = \frac{Y - \mu}{\sigma^2 V(\mu)} \tag{1}$$

  has several properties similar to **log-likelihood derivative (i.e., the score)**,

$$\begin{array}{lcl} \text{E}(U) & = & 0 \\ \text{Var}(U) & = & 1/\left(\sigma^2 V(\mu)\right) \\ -\text{E}\left(\frac{\partial U}{\partial \mu}\right) & = & 1/\left(\sigma^2 V(\mu)\right) \end{array}$$

- And the statistic:
$$Q(\mu|y) = \int_y^\mu u(y|y)dt = \int_y^\mu \frac{y-t}{\sigma^2 V(t)}dt \tag{2}$$
behaves like a **log-likelihood function**. We refer to this as **log quasi-likelihood**.

- The quasi-likelihood for the complete data:
$$Q(\mu|\mathbf{y}) = \sum Q(\mu_i|y_i)$$

- The **quasi-deviance function** for a single observation is
$$D(y|\mu) = -2\sigma^2 Q(\mu|y) = 2\int_y^\mu \frac{y-t}{V(t)}dt \tag{3}$$

- The total deviance
$$D(\mathbf{y}|\mu) = \sum D(y_i|\mu_i)$$
only depends on μ and $\mathbf{y}$, but not σ^2.

- The complete quasi-likelihood only depends multiplicately on σ^2, so that it does not affect the MLEs of $\mu(\beta)$ and hence β.

Example 1 *(**Quasi-Gaussian Likelihood**) When the variance function, $V(\mu) = 1$, then*
$$U = \frac{Y-\mu}{\sigma^2}$$
so that the quasi-likelihood becomes
$$Q(\mu|y) = \int_y^\mu \frac{y-t}{\sigma^2}dt = -\frac{(Y-\mu)^2}{2}$$
which is the same as the likelihood for a normal distribution.

Example 2 *(**Quasi-Poisson Likelihood**) When the variance function $V(\mu) = \mu$, then the quasi-score function:*
$$U = \frac{Y-\mu}{\mu\sigma^2}$$
so the quasi-likelihood is
$$Q(\mu|y) = \int_y^\mu \frac{y-t}{t\sigma^2} = y\log\mu - \mu$$
which is the same as the likelihood for a Poisson distribution.

1.1.1 Other Quasi-likelihoods

1.2 Quasi-likelihood estimation

The **quasi-score function** is $\partial Q(\mu|\mathbf{y})/\partial\beta$, which is
$$\mathbf{U}(\beta) = \mathbf{D^T V^{-1}}(\mathbf{Y}-\mu)/\sigma^{\mathbf{2}} = 0. \tag{4}$$
where

Table 1: Quasi-Likelihoods

$V(\mu)$	$Q(\mu\|y)$	Distribution	Canonical parameter	Range restritions
1	$-(y-\mu)^2/2$	Normal	μ	-
μ	$y\log\mu-\mu$	Poisson	$\log\mu$	$\mu>0$
μ^2	$-y/\mu-\log\mu$	Gamma	$-1/\mu$	$\mu>0, y\geq 0$
μ^3	$-y/2\mu^2+1/\mu$	Inverse Gaussian	$-1/2\mu^2$	$\mu>0, y\geq 0$
μ^ξ	$\mu_{-\xi}\left(\frac{\mu y}{1-\xi}-\frac{\mu^2}{2-\xi}\right)$	-	$\frac{1}{(1-\xi)\mu^{\xi-1}}$	$\mu>0, \xi\neq 0,1,2$
$\mu(1-\mu)$	$y\log(\mu/(1-\mu))+\log(1-\mu)$	Binomial	$\log(\mu/(1-\mu))$	$0<\mu<1, 0\leq y\leq 1$
$\mu+\mu^2/k$	$y\log\left(\frac{\mu}{k+\mu}\right)+k\log\left(\frac{k}{k+\mu}\right)$	Negative binomial	$\log\left(\frac{k}{k+\mu}\right)$	$\mu>0, y\geq 0$

- $\mathbf{D}\in\mathbb{R}^{n\times p}$ with $d_{ir}=\partial u_i/\partial\beta_r$
- $\mathrm{Cov}(\mathbf{U}(\beta))=-\mathrm{E}(\partial\mathbf{U}(\beta))/\partial\beta$, and is

$$\mathbf{i}_\beta=\mathbf{D^T V^{-1} D}/\sigma^{\mathbf{2}}. \tag{5}$$

  which is similar to the Fisher information for MLE.

- The asymptotic covariacne matrix of $\hat{\beta}$ is

$$\mathrm{Cov}(\hat{\beta})\approx\mathbf{i}_\beta^{-\mathbf{1}}=\sigma^2(\mathbf{D^T V^{-1} D})^{-1} \tag{6}$$

- Estimation of σ^2:

$$\hat{\sigma}^2=\frac{1}{n-p}\sum_i\frac{(Y_i-\mu_i)^2}{V_i(\hat{\mu}_i)}=X^2/(n-p) \tag{7}$$

1.3 Longitudinal Data Analysis

For longitudinal data, we can also use linear models with assumptions of correlations, when the number of observation per person, n_i, is small relative to the number of individuals m.
However, for nonlinear discrete longitundinal data, we need to consider the different approach.
There are three extensions of GLMs for longitudinal data:

- Marginal models
- Random Effects models
- Transition models

1.3.1 Marginal models

- Cross-sectional study
- Assumptions:

  (1) The marginal expectation of the response, $E(Y_{ij})=\mu_{ij}$, depends on the predictors, x_{ij}, by $h(\mu_{ij})=x'_{ij}\beta$, where $h(\cdot)$ is a known link function.

  (2) The marginal variance $Var(Y_{ij})=v(\mu_{ij})$ where $v(\cdot)$ is a known variance function and ϕ is a scale parameter.

(3) The correlation $\text{Corr}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \alpha)$ where $\rho(\cdot)$ is a known function.

- Marginal models are natural analogues for correlated data of GLMs for independent data.

1.4 Generalized Estimating Equations (GEEs)

Generalized estimating equations (GEEs) are a specific family of marginal models, which are used to model correlated data from

- **Longitudinal/repeated measures studies**: for same subjects, same measures, successive times. And the succesive measurements are expected to be correlated.
- Clustered/multi-level studies

1.4.1 Notations

For a set of repeated measurements y_{ij}, where $i = 1, \ldots, N$ for each subject; while $j = 1, \ldots, n_i$ be the times for subject i.
Similarly, for clustered data y_{ij}, where $i = 1, \ldots, N$ denotes clusters; while $j = 1, \ldots, n_i$ denotes measurements within cluster i.
In a normal linear model, for unit i

$$\begin{array}{lcl} E(y_i) & = & \mu_i = X_i\beta \\ y_i & \sim & N(\mu_i, V_i) \end{array}$$

where

- $X_i \in \mathbb{R}^{n \times p}$ is the design matrix;
- $\beta \in \mathbb{R}^p$ is the parameter vector;
- $V_i \in \mathbb{R}^{n_i \times n_i}$ is the variance-covariance matrix (e.g., $V_i = \sigma_i^2 I$ if measurements are independent).

For all units:

$$\begin{array}{lcl} E(\mathbf{y}) & = & \mu = \mathbf{X}\beta \\ \mathbf{y} & \sim & N(\mu, \mathbf{V}) \end{array}$$

where

$$\mu = (\mu_1, \ldots, \mu_N)^T, \mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^T, \mathbf{V} = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_N).$$

Then we need to estimate β and $\mathbf{V}$ in the model.
Similarly, we use the log-likelihood function:

$$l = (\mathbf{y} - \mu)^{\mathbf{T}} \mathbf{V}^{-\mathbf{1}} (\mathbf{y} - \mu)$$

and the score function:

$$U(\beta) = \frac{\partial}{\partial \beta} \mathbf{X}^{\mathbf{T}} \mathbf{V}^{-\mathbf{1}} (\mathbf{y} - \mu)$$

which can be solved using a set of score equations:

$$\mathbf{X}_{\mathbf{i}}^{\mathbf{T}} \mathbf{V}^{-\mathbf{1}} (\mathbf{y}_{\mathbf{i}} - \mathbf{X}_{\mathbf{i}} \beta) = \mathbf{0}$$

1.4.2 Marginal models for generlized linear model

Similarly,

$$E(Y_{ij}) = \mu_{ij}, g(\mu_{ij}) = \eta_{ij} = x_i\beta$$

and the score function becomes

$$U(\beta) = \mathbf{D_i^T V_i^{-1}(y_i - \mu_i)} = \mathbf{0}$$

where $\mathbf{D}_i$ is a matrix of derivatives with elements:

$$\frac{\partial \mu_i}{\partial \beta_k} = \frac{\partial \mu_i}{\partial \eta_k} x_{ik}$$

and $\mathbf{V}_i$ is diagonal with elements $\text{Var}(Y_{ij})$.

1.4.3 GEEs

Since Y_{ij}'s are not necessarily independent, if $\mathbf{R}_i \in \mathbb{R}^{n_i \times n_i}$ is the correlation matrix for cluster i, then the variance-covariance matrix $\mathbf{V}$ can be rewritten as:

$$\mathbf{V}_i = \mathbf{A_i^{1/2} R_i A_i^{1/2}}$$

where $\mathbf{A}_i$ is the diagonal matrix with elements $\text{Var}(Y_{ij})$. And thus

$$U(\beta) = \mathbf{D_i^T V_i^{-1}(y_i - \mu_i)} = 0$$

where $\mathbf{V}_i = \mathbf{A_i^{1/2} R_i A_i^{1/2}}$.

- $\mathbf{D}_i$ is the matrix of derivatives: $\frac{\partial \mu_i}{\partial \beta_k}$;
- $\mathbf{V}_i$ is the "working" variance-covariance matrix of Y_i;
- $\mathbf{A}_i = \text{diag}\{\text{var}(Y_{ik})\}$
- $\mathbf{R}_i$ is the correlation matrix for Y_i
- ϕ is an overdispersion parameter.

1.5 Random Effects Models

The previous model, either linear or generalized linear, can be used to estimate the "fixed" effects, which consist of specific and repeatable categories/variables that are representative of an entire population (e.g., species, gender, age). In a longitudinal study with repeated mesures, and in studies using hierarchical (nested) sampling, it is also possible to estimate effects associated with individuals sampled at random from the population of interest. These are "random" effects which convey information about the degree that individuals in a population differ but not how or why they differ.
Then how to differetiate fixed and random effects?

- Fixed effects can capture informations that are beyond the current analysis (a species of tree)
- Random effects contain only the information that are not beyond the current analysis (a group of tree within an observation)

- Fixed effects influence the *mean* of the response.
- Random effects only infludence the *variance* of the response.

Then why mixed-effects models? Mixed-effects models are particularly useful to deal with potential pseudoreplication and unbalanced designs. Including random effects can also account for variation that can mask patterns if we only consider fixed effects.

1.5.1 General specification of random-effects GLM

(1) Given U_i, the response $Y_{i1}, \ldots, Y_{in_i}$ are mutually independent and

$$f(y_{ij}|U_i) = \exp\left[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right]$$

(2) The conditional moments:

$$\begin{aligned} \mu_{ij} &= E(Y_{ij}|U_i) = b'(\theta_{ij}) \\ v_{ij} &= Var(Y_{ij}|U_i) = b''(\theta_{ij})\phi \end{aligned}$$

which satisfy:

$$\begin{aligned} h(\mu_{ij}) &= x'_{ij}\beta^* + d'_{ij}U_i \\ v_{ij} &= v(\mu_{ij})\phi \end{aligned}$$

where $h(\cdot)$ and $v(\cdot)$ are known link and variance functions, and d_{ij} is a subset of x_{ij}.

(3) The random effects, $U_i, i = 1, \ldots, m$ are mutually independent with a common underlying multivarite distribution, F.

1.5.2 Basic underlying random effects model

- There is natural heterogeneity across individuals in their regression coefficients (which can be represented by a probability distribution)
- Correlation among observations for one person arises from their sharing of unobservable variable, U_i.
- Also known as *latent variable* model

Contrast to the marginal models, the random effects model is especially useful when the objective is to make inference about individuals rather than the population average.

Example 3 *(**Correlation from random effects**)*

$$Y_{ij} = \mu + u_i + \epsilon_{ij}$$

where

- Y_{ij}*: response for unit* i *in repeated* j*;*
- μ*: average value for population.*
- u_i*: random effect with* $u_i \sim N(0, \sigma_u^2)$

- ϵ_{ij}: *error with* $\epsilon_{ij} \sim N(0, \sigma_e^2)$

therefore,

- $E(Y_{ij}) = \mu$
- $Var(Y_{ij}) = \sigma_u^2 + \sigma_e^2$
- *The covariance matrix:*

$$cov(Y_{ij}, Y_{km}) = \begin{cases} \sigma_u^2 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}$$

So $\mathbf{V}_i$ *is exchangeable with elements* $\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$*, which is called* ***intra-class correlation coefficient (ICC)****.*

2 Exercises

1. (20 points) Write down the quasi-score equation for
   (1) Quasi-Poisson distribution
   (2) Quasi-Binomial distribution
   (3) Quasi-Gamma distribution
   (4) Quasi-Inverse Gaussian distribution
   (5) Quasi-Negative binomial distribution.

2. (20 points) Assume that the random variable Y is Poisson distributed with probability mass function

$$P(Y = y|\lambda) = \frac{\lambda^y}{y!} \exp(-\lambda)$$

   (1) Show that the distribution of Y belongs to the exponential distribution family. That is show that the function can be rewritten as the form

$$\exp\{[\theta y - b(\theta)]/a(\phi) + c(y, \phi)\},$$

   and determine $\theta, b(\theta), a(\phi)$ and $c(y, \phi)$.

   Solution: We may rewrite the pmf as

$$P(Y = y|\lambda) = \frac{\lambda^y}{y!} \exp(-\lambda) = \exp\{y \log(\lambda) - \lambda - \log(y!)\}$$

   with $\theta = \log(\lambda), b(\theta) = \lambda = \exp(\theta), a(\phi) = 1$ and $c(y, \phi) = -\log(y!)$.

   Assume that $Y_1, \ldots, Y_n$ are independent with the Poisson distribution, and let $\mu_i = E(Y_i), i = 1, \ldots, n$

   (2) Explain what we mean by a generalized linear model (GLM) for Y_i with link function g, and determine the canonical link function.

Solution: A GLM for $Y_i, i = 1, \ldots, n$ with link function g, is specified by assuming that

- $Y_i \sim Poisson(\mu_i)$ and Y_i's are independent.
- Corresponding to each Y_i we have the linear predictor $\eta_i = \mathbf{x_i}\beta = \sum_{j=1}^{p} x_{ij}\beta_j$
- The mean $\mu_i = E(Y_i)$ is linked with the linear predictor by the link function $g(\mu_i) = \eta_i$. Here the link function g is a strictly increasing, differentiable function.

The canonical link function is when the linear predictor $\eta_i = \theta_i$, and then we have $g(\mu_i) = \theta_i$. and then we have

$$\log(\mu_i) = \log(\lambda_i) = \theta_i$$

so $g(\mu_i) = \log(\mu_i)$ is the canonical link function.

(3) Derive an expression for the log-likelihood function $L(\mu; \mathbf{y})$ where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the observed value of $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and $\mu = (\mu_1, \ldots, \mu_n)^T$.

Solution: The likelihood function is given by

$$L(\mu; \mathbf{y}) = \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} \exp(-\mu_i)$$

hence the log-likelihood function becomes

$$l(\mu; \mathbf{y}) = \log L(\mu; \mathbf{y}) = \sum_{i=1}^{n} \{y_i \log(\mu_i) - \mu_i - \log(y_i!)\}.$$

(4) Explain what we mean by a saturated model and determine the maximum of $L(\mu; \mathbf{y})$ for the saturated model.

Solution: For a saturated model, there are no constraint on the expected values, so there is a separated parameter μ_i for each observation y_i.
The log-likelihood obtains its maximum value when

$$\frac{\partial}{\partial \mu_i} l(\mu; \mathbf{y}) = 0$$

then we have

$$\frac{\partial}{\partial \mu_i} l(\mu; \mathbf{y}) = \frac{y_i}{\mu_i} - 1$$

so the log-likelihood takes its maximum value when $y_i = \mu_i$. Thus the ML estimates for the saturated model are $\tilde{\mu}_i = y_i$, and therefore the maximum value of the log-likelihood becomes

$$l(\mathbf{y}; \mathbf{y}) = \sum_{i=1}^{n} \{y_i \log(y_i) - y_i - \log(y_i!)\}$$

(5) Explain what we mean by the deviance $D(\mathbf{y};\hat{\mu})$ of a Poisson GLM, find an expression for the deviance, and discuss how it may be used.

Solution: For a Poisson GLm we have $a(\phi) = 1$. Then the deviance

$$D(\mathbf{y};\hat{\mu}) = -2\log\left(\frac{\text{max likelihood for actual model}}{\text{max likelihood for saturated model}}\right)$$

The deviance measures how far the log-likelihod of the model is from the maximum value of the log-likelihood. For a Poisson GLM the deviance is given by

$$D(\mathbf{y};\hat{\mu}) = -2\log\left(\frac{\prod_{i=1}^{n}(\hat{\mu}_i^{y_i}/y_i!)\exp(-\hat{\mu}_i)}{\prod_{i=1}^{n}(y_i^{y_i}/y_i!)\exp(-y_i)}\right) = 2\sum_{i=1}^{n}\left\{y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}$$

The deviances may be used for comparing nested models M_0 and M_1 using the likelihood-ratio test:

$$\begin{aligned} G^2(M_0|M_1) &= -2\log\left(\frac{\text{max likelihod for } M_0}{\text{max likelihood for } M_1}\right) \\ &= D(\mathbf{y};\hat{\mu}_{\mathbf{0}}) - D(\mathbf{y};\hat{\mu}_{\mathbf{1}}) \sim \chi^2_{df=p1-p0} \end{aligned}$$

3. (10 points) Mittlbock and Heinzl (2001) compare Poisson and logistic regression models for data in which the event rate is small so that the Poisson distribution provides a reasonable approximation to the Binomial distribution. An example is the number of deaths from coronary heart disease among British doctors.

Table 2: Ten years of deaths from coronary diseases

Age	Smokers		Non-smokers	
group	Deaths	Person-years	Deaths	Person-years
35-44	32	52407	2	18790
45-54	104	43248	12	10673
55-64	206	28612	28	5710
65-74	186	12663	28	2585
75-84	102	5317	31	1462

We can fit the model $Y_i \sim \text{Poisson}(\text{deaths}_i)$ with the following equation:

$$\log(\text{deaths}_i) = \log(\text{personyears}_i) + \beta_1 + \beta_2\text{smoke}_i + \beta_3\text{agecat}_i + \beta_4\text{agesq}_i + \beta_5\text{smkage}_i.$$

An alternative is $Y_i \sim \text{Bin}(\text{personyears}_i, \pi_i)$ with

$$\text{logit}(\pi_i) = \beta_1 + \beta_2\text{smoke}_i + \beta_3\text{agecat}_i + \beta_4\text{agesq}_i + \beta_5\text{smkage}_i.$$

Another version is based on a Bernoulli distribution $Z_j \sim \textit{Bernoulli}(\pi_i)$ for each doctor in group i with

$$Z_j = \begin{cases} 1 & j = 1,\ldots,\text{deaths}_i \\ 0 & j = \text{deaths}_i + 1,\ldots,\text{personyears}_i \end{cases}$$

and

$$\text{logit}(\pi_i) = \beta_1 + \beta_2 \text{smoke}_i + \beta_3 \text{agecat}_i + \beta_4 \text{agesq}_i + \beta_5 \text{smkage}_i.$$

(1) (5 points) Fit all three models. Verify that the β estimates are very similar.

(2) (5 points) Calculate the statistics D, X^2 and pseudo R^2 for all three models. Notice that the pseudo R^2 is much smaller for the Bernoulli model. This is prbably due to the reason that the Poisson and Binomial models are estimating the probability of deaths for each group (which is relatively easy) whereas the Bernoulli model is estimating the probability of deaths for an individual (which is much more difficult).

4. (30 points) Twenty-four patients with stroke are randomized to three different treatments:

- A: new OT intervention;
- B: special stroke unit in the same hospital;
- C: usual care in different hospital.

Each group has 8 patients.

The primary outcome is the meaurement of functional ability - Barthel index, which was measured weekly for 8 weeks.

```
stroke <- read.table("stroke.dat", header=TRUE)
stroke.long <- reshape(stroke, idvar=c("Subject", "Group"),
                  varying=3:10,timevar="Week", direction="long")
rownames(stroke.long) <- NULL
names(stroke.long)[4] <- "Ability"
stroke.long$Group <- factor(stroke.long$Group)
stroke.long$Subject <- factor(stroke.long$Subject)
```

(1) (5 points) **Pooled analysis ignoring correlation within patients**

$$Y_{ijk} = \alpha_j + \beta_j k + e_{ijk}$$

where j denotes for groups, k for times, and i for subjects. Here we use different intercepts and different slopes for groups. Assume that all Y_{ijk} are independent and of the same variance (i.e. ignoring the correlation between observations). Use multiple regression to compare α_j's and β_j's. Here we use the interaction term to model the different slopes between different groups.

(2) (5 points) **Data reduction** Fit a linear model for each patient:

$$Y_{ijk} = \alpha_{ij} + \beta_{ij} k + e_{ijk}$$

This also assume independence and constant variance.

(A) Use simple linear regression to estimate α_{ij} and β_{ij}.

(B) Perform ANOVA using estimates $\hat{\alpha}_{ij}$ as the observations and groups as levels of a factor in order to compare α_j's.

(C) Similarly compare β_j's using $\hat{\beta}_{ij}$ as the observations.

(3) (5 points) **Repeated-measures analysis using various variance-covariance structures** Fit

$$Y_{ijk} = \alpha_j + \beta_j k + e_{ijk}$$

where

- α_j and β_j are the parameters of interest.

Assume normality for e_{ijk} but try various forms for variance-covariance matrix.

For the stroke data, choose the **auto-regression structure** (e.g., AR(1)) as the appropriate model. And then use GEEs to fit the models.

Note: In R, you need to apply the `geepack::geeglm()` function.

(4) (5 points) Mixed/Random-effects model Use the model

$$Y_{ijk} = (\alpha_j + a_{ij}) + (\beta_j + b_{ij})k + e_{ijk}$$

where

- α_j and β_j are fixed effects for groups.
- $a_{ij} \sim N(0, \sigma_a^2), b_{ij} \sim N(0, \sigma_b^2)$ and $e_{ijk} \sim N(0, \sigma_e^2)$ are random effects and all are independent.
- Fit this mixed model and use estimates of fixed effects to compare α_j's and β_j's.

In R, you need to apply the `nlme::lme()` function.

(5) (10 points) Make a conclusion based on the above models and compare the advantages and disadvantage of the above models.

5. (20 points) Assume that $U_i \sim N(0, \sigma^2)$. Given $U_i = u_i$, the binary random variables $Y_{i1}, \ldots, Y_{in_i}$ are independent with

$$P(Y_{ij} = 1|U_i = u_i) = 1 - P(Y_{ij} = 0|U_i = u_i) = \Phi(\beta_0 + \beta_1 x_{ij} + u_i)$$

where $\Phi(\cdot)$ is the CDF for standard normal distribution, and x_{ij}'s are known.

(1) What is this model called? Describe one or more situations where such a model can be useful.

Solution: The model is a **generalized linear mixed model (GLMM)**. More specifically, it is a **probit-normal model for binary data with random intercept**.

The model can be used to study clustered binary data, e.g., the occurrence of a disease in schools of pupils (each school is a cluster). The effect of the random intercept u_i is to bring correlations for the observations within a same cluster.

A marginal model for Y_{ij}'s is given by:

$$P(Y_{ij} = 1) = 1 - P(Y_{ij} = 0) = \Phi(\gamma_0 + \gamma_1 x_{ij})$$

(2) Show how the parameters γ_0 and γ_1 may be expressed in terms of β_0, β_1 and σ^2.

Solution: To study the relation between the GLMM model and the marginal model, we need to derive the marginal proabbility. Let Z be a standard normal random variable that is dependent on U_i, then

$$\begin{aligned} P(Y_{ij} = 1|U_i = u_i) &= \Phi(\beta_0 + \beta_1 x_{ij} + u_i) \\ &= P(Z \le \beta_0 + \beta_1 x_{ij} + u_i) \\ &= P(Z - u_i \le \beta_0 + \beta_1 x_{ij}) \end{aligned}$$

Table 3: Data of stroke recovery

		Week							
Subject	Group	1	2	3	4	5	6	7	8
1	A	45	45	45	45	80	80	80	90
2	A	20	25	25	25	30	35	30	50
3	A	50	50	55	70	70	75	90	90
4	A	25	25	35	40	60	60	70	80
5	A	100	100	100	100	100	100	100	100
6	A	20	20	30	50	50	60	85	95
7	A	30	35	35	40	50	60	75	85
8	A	30	35	45	50	55	65	65	70
9	B	40	55	60	70	80	85	90	90
10	B	65	65	70	70	80	80	80	80
11	B	30	30	40	45	65	85	85	85
12	B	25	35	35	35	40	45	45	45
13	B	45	45	80	80	80	80	80	80
14	B	15	15	10	10	10	20	20	20
15	B	35	35	35	45	45	45	50	50
16	B	40	40	40	55	55	55	60	65
17	C	20	20	30	30	30	30	30	30
18	C	35	35	35	40	40	40	40	40
19	C	35	35	35	40	40	40	45	45
20	C	45	65	65	65	80	85	95	100
21	C	45	65	70	90	90	95	95	100
22	C	25	30	30	35	40	40	40	40
23	C	25	25	30	30	30	30	35	40
24	C	15	35	35	35	40	50	65	65

If we let $f_{U_i}(u_i)$ denote the density of U_i, then

$$\begin{aligned} P(Y_{ij}=1) &= \int_{-\infty}^{\infty} P(Y_{ij}=1|U_i=u_i) f_{U_i}(u_i) du_i \\ &= \int_{-\infty}^{\infty} P(Z-u_i \le \beta_0+\beta_1 x_{ij}) f_{U_i}(u_i) du_i \\ &= \int_{-\infty}^{\infty} P(Z-U_i \le \beta_0+\beta_1 x_{ij}|U_i=u_i) f_{U_i}(u_i) du_i \\ &= P\left(Z-U_i \le \beta_0+\beta_1 x_{ij}\right). \end{aligned}$$

Since $Z \sim N(0,1)$ and $U_i \sim N(0,\sigma^2)$, therefore

$$Z-U_i \sim N(0,1+\sigma^2) \Rightarrow \frac{Z-U_i}{\sqrt{1+\sigma^2}} \sim N(0,1).$$

That is

$$P(Y_{ij}=1) = P\left(\frac{Z-U_i}{\sqrt{1+\sigma^2}} \le \frac{\beta_0+\beta_1 x_{ij}}{\sqrt{1+\sigma^2}}\right) = \Phi\left(\frac{\beta_0+\beta_1 x_{ij}}{\sqrt{1+\sigma^2}}\right)$$

Therefore we can get

$$\gamma_j = \beta_j/\sqrt{1+\sigma^2}, j=0,1$$

(3) Comment on the marginal model and the random-effects models for such a clustered binary data, based on the above questions.

Solution: The regression coeffcient γ_1 for the marginal model is the population-level effect on the probit scale of one unit's increase in the covariate x_i without consideration of clusters; while the regression coefficient β_1 is the effect of one unit's increase of the covariate when considering the cluster with the random intercept.
The regression coefficient γ_1 is closer to 0 than β_1, with the ratio of $\gamma_1/\beta_1 = 1/\sqrt{1+\sigma^2}$. Thus the larger the variance of the random intercept in the GLMM, the more the two regression coefficients will differ.

6. (20 points) As we know, the saturated model has a separated parameter μ_i with no contraint. Let $\hat{\theta}_i$ and $\tilde{\theta}_i$ denote the parameter under model ω and Ω (saturated model), respectively. The likelihood ratio test (LRT) criterion to compare the two models in the exponential family has the form:

$$-2\log\lambda = -2\log\frac{L(\hat{\theta})}{L(\tilde{\theta})} = 2\left[l_\Omega(\tilde{\theta})\right] = 2\sum_{i=1}^{n}\frac{y_i(\tilde{\theta}_i-\hat{\theta}_i)-(b(\tilde{\theta}_i)-b(\hat{\theta}_i))}{a_i(\phi)}$$

assuming that $a_i(\phi) = \phi/w_i (w_i = 1)$ for known prior weight w_i.

In GLM, the deviance for model ω can be defined as

$$D(\omega) = 2a(\phi)\left[l_\Omega(\tilde{\theta}) - l_\omega(\hat{\theta})\right] = 2\sum_{i=1}^{n}\left[y_i(\tilde{\theta}_i-\hat{\theta}_i)-(b(\tilde{\theta})-b(\hat{\theta}))\right].$$

Thus

$$-2\log\lambda = D(\omega)/a(\phi)$$

is called the **scaled deviance**.

In order to compare two nested models:

$$-2\log\lambda = -2\log\left[\frac{L_{\omega_1(\theta_1)}}{L_{\omega_2}(\theta_2)}\right] = 2\{[l_\Omega(\theta)-l_{\omega_1}(\theta_1)]-[l_\Omega(\theta)-l_{\omega_2}(\theta_2)]\} = \frac{D(\omega_1)-D(\omega_2)}{a(\phi)} \overset{n\to\infty}{\sim} \chi^2_{p_2-p_1}$$

Here the scale parameter $a(\phi)$ is either known or estimated using the larger model, ω_2.

Write down the deviance for

(1) Normal distribution.

Solution: For normal distribution, the pdf

$$\begin{aligned} f(y) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= \exp\left[\frac{y\mu-\frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \end{aligned}$$

then we have

$$\theta = \mu, b(\theta) = \frac{1}{2}\theta^2$$

Hence for the saturated model:

$$\tilde{\theta} = y, b(\tilde{\theta}) = \frac{1}{2}y^2$$

and for current model:

$$\hat{\theta} = \hat{\mu}, b(\hat{\theta}) = \frac{1}{2}\hat{\mu}^2$$

hence the deviance become:

$$\begin{aligned} D(\omega) &= 2\left[y(\tilde{\theta}-\hat{\theta}) - (b(\tilde{\theta}) - b(\hat{\theta}))\right] \\ &= 2\left[y(y-\hat{\mu}) - (\tfrac{1}{2}y^2 - \tfrac{1}{2}\hat{\mu}^2)\right] \\ &= 2\left(\tfrac{1}{2}y^2 - y\hat{\mu} + \tfrac{1}{2}\hat{\mu}^2\right) \\ &= (y-\hat{\mu})^2 \end{aligned}$$

which is the residual sum of squares (RSS).

(2) Poisson distribution.

Solution: Since for Poisson distribution

$$\theta = \log\mu, b(\theta) = \mu = \exp(\theta)$$

therefore

$$\tilde{\theta} = \log y, b(\tilde{\theta}) = y$$
$$\hat{\theta} = \log\hat{\mu}, b(\hat{\theta}) = \hat{\mu}$$

therefore the deviance

$$D(\omega) = 2\left[y(\tilde{\theta}-\hat{\theta}) - (b(\tilde{\theta}) - b(\hat{\theta}))\right] = 2\left[y\log(y/\hat{\mu}) - (y-\hat{\mu})\right]$$

(3) Binomial distribution.

Solution: For binomial distribution:

$$\theta = \log(\frac{\pi}{1-\pi}), b(\theta) = -n\log(1-\pi)$$

hence

$$\tilde{\theta} = \log(\frac{y/n}{1-y/n}), b(\tilde{\theta}) = -n\log(1-y/n)$$

$$\hat{\theta} = \log(\frac{\hat{\mu/n}}{1-\hat{\mu}/n}), b(\hat{\theta}) = -n\log(1-\hat{\mu}/n)$$

thus the deviance:

$$\begin{aligned} D(\omega) &= 2\left[y(\tilde{\theta}-\hat{\theta}) - (b(\tilde{\theta}) - b(\hat{\theta}))\right] \\ &= 2\left[y\left(\log\tfrac{y/n}{1-y/n} - \log\tfrac{\hat{\mu}/n}{1-\hat{\mu}/n}\right) + n\log(1-y/n) - n\log(1-\hat{\mu}/n)\right] \\ &= 2\left[y\log(y/\hat{\mu}) - y\log\tfrac{n-y}{n-\hat{\mu}} + n\log\tfrac{n-y}{n} - n\log\tfrac{n-\hat{\mu}}{n}\right] \\ &= 2\{y\log(y/\hat{\mu}) + (n-y)\log[(n-y)/(n-\hat{\mu})]\} \end{aligned}$$

where $\hat{\mu} = n\hat{\pi}$.

7. (20 points) The measurement of *left ventricular volume* of the heart is important for studies of cardiac physiology and clinical management of patients with heart disease. An indirect way of measuring the volume, y, involves a measurement called *parallel conductance volume*, x. Boltwood et al. (1989) found an approximately linear association between y and x in a study of dogs under various "load" conditions. The results, reported by Glantz and Slinker (1990), are shown in the following table.

Table 4: Measurements of left ventricular volume and parallel conductance volume

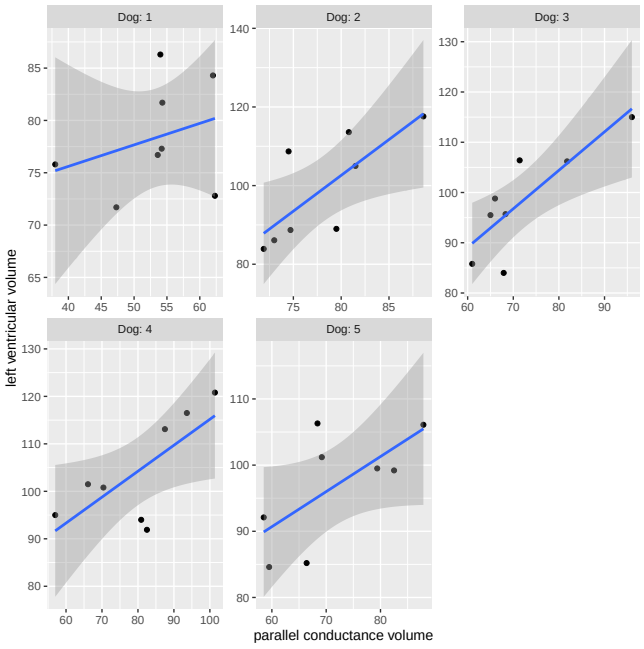
		Condition							
Dog		1	2	3	4	5	6	7	8
1	y	81.70	84.30	72.80	71.70	76.70	75.80	77.30	86.30
	x	54.30	62.00	62.30	47.30	53.60	38.00	54.20	54.00
2	y	105.00	113.60	108.70	83.90	89.00	86.10	88.70	117.60
	x	81.50	80.80	74.50	71.90	79.50	73.00	74.70	88.60
3	y	95.50	95.70	84.00	85.80	98.80	106.20	106.40	115.00
	x	65.00	68.30	67.90	61.00	66.00	81.80	71.40	96.00
4	y	113.10	116.50	100.80	101.50	120.80	95.00	91.90	94.00
	x	87.50	93.60	70.40	66.10	101.40	57.00	82.50	80.90
5	y	99.50	99.20	106.10	85.20	106.30	84.60	92.10	101.20
	x	79.40	82.50	87.90	66.40	68.40	59.50	58.50	69.20

(1) (**EDA**) Conduct a exploratory analysis of these data.

Solution:

- Use `pairs()` to draw pairwise scatter plot;
- Draw scatterplot of y versus x;
- Compute the overall and dog-wise Pearson's correlation coefficient between x and y.

```
# load the data
load("data/ventricular.RData")
library(ggplot2)
f <- ggplot(ventricular, aes(x, y)) + geom_point()
f <- f + geom_smooth(method="lm")
f <- f + facet_wrap(~Dog, scales="free", labeller=label_both)
f + xlab("parallel conductance volume") + ylab("left ventricular volume")
```



(2) (**Pooled analysis**) Let (Y_{jk}, x_{jk}) denote the k-th measurement $(k = 1, \ldots, 8)$ on dog j $(j = 1, \ldots, 5)$. Fit the linear model:

$$E(Y_{jk}) = \mu = \alpha + \beta x_{jk}, Y \sim N(\mu, \sigma^2).$$

Solution: Conduct the general linear regression on the whole data set:

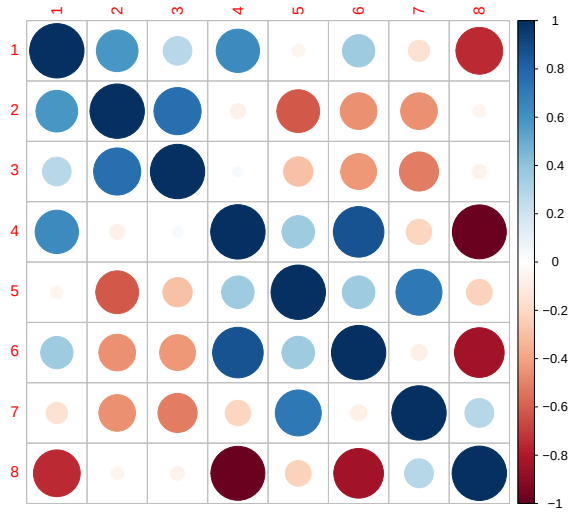
```r
# do regression analysis
m1 <- lm(y ~ x, data=ventricular)
#plot(m1) # diagnosis
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ x, data = ventricular)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8908  -5.3370   0.9632   5.9919  12.9169
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.76808    6.61726   6.161 3.43e-07 ***
## x            0.76923    0.09156   8.401 3.42e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.911 on 38 degrees of freedom
## Multiple R-squared:   0.65,Adjusted R-squared:  0.6408
## F-statistic: 70.58 on 1 and 38 DF,  p-value: 3.416e-10
```

```r
# compute the Pearson's residuals
rs <- resid(m1, type="pearson")
rms <- matrix(rs, ncol=8, byrow=TRUE)

# compute the within-dog residual correlations
rcor <- cor(rms)

# plot the bubble plot of the correlation
library(corrplot)
corrplot(rcor)
```



(3) (**Data reduction analysis**) Fit a linear model based on the data reduction approach.

Solution: Fit a linear regression on each dog separately.

(4) (**Random-effects model**) Fit a suitable random effects model by specifying the fixed effects and random effects.

Solution: use `nlme::lme()` to fit the model:

```
library(nlme)
# only x in the fixed effect
lme1 <- lme(y ~ x, data=ventricular, random=~1|Dog/condition)
# included Dog into fixed effect
lme2 <- lme(y ~ x + Dog, data=ventricular, random=~1|Dog/condition)
# includes condition into the fixed effect
lme3 <- lme(y ~ x + condition, data=ventricular, random=~1|Dog/condition)
# included both into the fixed effect
lme4 <- lme(y ~ x + Dog + condition, data=ventricular, random=~1|Dog/condition)
```

(5) (**GEE**) Fit a clustered model using a GEE.

Solution:

```
library(geepack)
gee1 <- geeglm(y ~ x + condition, data=ventricular,
        family=gaussian, id=Dog, waves=condition,
        corstr="exchangeable", std.err="san.se")
gee2 <- geeglm(y ~ x + Dog, data=ventricular,
        family=gaussian, id=condition, waves=condition,
        corstr="exchangeable", std.err="san.se")
```

(6) Compare the results obtained from each approach. Which method(s) do you prefer? Why?