Biostat 200C Homework 5

Due June 3 @ 11:59PM

```
        library(faraway)
        library(tidyr)
        library(ggplot2)
        library(lme4)
```

```
Loading required package: Matrix


Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```
        library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
        library(reshape2)
```

```
Attaching package: 'reshape2'

The following object is masked from 'package:tidyr':

    smiths
```

```
        library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```
library(geepack)
```

```
Attaching package: 'geepack'

The following object is masked from 'package:faraway':

    ohio
```

```
library(faraway)
```

Q1. Balanced one-way ANOVA random effects model

Consider the balanced one-way ANOVA random effects model with a levels and n observations in each level

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, n.$$

where α_i are iid from $N(0, \sigma_\alpha^2)$, ϵ_{ij} are iid from $N(0, \sigma_\epsilon^2)$.

1. Derive the ANOVA estimate for μ, σ_α^2, and σ_ϵ^2. Specifically show that

$$\mathbb{E}(\bar{y}_{\cdot\cdot}) = \mathbb{E}\left(\frac{\sum_{ij} y_{ij}}{na}\right) = \mu$$
$$\mathbb{E}(\text{SSE}) = \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2\right] = a(n-1)\sigma_\epsilon^2$$
$$\mathbb{E}(\text{SSA}) = \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2\right] = (a-1)(n\sigma_\alpha^2 + \sigma_\epsilon^2),$$

   which can be solved to obtain ANOVA estimate

$$\widehat{\mu} = \frac{\sum_{ij} y_{ij}}{na},$$
$$\widehat{\sigma}_\epsilon^2 = \frac{\text{SSE}}{a(n-1)},$$
$$\widehat{\sigma}_\alpha^2 = \frac{\text{SSA}/(a-1) - \widehat{\sigma}_\epsilon^2}{n}.$$

Answer:

$E(\bar{y}_{\cdot\cdot}) = E(\frac{\sum_{ij} y_{ij}}{na}) = \frac{\sum_{ij} E(y_{ij})}{na} = \frac{\sum_{ij} \mu}{na} = \frac{na\mu}{na} = \mu$

$\vec{Y} = \mu\vec{\mathbb{1}}_{na} + Z\vec{\alpha} + \vec{\epsilon}$ where $Z = I_a \otimes \mathbb{1}_{n\times n}$ and $\epsilon \sim N(0, \sigma_\epsilon^2 I_{na})$

$Cov(\vec{Y}) = ZZ'\sigma_a^2 + \sigma_\epsilon^2 I_{na}$

define

$$y = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in} \end{bmatrix}$$

$\bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} = \frac{1}{n}\mathbb{1}_{\mathbb{n}}' y_i$

$y_i - \bar{y}_{i.} = y_i - \frac{1}{n}\mathbb{1}_{\mathbb{n}}' y_i = (I_n - \frac{1}{n}\mathbb{1}_{\mathbb{n}}\mathbb{1}_{\mathbb{n}}')y_i$

Then we can rewrite it as:

$SSE = \vec{y}' A_1 \vec{y}$

$$A_1 = \begin{bmatrix} I_n - n^{-1}\mathbb{1}_{\mathbb{n}}\mathbb{1}_{\mathbb{n}}' & & \\ & \ddots & \\ & & I_n - n^{-1}\mathbb{1}_{\mathbb{n}}\mathbb{1}_{\mathbb{n}}' \end{bmatrix}$$

$$\begin{aligned} E(SSE) &= E(\vec{y}' A_1 \vec{y}) \\ &= E(trA_1 \vec{y}\vec{y}') \\ &= trA_1 Cov(\vec{y}) \\ &= trA_1(ZZ'\sigma_a^2 + \sigma_\epsilon^2 I_{na}) + \mu^2 trA_1 \mathbb{1}_{\mathbb{na}}\mathbb{1}_{\mathbb{na}}' \\ &= 0 + a(n-1)\sigma_\epsilon^2 + 0 \\ &= a(n-1)\sigma_\epsilon^2 \end{aligned}$$

$SSA = \vec{y}' A_0 \vec{y}$

where $A_0 = I_{na} - \frac{1}{an}\mathbb{1}_{\mathbb{na}}\mathbb{1}_{\mathbb{na}}'$

$$\begin{aligned} E(SST) &= trA_0(\sigma_a^2 ZZ' + \sigma_\epsilon^2 I_{na}) + \mu^2 trA_0 \mathbb{1}_{\mathbb{na}}\mathbb{1}_{\mathbb{na}}' \\ &= n(a-1)\sigma_a^2 + \sigma_\epsilon^2(na-1) + 0 \\ &= n(a-1)\sigma_a^2 + \sigma_\epsilon^2(na-1) \end{aligned}$$

$$\mathbb{E}(\text{SSA}) = \mathbb{E}(\text{SST}) - \mathbb{E}(\text{SSE}) = (a-1)(n\sigma_\alpha^2 + \sigma_\epsilon^2).$$

2. Derive the MLE estimate for μ, σ_α^2, and σ_ϵ^2. Hint: write down the log-likelihood and find the maximizer.

The log-likelihodd is

$$\begin{aligned} \ell(\mu, \sigma_\alpha^2, \sigma_\epsilon^2) = & -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det(\sigma_\alpha^2 \mathbf{Z}\mathbf{Z}^T + \sigma_\epsilon^2 \mathbf{I}) - \frac{1}{2}(\mathbf{y} - \mathbf{1}_{na}\mu)^T(\sigma_\alpha^2 \mathbf{Z}\mathbf{Z}^T + \sigma_\epsilon^2 \mathbf{I})^{-1}(\mathbf{y} - \mathbf{1}_{na}\mu) \\ = & \sum_i -\frac{1}{2}\log\det(\sigma_\alpha^2 \mathbf{1}_n \mathbf{1}_n^T + \sigma_\epsilon^2 \mathbf{I}_n) - \frac{1}{2}(\mathbf{y}_i - \mathbf{1}_n\mu)^T(\sigma_\alpha^2 \mathbf{1}_n \mathbf{1}_n^T + \sigma_\epsilon^2 \mathbf{I}_n)^{-1}(\mathbf{y}_i - \mathbf{1}_n\mu). \end{aligned}$$

By Woodbury formula

$$(\sigma_\alpha^2 \mathbf{1}_n \mathbf{1}_n^T + \sigma_\epsilon^2 \mathbf{I}_n)^{-1} = \sigma_\epsilon^{-2} \mathbf{I}_n - \frac{\sigma_\epsilon^{-2}\sigma_\alpha^2}{\sigma_\epsilon^2 + n\sigma_\alpha^2} \mathbf{1}_n \mathbf{1}_n^T$$
$$\det(\sigma_\alpha^2 \mathbf{1}_n \mathbf{1}_n^T + \sigma_\epsilon^2 \mathbf{I}_n) = \sigma_\epsilon^{2n}(1 + n\sigma_\alpha^2/\sigma_\epsilon^2).$$

Let $\lambda = \sigma_\alpha^2/\sigma_\epsilon^2$, then the log-likelihood is

$$\begin{aligned} \ell(\mu, \sigma_\alpha^2, \sigma_\epsilon^2) =& -\frac{na}{2}\log \sigma_\epsilon^2 - \frac{a}{2}\log(1+n\lambda) - \frac{\sigma_\epsilon^{-2}}{2}\text{SST}(\mu) + \frac{\sigma_\epsilon^{-2}}{2}\frac{n\lambda}{1+n\lambda}\text{SSA}(\mu) \\ =& -\frac{na}{2}\log \sigma_\epsilon^2 - \frac{a}{2}\log(1+n\lambda) - \frac{\sigma_\epsilon^{-2}}{2}\frac{\text{SST}(\mu) + n\lambda\text{SSA}}{1+n\lambda}. \end{aligned}$$

Setting derivative with respect to μ to 0 yields

$$\hat{\mu} = \bar{y}_{\cdot\cdot}.$$

Setting derivative with respect to σ_ϵ^2 to 0 yields equation

$$\sigma_\epsilon^2 = \frac{\text{SST} - \frac{n\lambda}{1+n\lambda}\text{SSA}}{na} = \frac{\text{SST} + n\lambda\text{SSE}}{na(1+n\lambda)}.$$

Substitution of the above expression into the log-likelihood shows we need to maximize

$$\begin{aligned} & -\frac{na}{2}\log\left(\text{SST} - \frac{n\lambda}{1+n\lambda}\text{SSA}\right) - \frac{a}{2}\log(1+n\lambda) \\ =& -\frac{na}{2}\log\left(\text{SST} + n\lambda\text{SSE}\right) + \frac{(n-1)a}{2}\log(1+n\lambda). \end{aligned}$$

Setting derivative to 0 gives the maximizer

$$\hat{\lambda} = \frac{n-1}{n}\frac{\text{SST}}{\text{SSE}} - 1.$$

Thus

$$\hat{\sigma}_\epsilon^2 = \frac{\text{SST} - \frac{n\hat{\lambda}}{1+n\hat{\lambda}}\text{SSA}}{na} = \frac{\text{SSE}}{(n-1)a}$$

(same as ANOVA estimate) and

$$\hat{\sigma}_\alpha^2 = \frac{\text{SSA}}{an} - \frac{\text{SSE}}{an(n-1)}.$$

$$L(\mu, \sigma_\alpha^2, \sigma_\epsilon^2) = \prod_{i=1}^a \prod_{j=1}^n \frac{1}{\sqrt{2\pi(\sigma_\alpha^2 + \sigma_\epsilon^2)}} \exp\left(-\frac{(y_{ij} - \mu)^2}{2(\sigma_\alpha^2 + \sigma_\epsilon^2)}\right)$$

$$\ell(\mu, \sigma_\alpha^2, \sigma_\epsilon^2) = -\frac{an}{2}\log(2\pi) - \frac{an}{2}\log(\sigma_\alpha^2 + \sigma_\epsilon^2) - \frac{1}{2(\sigma_\alpha^2 + \sigma_\epsilon^2)} \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \mu)^2$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma_\alpha^2 + \sigma_\epsilon^2} \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \mu)$$

Setting this to zero:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \mu) = 0$$

$$\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{n} \mu$$

$$\mu = \frac{1}{an} \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}$$

So the MLE for μ is:

$$\hat{\mu} = \frac{1}{an} \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}$$

Similarly, we can derive the MLE for σ_α^2 and σ_ϵ^2 with the same procedures:

$$\hat{\sigma}_\epsilon^2 = \frac{\text{SSE}}{a(n-1)}$$

$$\hat{\sigma}_\alpha^2 = \frac{\text{SSA}/(a-1) - \hat{\sigma}_\epsilon^2}{n}$$

Where:

$$\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2$$

$$\text{SSA} = \sum_{i=1}^{a} n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

3. (**Optional**) Derive the REML estimate for μ, σ_α^2, and σ_ϵ^2.

4. For all three estimates, check that your results match those we obtained using R for the `pulp` example in class.

Answer:

```
data(pulp)
```

```
mean(pulp$bright)
```

```
[1] 60.4
```

```
(aovmod <- aov(bright ~ operator, data = pulp) %>%
  summary())
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
operator     3   1.34  0.4467   4.204 0.0226 *
Residuals   16   1.70  0.1062
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(aovmod[1][[1]][[3]][1] - aovmod[1][[1]][[3]][2]) / 5
```

```
[1] 0.06808333
```

```
aovmod[1][[1]][[3]][2]
```

```
[1] 0.10625
```

```
overall_mean <- mean(pulp$bright)
factor_means <- pulp %>%
  group_by(operator) %>%
  summarise(mean_bright = mean(bright))

SSE <- pulp %>%
  group_by(operator) %>%
  summarise(sum_sq = sum((bright - mean(bright))^2)) %>%
  summarise(SSE = sum(sum_sq)) %>%
  pull(SSE)

SSA <- factor_means %>%
  summarise(SSA = sum(n() * (mean_bright - overall_mean)^2)) %>%
  pull(SSA)

a <- nlevels(pulp$operator)
n <- nrow(pulp) / a
sigma2_epsilon <- SSE / (a * (n - 1))

sigma2_alpha <- (SSA / (a - 1) - sigma2_epsilon) / n

# Print results
cat("SSE:", SSE, "\n")
```

```
SSE: 1.7
```

```
cat("SSA:", SSA, "\n")
```

```
SSA: 1.072
```

```
        cat("sigma^2_epsilon:", sigma2_epsilon, "\n")
```

```
sigma^2_epsilon: 0.10625
```

```
        cat("sigma^2_alpha:", sigma2_alpha, "\n")
```

```
sigma^2_alpha: 0.05021667
```

σ_ϵ^2 does not match.

Q2. Estimation of random effects

1. Assume the conditional distribution

$$\mathbf{y} \mid \boldsymbol{\gamma} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2\mathbf{I}_n)$$

and the prior distribution

$$\boldsymbol{\gamma} \sim N(\mathbf{0}_q, \boldsymbol{\Sigma}).$$

Then by the Bayes theorem, the posterior distribution is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \boldsymbol{\gamma}) \times f(\boldsymbol{\gamma})}{f(\mathbf{y})},$$

where f denotes corresponding density. Show that the posterior distribution is a multivariate normal with mean

$$\mathbb{E}(\boldsymbol{\gamma} \mid \mathbf{y}) = \boldsymbol{\Sigma}\mathbf{Z}^T(\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Answer:

Assume the conditional distribution

$$y \mid \gamma \sim N(X\beta + Z\gamma, \sigma^2 I_n)$$

and the prior distribution

$$\gamma \sim N(0, \Sigma).$$

Then by Bayes' theorem, the posterior distribution is

$$f(\gamma \mid y) = \frac{f(y \mid \gamma) \times f(\gamma)}{f(y)},$$

where f denotes the corresponding density.

The likelihood function is

$$f(y \mid \gamma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta - Z\gamma)^\top (y - X\beta - Z\gamma)\right).$$

The prior density is

$$f(\gamma) = \frac{1}{(2\pi)^{q/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\gamma^\top \Sigma^{-1}\gamma\right).$$

Therefore, the posterior distribution is proportional to the product of these densities:

$$f(\gamma \mid y) \propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta - Z\gamma)^\top (y - X\beta - Z\gamma)\right) \times \exp\left(-\frac{1}{2}\gamma^\top \Sigma^{-1}\gamma\right).$$

Combining the exponents, we get

$$f(\gamma \mid y) \propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}(y - X\beta - Z\gamma)^\top (y - X\beta - Z\gamma) + \gamma^\top \Sigma^{-1}\gamma\right]\right).$$

Expanding the term inside the exponent:

$$\frac{1}{\sigma^2}(y - X\beta - Z\gamma)^\top (y - X\beta - Z\gamma) = \frac{1}{\sigma^2}\left[(y - X\beta)^\top (y - X\beta) - 2(y - X\beta)^\top Z\gamma + \gamma^\top Z^\top Z\gamma\right],$$

so the posterior can be rewritten as:

$$f(\gamma \mid y) \propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}(y - X\beta)^\top (y - X\beta) - \frac{2}{\sigma^2}(y - X\beta)^\top Z\gamma + \left(\frac{1}{\sigma^2}Z^\top Z + \Sigma^{-1}\right)\gamma^\top \gamma\right]\right).$$

To simplify, complete the square for γ:

$$\left(\frac{1}{\sigma^2}Z^\top Z + \Sigma^{-1}\right)\gamma^\top \gamma - \frac{2}{\sigma^2}(y - X\beta)^\top Z\gamma = \left(\gamma - \hat{\gamma}\right)^\top \left(\frac{1}{\sigma^2}Z^\top Z + \Sigma^{-1}\right)\left(\gamma - \hat{\gamma}\right) + \text{constant},$$

where

$$\hat{\gamma} = \left(\frac{1}{\sigma^2}Z^\top Z + \Sigma^{-1}\right)^{-1} \frac{1}{\sigma^2}Z^\top (y - X\beta).$$

Thus, the posterior distribution is

$$\gamma \mid y \sim N\left(\hat{\gamma}, \left(\frac{1}{\sigma^2}Z^\top Z + \Sigma^{-1}\right)^{-1}\right).$$

The mean of the posterior distribution is

$$E(\gamma \mid y) = \hat{\gamma} = \Sigma Z^\top \left(Z\Sigma Z^\top + \sigma^2 I\right)^{-1}(y - X\beta).$$

2. For the balanced one-way ANOVA random effects model, show that the posterior mean of random effects is always a constant (less than 1) multiplying the corresponding fixed effects estimate.

Answer:

From the previous derivation, we know the posterior distribution of α_i given the data (y) is:

$$\alpha_i \mid y \sim N\left(\hat{\alpha}_i, \left(\frac{1}{\sigma^2} Z^\top Z + \Sigma^{-1}\right)^{-1}\right),$$

where $\hat{\alpha}_i$ is the posterior mean.

For the balanced one-way ANOVA random effects model, the posterior mean of α_i is:

$$E(\alpha_i \mid y) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{n}}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}),$$

where $\bar{y}_{i\cdot}$ is the mean of the observations in the (i)-th group, and $\bar{y}_{\cdot\cdot}$ is the overall mean of all observations.

The fixed effects estimate for α_i is simply $\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$. Thus, the posterior mean of the random effects is:

$$E(\alpha_i \mid y) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{n}}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}).$$

Let $\lambda = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{n}}$

Clearly, λ is a constant that is less than 1 because $\sigma_\alpha^2 > 0$ and $\sigma_\epsilon^2 > 0$

Therefore,

$$E(\alpha_i \mid y) = \lambda(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}),$$

where λ is the shrinkage factor which is always less than 1.

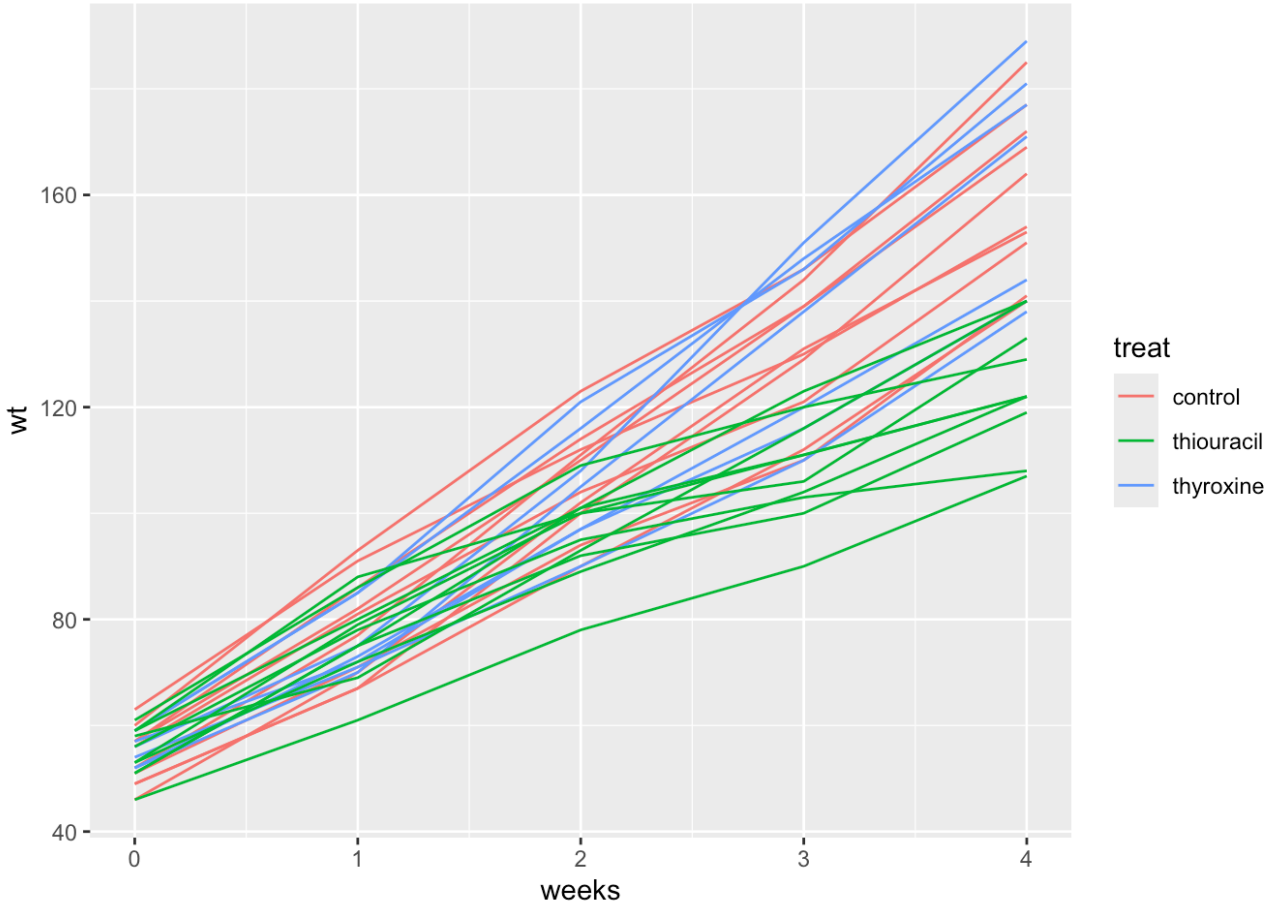
Q3. ELMR Exercise 11.1 (p251)

1. The `ratdrink` data consist of five weekly measurements of body weight for 27 rats. The first 10 rats are on a control treatment while 7 rats have thyroxine added to their drinking water. Ten rats have thiouracil added to their water.
   (a) Plot the data showing how weight increases with age on a single panel, taking care to distinguish the three treatment groups. Now create a three-panel plot, one for each group. Discuss what can be seen.
   (b) Fit a linear longitudinal model that allows for a random slope and intercept for each rat. Each group should have a different mean line. Give interpretation for the following estimates:
      i. The fixed effect intercept term.
      ii. The interaction between thiouracil and week.
      iii. The intercept random effect SD.
   (c) Check whether there is a significant treatment effect.
   (d) Construct diagnostic plots showing the residuals against the fitted values and a QQ plot of the residuals. Interpret.
   (e) Construct confidence intervals for the parameters of the model. Which random effect terms may not be significant? Is the thyroxine group significantly different from the control group?

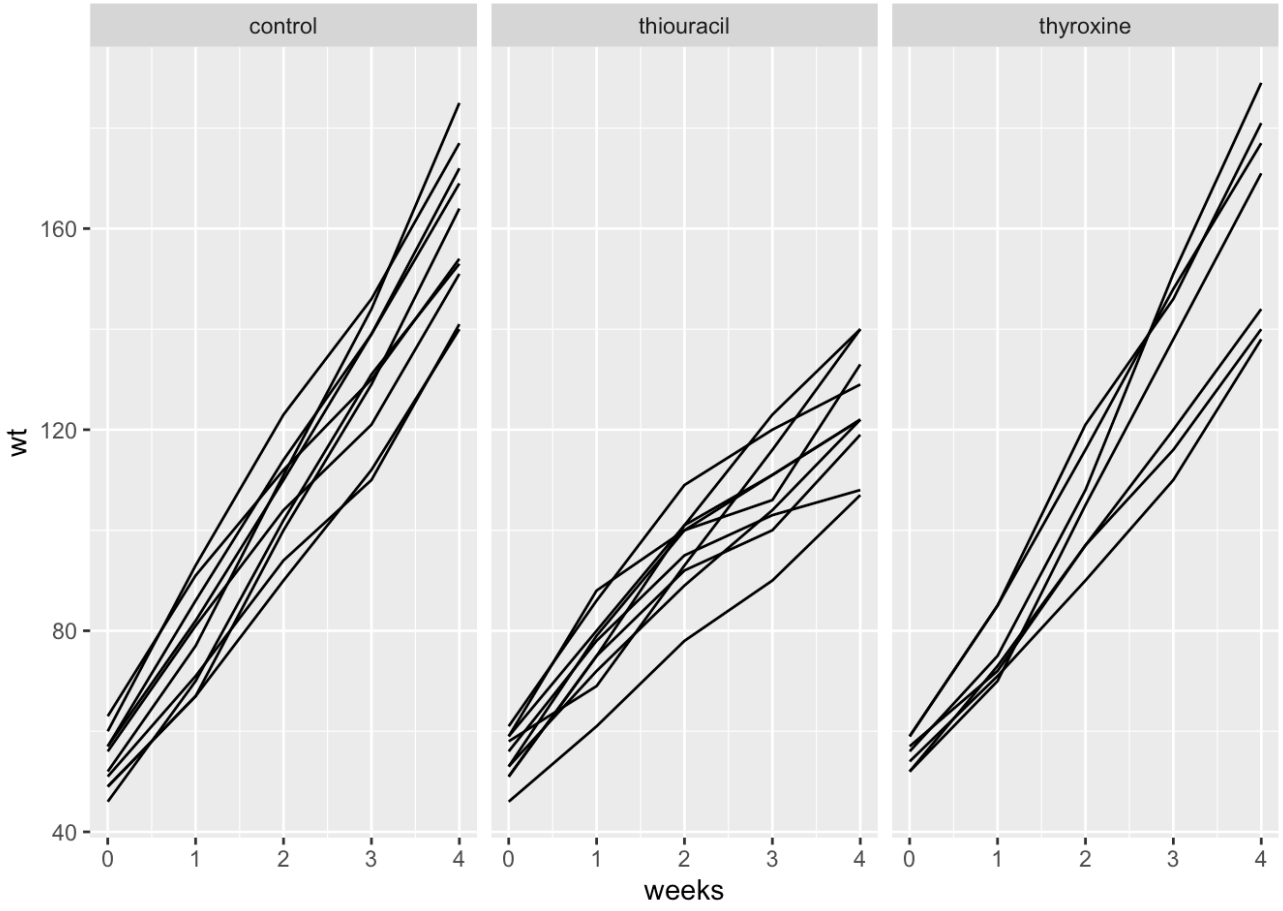
(a) Answer:

```
data(ratdrink)
help(ratdrink)
```

```
ratdrink |>
  ggplot() +
  geom_line(mapping = aes(x = weeks, y = wt, group = subject, color = treat))
```




```
ratdrink |>
  ggplot() +
  geom_line(mapping = aes(x = weeks, y = wt, group = subject)) +
  facet_wrap(~ treat)
```



We can see the weight increases with time for all three groups. The treatment group with added thiouracil seems to increase slower tha the control group and the other treatment group.

(b) Answer:

```
mmod <- lmer(wt ~ weeks * treat + (weeks | subject), data = ratdrink)
summary(mmod)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: wt ~ weeks * treat + (weeks | subject)
   Data: ratdrink

REML criterion at convergence: 878.7

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.83136 -0.54991  0.04003  0.58230  2.03660

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 32.49    5.700
          weeks       14.14    3.760    -0.13
 Residual             18.90    4.348
Number of obs: 135, groups:  subject, 27

Fixed effects:
```

```
                     Estimate Std. Error t value
(Intercept)           52.8800     2.0937  25.256
weeks                 26.4800     1.2661  20.915
treatthiouracil        4.7800     2.9610   1.614
treatthyroxine        -0.7943     3.2628  -0.243
weeks:treatthiouracil -9.3700     1.7905  -5.233
weeks:treatthyroxine   0.6629     1.9730   0.336

Correlation of Fixed Effects:
            (Intr) weeks  trtthr trtthy wks:trtthr
weeks       -0.250
treatthircl -0.707  0.177
treatthyrxn -0.642  0.160  0.454
wks:trtthrc  0.177 -0.707 -0.250 -0.113
wks:trtthyr  0.160 -0.642 -0.113 -0.250  0.454
```

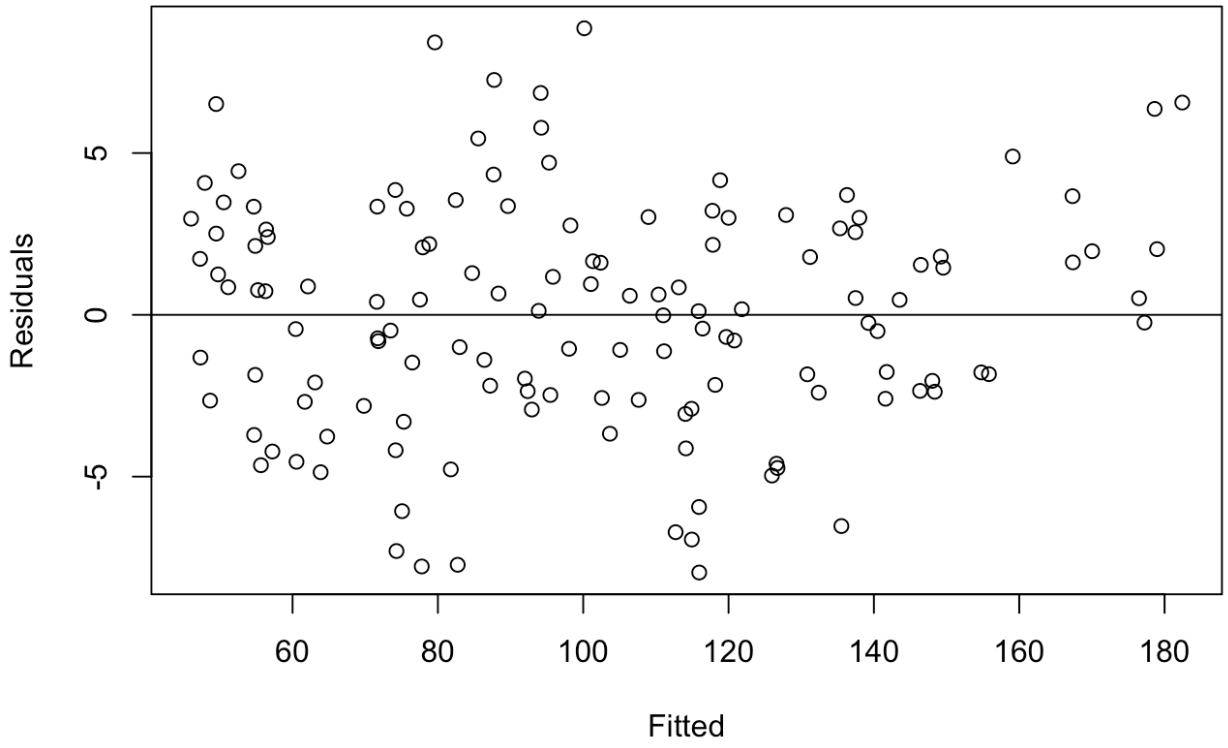
- For control group at weeks 0, the expected weight is 52.88 for a typical individual.
- For every one week increase, the rate of change for expected weight decreases -9.3700 compared to control group for typical individuals.
- The SD of intercept of random effect is 5.7 which might suggests there is slightly large differences between different subjects.

(c) Answer:

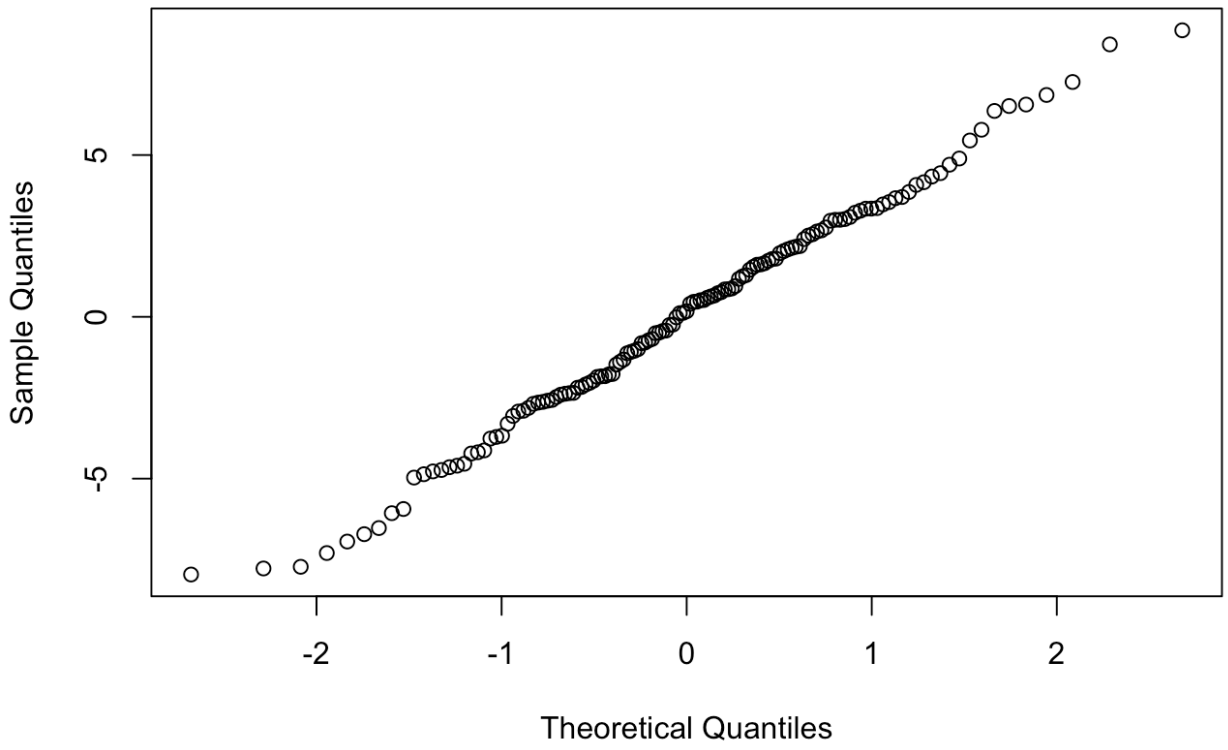
Since the 95% confidence interval of all treatments contains 0, we do not have enough evidence to conclude there is significant difference of effect between treatments and control group.

(d) Answer:

```
plot(resid(mmod) ~ fitted(mmod), xlab = "Fitted", ylab = "Residuals")
abline(h = 0 )
```




```
qqnorm(resid(mmod), main = "")
```



The residuals are equally scattered about the x-axis with no evidence of homoscedasticity. Assumptions of linearity and equal variance appear reasoanbly satisfied. QQ plot shows the residuals roughly follow a normal distribution.

(e) Answer:

```
confint(mmod)
```

```
Computing profile confidence intervals ...
```

	2.5 %	97.5 %
.sig01	3.4506344	7.6654748
.sig02	-0.5261017	0.3794396
.sig03	2.6064086	4.8687660
.sigma	3.7555591	5.1142342
(Intercept)	48.8692859	56.8907140
weeks	24.0547425	28.9052574
treatthiouracil	-0.8920062	10.4520061
treatthyroxine	-7.0445322	5.4559606
weeks:treatthiouracil	-12.7998320	-5.9401680
weeks:treatthyroxine	-3.1166337	4.4423479

Q4. ELMR Exercise 13.1 (p295)

1. The `ohio` data concern 536 children from Steubenville, Ohio and were taken as part of a study on the effects of air pollution. Children were in the study for 4 years from ages 7 to 10. The response was whether they wheezed or not. The variables are:

   resp an indicator of wheeze status (1 = yes, 0 = no)

   id an identifier for the child

   age 7 yrs = −2, 8 yrs = −1, 9 yrs = 0, 10 yrs = 1

   smoke an indicator of maternal smoking at the first year of the study (1 = smoker, 0 = nonsmoker)

   (a) Do any of the mothers in the study change their smoking status during the period of observation?

   (b) Construct a table that shows proportion of children who wheeze for 0, 1, 2, 3 or 4 years broken down by maternal smoking status.

   (c) Make plot which shows how the proportion of children wheezing changes by age with a separate line for smoking and nonsmoking mothers.

   (d) Group the data by child to count the total (out of four) years of wheezing. Fit a binomial GLM to this response to check for a maternal smoking effect. Does this prove there is a smoking effect or could there be another plausible explanation?

   (e) Fit a model for each individual response using a GLMM fit using penalized quasi-likelihood. Describe the effects of age and maternal smoking. How do the odds of wheezing change numerically over time?

   (f) Now fit the same model but using adaptive Gaussian-Hermit quadrature. Compare to the previous model fit.

   (g) Use `INLA` to fit the same model. What does this model say about the effect of age and maternal smoking?

   (h) Use `STAN` to fit the same model. Check the MCMC diagnostics and again discuss the age and maternal smoking effects.

   (i) Fit the model using GEE. Use an autoregressive rather than exchangeable error structure. Compare the results to the previous model fits. In your model, what indicates that a child who already wheezes is likely to continue to wheeze?

   (j) What is your overall conclusion regarding the effect of age and maternal smoking? Can we trust the GLM result or are the GLMM models preferable?

(a) Answer:

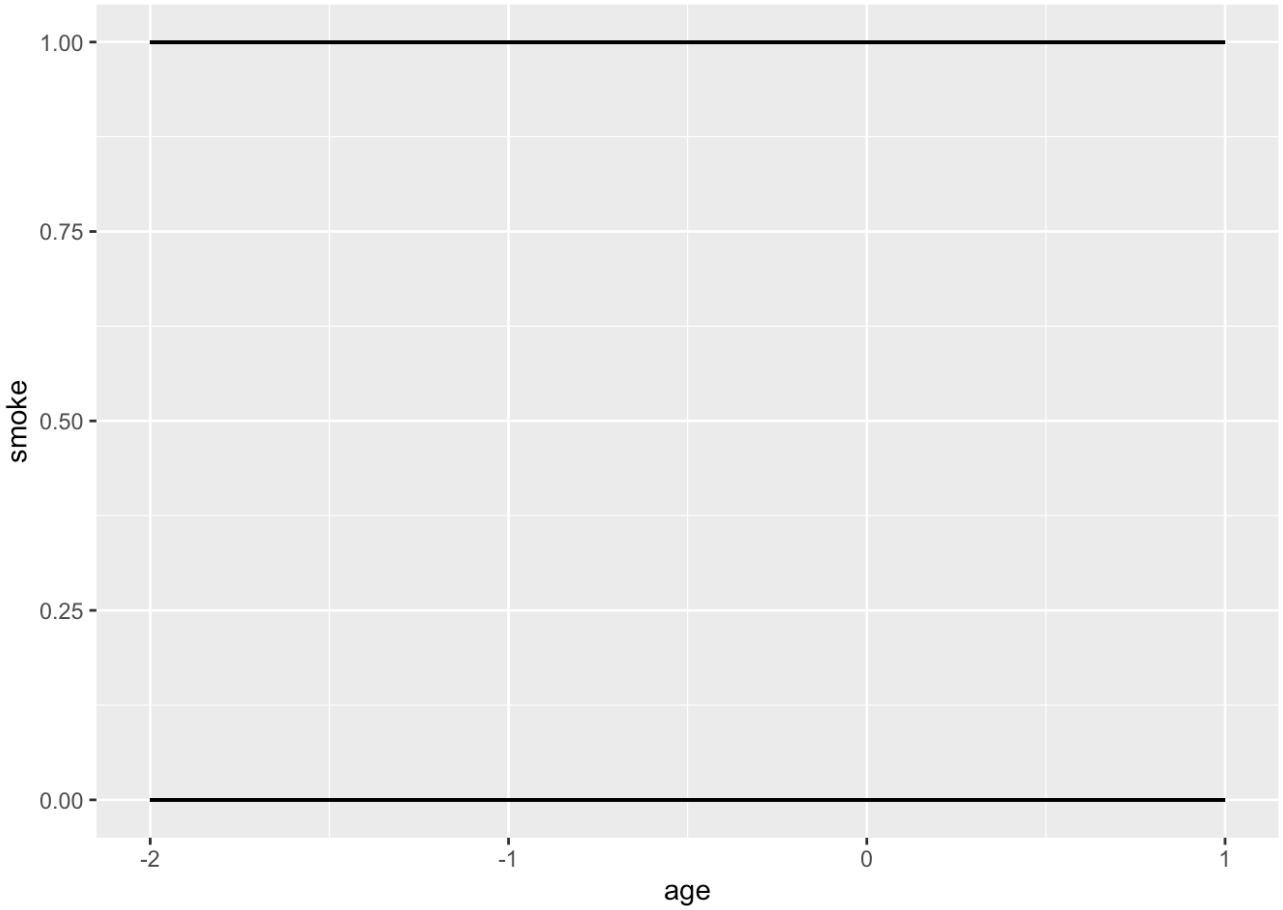
```
data(ohio)
help(ohio)
```

```
Help on topic 'ohio' was found in the following packages:

  Package               Library
  faraway               /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/library
  geepack               /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/library


Using the first match ...
```

```
ohio |>
  ggplot() +
  geom_line(mapping = aes(x = age, y = smoke, group = id))
```



No, no mother changed their smoke status during the study.

(b) Answer:

```
wheeze_count <- ohio %>%
  group_by(id, smoke) %>%
  summarize(t = sum(resp), .groups = 'drop')

wheeze_table <- table(wheeze_count$smoke, wheeze_count$t)
```

```
wheeze_table <- as.data.frame.matrix(wheeze_table)


wheeze_table$total <- rowSums(wheeze_table)
wheeze_table <- wheeze_table %>%
  mutate(across(everything(), ~ ./total)) %>%
  dplyr::select(-total)


print(wheeze_table)
```

```
          0         1          2          3          4
0 0.6771429 0.1857143 0.07142857 0.03428571 0.03142857
1 0.6310160 0.1711230 0.10160428 0.05882353 0.03743316
```

(c) Answer:

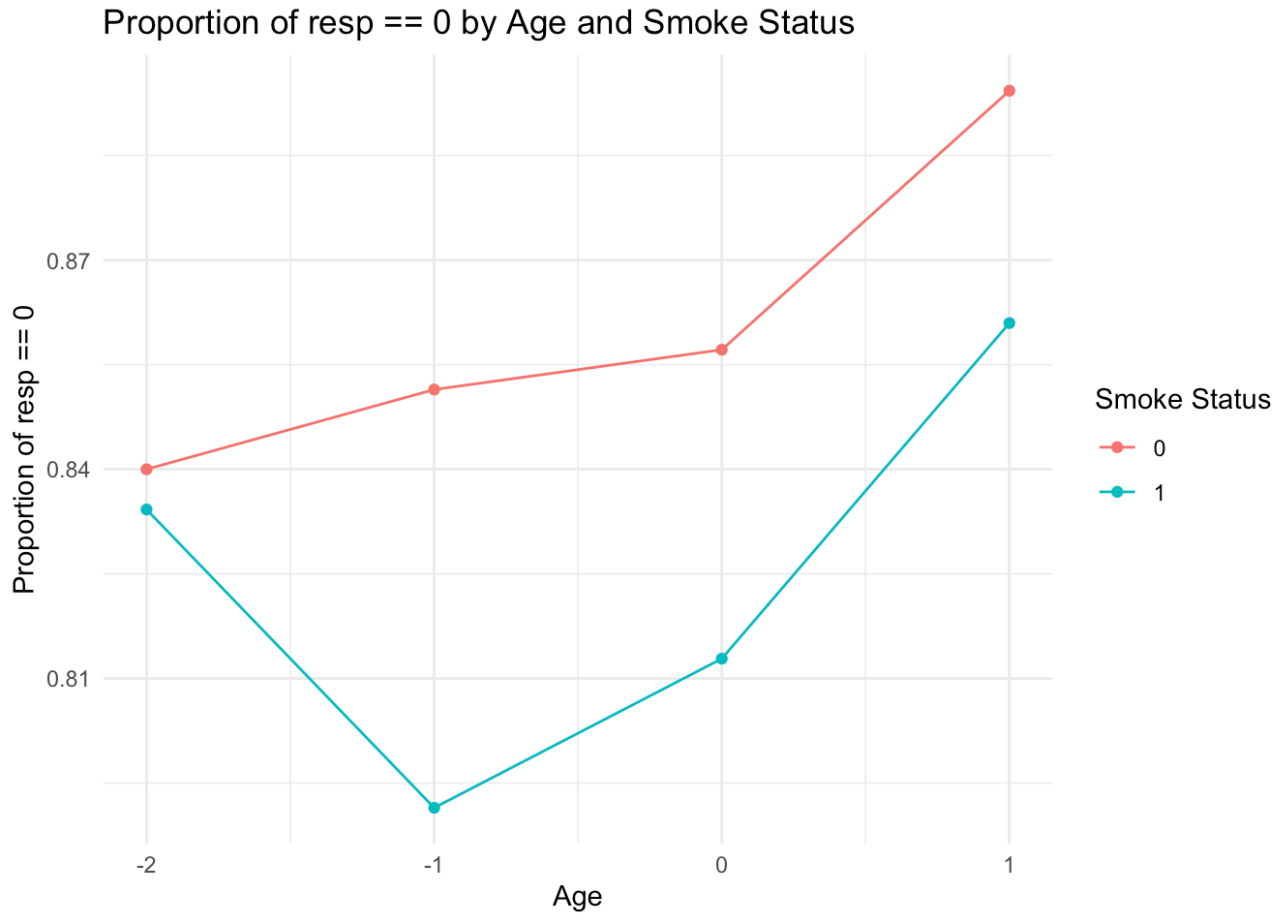
```
proportion_table <- ohio |>
  group_by(age, smoke) |>
  summarise(proportion_resp0 = mean(resp == 0)) |>
  ungroup()
```

```
`summarise()` has grouped output by 'age'. You can override using the `.groups`
argument.
```

```
proportion_table_wide <- dcast(proportion_table, age ~ smoke, value.var = "proportio
```

```
proportion_table$smoke <- as.factor(proportion_table$smoke)
```

```
ggplot(proportion_table, aes(x = age, y = proportion_resp0, color = smoke, group = s
  geom_line() +
  geom_point() +
  labs(title = "Proportion of resp == 0 by Age and Smoke Status",
       x = "Age",
       y = "Proportion of resp == 0",
       color = "Smoke Status") +
  theme_minimal()
```



(d) Answer:

```
#remove age
binomial_df <- ohio |>
  group_by(id) |>
  summarise(total_resp = sum(resp),
            smoke = first(smoke)) |>
  mutate(smoke = as.factor(smoke))
```

```
glmmod <- glm(cbind(total_resp, 4 - total_resp) ~ smoke, data = binomial_df, family

summary(glmmod)
```

```
Call:
glm(formula = cbind(total_resp, 4 - total_resp) ~ smoke, family = binomial,
    data = binomial_df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.82124    0.07719 -23.595   <2e-16 ***
smoke1       0.27156    0.12334   2.202   0.0277 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1045.3  on 536  degrees of freedom
Residual deviance: 1040.5  on 535  degrees of freedom
AIC: 1337.9

Number of Fisher Scoring iterations: 4
```

No, we cannot prove it. Since we are assuming within each group(`id`), the observations are independent, which is not true in this case. The assumption is violated hence the inference is not valid.

(e) Answer:

```
ohio$id <- as.factor(ohio$id)
ohio$smoke <- as.factor(ohio$smoke)
```

```
modpql <- glmmPQL(resp ~ age + smoke,
                  random = ~ 1 | id,
                  family = binomial,
                  data   = ohio)
```

```
iteration 1

iteration 2

iteration 3

iteration 4

iteration 5

iteration 6

iteration 7

iteration 8
```

```
summary(modpql)
```

```
Linear mixed-effects model fit by maximum likelihood
  Data: ohio
  AIC BIC logLik
   NA  NA     NA

Random effects:
 Formula: ~1 | id
        (Intercept)  Residual
StdDev:    2.057175 0.6355563

Variance function:
 Structure: fixed weights
 Formula: ~invwt
```

```
Fixed effects:  resp ~ age + smoke
                Value  Std.Error   DF    t-value p-value
(Intercept) -2.7658365 0.14218299 1610 -19.452654  0.0000
age         -0.1815756 0.04365164 1610  -4.159652  0.0000
smoke1       0.3251839 0.23131699  535   1.405793  0.1604
 Correlation:
       (Intr) age
age     0.197
smoke1 -0.591 -0.003

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.6145143 -0.2829352 -0.2583719 -0.2154580  3.4443795

Number of Observations: 2148
Number of Groups: 537
```

– age`: For one year increases in age, the log odds of wheezing decreases by 0.1816 for a typical individual.

– smoke`: The log odds of wheezing for a smoker is 0.3252 higher than for a non-smoker of the same age for typical individuals.

For one year increase in age, the odds of wheezing decrease by 16.61% for a typical individual.

(f) Answer:

```r
modgh <- glmer(resp ~ age + smoke + (1 | id),
               nAGQ     = 25,
               family   = poisson,
               data     = ohio)
summary(modgh)
```

```
Generalized linear mixed model fit by maximum likelihood (Adaptive
  Gauss-Hermite Quadrature, nAGQ = 25) [glmerMod]
 Family: poisson  ( log )
Formula: resp ~ age + smoke + (1 | id)
   Data: ohio

     AIC      BIC   logLik deviance df.resid
  1145.4   1168.1   -568.7   1137.4     2144

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.7074 -0.2760 -0.2510 -0.2281  2.5872

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 1.153    1.074
Number of obs: 2148, groups:  id, 537

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.58178    0.13462 -19.179   <2e-16 ***
```

```
age         -0.09596    0.04973  -1.929   0.0537 .
smoke1       0.24806    0.16264   1.525   0.1272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
       (Intr) age
age     0.229
smoke1 -0.479  0.000
```

Compared with GLMM fit using penalized quasi-likelihood, this model gives a different estimates of the coefficient. The signs of the estimates are the same but the values are different. This model also indicates that age is not a significant predictor of wheezing while GLMM fit using penalized quasi-likelihood indicates that age is a significant predictor of wheezing.

(g) Answer:

```r
library(INLA)
```

```
Loading required package: sp

This is INLA_24.05.01-1 built 2024-05-01 18:56:18 UTC.
 - See www.r-inla.org/contact-us for how to get help.
 - List available models/likelihoods/etc with inla.list.models()
 - Use inla.doc(<NAME>) to access documentation
```

```r
inla_mod <- inla(resp ~ age + smoke + f(id, model = "iid"),
                 family = "binomial",
                 data = ohio,
                 verbose = TRUE)
summary(inla_mod)
```

```
Time used:
    Pre = 0.697, Running = 0.413, Post = 0.0259, Total = 1.14
Fixed effects:
             mean    sd 0.025quant 0.5quant 0.975quant   mode kld
(Intercept) -2.946 0.201     -3.360   -2.938     -2.574 -2.938   0
age         -0.173 0.063     -0.297   -0.173     -0.050 -0.173   0
smoke1       0.385 0.239     -0.083    0.384      0.858  0.384   0

Random effects:
  Name    Model
    id IID model

Model hyperparameters:
                  mean    sd 0.025quant 0.5quant 0.975quant  mode
Precision for id 0.276 0.047      0.196    0.272      0.378 0.265

Marginal log-Likelihood:  -834.98
 is computed
Posterior summaries for the linear predictor and the fitted values are computed
(Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

This model indicates that age is a significant predictor of wheezing while smoke is not a significant predictor of wheezing using 95% confidence interval.

The model says that one year increase in age, the log odds of wheezing decrease by 0.73 for a typical individual. It also says for a smoker, the log odds of wheezing is 0.385 higher than for a non-smoker of the same age for typical individuals.

(i) Answer:

```
        modgeep <- geeglm(resp ~ age + smoke,
                          id        = id,
                          corstr    = "ar1",
                          scale.fix = TRUE,
                          data      = ohio,
                          family    = binomial(link = "logit"))
        summary(modgeep)
```

```
Call:
geeglm(formula = resp ~ age + smoke, family = binomial(link = "logit"),
    data = ohio, id = id, corstr = "ar1", scale.fix = TRUE)

 Coefficients:
            Estimate  Std.err    Wald Pr(>|W|)
(Intercept) -1.90262  0.11533 272.174   <2e-16 ***
age         -0.11490  0.04544   6.394   0.0115 *
smoke1       0.23340  0.18137   1.656   0.1981
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation structure = ar1
Scale is fixed.

  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.5007 0.03799
Number of clusters:   537  Maximum cluster size: 4
```

The estimate of the alpha is 0.501. This reflects a moderate level of persistence in wheezing over time, indicating that children who wheeze at one time point have a moderate likelihood of wheezing at the next.

(j) Answer:

Smoker will have larger odds of wheezing than non-smoker of the same age. Increasing in age will have a negative effect on the odds of wheezing. The GLMM model is prefereable since it accounts for the correlation within the same individual.