TECHNICAL UNIVERSITY OF DENMARK
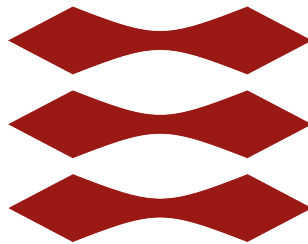
EXAM ASSIGNMENT

## 02429 Analysis of correlated data: Mixed linear models

Student:
Michella Ravn Søndergaard    s184005

DTU - Kongens Lyngby (Denmark)

December 5, 2025

# 1 Radon level in American households (Minnesota)

In the exam assignment, a dataset of radon measurements is given with 919 homes in 85 counties in Minnesota, the variables of interest are in Table 1, corresponding R code is in Appendix A.1.

*Table 1: Variables used in the analysis*

| Variable | Type | Description | Range of values |
|---|---|---|---|
| $y$ | Continuous | Log radon measurement for house $i = 1, \ldots, 919$ | $[-2.30, 3.76]$ |
| $x$ | Binary | Floor of measurement {basement, ground floor} | $\{0, 1\}$ |
| $u$ | Continuous | Log uranium measurement in county $j = 1, \ldots, 85$ | $[-0.88, 0.53]$ |

The objective is to model radon concentrations and make inferential statements at county-level. The data is hierarchical with houses within counties. Therefore, **county** is modeled as random.

## 1.1 Exploratory analysis

In the preliminary exploratory analysis the county data is grouped into regions, two counties does not have any data, see Figure 1a. R code and county numbers on map can be found in appendix A.2.
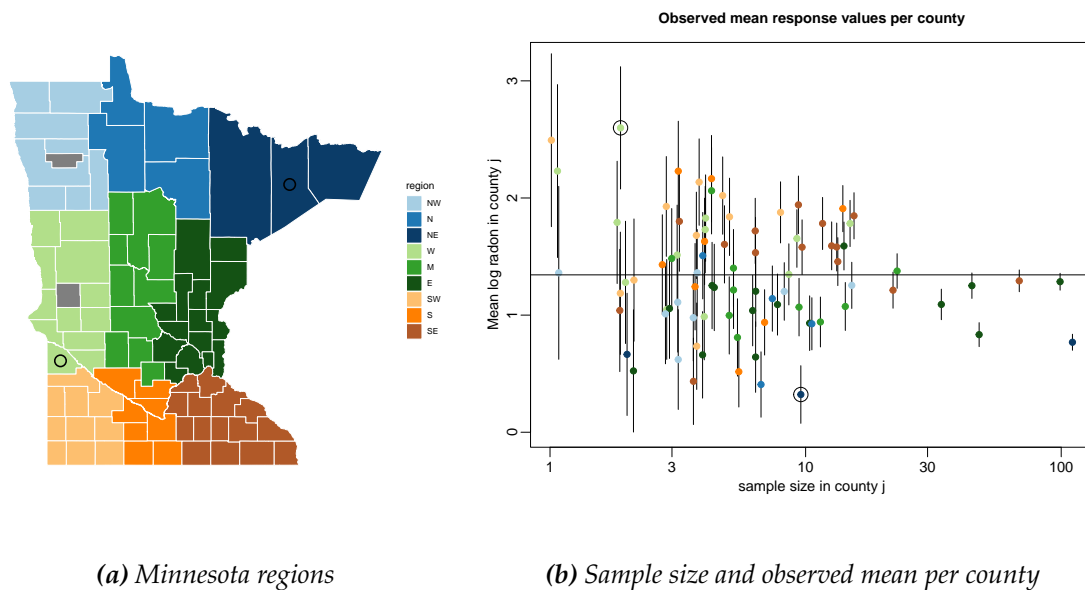


*(a) Minnesota regions*



*(b) Sample size and observed mean per county*

*Figure 1: County regions and sample size*

The data is unbalanced with sample sizes in range from 1 to more than 100 with most counties within 3-12, Figure 1b. Smaller sample size produces larger standard error and opposite for large sample size, these are calculated using equal variance. The maximum radon measurement is in the W region, county Lac Qui Parle with only two observations, and the minimum is in the NE region, county Lake with sample size eight. The observed means ordered in regions, Figure 2, the northern regions are generally low and not as dispersed as the middle and southern. Another interesting perspective could be to model region as a fixed effect, however this is out of scope for this project.
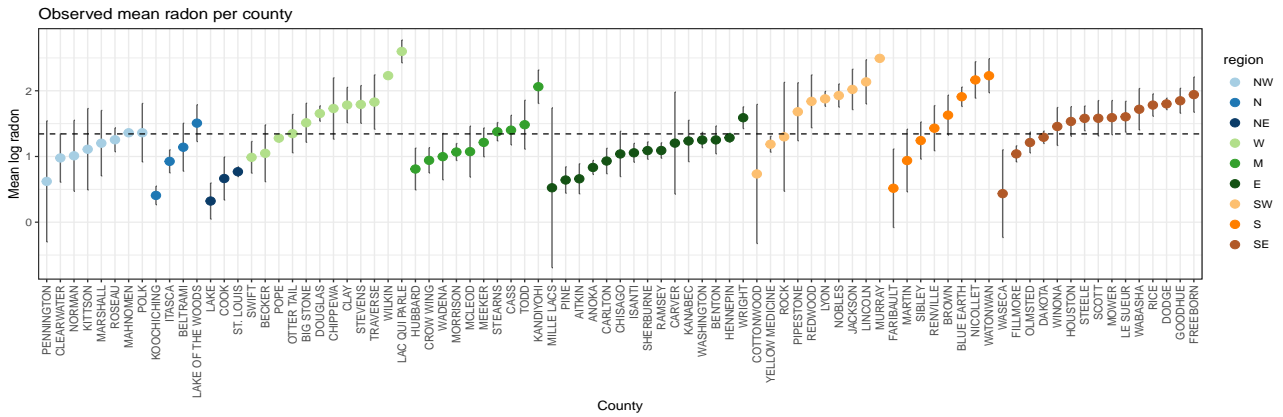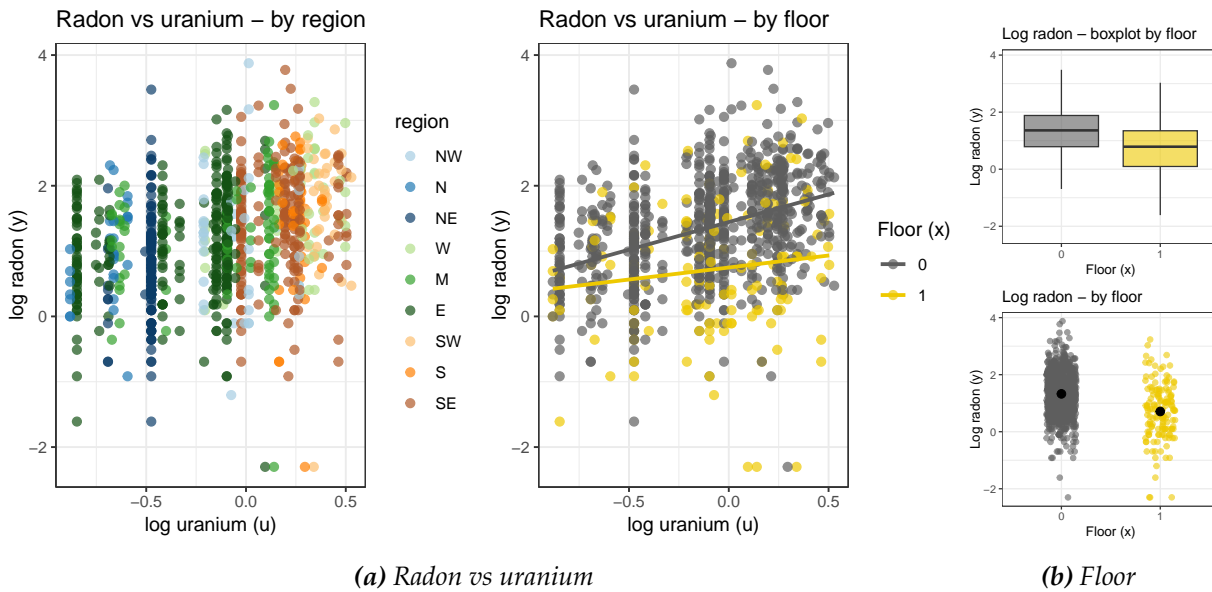
**Figure 2:** *Observed mean radon per county, ordered*

Uranium is a county measure, therefore, the scatterplot displays a banded structure, Figure 3a. The northern regions have lower uranium whereas the southern regions have higher. Ground floor generally yields lower radon concentrations and is less susceptible for higher concentrations of uranium. Figure 3b also indicate lower radon concentration for ground floor measurements. Notice the imbalance; there are more than 5 times as many measurements in basement than at ground level.



**(a)** *Radon vs uranium*

**(b)** *Floor*

**Figure 3:** *Floor, uranium and radon*

## 1.2   Modelling

Proposed initial multi-level, mixed model, recall variable names from Table 1

$$m0: \qquad y_{ij} = \mu + \alpha \cdot u_j + \beta(x_i) + \gamma(x_i) \cdot u + d_j + \varepsilon_{ij} \qquad (1)$$

$$d_j \sim N\left(0, \sigma_{county}^2\right), \quad \varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

$$\text{Cov}(d_j, \varepsilon_{ij}) = 0, \quad i = 1, \ldots, 919, j = 1, \ldots, 85$$

where $\delta_j$ is the random effect of county and $\varepsilon_{ij}$ is the error.

Starting from the proposed model and using backwards elimination at a 0.05 significance level. Firstly the random effect of county is tested using restricted maximum likelihood, (REML)-based test, using the function `ranova` in R, see appendix A.3. The random effect was found to be significant with a p-value of 0.005, meaning substantial heterogeneity between counties. Assessing the fixed effects a maximum likelihood (ML)-based test is performed, using `drop1`. The interaction was found significant with a p-value of 0.029. The diagnostics plots in Figure 4 to check model assumptions.
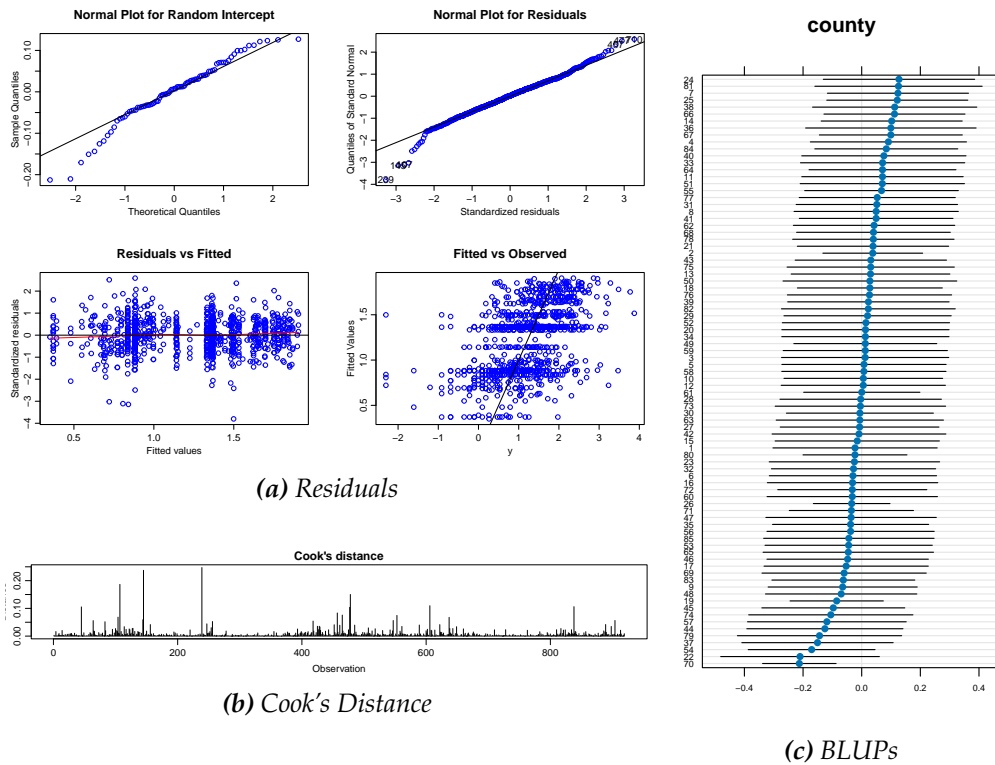


*(a) Residuals*



*(b) Cook's Distance*



*(c) BLUPs*

*Figure 4: Model diagnostics*

From Figure 4a the normality assumption for random effect and residuals generally hold with a heavy lower tail. No underlying structure in residuals vs fitted; the bands are from the the data itself. Similar for Best Linear Unbiased Predictions, BLUPs in Figure 4c, the range is around [-0,4, 0.4] and are roughly symmetric with a heavier lower tail. The two outmost smallest counties are 70, St. Louis (NE) and 22, Faribault (S). **Cooks distance**, Figure 4b is investigated for influential observations and potential outliers, the observations with a cooks distance above 0.2 is in Table 12 in Appendix A.3. Although having very low radon concentrations, they were kept in the data. Performance statistics in Table 2

*Table 2: Model performance statistics*

| Model | # Params | AIC | BIC | ICC | $R^2$ |
|-------|----------|---------|---------|------|-------|
| $m_0$ | 4 | 2142.91 | 2171.85 | 0.04 | 0.21 |

The $R^2$ can at first glance seem quite low, with the mixed model explaining 21% of the variance within the data, however comparing results with Gelman it is reasonable, the argument is that radon

measurements are extremely noisy. Also the intraclass correlation is small at 4%, but significant.

The model parameters, Table 3 show both the raw log scale values and back-transformed; the back-transformed model; $y_{ij} = \log(R_{ij})$, is

$$R_{ij} = \exp\left(\mu + \alpha u_j + \beta\left(x_i\right) + \gamma\left(x_i\right)u_j + d_j + \varepsilon_{ij}\right) \tag{2}$$

$$= \exp(\mu)\exp\left(\alpha u_j\right)\exp\left(\beta\left(x_i\right)\right)\exp\left(\gamma\left(x_i\right)u_j\right)\exp(d_j)\exp\left(\varepsilon_{ij}\right). \tag{3}$$

*Table 3: Fixed-effect estimates with CI*

| Parameter | Estimate | 2.5% | 97.5% | Back-transformed | 2.5% (BT) | 97.5% (BT) |
|---|---|---|---|---|---|---|
| $\mu$ | 1.47 | 1.40 | 1.55 | 4.37 | 4.05 | 4.71 |
| $u$ | 0.80 | 0.60 | 0.99 | 2.22 | 1.83 | 2.68 |
| $x_1$ | -0.71 | -0.85 | -0.57 | 0.49 | 0.43 | 0.57 |
| $u : x_1$ | -0.42 | -0.79 | -0.04 | 0.66 | 0.45 | 0.96 |

The baseline radon level (at $u = 0 \rightarrow$ uranium=1 $ppm$) in a basement of a house in a random county is 4.37 $\frac{pCi}{L}$ and one unit increase in the uranium level increases the radon with a factor 2.22. Ground level measurements approximately halves the radon concentration. Lastly interaction factor of 0.66 means the uranium effect is weaker on ground floor.

## 1.3 Results

Predicting the mean response for each county and comparing with the observed is in Figure 5. Overall the predictions seem more narrow and tend to overestimate some in the middle/eastern regions and underestimate in the southern regions.
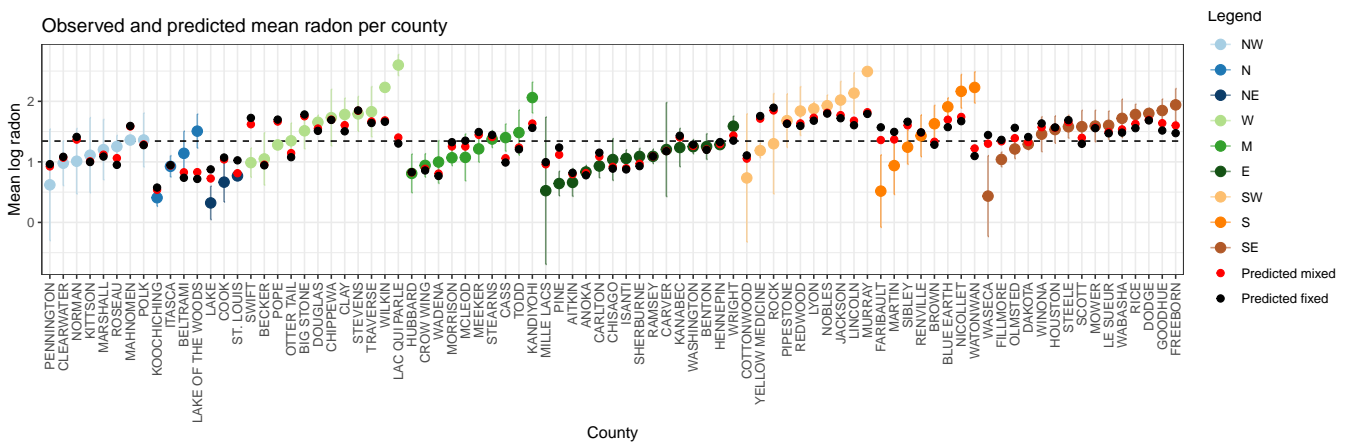


*Figure 5: Observed and predicted mean radon per county, ordered*

The predictions from the mixed model, using the known BLUPs, see Appendix A.4. Predicting an unknown county can be done using the fixed effect, i.e. $E(d_j) = 0$, this is simulated from the data and in Figure 5.

Using the mixed model for known data performs as expected better than for unknown. Estimated marginal means and confidence interval for log scale and back transformed (BT) is in Table 4. As expected higher level of uranium means higher concentration of radon and ground floor has approximately half the radon concentration as basement, and is less affected by uranium, Figure 6

| $x$ | $u$ | emmean | CI | emmean BT | CI BT |
|---|---|---|---|---|---|
| 0 | -0.47 | 1.10 | [0.98, 1.21] | 2.99 | [2.67, 3.35] |
| 1 | -0.47 | 0.58 | [0.39, 0.77] | 1.79 | [1.48, 2.17] |
| 0 | -0.10 | 1.40 | [1.32, 1.47] | 4.04 | [3.75, 4.36] |
| 1 | -0.10 | 0.73 | [0.60, 0.86] | 2.07 | [1.82, 2.36] |
| 0 | 0.18 | 1.62 | [1.53, 1.71] | 5.05 | [4.62, 5.53] |
| 1 | 0.18 | 0.83 | [0.67, 0.99] | 2.30 | [1.96, 2.70] |

**Table 4:** *Estimated marginal means with direct back transformation, BT, with 95% CI*



**Figure 6:** *Emmeans*



*(a) Intercepts*

*(b) Slopes*
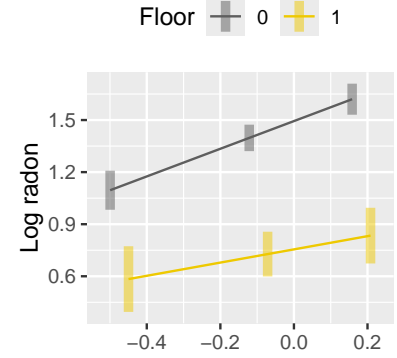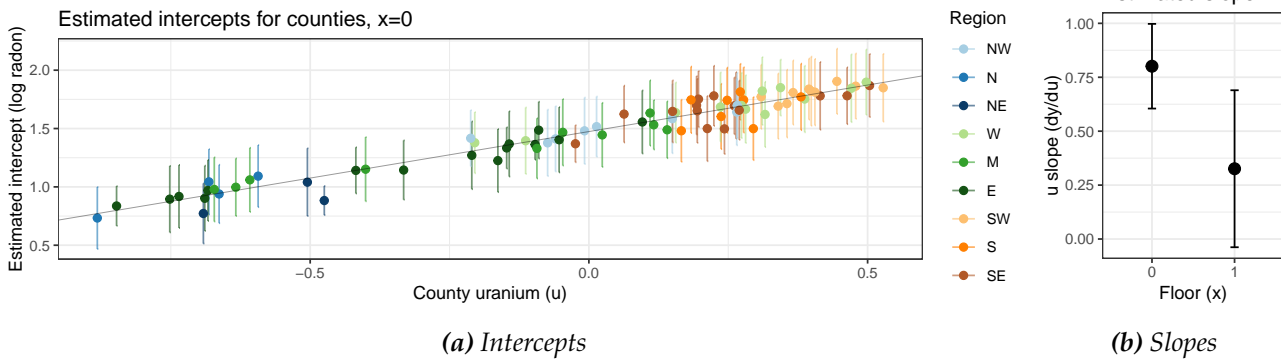
**Figure 7:** *Estimated intercepts and slope, log scale*

For each county, the intercepts at $x = 0$ can be calculated $ri_j = \mu + \alpha \cdot u_j + d_j$, with CI from posterior variance $\text{Var}(d_j|y_{ij})$, with fixed effect line $\mu + \alpha \cdot u_j$ shown in Figure 7a, the random intercept generally follows this line. Meaning that the variance is largely explained by uranium itself. The estimated slopes in Figure 7b show higher value, i.e. steeper slope, and tighter CI for basement. Both slopes are statistically different from zero, tested via `emtrends`, with stronger evidence for basement than ground level, R code is in Appendix A.4.

## 1.4   Conclusion

The proposed mixed model confirmed significant county heterogeneity and a significant floor–uranium interaction, with uranium having a weaker effect on ground-floor measurements. Although radon is highly variable, $R^2 = 0.21$, parameter estimates indicate that basement radon levels are highest with baseline $4.37 \frac{pCi}{L}$ at uranium level $1\,ppm$, and scales with uranium with a unit increase in uranium multiplies radon with 2.22. Radon concentration is roughly halved if measured at ground floor with smaller effect of uranium; one unit increase in uranium multiplies radon with 1.47.

# 2 Rats

A dataset for weight gain of 27 rats, in groups of 10, 7 and 10 is given. The data is formatted such that $y_t$ is the cumulative weight at time $t$, corresponding R code is in Appendix B.1.

*Table 5: Variables used in the rats growth analysis*

| Variable | Type | Description | Range of values |
|---|---|---|---|
| $y_t$ | Continuous | Body weight at week $t$ (response) | [61, 191] g |
| $y_0$ | Continuous | Initial weight before treatment (baseline covariate) | [46, 63] g |
| Trt | Categorical | Treatment group (control, thyroxin, thiouracil) | $\{1, 2, 3\}$ |
| week | Continous | Measurement week after treatment | $\{1, 2, 3, 4\}$ |
| Rat | Categorical | Rat identifier (random effect) | 27 levels |

The objective is to inferences of the treatment effect on the mean response, weight profile.

## 2.1 Exploratory analysis

In the preliminary exploratory analysis the weight curves is shown individually for each rat and mean for each treatment in Figure 8
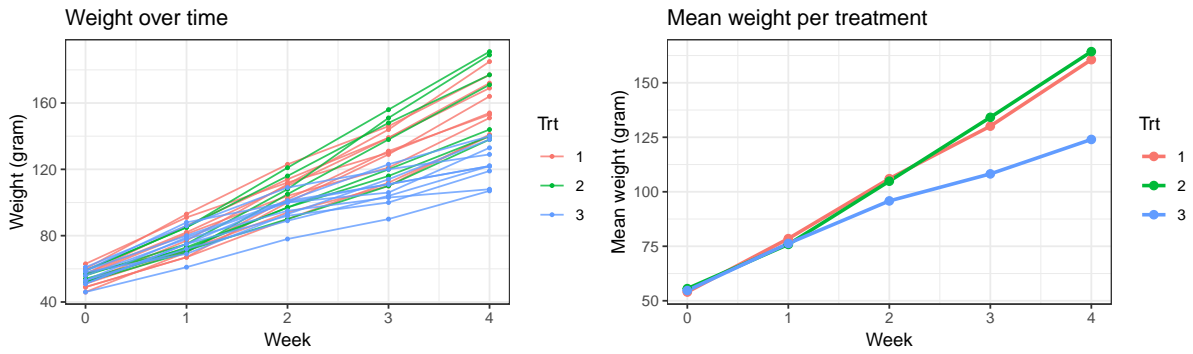


*Figure 8: Individual and mean weight curves*

The weight is measured each week, so the timestamps are equidistant, and what is observed is that in general the weight increases for all of the rats with the measurements fanning out over the weeks. From the mean profiles it seems that treatment 3 (thiouracil) does so at a slower rate already from week 2. R code can be found in Appendix B.2

## 2.2 Modelling

Proposed initial mixed model, with compound symmetry, recall variables from Table 5,

$$m_1: \quad y_{itk} = \mu + \alpha \cdot \text{week}_t + \beta(Trt_k) + \gamma(Trt_k) \cdot \text{week}_t + \delta \cdot y_{0i} + r_i + \varepsilon_{itk}, \quad (4)$$

$$r_i \sim N\left(0, \sigma_{\text{Rat}}^2\right), \varepsilon_{itk} \sim N\left(0, \sigma^2\right),$$

$$\text{Cov}(r_i, \varepsilon_{itk}) = 0, i = 1, \ldots, 27, \quad t = 1, \ldots, 4, \quad k = 1, 2, 3.$$

where $r_i$ is the random effect of rat and $\varepsilon_{itk}$ is the error.

Another model $m_2$ could be considering time as factor, i.e. $\alpha(\text{week}_t)$ and $\gamma(Trt_k), \text{week}_t$, when using anova test, the $m_1$ time as numeric is preferred, see R code in Appendix B.3. Another model considered is random slope

$$m_3: \quad y_{itk} = \mu + \alpha \cdot \text{week}_t + \beta(\text{Trt}_k) + \gamma(\text{Trt}_k) \cdot \text{week}_t + \delta \cdot y_{0i} + r_0(\text{Rat}_i) + r_1(\text{Rat}_i)\,\text{week}_t + \varepsilon_{itk}, \quad (5)$$

$$r_0(i), r_1(i) \sim N\left(0, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}\right),$$

$$\varepsilon_{itk} \sim N(0, \sigma^2), \quad i = 1, \ldots, 27, \ t = 1, \ldots, 4., k = 1, 2, 3.$$

The random slope is significant, tested using `ranova`. In the analysis the residuals vs fitted showed both funnel shape and curvature, see Appendix B.3, thus leading to $m_4$ and $m_5$ with curvature added

$$m_4: \quad y_{itk} = \mu + \alpha_1 \cdot \text{week}_t + \alpha_2 \cdot \text{week}_t^2 + \beta(\text{Trt}_k) + \gamma(\text{Trt}_k) \cdot \text{week}_t$$
$$+ \delta \cdot y_{0i} + r_0(\text{Rat}_i) + r_1(\text{Rat}_i)\,\text{week}_t + \varepsilon_{itk}, \quad (6)$$

$$m_5: \quad y_{itk} = \mu + \alpha_1 \cdot \text{week}_t + \alpha_2 \cdot \text{week}_t^2 + \alpha_3 \cdot \text{week}_t^3 + \beta(\text{Trt}_k) + \gamma(\text{Trt}_k) \cdot \text{week}_t$$
$$+ \delta \cdot y_{0i} + r_0(\text{Rat}_i) + r_1(\text{Rat}_i)\,\text{week}_t + \varepsilon_{itk}, \quad (7)$$

Rather complex models, but nevertheless significant when testing with `drop1` in Appendix B.3 is inspected. To accommodate for the funnel shape, a boxcox transformation is inspected leading to a log transformation of the response. Thus the models $m_6$ in the format of $m_3$ w. log transformed response, $m_7$ in the format of $m_4$ w. log transformed response and lastly $m_8$ in the format of $m_5$ w. log transformed response, their expressions can be found in Appendix B.3 Performance statistics for comparing these 8 models can be seen in Table 6

*Table 6: Model performance statistics for the rats data*

| Model | # Params | AIC | BIC | Marginal $R^2$ | Conditional $R^2$ |
|-------|----------|-----|-----|----------------|-------------------|
| $m_1$ | 8 | 765.27 | 789.40 | 0.88 | 0.95 |
| $m_2$ | 14 | 743.83 | 784.06 | 0.88 | 0.95 |
| $m_3$ | 8 | 673.91 | 703.41 | 0.88 | 0.99 |
| $m_4$ | 9 | 676.61 | 708.80 | 0.88 | 0.99 |
| $m_5$ | 10 | 668.54 | 703.41 | 0.88 | 0.99 |
| $m_6$ | 8 | -220.38 | -190.87 | 0.89 | 0.97 |
| $m_7$ | 9 | -259.15 | -226.97 | 0.90 | 0.99 |
| $m_8$ | 10 | -265.06 | -230.19 | 0.90 | 0.99 |

Now as the complexity of the model grows, the better fit and lower AIC, BIC. However, an appropriate model is not only overly fitted to the data but also interpretable. Log transforming the response greatly impacted AIC, BIC. The marginal and conditional $R^2$ describe how much of the total variation is explained by the fixed effects alone and by the full mixed model, respectively and all models have high $R^2$ especially the conditional is positively affected by the random slope for $m_3$. All 8 models'

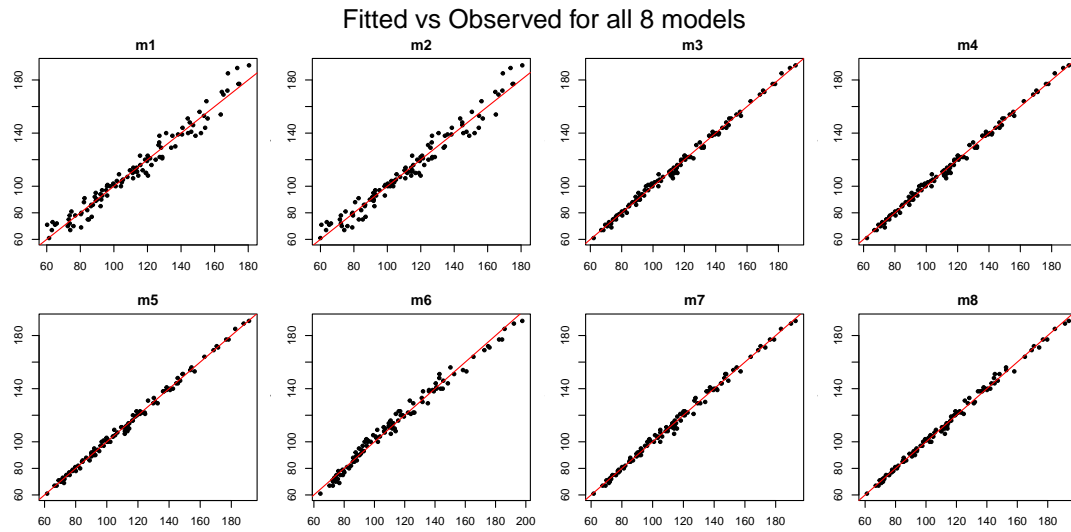fitted vs observed are illustrated in Figure 9.



**Figure 9:** *Fitted vs Observed*

The models $m_5$ and $m_8$ have the tightest alignment to the observations, however simple random slope $m_3$ performs comparably. Therefore **the recommended model is** $m_3$. The diagnostics plot for $m_3$.
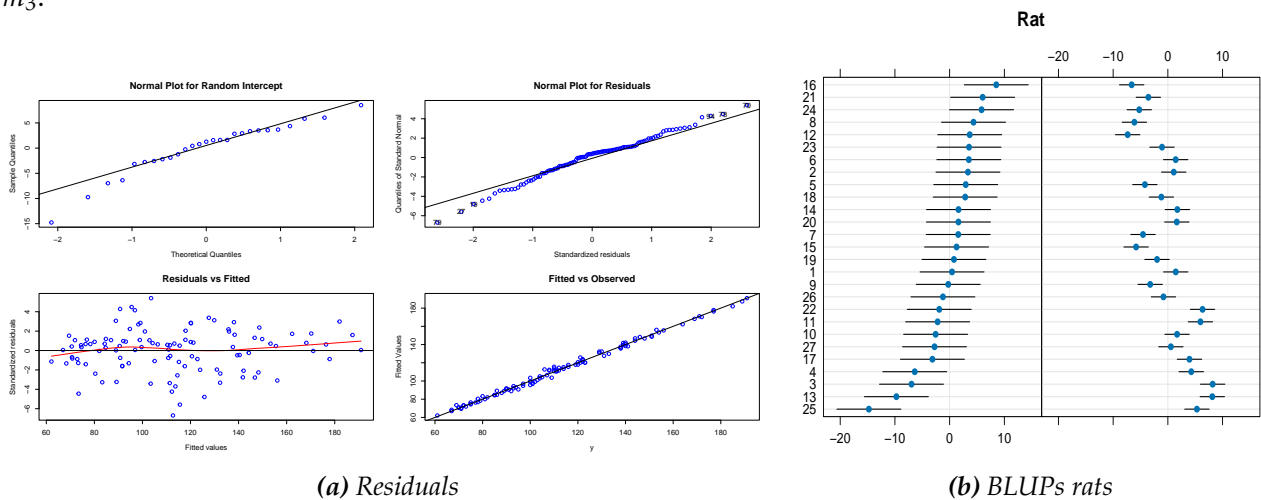


*(a) Residuals*



*(b) BLUPs rats*

**Figure 10:** *Model diagnostics*

The residuals vs fitted in Figure 10a show some funnel behavior and some slight curvature. Overall the normality assumptions holds for residuals and random effects. With random intercept and random slope shown in Figure 10b. Overall the normality assumption holds, the intercepts ranges roughly between [-20,15], the heavier lower tail for intercept indicate some rats' trajectories lie below average and their positive random slope indicate a higher growth rate. Conversely for the rats that start at higher baseline weight, they grow at a slower pace than treatment average, the slopes are roughly balanced ranging [-10,10].

8

The measurements that are closer in time for each rat theoretically should be more correlated than those further apart. However with 4 week measurements, there are only 3 distances. Also as the random slope is significant, the time dependence for the individual rats/within treatment is already in the model, which may explain why it is not possible to estimate correlation structures from $m_3$. Variograms for random intercept only, $m_1$ base, can be seen in Appendix B.3. When comparing the random intercept models with new correlations structures with the former models the AIC, BIC are similar.

Since model $m_3$ fit the data structures well and is easily interpreted, it is used for inference, the model parameters and treatment specific equations in Table 7.

*Table 7: $m_3$ parameter estimates with 95% CI*

| Parameter | Estimate | 2.5% | 97.5% |
|-----------|----------|------|-------|
| Intercept | -33.09 | -54.81 | -11.34 |
| week | 27.04 | 23.96 | 30.12 |
| Trt2 | -7.51 | -14.14 | -0.85 |
| Trt3 | 9.91 | 3.90 | 15.92 |
| y0 | 1.56 | 1.17 | 1.96 |
| week:Trt2 | 2.42 | -2.39 | 7.22 |
| week:Trt3 | -11.49 | -15.85 | -7.13 |

*Table 8: Trt-specific fixed regression equations*

Trt 1:  $\hat{y} = -33.09 + 27.04\,\text{week} + 1.56\,y_0$,

Trt 2:  $\hat{y} = -40.60 + 29.46\,\text{week} + 1.56\,y_0$,

Trt 3:  $\hat{y} = -23.18 + 15.55\,\text{week} + 1.56\,y_0$.

This indicate that the rats for treatment 2 starts lower than control and treatment 3 higher. The slopes indicate that control and treatment 2 grows at similar rate whereas treatment 3 is slower.

## 2.3   Results

The estimated marginal means for the treatments in Table 9 and plotted in Figure 11

| Trt | Week | EMMean | 95% CI |
|-----|------|--------|--------|
| **Week = 1.0** | | | |
| 1 | 1.0 | 79.3 | [76.1, 82.4] |
| 2 | 1.0 | 74.2 | [70.4, 77.9] |
| 3 | 1.0 | 77.7 | [74.6, 80.8] |
| **Week = 2.5** | | | |
| 1 | 2.5 | 119.8 | [114.0, 125.6] |
| 2 | 2.5 | 118.4 | [111.4, 125.3] |
| 3 | 2.5 | 101.0 | [95.2, 106.8] |
| **Week = 4.0** | | | |
| 1 | 4.0 | 160.4 | [150.0, 170.8] |
| 2 | 4.0 | 162.6 | [150.2, 174.9] |
| 3 | 4.0 | 124.3 | [114.0, 134.7] |

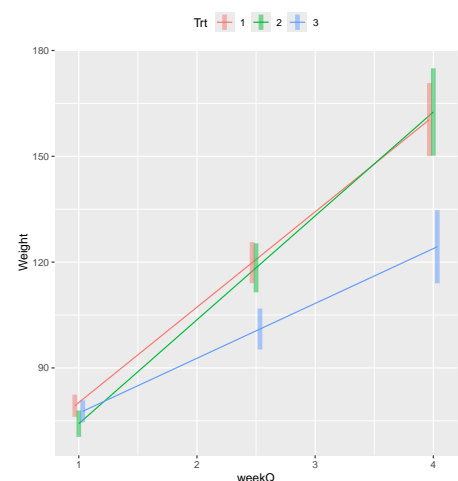*Table 9: Estimated marginal means (m3)*



*Figure 11: Interaction plot of mean weight by week and treatment*

9

The emmeans and trajectories indicate that treatment 3, thiouracil, differs from the control but treatment 2 lies almost directly aligned with the control. The slopes are tested to see if the treatments differ. in Table 10 we see the growth rate for control in $27\frac{g}{week}$ and for thyroxin it is $29.5\frac{g}{week}$ and lastly thiouracil $15.6\frac{g}{week}$.

| Trt | Slope | SE | 95% CI | $p$ |
|---|---|---|---|---|
| 1 | 27.0 | 1.61 | [23.7, 30.4] | < 0.0001 |
| 2 | 29.5 | 1.92 | [25.5, 33.4] | < 0.0001 |
| 3 | 15.6 | 1.61 | [12.2, 18.9] | < 0.0001 |

*Table 10: Estimated slopes (model $m_3$)*

| Contrast | Est. | SE | 95% CI | $p$ |
|---|---|---|---|---|
| 1–2 | -2.42 | 2.51 | [−8.68, 3.85] | 0.6062 |
| 1–3 | 11.49 | 2.28 | [ 5.81, 17.17] | 0.0001 |
| 2–3 | 13.91 | 2.51 | [ 7.64, 20.17] | < 0.0001 |

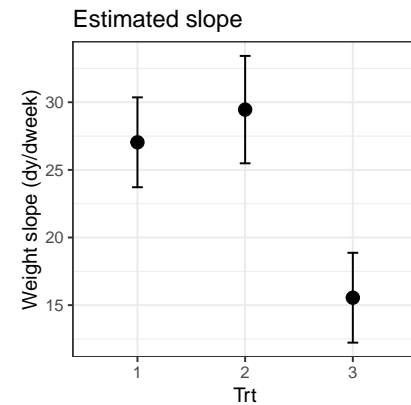*Table 11: Pairwise slope differences (Tukey-adjusted)*



*Figure 12: Estimated slopes by treatment*

From Table 11 it can be concluded that the difference in slopes between; 1 (control) and 2 (thyroxin) is insignificant, but difference between 1 (control) and 3 (thiouracil) is significant and also between the two treatments. The estimated difference between 1 and 3 can be interpreted as **rats with treatment 3 grows** $11.5\frac{g}{week}$ **less than rats with control treatment**. R code can be found in Appendix B.4

## 2.4 Conclusion

Based on the model comparison, the simpler random–slope model $m3$ was recommended since the model captured the main structure in the data, and performed comparably to more complex models while still pertaining simplicity and interpretability. The post hoc analysis showed that treatment 2 (thyroxin) does not differ from the control treatment, whereas treatment 3 (thiouracil) exhibits a lower weekly growth rate. The estimated marginal means confirm that thiouracil–treated rats has lower weight growth and exhibited lower mean weight from week 2 onward. The slope differences indicate that these rats grow approximately $11.5\frac{g}{week}$ less than rats in the control group. Thus, the primary treatment effect presents through a reduced growth trajectory for treatment 3, while treatment 2 shows a response pattern similar to the control.

# Appendix A: Radon

All R code can also be found in `https://github.com/MichellaRavn/02429FinalAssignment`

## A.1   Read and prepare data

```r
# Libraries
library(sf)
library(tigris)
library(ggplot2)
library(patchwork)
library(lmerTest)
library(MASS)
library(lattice)
library(performance)
library(predictmeans) #residplot
library(emmeans)

# Load the data
df<-read.table("Data/Radon_MN.csv",header = T)

# Cleaning the data
df$county.name <- toupper(trimws(df$county.name))
df$county.name[df$county.name == "ST LOUIS"] <- "ST. LOUIS"
names(df)[names(df) == "u.full"] <- "u"
df$x <- as.factor(df$x)
df$county <- as.factor(df$county)
```

## A.2   Exploratory analysis

```r
# Geographic placement of the counties
SW<-c("ROCK", "NOBLES","JACKSON","COTTONWOOD","MURRAY","PIPESTONE","LINCOLN","LYON",
    "REDWOOD","YELLOW MEDICINE")
W<-c("LAC QUI PARLE","CHIPPEWA","SWIFT","BIG STONE","STEVENS","WILKIN","TRAVERSE","GRANT",
    "CLAY","OTTER TAIL","BECKER", "POPE","DOUGLAS") #13
NW<-c("NORMAN","MAHNOMEN","POLK","RED LAKE","PENNINGTON","MARSHALL","KITTSON","ROSEAU",
    "CLEARWATER") # 9
N<-c("LAKE OF THE WOODS","KOOCHICHING","BELTRAMI","ITASCA") # 4
NE<-c("ST. LOUIS","LAKE","COOK") # 3
E<-c("CARLTON","PINE","CHISAGO","ISANTI","AITKIN","KANABEC","MILLE LACS","CHISAGO",
    "BENTON","SHERBURNE","ANOKA","WASHINGTON","RAMSEY","HENNEPIN","WRIGHT","CARVER")
SE<-c("DAKOTA","GOODHUE","WABASHA","WINONA","HOUSTON","FILLMORE","OLMSTED","MOWER",
    "FREEBORN","DODGE","STEELE","RICE","WASECA","SCOTT","LE SUEUR")
S<-c("MARTIN","FARIBAULT","WATONWAN","BLUE EARTH","BROWN","NICOLLET","SIBLEY","RENVILLE")
M<-c("KANDIYOHI","MCLEOD","MEEKER","STEARNS","TODD","MORRISON","CROW WING","WADENA",
    "CASS","HUBBARD")

regions <- list(
  SW = SW, W = W, NW = NW,
  N = N, NE = NE, E = E,
  SE = SE, S = S, M = M)

# Assigning region to each county
lookup <- rep(names(regions), sapply(regions, length))
names(lookup) <- unlist(regions)
df$region <- lookup[df$county.name]

# Storing color values
# Blue scale (north)
```

```r
51 col_NW <- "#A6CEE3"    # light blue
52 col_N  <- "#1F78B4"    # medium blue
53 col_NE <- "#0B3C68"    # dark blue
54
55 # Green scale (middle)
56 col_W <- "#B2DF8A"     # light green
57 col_M <- "#33A02C"     # medium green
58 col_E <- "#145214"     # dark green
59
60 # Orange scale (south)
61 col_SW <- "#FDBF6F"    # light orange
62 col_S  <- "#FF7F00"    # medium orange
63 col_SE <- "#B15928"    # dark orange
64
65 # Combine into named vector
66 region_colors <- c(
67   NW = col_NW,
68   N  = col_N,
69   NE = col_NE,
70   W  = col_W,
71   M  = col_M,
72   E  = col_E,
73   SW = col_SW,
74   S  = col_S,
75   SE = col_SE)
76
77 region.order <- c("NW", "N", "NE", "W", "M", "E", "SW", "S", "SE")
78
79 # Provided code:
80 # Counties
81 (J <- length(unique(df$county.name)))
82
83 # Mean of response
84 (ybarbar = mean(df$y))
85
86 (sample.size <- as.vector(table(df$county)))
87 (sample.size.jittered <- sample.size*exp(runif (J, -.1, .1)))
88 (cty.mns = tapply(df$y,df$county,mean))
89 (cty.vars = tapply(df$y,df$county,var))
90 (cty.sds = mean(sqrt(cty.vars[!is.na(cty.vars)]))/sqrt(sample.size))
91 (cty.sds.sep = sqrt(tapply(df$y,df$county,var)/sample.size))
92
93 # Creating a map plot
94 options(tigris_use_cache = TRUE)
95
96 mn_map <- counties(state = "MN", year = 2020)
97 mn_map <- st_as_sf(mn_map)
98 mn_map$NAME <- toupper(trimws(mn_map$NAME))
99
100 names(mn_map)[names(mn_map) == "NAME"] <- "county.name"
101
102 county_region_map <- unique(df[, c("county.name", "region")])
103 county_number_map <- unique(df[, c("county.name", "county")])
104
105 mn_map_joined <- merge(
106   mn_map,
107   merge(county_region_map, county_number_map, by="county.name"),
108   by = "county.name",
109   all.x = TRUE
```

```
110 )
111
112 mn_map_joined$region <- factor(mn_map_joined$region,
113                                levels = region.order)
114
115 centroids <- st_centroid(mn_map_joined)
116 highlight <- centroids[centroids$county %in% c(36, 37), ]
117
118 ggplot() +
119   geom_sf(data = mn_map_joined, aes(fill = region), color = "white", size = 0.2) +
120
121   # Min and max measurements:
122   #geom_sf(
123   #  data = highlight,
124   #  color = "black",
125   #  size = 5,
126   #  shape = 21,
127   #  stroke = 1.2
128   #) +
129
130   # Uncomment for numbers added
131   geom_sf_text(
132     data = centroids,
133     aes(label = county),
134     size = 3,
135     color = "black"
136   ) +
137
138   scale_fill_manual(values = region_colors,
139                     breaks = region.order ) +
140   theme_void()
```
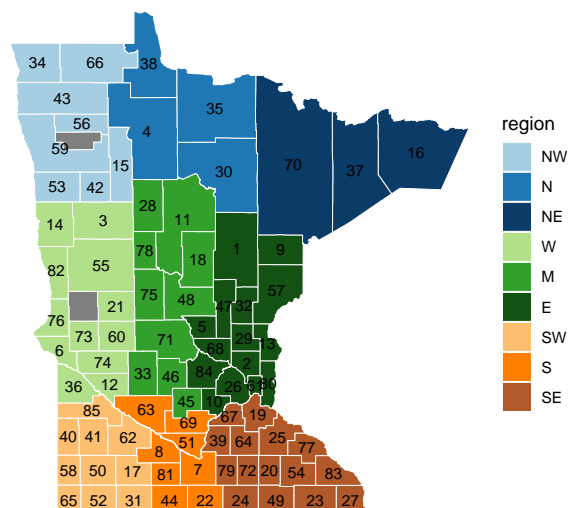


*Figure 13: Minnesota Conties with numbers*

```r
141 # Plotting means
142 # County names and regions aligned to tapply indexing for plotting
143 county_names <- unique(df$county.name)#tapply(df£county.name, df£county, function(x) x[1])
144 county_regions <- tapply(df$region, df$county, function(x) x[1])
145
146 plot_df <- data.frame(
147   county = county_names,
148   region = factor(county_regions, levels = region.order),
149   mean_y = as.numeric(cty.mns),
150   sd_y   = as.numeric(cty.sds.sep))
151 plot_df$lower <- plot_df$mean_y - plot_df$sd_y
152 plot_df$upper <- plot_df$mean_y + plot_df$sd_y
153 plot_df <- plot_df[order(plot_df$region, plot_df$mean_y), ]
154 plot_df$county.order <- factor(plot_df$county, levels = plot_df$county)
155
156
157 ggplot(plot_df, aes(x = county.order, y = mean_y, color = region)) +
158
159   # Error bars first (behind points)
160   geom_errorbar(aes(ymin = lower, ymax = upper),
161                 width = 0.0,        # no horizontal bar
162                 color = "black",    # or same color as points if preferred
163                 alpha = 0.6,        # slightly transparent
164                 linewidth = 0.5) +
165
166   # Points on top
167   geom_point(size = 3) +
168
169   scale_color_manual(values = region_colors) +
170   labs(
171     title = "Observed mean radon per county",
172     x = "County",
173     y = "Mean log radon"
174   ) +
175   theme_bw() +
176   theme(
177     axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)
178   )
```

```r
179 # The given plot about samplesize with a few alterations
180 par(mfrow=c(1,1))
181
182 # First an empty plot
183 plot(sample.size.jittered, cty.mns,
184      type = "n",
185      xlab="sample size in county j",
186      ylab="Mean log radon in county j",
187      log="x", cex.lab=.9, cex.axis=1, mgp=c(1.5,.5,0),
188      ylim=c(0,3.2), yaxt="n", xaxt="n")
189
190 axis(1, c(1,3,10,30,100), cex.axis=.9, mgp=c(1.5,.5,0))
191 axis(2, seq(0,3), cex.axis=.9, mgp=c(1.5,.5,0))
192
193 # Draw error bars
194 for (j in 1:J){
195   lines(rep(sample.size.jittered[j],2),
196         cty.mns[j] + c(-1,1)*cty.sds[j], lwd=.5)
197 }
198
```

```r
199 # Adding points after to have error bars behind
200 points(sample.size.jittered, cty.mns,
201       pch = 20, cex = 1.3,
202       col = region_colors[ df$region[ match(names(cty.mns), df$county) ] ])
203
204 # Horizontal mean line
205 abline(h = mean(cty.mns), lwd = .5)
206
207 title("Observed mean response values per county", cex.main=.9, line=1)
208
209 # The max and min
210 cty.mns[cty.mns==max(cty.mns)]
211 cty.mns[cty.mns==min(cty.mns)]
212
213 county_names[cty.mns==max(cty.mns)] # Lac Qui Parle
214 county_names[cty.mns==min(cty.mns)] # Lake
215
216 # highlight max/min if needed
217 points(sample.size.jittered[36], cty.mns[36], cex=2)
218 points(sample.size.jittered[37], cty.mns[37], cex=2)
```

```r
219 ## Uranium plots
220 floor_colors <- c(
221   "1" = "gold2",    # ground floor
222   "0" = "grey37"      # basement
223 )
224
225 # Plot 1: Region
226 p1 <- ggplot(df, aes(x = u, y = y, color = region)) +
227   geom_point(alpha = 0.7, size = 2) +
228   geom_smooth(method = "lm", se = FALSE, linewidth = 1) +
229   scale_color_manual(values = region_colors, breaks=region.order) +
230   labs(
231     title = "Radon vs uranium - by region",
232     x = "log uranium (u)",
233     y = "log radon (y)"
234   ) +
235   theme_bw()
236
237 # Plot 2: Floor
238 p2 <- ggplot(df, aes(x = u, y = y, color = factor(x))) +
239   geom_point(alpha = 0.7, size = 2) +
240   geom_smooth(method = "lm", se = FALSE, linewidth = 1) +
241   scale_color_manual(values = floor_colors, name = "Floor (x)") +
242   labs(
243     title = "Radon vs uranium - by floor",
244     x = "log uranium (u)",
245     y = "log radon (y)"
246   ) +
247   theme_bw()
248
249 # Combine
250 p1 | p2
251
252 # Investigate imbalance
253 table(df$x)
254
255 ## Investigate floor status
256 p_strip <- ggplot(df, aes(x = x, y = y, color = x)) +
```

```
257   geom_jitter(width = 0.15, alpha = 0.6, size = 1.8) +
258   stat_summary(fun = mean, geom = "point",
259                size = 3, shape = 21, fill = "black", color = "black") +
260   scale_color_manual(values = floor_colors) +
261   labs(
262     title = "Log radon - by floor",
263     x = "Floor (x)",
264     y = "Log radon (y)"
265   ) +
266   theme_bw() +
267   theme(
268     axis.text.x = element_text(angle = 0, hjust = 0.5),
269     legend.position = "none"
270   )
271
272 p_box <- ggplot(df, aes(x = x, y = y, fill = x)) +
273   geom_boxplot(alpha = 0.6, outlier.shape = NA) +
274   scale_fill_manual(values = floor_colors) +
275   labs(
276     title = "Log radon { boxplot by floor",
277     x = "Floor (x)",
278     y = "Log radon (y)"
279   ) +
280   theme_bw() +
281   theme(
282     axis.text.x = element_text(angle = 0, hjust = 0.5),
283     legend.position = "none"
284   )
285
286 p_box / p_strip
```

## A.3   Model selection w. output

```
287 m0 <- lmer(y ~ u:x + u + x + (1|county), data=df, REML=TRUE)
288 ranova(m0)
```

```
ANOVA-like table for random-effects: Single term deletions

Model:
y ~ u + x + (1 | county) + u:x
             npar  logLik    AIC    LRT Df Pr(>Chisq)
<none>          6 -1065.5 2142.9
(1 | county)    5 -1069.5 2148.9 8.0178  1   0.004632 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
10 m0 <- update(m0,REML=F)
11 drop1(m0) # interaction significant
```

```
Single term deletions using Satterthwaite's method:

Model:
y ~ u:x + u + x + (1 | county)
    Sum Sq Mean Sq NumDF  DenDF F value Pr(>F)
u:x 2.7474  2.7474     1 894.84  4.8008 0.0287 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing random slope

```
9 m_rs<-lmer(y ~ u:x + u + x + ( 1 + x | county), data = df)
10 ranova(mrs) # random intercept not significant
```

```
ANOVA-like table for random-effects: Single term deletions

Model:
y ~ u + x + (1 + x | county) + u:x
                      npar  logLik    AIC     LRT Df Pr(>Chisq)
<none>                   8 -1066.1 2148.2
x in (1 + x | county)   6 -1065.5 2142.9 -1.3294  2          1
```

```
8 residplot(m0)
9 AIC(m0)
10 BIC(m0)
11 r2_nakagawa(m0)
12
13 var_county   <- as.numeric(VarCorr(m0)$county)         # 9.323
14 var_resid <- sigma(m0)^2                                # 22.875
15 icc <- var_county / (var_county + var_resid)
16 icc
17 summary(m0)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: y ~ u * x + u + x + (1 | county)
   Data: df

REML criterion at convergence: 2130.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.0214 -0.6273  0.0235  0.6222  3.4198

Random effects:
 Groups   Name        Variance Std.Dev.
 county   (Intercept) 0.02381  0.1543
 Residual             0.57315  0.7571
Number of obs: 919, groups:  county, 85

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)   1.47409    0.03794  57.64818  38.852  < 2e-16 ***
u             0.79718    0.09773  54.00771   8.157  5.5e-11 ***
x1           -0.70959    0.07126 909.47589  -9.958  < 2e-16 ***
u:x1         -0.41669    0.19092 890.08581  -2.183   0.0293 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
     (Intr) u      x1
u     0.173
x1   -0.370 -0.104
u:x1 -0.104 -0.358  0.268
```

```
32 cd <- cooks.distance(m0)
33 plot(cd, type = "h", main = "Cook's distance", xlab = "Observation", ylab = "Distance")
34 df[cd>0.2,c("y","pred","county","county.name","region","u","x")]
```

***Table 12:*** *Observations with Cook's distance greater than 0.2*

| Observation | y | Pred | County | County name | Region | u | x |
|---|---|---|---|---|---|---|---|
| 145 | −2.30 | 0.84 | 17 | COTTONWOOD | SW | 0.34 | 1 |
| 239 | −2.30 | 1.50 | 22 | FARIBAULT | S | 0.30 | 0 |

County 17, Cottonwood, and 22, Faribault have quite low observed radon concentrations, y, but are predicted at a higher concentration. A further analysis can be made with removing these and seeing the new model fit.

```
35 # Random effect
36 dotplot(ranef(m0, condVar = TRUE),
37         strip = FALSE,
38         ylab = "",
39         scales = list(cex = 0.5))   # smaller labels


40 ## Model parameters
41 # Raw
42 est  <- fixef(m0)
43 ci   <- confint(m0, parm = names(est))  # 2.5% and 97.5%
44
45 # Back-transform (exp)
46 est_bt <- exp(est)
47 ci_bt  <- exp(ci)
48
49 #--- Combine into table ---
50 param_table <- data.frame(
51   Parameter = names(est),
52   Estimate_raw = round(est, 2),
53   CI_low_raw = round(ci[,1], 2),
54   CI_high_raw = round(ci[,2], 2),
55   Estimate_bt = round(est_bt, 2),
56   CI_low_bt = round(ci_bt[,1],2),
57   CI_high_bt = round(ci_bt[,2], 2)
58 )
59
60 param_table
```

## A.4   Results

```
61 ### Emmeans
62 # 25%, 50%, 75% quantiles
63 u_vals <-quantile(df$u, probs = c(0.25, 0.50, 0.75))
64
65 # emmeans table
66 em <- emmeans(m0, "x", by = "u", at = list(u = u_vals),adjust="tukey")
67 em
68
69 # Back-transform to radon scale
70 em_bt <- transform(
71   as.data.frame(em),
72   em_bt      = exp(emmean),
73   lower_bt   = exp(lower.CL),
74   upper_bt   = exp(upper.CL)
```

```r
75 )
76 em_bt
77
78 emmip(m0,x ~ u, at = list(u = u_vals) , CIs = TRUE) + theme(legend.position="top")+
79   #ggtitle("Emmeans interaction")+
80   ylab("Log radon") +
81   labs(color="Floor") +
82   scale_color_manual(values=floor_colors)
83
84 emtrends(m0, pairwise ~x, var="u",infer=TRUE,adjust="tukey")
85 # low change in radon c for higher val of u,
86 # significant slopes, strong evidence for x=0
87 # significant difference
88
89 slopes <- emtrends(m0, ~ x, var = "u")   # dy/du
90 slopes_df <- as.data.frame(slopes)
91
92 ggplot(slopes_df, aes(x = x, y = u.trend)) +
93   geom_point(size = 3) +
94   geom_errorbar(aes(ymin = lower.CL, ymax = upper.CL), width = 0.1) +
95   labs(
96     title = "Estimated slope",
97     x = "Floor (x)",
98     y = "u slope (dy/du)"
99   ) +
100   theme_bw()
```

```r
101 ### Estimated intercept
102 # Random effects with conditional variance
103 re   <- ranef(m0, condVar = TRUE) # condVar to get posterior variance
104
105 # random intercept deviations
106 ri   <- re$county
107 ri$county <- rownames(ri)
108 u_cty <- tapply(df$u, df$county, mean)    # same value of u in each county
109 ri$u <- u_cty[ri$county]
110
111
112 # intercept params
113 beta0 <- fixef(m0)["(Intercept)"]
114 alpha<- fixef(m0)["u"]
115
116 # county-specific regression intercepts = mu + alpha * u + b0_j
117 ri$intercept <- beta0 + ri[,"(Intercept)"] + alpha * ri$u
118
119 # extract posterior variances, store standard error
120 post_var <- attr(ranef(m0, condVar = TRUE)$county, "postVar")[1,1,]
121 ri$SE <- sqrt(post_var)
122 ri$lower <- ri$intercept - qnorm(1-(0.05/2)) * ri$SE
123 ri$upper <- ri$intercept + qnorm(1-(0.05/2)) * ri$SE
124
125 # adding region for coloring
126 ri$region <- df$region[ match(ri$county, df$county) ]
127 ri$region <- factor(ri$region, levels = region.order)
128
129 ggplot(ri, aes(x = u, y = intercept, color = region)) +
130
131   # error bars first (behind points)
132   geom_errorbar(aes(ymin = lower, ymax = upper), width = 0, alpha = 0.6) +
```

```
133
134   # colored points
135   geom_point(size = 2) +
136
137   geom_abline(
138     intercept = beta0, slope = alpha,
139     color     = "black", alpha=0.4, linewidth = 0.2) +
140
141   scale_color_manual(values = region_colors, name = "Region") +
142
143   labs(
144     title = "Estimated intercepts for counties, x=0",
145     x = "County uranium (u)",
146     y = "Estimated intercept (log radon)"
147   ) +
148   theme_bw()
```

```
149 # Predictions
150 # Mixed predictions
151 df$m0 <- predict(m0)
152 # Fixed predictions E(r_i)=0
153 df$m0_fix <- predict(m0,re.form=NA)
```

# Appendix B: Rats

## B.1   Read and prepare data

```
154 library(ggplot2)
155 library(patchwork)
156 library(lmerTest)
157 library(MASS)
158 library(lattice)
159 library(predictmeans)
160 library(plyr)
161 library(dplyr)
162 library(gridExtra)
163 library(emmeans)
164
165 # Load the data
166 rats <- read.table("Data/Rats_Box.txt", header = TRUE, sep = ",")
167 treat_names <- c("Control", "Thyroxin", "Thioracil")
168
169 # Finding weight parameter, cumulative weight gain
170 rats <- within(rats, {
171   y1 <- y0 + y1
172   y2 <- y1 + y2
173   y3 <- y2 + y3
174   y4 <- y3 + y4
175 })
176
177 # Reshaping to long format
178 rats_long <- reshape(
179   rats,
180   varying = list(c("y1","y2","y3","y4")),
181   v.names = "y",
182   timevar = "week",
183   times = 1:4,
184   direction = "long"
185 )
186
187 rats_long <- rats_long[order(rats_long$Rat, rats_long$week), ]
188 row.names(rats_long) <- NULL
189
190 # Now storing the correct format
191 rats <- rats_long
192
193 # make treatment and cage factors
194 rats$Trt <- as.factor(rats$Trt)
195 rats$Rat <- as.factor(rats$Rat)
196 rats$y <- as.numeric(rats$y)
197
198 # Data structure
199 str(rats)
200
201 rats$weekQ <- as.numeric(rats$week)
202 rats$weekF <- as.factor(rats$week)
```

## B.2   Exploratory analysis

```
203 # for plotting
204 baseline <- rats |>
```

```
205   distinct(Rat, Trt, y0) |>
206   mutate(
207     weekQ = 0,
208     weekF = factor(0),
209     y     = y0
210   )
211
212 rats_plot <- bind_rows(
213   rats[, c("Rat", "Trt", "weekQ", "weekF", "y")],
214   baseline[, c("Rat", "Trt", "weekQ", "weekF", "y")]
215 )
216
217 # Individual profiles
218 p1 <- ggplot(rats_plot, aes(x = weekQ, y = y, group = Rat, colour = Trt)) +
219   geom_line(alpha = 0.8) +
220   geom_point(size = 0.7) +
221   labs(
222     title = "Weight over time",
223     x = "Week",
224     y = "Weight (gram)"
225   ) +
226   theme_bw()
227
228 # Mean profiles
229 mns <- rats_plot |>
230   group_by(Trt, weekQ) |>
231   summarise(y = mean(y), .groups = "drop")
232
233 p2 <- ggplot(mns, aes(x = weekQ, y = y, group = Trt, colour = Trt)) +
234   geom_line(size = 1) +
235   geom_point(size = 2) +
236   labs(
237     title = "Mean weight per treatment",
238     x = "Week",
239     y = "Mean weight (gram)"
240   ) +
241   theme_bw()
242 dev.off()   #
243
244 p1 / p2
```

## B.3   Model selection w. output

```
245 # models
246 m1 <- lmer(y ~ weekQ + Trt + weekQ:Trt + y0 + (1 | Rat), data = rats)
247 m2 <- lmer(y ~ weekF + Trt + weekF:Trt + y0 + (1 | Rat), data = rats)
248 m3 <- lmer(y ~ weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat), data = rats)
249 m4 <- lmer(y ~ I(weekQ^2) + weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat), data = rats)
250 m5 <- lmer(y ~ I(weekQ^3) + I(weekQ^2) + weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat), data = rats)
```

```
251 ranova(m1)
```

```
ANOVA-like table for random-effects: Single term deletions

Model:
y ~ weekQ + Trt + y0 + (1 | Rat) + weekQ:Trt
          npar  logLik   AIC    LRT Df Pr(>Chisq)
<none>       9 -373.63 765.27
```

```
(1 | Rat)     8 -394.91 805.83 42.564   1   6.84e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10 `ranova(m3)`
```
ANOVA-like table for random-effects: Single term deletions

Model:
y ~ weekQ + Trt + y0 + (1 + weekQ | Rat) + weekQ:Trt
                            npar  logLik    AIC    LRT Df Pr(>Chisq)
<none>                        11 -325.95 673.91
weekQ in (1 + weekQ | Rat)     9 -373.63 765.27 95.357  2  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10 `ranova(m1)`
11 `ranova(m2)`
12 `ranova(m3)`
13 `ranova(m4)`
14 `ranova(m5)`
15
16 `anova(m1,m2)` *# prefers m1, time as a numeric*
```
refitting model(s) with ML (instead of REML)
Data: rats
Models:
m1: y ~ weekQ + Trt + weekQ:Trt + y0 + (1 | Rat)
m2: y ~ weekF + Trt + weekF:Trt + y0 + (1 | Rat)
   npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
m1    9 782.58 806.72 -382.29    764.58
m2   15 790.81 831.04 -380.40    760.81 3.7754  6      0.707
```

9 `anova(m1,m3)` *# prefers m3, random slope*
```
refitting model(s) with ML (instead of REML)
Data: rats
Models:
m1: y ~ weekQ + Trt + weekQ:Trt + y0 + (1 | Rat)
m3: y ~ weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat)
   npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
m1    9 782.58 806.72 -382.29    764.58
m3   11 689.17 718.67 -333.58    667.17 97.418  2  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

11 `anova(m1,m4)` *# quadratic is significant*
```
refitting model(s) with ML (instead of REML)
Data: rats
Models:
m1: y ~ weekQ + Trt + weekQ:Trt + y0 + (1 | Rat)
m4: y ~ I(weekQ^2) + weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat)
   npar    AIC    BIC  logLik -2*log(L)   Chisq Df Pr(>Chisq)
m1    9 782.58 806.72 -382.29    764.58
m4   12 691.16 723.35 -333.58    667.16 97.419  3  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
11 anova(m1,m5) # prefer m5, cubic - consider complexity
```

```
refitting model(s) with ML (instead of REML)
Data: rats
Models:
m1: y ~ weekQ + Trt + weekQ:Trt + y0 + (1 | Rat)
m5: y ~ I(weekQ^3) + I(weekQ^2) + weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat)
   npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
m1    9 782.58 806.72 -382.29    764.58
m5   13 682.82 717.69 -328.41    656.82 107.76  4  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
11 m3 <- update(m3,REML=F)
12 drop1(m3) # Interaction significant
```

```
Single term deletions using Satterthwaite's method:

Model:
y ~ weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat)
          Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
y0        549.99  549.99     1    27  65.231 1.120e-08 ***
weekQ:Trt 368.27  184.14     2    27  21.839 2.281e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
10 m3_lm <- lm(y ~ weekQ + Trt + weekQ:Trt + y0, data = rats)
11 m5_lm <- lm(y ~ I(weekQ^3) + I(weekQ^2) + weekQ + Trt + weekQ:Trt + y0, data = rats)
12
13 # Checking for transformation (no indication, but making sure)
14 aux <- boxcox(m5_lm, lambda = seq(-1, 2, by = 0.05))
15 (lambda<-aux$x[which.max(aux$y)]) # 0 is in the CI, 1 is not = log transform
16
17 # Log transforming
18 rats$logy <- log(rats$y)
19 rats$logy0 <- log(rats$y0)
20
21
22 m6 <- lmer(logy ~ weekQ + Trt + weekQ:Trt + logy0 + (1 + weekQ | Rat), data = rats)
23 m7 <- lmer(logy ~ weekQ^2 + weekQ + Trt + weekQ:Trt + logy0 + (1 + weekQ | Rat), data = rats)
24 m8 <- lmer(logy ~ weekQ^3+weekQ^2 + weekQ + Trt + weekQ:Trt + logy0 + (1 + weekQ | Rat), data = rats)
25
26 anova(m4,m5)
```

```
refitting model(s) with ML (instead of REML)
Data: rats
Models:
m4: y ~ I(weekQ^2) + weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat)
m5: y ~ I(weekQ^3) + I(weekQ^2) + weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat)
   npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
m4   12 691.16 723.35 -333.58    667.16
m5   13 682.82 717.69 -328.41    656.82 10.341  1   0.001301 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
11 anova(m5,m6) # prefer m6
```

```
refitting model(s) with ML (instead of REML)
Data: rats
Models:
m6: logy ~ weekQ + Trt + weekQ:Trt + logy0 + (1 + weekQ | Rat)
m5: y ~ I(weekQ^3) + I(weekQ^2) + weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat)
   npar     AIC     BIC  logLik -2*log(L) Chisq Df Pr(>Chisq)
m6   11 -261.58 -232.08  141.79   -283.58
m5   13  682.82  717.69 -328.41    656.82     0  2           1
```

```
 9 AIC(m1,m2,m3,m4,m5,m6,m7,m8)
10 BIC(m1,m2,m3,m4,m5,m6,m7,m8)
11
12 r2_nakagawa(m1)
13 r2_nakagawa(m2)
14 r2_nakagawa(m3)
15 r2_nakagawa(m4)
16 r2_nakagawa(m5)
17 r2_nakagawa(m6)
18 r2_nakagawa(m7)
19 r2_nakagawa(m8)
```

```
20 # Comparing fitted vs observed
21 par(mfrow=c(2,4), mar=c(3,3,2,1), oma=c(0,0,3,0))
22
23 plot(predict(m1, level=1), rats$y, pch=20,
24     main="m1 ")
25 abline(0,1, col="red")
26
27 plot(predict(m2, level=1), rats$y, pch=20,
28     main="m2")
29 abline(0,1, col="red")
30
31 plot(predict(m3, level=1), rats$y, pch=20,
32     main="m3 ")
33 abline(0,1, col="red")
34
35 plot(predict(m4, level=1), rats$y, pch=20,
36     main="m4")
37 abline(0,1, col="red")
38
39 plot(predict(m5, level=1), rats$y, pch=20,
40     main="m5")
41 abline(0,1, col="red")
42
43 plot(exp(predict(m6, level=1)), rats$y, pch=20,
44     main="m6")
45 abline(0,1, col="red")
46
47 plot(exp(predict(m7, level=1)), rats$y, pch=20,
48     main="m7")
49 abline(0,1, col="red")
50
51 plot(exp(predict(m8, level=1)), rats$y, pch=20,
52     main="m8")
53 abline(0,1, col="red")
54
```

```
55 # Add outer title
56 mtext("Fitted vs Observed for all 8 models", outer=TRUE, cex=1.5)
57
```

```
58 residplot(m1) # looks like adding curvature is needed
59 residplot(m2)
60 residplot(m3) # funnel shaped
61 residplot(m4) # almost identical to m3
62 residplot(m5) # slightly more flat res vs fit
63 # Log models
64 residplot(m6)
65 residplot(m7)
66 residplot(m8)
```

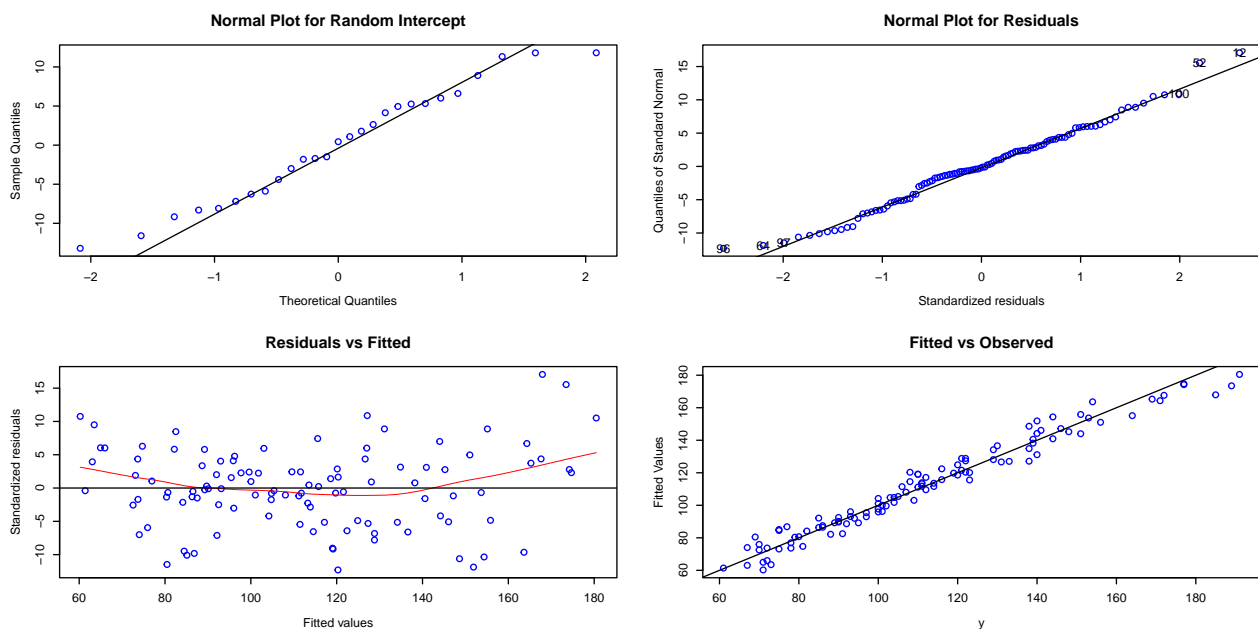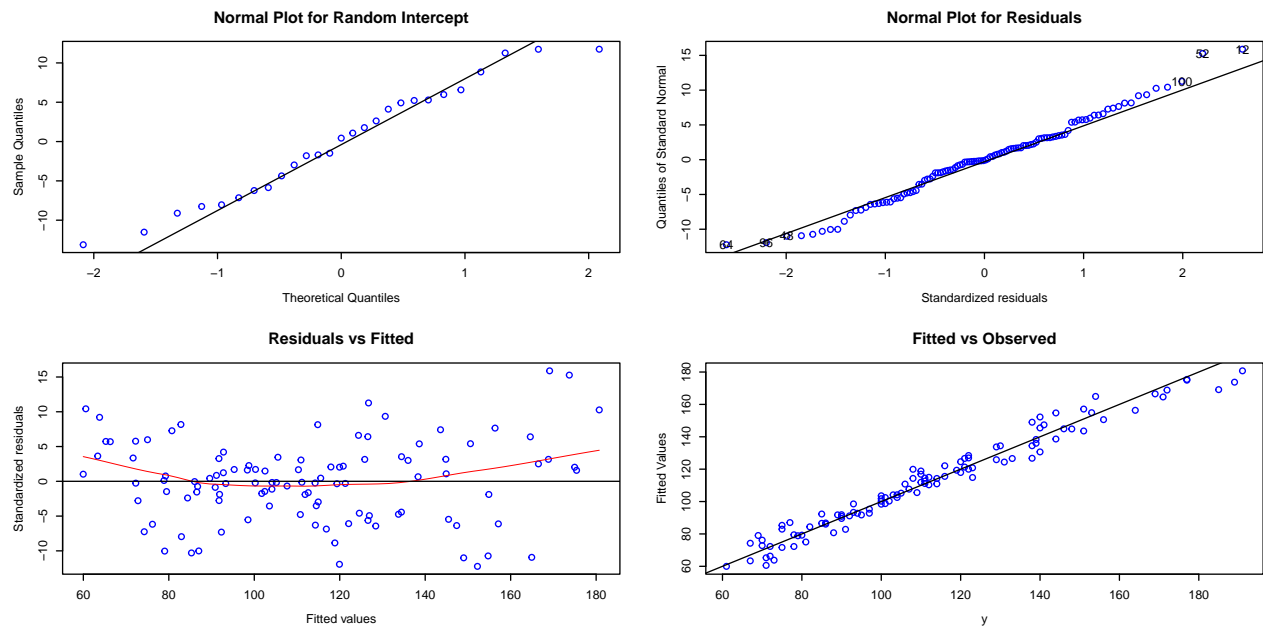## *m1*



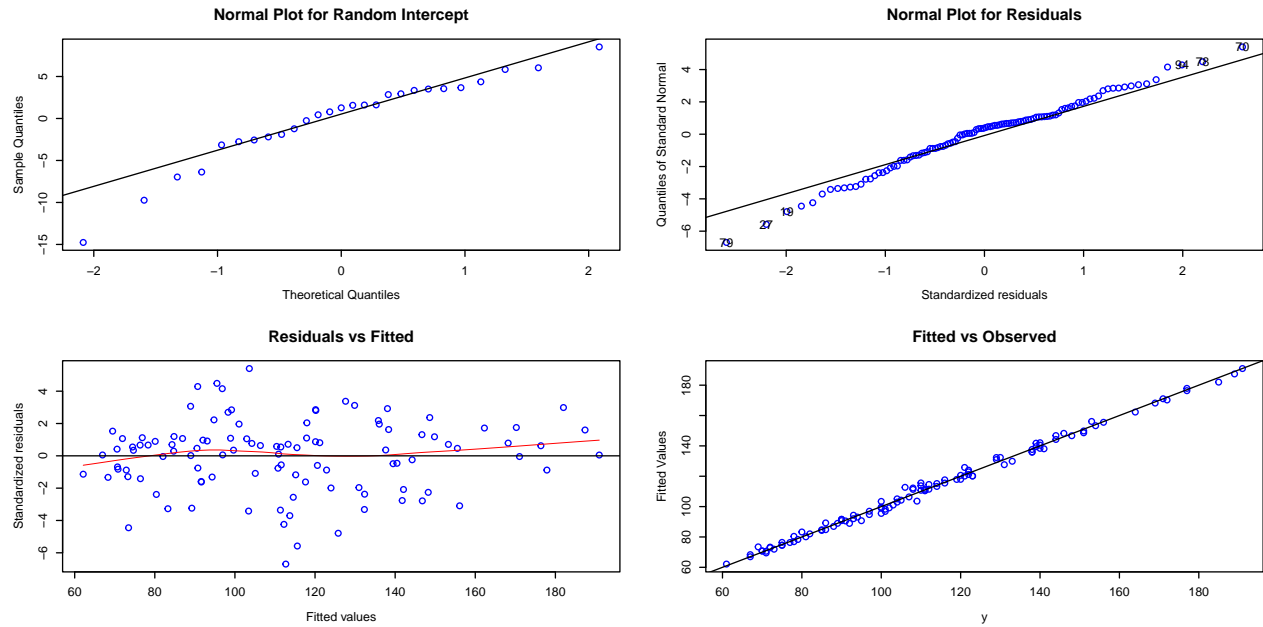***Figure 14:*** *Model diagnostics m1*

## *m2*



*Figure 15:* Model diagnostics m2

## *m3*



*Figure 16:* Model diagnostics m3
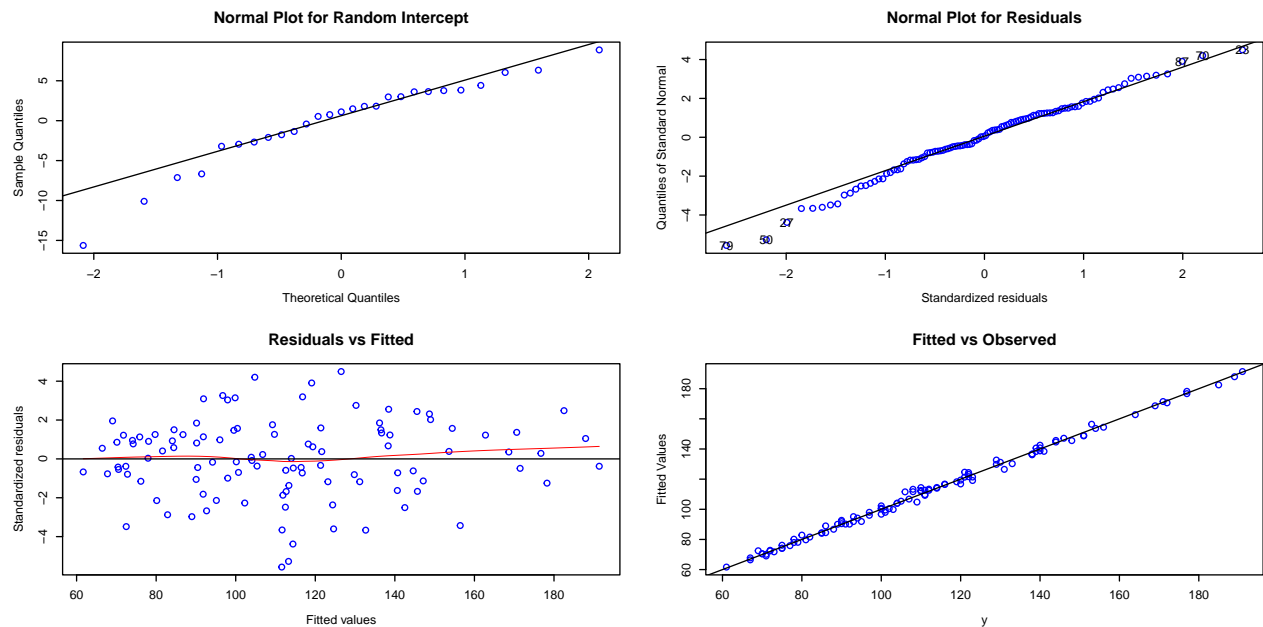
## *m4*



*Figure 17: Model diagnostics m4*

## *m5*



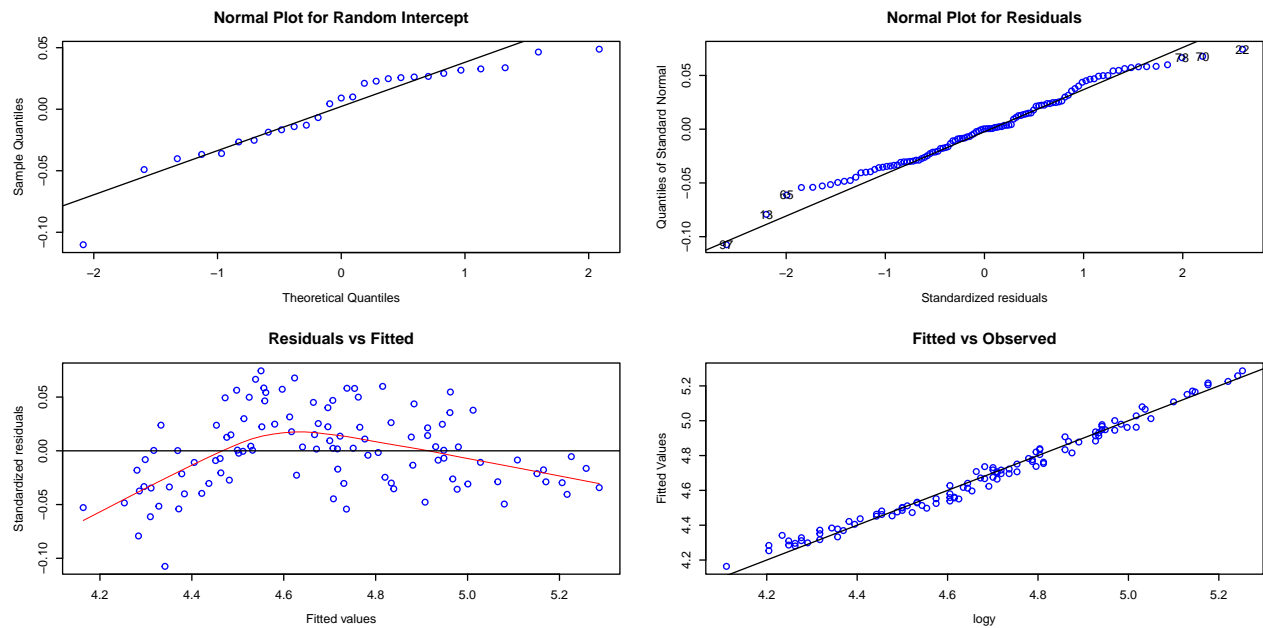*Figure 18: Model diagnostics m5*

## *m6*



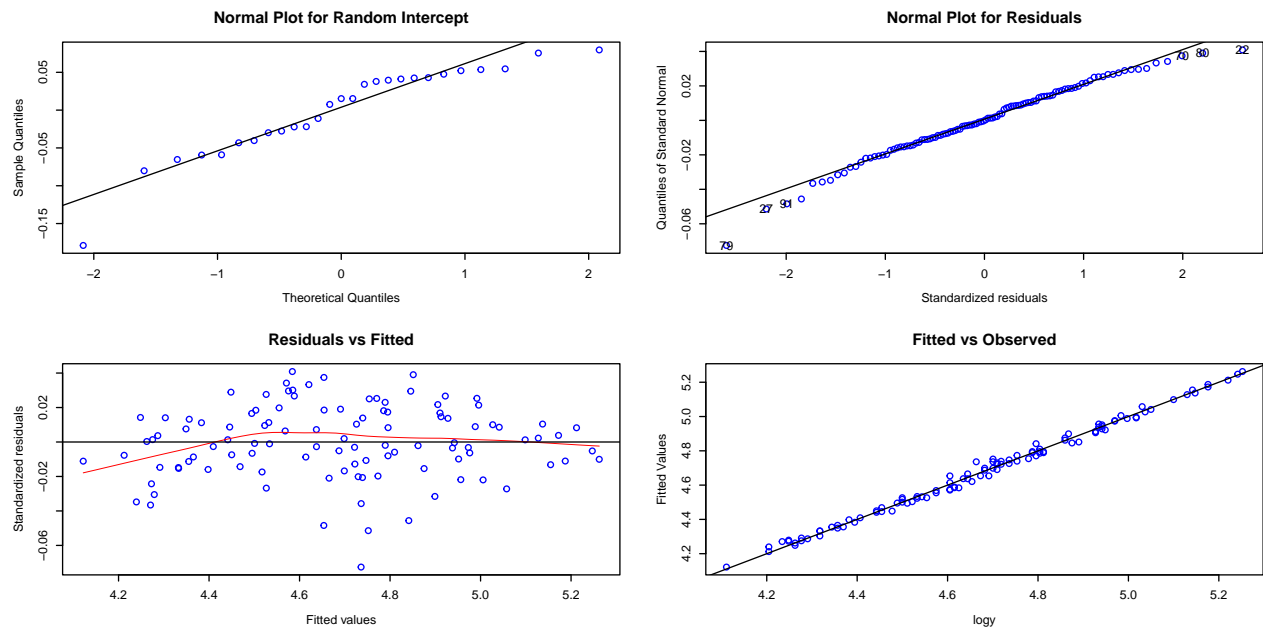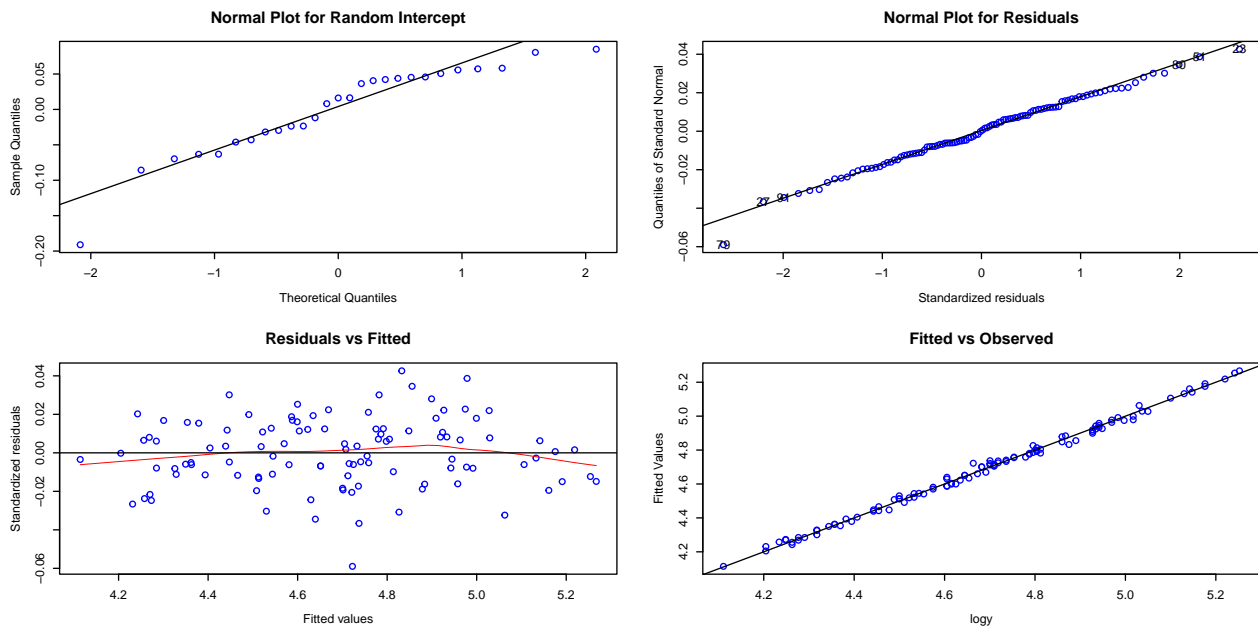*Figure 19: Model diagnostics m6*

## *m7*



*Figure 20: Model diagnostics m7*

## *m8*



*Figure 21: Model diagnostics m8*

67 `summary(m3)`

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: y ~ weekQ + Trt + weekQ:Trt + y0 + (1 + weekQ | Rat)
   Data: rats

REML criterion at convergence: 651.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.3066 -0.4466  0.1335  0.3912  1.8599

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Rat      (Intercept) 37.403   6.116
          weekQ       24.214   4.921    -0.74
 Residual             8.432    2.904
Number of obs: 108, groups:  Rat, 27

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept) -33.0875    11.5262  23.9499  -2.871  0.00843 **
weekQ        27.0400     1.6094  23.9982  16.802 8.98e-15 ***
Trt2         -7.5100     3.5019  23.8513  -2.145  0.04239 *
Trt3          9.9074     3.1673  23.6826   3.128  0.00462 **
y0            1.5609     0.2094  23.0002   7.454 1.41e-07 ***
weekQ:Trt2    2.4171     2.5080  23.9982   0.964  0.34477
weekQ:Trt3  -11.4900     2.2760  23.9982  -5.048 3.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
```

```
            (Intr) weekQ   Trt2    Trt3    y0      wkQ:T2
weekQ       -0.143
Trt2        -0.032   0.472
Trt3        -0.092   0.522   0.456
y0          -0.981   0.000  -0.094  -0.046
weekQ:Trt2   0.092  -0.642  -0.736  -0.335   0.000
weekQ:Trt3   0.101  -0.707  -0.334  -0.738   0.000   0.454
```

Variogram

```
38 ## Correlation structures
39 cor_structures <- list(
40   COMP   = corCompSymm(form = ~ weekQ | Rat),
41   GAUS   = corGaus(form = ~  weekQ | Rat),
42   EXP    = corExp(form = ~ weekQ | Rat),
43   AR1    = corAR1(form = ~  weekQ | Rat), # Serial, order 1
44   CAR1   = corCAR1(form = ~ weekQ | Rat), # Continous AR1
45   LIN    = corLin(form = ~ weekQ | Rat),
46   RATIO  = corRatio(form = ~ weekQ | Rat),
47   SPHER  = corSpher(form = ~ weekQ | Rat)
48 )
49
50 # Fit one model for each structure
51 M <- list()
52
53 for (nm in names(cor_structures)) {
54   M[[nm]] <- try(lme(
55     logy ~ weekQ + Trt + weekQ:Trt + y0 ,
56     random = ~ 1  | Rat, # random slope does not converge
57     correlation = cor_structures[[nm]],
58     data = rats,
59     control = lmeControl(msMaxIter = 5000, niterEM = 50)) ) }
60 # Check names
61 names(M)
62
63 V <- list()
64 for (nm in names(M)) {
65
66   var_obj <- Variogram(M[[nm]], form = ~ weekQ | Rat, data = rats)
67
68   V[[nm]] <- plot(
69     var_obj,
70     main = paste("Variogram -", nm),
71     pch = 16,
72     cex = 1,
73     ylim = c(0, 2),
74     smooth = FALSE,
75     xlab = "Distance (week)" )}
76
77 # Display all in a grid
78 grid.arrange(
79   grobs = V,
80   ncol = 4)
```
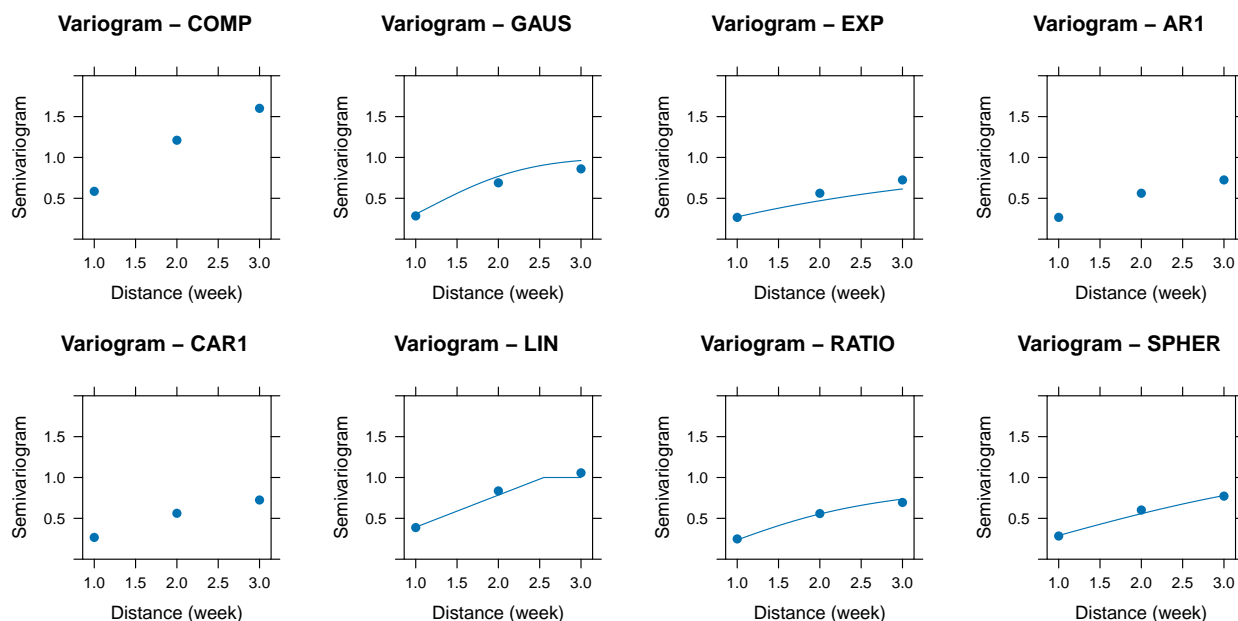
*Figure 22: Variograms using m1, random intercept model*

```
81  ## BLUPS
82  dotplot(ranef(m3, condVar = TRUE),
83          strip = FALSE,
84          ylab = "",
85          scales = list(cex = 0.8))   # smaller labels
```

## B.4   Results

```
86  ### Emmeans
87  wvals <- c(1,2.5,4)
88
89  # emmeans table
90  em <- emmeans(m3, "Trt", by = "weekQ", at = list(weekQ=wvals))
91  em
92
93  emmip(m3,Trt ~ weekQ, at = list(weekQ = wvals) , CIs = TRUE) + theme(legend.position="top")+
94    #ggtitle("Emmeans interaction")+
95    ylab("Weight") +
96    labs(color="Trt")
97
98  emtrends(m3, pairwise ~Trt, var="weekQ",infer=TRUE, adjust="Tukey")
99
100
101 slopes <- emtrends(m3, ~ Trt, var = "weekQ")
102 slopes_df <- as.data.frame(slopes)
103 ggplot(slopes_df, aes(x = Trt, y = weekQ.trend)) +
104   geom_point(size = 3) +
105   geom_errorbar(aes(ymin = lower.CL, ymax = upper.CL), width = 0.1) +
106   labs(
107     title = "Estimated slope",
108     x = "Trt",
109     y = "Weight slope (dy/dweek)"
110   ) +
```

```
111    theme_bw()
```

```
112 # Predictions
113 # Mixed predictions
114 rats$m3 <- predict(m3)
115 # Fixed predictions
116 rats$m3_fix <- predict(m3,re.form=NA)
```