Project 3

olarr@ntnu.no: 10031 ; tizianom@ntnu.no: 10006

Problem 1

a)

Our goal is to write

$$y_{i,j} = \beta_0 + \beta_1 x_{i,j} + \gamma_{0,i} + \gamma_{1,i} x_{i,j} + \epsilon_{i,j} \qquad (1)$$

where $\gamma_i = (\gamma_{0,i}, \gamma_{1,i})^T$ are iid $N_2(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau_0^2 & \tau_{01}\\ \tau_{01} & \tau_1^2\end{bmatrix})$ and $\epsilon_{i,j}$ are iid $N(0, \sigma^2)$, as

$$y = X\beta + U\gamma + \epsilon$$

Here $i = 1, ..., m$ are the number of clusters, and $j = 1, ..., n_i$ are the number of observations in each cluster.

We start with writing $y_i = X_i\beta + U_i\gamma_i + \epsilon_i$

y_i is a $n_i \times 1$ vector containing the responses.

X_i is a $n_i \times p$ vector containing p covariates including the intercept for every observation in a cluster.

U_i is a $n_i \times (q+1)$ vector containing the random terms for every observation in a cluster.

β is a $p \times 1$ vector containing the regression coefficients for each covariates in X_i.

γ_i is a $(q+1) \times 1$ vector containing the regression coefficients for each random terms in U_i.

ϵ_i is a $n_i \times 1$ vector containing the error terms.

If we now write

$$y = \begin{bmatrix}y_1\\ \vdots \\ y_m\end{bmatrix} \quad X = \begin{bmatrix}X_1\\ \vdots \\ X_m\end{bmatrix} \quad U = \begin{bmatrix}U_1 & \dots & 0\\ \vdots & \ddots & \vdots \\ 0 & \dots & U_m\end{bmatrix} \quad \gamma = \begin{bmatrix}\gamma_1\\ \vdots \\ \gamma_m\end{bmatrix} \quad \epsilon = \begin{bmatrix}\epsilon_1\\ \vdots \\ \epsilon_m\end{bmatrix}$$

then (1) can be written as

$$y = X\beta + U\gamma + \epsilon$$

b)

```
mylmm = function(y, x, group, REML) {
    m <- nlevels(group)
    X = model.matrix(~1 + x)
    # Computing U
    U <- list()  # set up an empty list
    for (i in 1:m) {
        # Construct and assign i'th block to i'th list
        # element
        U[[i]] <- cbind(1, x[group == unique(group)[i]])
    }
    # Change the list to a block diagonal matrix
    U <- bdiag(U)
    # G(theta)
    getG <- function(theta) {
        gg <- matrix(c(theta[1], theta[3] * sqrt(theta[1] * theta[2]),
            theta[3] * sqrt(theta[1] * theta[2]), theta[2]),
            nrow = 2, byrow = T)
        gm <- list()
        for (i in 1:m) {
            gm[[i]] = gg
        }
        return(bdiag(gm))
    }
    # R(theta)
    getR <- function(theta) {
        return(diag(rep(theta[4], length(y))))
    }
    # V(theta)
    getV <- function(theta) {
        return(as.matrix(U %*% getG(theta) %*% t(U) + getR(theta)))
    }
    # beta(theta)
    getBeta <- function(theta) {
        return(solve(t(X) %*% solve(getV(theta)) %*% X) %*% t(X) %*%
            solve(getV(theta)) %*% y)
    }
    # log likelihood, REML if true
    loglik <- function(theta) {
        V <- getV(theta)
        betah <- getBeta(theta)
        l <- -0.5 * (determinant(V)$modulus + t(y - X %*% betah) %*%
            solve(V) %*% (y - X %*% betah))
```

```
        if (REML)
            l <- l - 0.5 * determinant(t(X) %*% solve(V) %*%
                X)$modulus
        l
    }
    opt <- optim(c(1, 1, 1, 1), fn = loglik, method = "L-BFGS-B",
        control = list(fnscale = -1), lower = c(0, 0, -1, 0),
        upper = c(Inf, Inf, 1, Inf))
    theta <- opt$par
    thetam <- matrix(c(theta[1], sqrt(theta[1]), NaN, theta[2],
        sqrt(theta[2]), theta[3]), nrow = 2, ncol = 3, byrow = TRUE,
        dimnames = list(c("(Intercept)", "Days"), c("Variance",
            "Std.Dev.", "Corr.")))
    betas <- getBeta(theta)
    var <- solve(t(X) %*% solve(getV(theta)) %*% X)
    sdev <- sqrt(diag(var))
    corr <- var[1, 2]/(sdev[1] * sdev[2])
    corrm <- matrix(corr, byrow = TRUE, dimnames = list(c("Days"),
        c("(Intr)")))
    fixedeff <- matrix(c(betas[1], sdev[1], betas[2], sdev[2]),
        nrow = 2, ncol = 2, byrow = TRUE, dimnames = list(c("(Intercept)",
            "Days"), c("Estimate", "Std_Error")))
    list(`Random effects, my function:` = thetam, `Fixed Effects, my function:` = fixede
        `Correlation of Fixed Effects, my function:` = corrm)
}
```

c)

```
rsleepmodF = lme4::lmer(Reaction ~ 1 + Days + (1 + Days | Subject),
    REML = F, data = sleepstudy)
sleepmodF <- mylmm(y = sleepstudy$Reaction, x = sleepstudy$Days,
    group = sleepstudy$Subject, REML = F)
list(`Random effects, R function:` = VarCorr(rsleepmodF), `Fixed Effects, R function` =
    1:2], `Correlation of Fixed Effects, R function:` = cov2cor(vcov(rsleepmodF))[1,
    2])
```

```
## $`Random effects, R function:`
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 23.7798
##           Days         5.7168  0.081
##  Residual             25.5919
##
## $`Fixed Effects, R function`
##              Estimate Std. Error
## (Intercept) 251.40510   6.632123
## Days         10.46729   1.502230
##
## $`Correlation of Fixed Effects, R function:`
## [1] -0.1375604
```

```
sleepmodF
```

```
## $`Random effects, my function:`
##              Variance  Std.Dev.      Corr.
## (Intercept) 565.52507 23.780771        NaN
## Days         32.68248  5.716859 0.08131161
##
## $`Fixed Effects, my function:`
##              Estimate Std_Error
## (Intercept) 251.40510  6.632318
## Days         10.46729  1.502242
##
## $`Correlation of Fixed Effects, my function:`
##          (Intr)
## Days -0.1375578
```

```
rsleepmodT = lme4::lmer(Reaction ~ 1 + Days + (1 + Days | Subject),
    REML = T, data = sleepstudy)
```

```
sleepmodT <- mylmm(y = sleepstudy$Reaction, x = sleepstudy$Days,
    group = sleepstudy$Subject, REML = T)
list(`Random effects, R function:` = VarCorr(rsleepmodT), `Fixed Effects, R function` =
    1:2], `Correlation of Fixed Effects, R function:` = cov2cor(vcov(rsleepmodT))[1,
    2])
```

```
## $`Random effects, R function:`
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 24.7407
##           Days         5.9221  0.066
##  Residual             25.5918
##
## $`Fixed Effects, R function`
##              Estimate Std. Error
## (Intercept) 251.40510   6.824597
## Days         10.46729   1.545790
##
## $`Correlation of Fixed Effects, R function:`
## [1] -0.1375519
```

```
sleepmodT
```

```
## $`Random effects, my function:`
##              Variance  Std.Dev.      Corr.
## (Intercept) 612.15836 24.741834        NaN
## Days         35.07508  5.922422 0.06553024
##
## $`Fixed Effects, my function:`
##              Estimate Std_Error
## (Intercept) 251.40510  6.824840
## Days         10.46729  1.545851
##
## $`Correlation of Fixed Effects, my function:`
##          (Intr)
## Days -0.1375541
```

d)

The difference between ML and REML estimates is that REML removes the bias of the ML estimator by integrating over the vector of regression coefficients. In other words, the REML removes all the information from the mean estimator The REML estimator is not generally unbiased

Problem 2

```
data(ohio)
```

a)

```
modglmm <- glmer(resp ~ age + smoke + (1 | id), family = binomial(link = "logit"),
    ohio)
modglm <- glm(resp ~ age + smoke, family = binomial(link = "logit"),
    ohio)
list(`GLMM:` = summary(modglmm)$coefficients, `GLM:` = summary(modglm)$coefficients)
```

```
## $`GLMM:`
##               Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -3.3739539 0.27497502 -12.270038 1.311918e-34
## age         -0.1767645 0.06796698  -2.600741 9.302259e-03
## smoke        0.4147806 0.28704052   1.445024 1.484510e-01
##
## $`GLM:`
##               Estimate Std. Error    z value      Pr(>|z|)
## (Intercept) -1.8837347 0.08384302 -22.467399 8.651082e-112
## age         -0.1134128 0.05408199  -2.097052  3.598895e-02
## smoke        0.2721386 0.12347306   2.204032  2.752210e-02
```

Without "id" as a random intercept, the std. error decreases for every parameter. The age std. error doesn't change as much as the (intercept) and smoke std. errors.

b)

```
cat("Conditional odds:", exp(summary(modglmm)$coefficients[3,
    1]), "\nMarginal odds:", exp((1 + 0.6 * (summary(modglmm)$coefficients[3,
    2])^2) * summary(modglm)$coefficients[3, 1]))
```

```
## Conditional odds: 1.514038
## Marginal odds: 1.330549
```

Here we see that the conditional odds of wheezing changes by a factor of 1.514038, and the marginal odds of wheezing changes by a factor of 1.330549 when looking at the effect of maternal smoking.

c)

```
modglmmage <- glmer(resp ~ age + smoke + (1 + age | id), family = binomial(link = "logit
    ohio)
modglmmsmoke <- glmer(resp ~ age + smoke + (1 + smoke | id),
    family = binomial(link = "logit"), ohio)
lrtage <- head(2 * (logLik(modglmmage) - logLik(modglmm)))
lrtsmoke <- head(2 * (logLik(modglmmsmoke) - logLik(modglmm)))

pvalage <- 0.5 * pchisq(lrtage, 1, lower.tail = F) + 0.5 * pchisq(lrtage,
    2, lower.tail = F)
pvalsmoke <- 0.5 * pchisq(lrtsmoke, 1, lower.tail = F) + 0.5 *
    pchisq(lrtsmoke, 2, lower.tail = F)
cat("p-value of modglmmage:", pvalage, "\np-value of modglmmsmoke:",
    pvalsmoke)
```

```
## p-value of modglmmage: 0.2795205
## p-value of modglmmsmoke: 0.9805645
```

We see that the p-value for the age model is VERY large. Hence there is no clear evidence in the data for variation in the effect of age between different children.

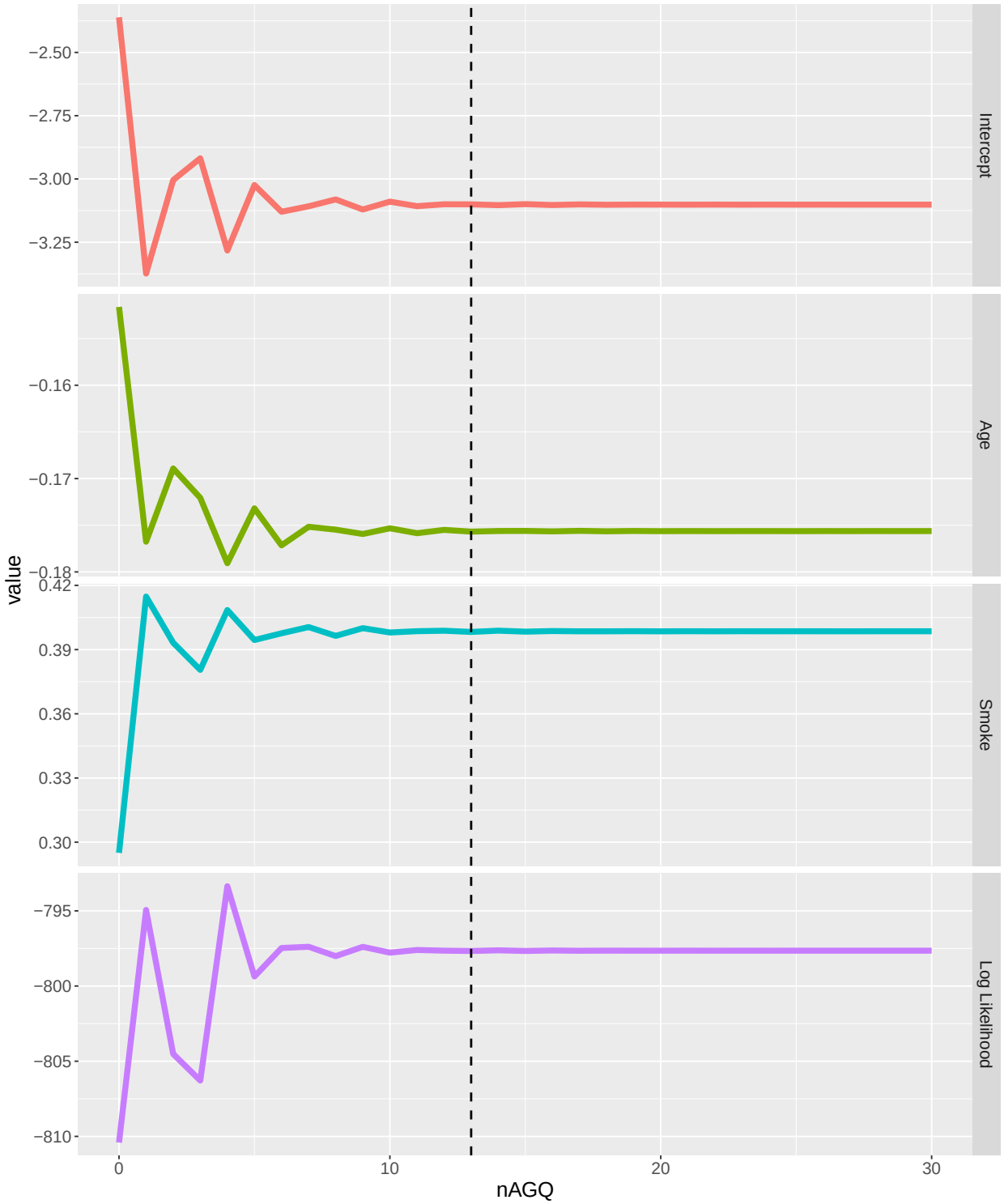
```
alpha <- 0.05
f <- function(x) {
    0.5 * pchisq(x, 1, lower.tail = T) + 0.5 * pchisq(x, 2, lower.tail = T) -
        (1 - alpha)
}
crit <- uniroot(f, c(0, 100))$root
cat("The critical value is", crit, "\nLRT of age model:", lrtage,
    "\nLRT of model:", lrtsmoke)
```

```
## The critical value is 5.138379
## LRT of age model: 1.886229
## LRT of model: 0.00224032
```

These LRT values are less than the critical, so we do NOT reject the null-hypothesis. This means that there is no variation in the effect of age between different children. This coincides with the p-value result.

d)

```
nAGQ <- seq(0, 30)
nAGQ <- append(append(append(nAGQ, nAGQ), nAGQ), nAGQ)
inter <- c()
ag <- c()
smok <- c()
logl <- c()
for (i in 0:30) {
    adap <- glmer(resp ~ age + smoke + (1 | id), family = binomial(link = "logit"),
        ohio, nAGQ = i)
    test <- summary(adap)$coefficients[, 1]
    inter[i + 1] <- test[1]
    ag[i + 1] <- test[2]
    smok[i + 1] <- test[3]
    logl[i + 1] <- head(logLik(adap))
}
value <- rep(0, 31 * 4)
value[1:length(inter)] <- inter
value[(length(inter) + 1):(length(inter) * 2)] <- ag
value[(length(inter) * 2 + 1):(length(inter) * 3)] <- smok
value[(length(inter) * 3 + 1):(length(inter) * 4)] <- logl
type <- rep(0, 31 * 4)
type[1:length(inter)] <- rep("A Intercept", 31)
type[(length(inter) + 1):(length(inter) * 2)] <- rep("B Age",
    31)
type[(length(inter) * 2 + 1):(length(inter) * 3)] <- rep("C Smoke",
    31)
type[(length(inter) * 3 + 1):(length(inter) * 4)] <- rep("D Log Likelihood",
    31)
df <- data.frame(nAGQ, value, type)
ggplot(df, aes(nAGQ, value, color = type)) + geom_line(lwd = 2) +
    facet_grid(type ~ ., scales = "free_y", labeller = labeller(type = c(`A Intercept` =
        `B Age` = "Age", `C Smoke` = "Smoke", `D Log Likelihood` = "Log Likelihood"))) +
    theme(legend.position = "none", text = element_text(size = 15)) +
    geom_vline(xintercept = 13, linetype = "dashed", size = 0.75)
```



Here we see that all the estimates and the log likelihood converges after we use more than 13 quadrature points, so in conclusion, the optimal amount of quadrature points using adaptive Gauss Hermite quadrature is 13 points. The dashed line is 13 quadrature points

e)

First we need to show the conditional and marginal expectations.

Model:

$$\begin{aligned} y_{ij}|\gamma_i &\sim Bernoulli(\pi_{ij}) \\ probit(\pi_{ij}) &= x_{ij}^T\beta + \gamma_i \end{aligned}$$

Conditional expectation:

$$\begin{aligned} E[y_{ij}|\gamma_i] &= \pi_{ij} \\ &= \Phi(x_{ij}^T\beta + \gamma_i) \end{aligned}$$

Marginal expectation:

$$\begin{aligned} E[y_{ij}] &= E[y_{ij}|\gamma_i] \\ &= E[\Phi(x_{ij}^T\beta + \gamma_i)] \\ &= \int \Phi(x_{ij}^T\beta + \gamma_i) f(\gamma_i) d\gamma_i \\ &= \int P(Z \leq x_{ij}^T\beta + \gamma_i|\gamma_i) f(\gamma_i) d\gamma_i \ , \quad Z \sim N(0,1) \\ &\overset{LTP}{=} P(Z \leq x_{ij}^T\beta + \gamma_i) \\ &= P\left(\frac{Z-\gamma_i}{\sqrt{1+\tau^2}} \leq \frac{x_{ij}^T\beta}{\sqrt{1+\tau^2}}\right) \ , \quad \frac{Z-\gamma_i}{\sqrt{1+\tau^2}} \sim N(0,1) \\ &= \Phi\left(\frac{x_{ij}^T\beta}{\sqrt{1+\tau^2}}\right) \end{aligned}$$

So then the covariance becomes:

$$\begin{aligned} Cov[y_{ij}, y_{ik}] &\overset{LTC}{=} Cov\Big[E[y_{ij}\Big|\gamma_i], E[y_{ik}\Big|\gamma_i]\Big] + \overbrace{E[Cov[y_{ij}, y_{ik}|\gamma_i]]}^{\approx 0} \\ &= Cov\Big[\Phi(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \gamma_i), \Phi(\mathbf{x}_{ik}^T\boldsymbol{\beta} + \gamma_i)\Big] \\ &\overset{LTC}{=} E\Big[\Phi(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \gamma_i)\Phi(\mathbf{x}_{ik}^T\boldsymbol{\beta} + \gamma_i)\Big] - E[y_{ij}]E[y_{ik}] \\ &= \underbrace{E\Big[\Phi(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \gamma_i)\Phi(\mathbf{x}_{ik}^T\boldsymbol{\beta} + \gamma_i)\Big]}_{(*)} - \Phi\left(\frac{\mathbf{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right)\Phi\left(\frac{\mathbf{x}_{ik}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right) \end{aligned}$$

Now we want to find the relation between the expression in $(*)$, where $\gamma_i \sim N(0, \tau^2)$, and the cumulative density function $F(x) = P(X \leq x)$ of the bivariate normal distribution.

$$
\begin{aligned}
(*) &\overset{\perp\!\!\!\perp}{=} E[P(Z_1 \le x_{ij}^T\beta + \gamma_i|\gamma_i)P(Z_2 \le x_{ik}^T\beta + \gamma_i|\gamma_i)] \\
&= [Conditional \;\; independence] \\
&= E\Big[P(Z_1 \le x_{ij}^T\beta + \gamma_i \cap Z_2 \le x_{ik}^T\beta + \gamma_i|\gamma_i)\Big] \\
&= \int P(Z_1 \le x_{ij}^T\beta + \gamma_i \cap Z_2 \le x_{ik}^T\beta + \gamma_i|\gamma_i) f(\gamma_i)\, d\gamma_i \\
&\overset{LTP}{=} P(Z_1 \le x_{ij}^T\beta + \gamma_i \cap Z_2 \le x_{ik}^T\beta + \gamma_i) \\
&= P\left(\frac{\mathrm{Z}_1 - \gamma_i}{\sqrt{1+\tau^2}} \le \frac{\mathbf{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}} \cap \frac{\mathrm{Z}_2 - \gamma_i}{\sqrt{1+\tau^2}} \le \frac{\mathbf{x}_{ik}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right) \sim N_2(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \frac{\tau^2}{1+\tau^2}\\ \frac{\tau^2}{1+\tau^2} & 1\end{bmatrix})
\end{aligned}
$$

Because $Var[\frac{\mathrm{Z}_1-\gamma_i}{\sqrt{1+\tau^2}}] = \frac{Var[\mathrm{Z}_1-\gamma_i]}{1+\tau^2} = \frac{1+\tau^2}{1+\tau^2} = 1$, and the same using $j = k$, and because $Cov[y_{ij}, y_{ik}] = \frac{1}{1+\tau^2}Cov[Z_1 - \gamma_i, Z_2 - \gamma_i] = \frac{\tau^2}{1+\tau^2}$

Substituting the value found just above in the previous equation, we get:

$$
Cov[y_{ij}, y_{ik}] = P\left(\frac{\mathrm{Z}_1 - \gamma_i}{\sqrt{1+\tau^2}} \le \frac{\mathbf{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}} \cap \frac{\mathrm{Z}_2 - \gamma_i}{\sqrt{1+\tau^2}} \le \frac{\mathbf{x}_{ik}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right) - \Phi\left(\frac{\mathbf{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right)\Phi\left(\frac{\mathbf{x}_{ik}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right)
$$

Since $y_{ij} \sim Bernoulli(\pi_{ij})$, we know

$$
\begin{aligned}
Var[y_{ij}] &= p(1-p) \\
&= E[y_{ij}](1 - E[y_{ij}]) \\
&= \Phi\left(\frac{\mathbf{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right)(1 - \Phi\left(\frac{\mathbf{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right))
\end{aligned}
$$

Now we can obtain the estimate of the intraclass correlation between wheezing statuses y_{ij}, y_{ik}:

$$
\begin{aligned}
Corr[y_{ij}, y_{ik}] &= \frac{Cov[y_{ij}, y_{ik}]}{\sqrt{Var[y_{ij}] \cdot Var[y_{ik}]}} \\
&= \frac{P\left(\frac{\mathrm{Z}_1-\gamma_i}{\sqrt{1+\tau^2}} \le \frac{\mathbf{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}} \cap \frac{\mathrm{Z}_2-\gamma_i}{\sqrt{1+\tau^2}} \le \frac{\mathbf{x}_{ik}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right) - \Phi\left(\frac{\mathbf{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right)\Phi\left(\frac{\mathbf{x}_{ik}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right)}{\sqrt{\Phi\left(\frac{\mathbf{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right)(1 - \Phi\left(\frac{\mathbf{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right)) \cdot \left[\Phi\left(\frac{\mathbf{x}_{ik}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right)(1 - \Phi\left(\frac{\mathbf{x}_{ik}^T\boldsymbol{\beta}}{\sqrt{1+\tau^2}}\right))\right]}}
\end{aligned}
$$

```r
modglmmprobit <- glmer(resp ~ age + smoke + (1 | id), family = binomial(link = "probit"),
    ohio)
coef <- summary(modglmmprobit)$coefficients
intercept <- coef[1, 1]
age <- coef[2, 1]
smoke <- coef[3, 1]
beta <- c(intercept, age, smoke)
tau2 <- as.numeric(matrix(VarCorr(modglmmprobit)))
xi7 <- c(1, -2, 0)
xi10 <- c(1, 1, 0)
z1 <- as.numeric(t(xi7) %*% beta)/(sqrt(1 + tau2))
z2 <- as.numeric(t(xi10) %*% beta)/(sqrt(1 + tau2))
sigm <- cbind(c(1, tau2/(1 + tau2)), c(tau2/(1 + tau2), 1))
cov <- pmvnorm(lower = c(-Inf, -Inf), upper = c(z1, z2), keepAttr = F,
    sigma = sigm) - pnorm(z1) * pnorm(z2)
var7 <- pnorm(z1) * (1 - pnorm(z1))
var10 <- pnorm(z2) * (1 - pnorm(z2))
corr <- cov/sqrt(var7 * var10)
cat("The intraclass correlation is", corr)
```

```
## The intraclass correlation is 0.6764836
```