Name: Yuxuan Zhang

Date: October 22, 2024

HW5

1. (a)
   - The macro-explanatory variable is w_j, as it is indexed by j, meaning it varies between groups but is constant within a group.
   - The micro-explanatory variable is $x_{i,j}$, as it is indexed by both i and j, meaning it varies across individuals within a group.
   - The fixed effect parameters are $\beta_0, \beta_1, \beta_2$. β_0 is the fixed intercept, β_1 is the fixed slope for the micro-level variable $x_{i,j}$, and β_2 is the fixed slope for the macro-level variable w_j.
   - The random effect parameters are $\alpha_{0,j}$, and $\alpha_{1,j}$, where $\alpha_{0,j}$ is the random intercept, and $\alpha_{1,j}$ is the random slope.
   - σ^2 is the variance of the residual errors $\epsilon_{i,j}$, ψ_0^2 is the variance of the random intercept $\alpha_{0,j}$, ψ_1^2 is the variance of the random slope $\alpha_{1,j}$.
   - The covariance between $\alpha_{0,j}$ and $\alpha_{1,j}$ is ψ_{01}.

   (b) The variable w_j is a group-level random variable that does not vary within a group. Adding $a_{2,j}w_j$ is unnecessary as all effects should've been already covered by $\beta_j w_j$.

   (c) Notice that under M_0 assumption, since $\psi_1^2 = 0$, the random slope $a_{1,j} = 0$ since $\mathbb{E}[a_{1,j}] = 0$. Then, we have the following null and alternative hypotheses:

   - Null hypothesis: The random effect slope, $a_{1,j}$ is 0.
   - Alternative hypothesis: The random effect slope, $a_{1,j}$, is not zero.

   From the hypotheses stated above, M_0 has 1 random effect coefficient, while M_1 has 2 random effect coefficients. Therefore,

$$\lambda(y) = \begin{cases} X_1 \text{ with probability } 1/2 \\ X_2 \text{ with probability } 1/2 \end{cases}$$

   where $X_1 \sim \chi_1^2$ and $X_2 \sim \chi_2^2$. Hence, the distribution of the LRT statistic under the null hypothesis is

$$\lambda | M_0 \sim \frac{1}{2}\left(\chi_1^2 + \chi_2^2\right).$$

   To obtain a p-value, we will need the LRT statistic, which is

$$\lambda(y) = 2 \cdot (\log_{M_1}(y) - \log_{M_2}(y)).$$

   Then, the p-value is

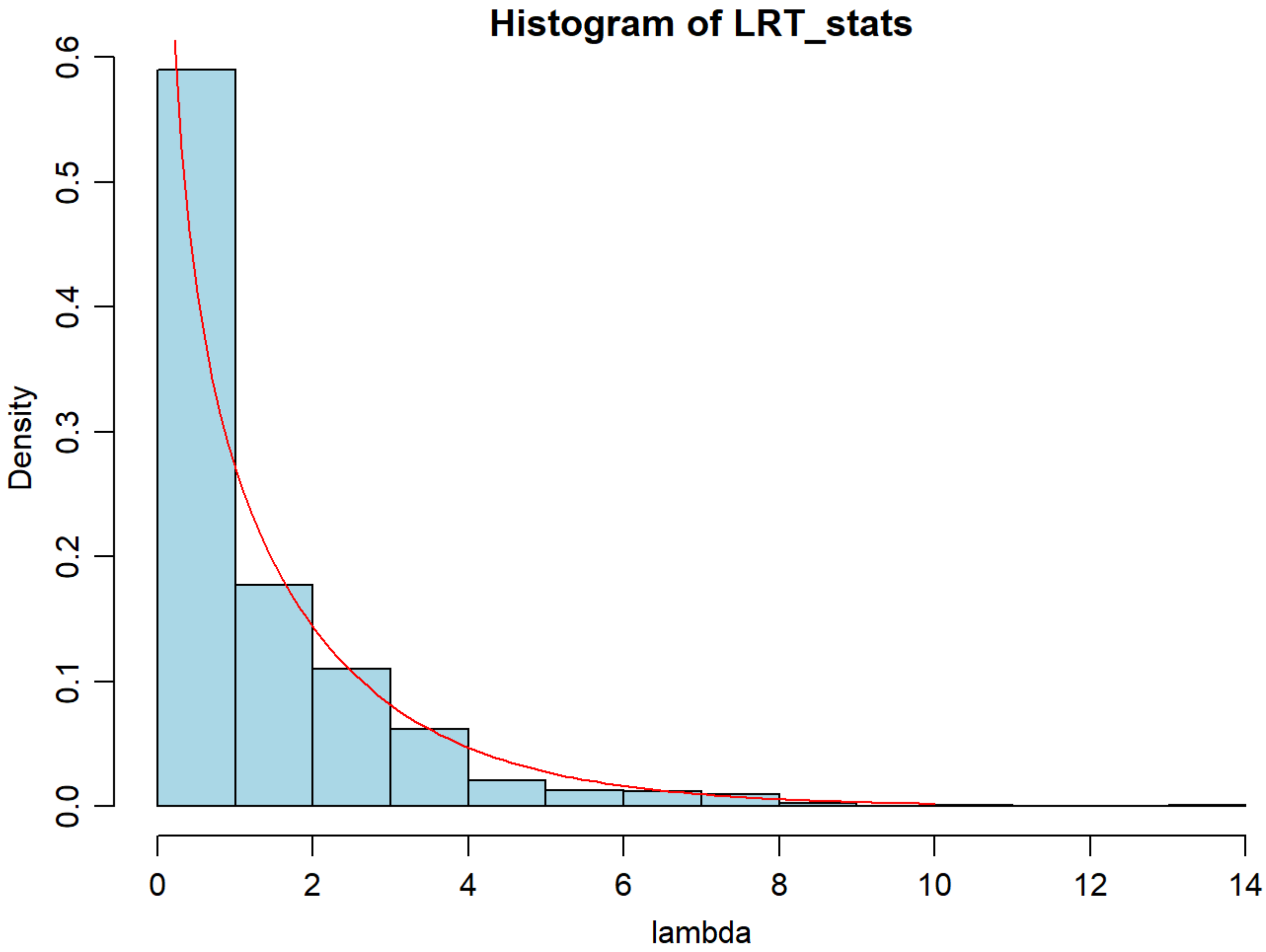
$$\text{p-value} = \frac{1}{2}\Pr(\chi_1^2 \geq \lambda) + \frac{1}{2}\Pr(\chi_2^2 \geq \lambda).$$

   (d)

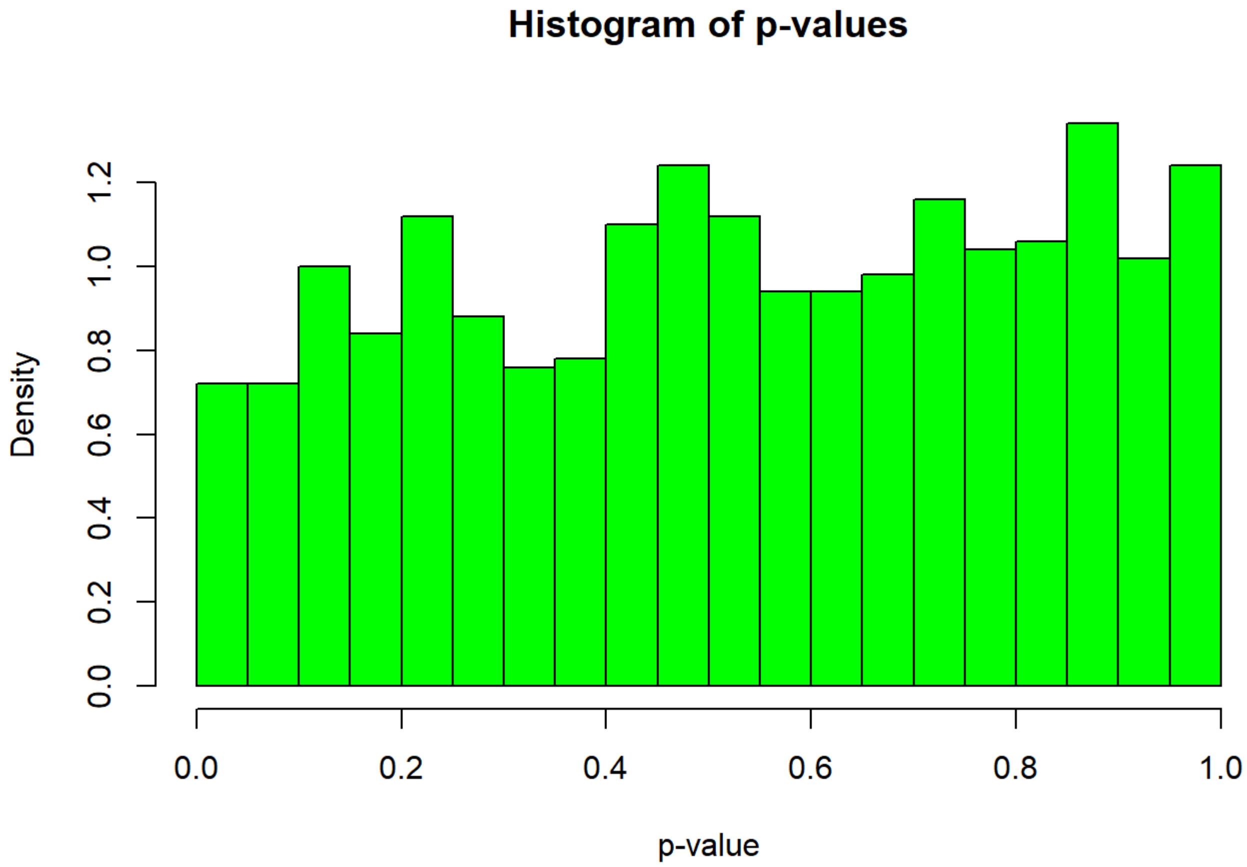
Q1 (d)

```
mpar<-function(...){par(mar=c(3,3,1,1),mgp=c(2,.75,0),tck=-.025,...)}
library(lme4)
library(ggplot2)
set.seed(610)
m <- 20
n <- 20
num_simulations <- 1000

beta0 <- 1
beta1 <- 1
beta2 <- 1
sigma2 <- 1
psi0_sq <- 1
psi1_sq <- 0
psi01 <- 0

LRT_stats <- numeric(num_simulations)
p_values <- numeric(num_simulations)

for (s in 1:num_simulations) {
  w_j <- rnorm(m)
  a0_j <- rnorm(m, mean = 0, sd = sqrt(psi0_sq))
  x_ij <- rnorm(m * n)
  epsilon_ij <- rnorm(m * n, mean = 0, sd = sqrt(sigma2))

  group <- rep(1:m, each = n)
  y <- beta0 + beta1 * x_ij + beta2 * rep(w_j, each = n) + rep(a0_j, each = n) + epsilon_ij
  data_sim <- data.frame(y = y, x = x_ij, w = rep(w_j, each = n), group = factor(group))

  full_model <- lmer(y ~ x + w + (x | group), data = data_sim, REML = FALSE)
  restricted_model <- lmer(y ~ x + w + (1 | group), data = data_sim, REML = FALSE)

  lambda <- 2 * (logLik(full_model) - logLik(restricted_model))
  LRT_stats[s] <- as.numeric(lambda)

  p_values[s] <- 0.5 * (1 - pchisq(lambda, df = 1)) + 0.5 * (1 - pchisq(lambda, df = 2))
}
```

```
mpar()
hist(LRT_stats, col="lightblue", prob=TRUE, xlab="lambda", nclass=15)
xs <- seq(0, 10, length=100)
lines(xs, 0.5 * dchisq(xs, df=1) + 0.5 * dchisq(xs, df=2), col="red")
```



```
hist(p_values, breaks = 30, probability = TRUE, col = "green",
main = "Histogram of p-values",
xlab = "p-value")
```



From the overlay plot, we see that the histogram of the LRT statistics fits well with the null distribution $\lambda \sim \frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$. The LRT statistics are positively skewed, with most of the values concentrated near 0, gradually tapering off, which aligns with the theoretical mixed chi-square distribution

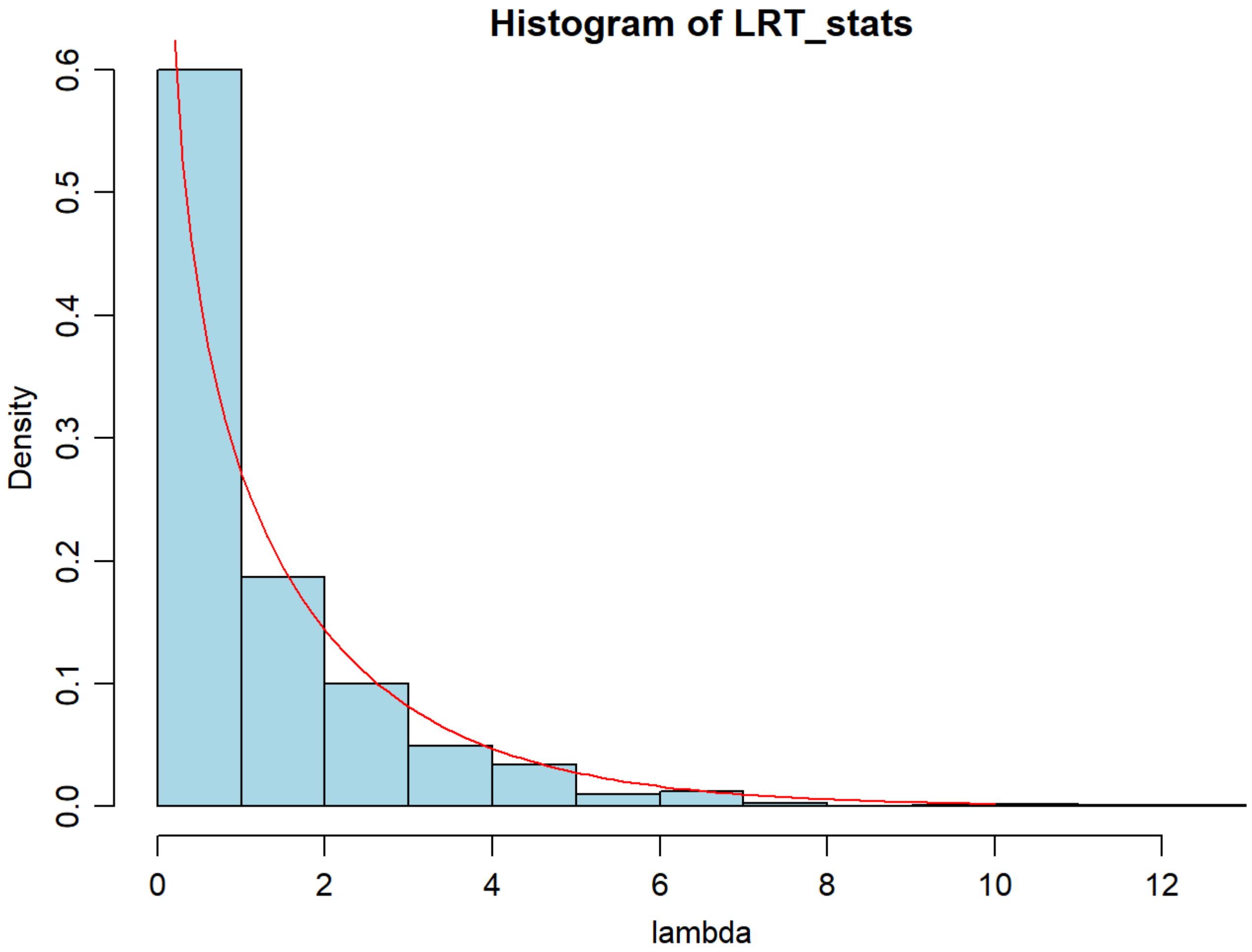
The histogram of p-values appears to be uniformly distributed between 0 and 1. Though we do see that the histogram is centered slightly towards larger p-values. This is expected since with fewer groups, the random effects are not estimated as precisely, leading to higher variability in the model fits. This increased variability causes the likelihood ratio test to be less sensitive, resulting in p-values that are often higher and tend to cluster towards larger values.

Q1 (e)

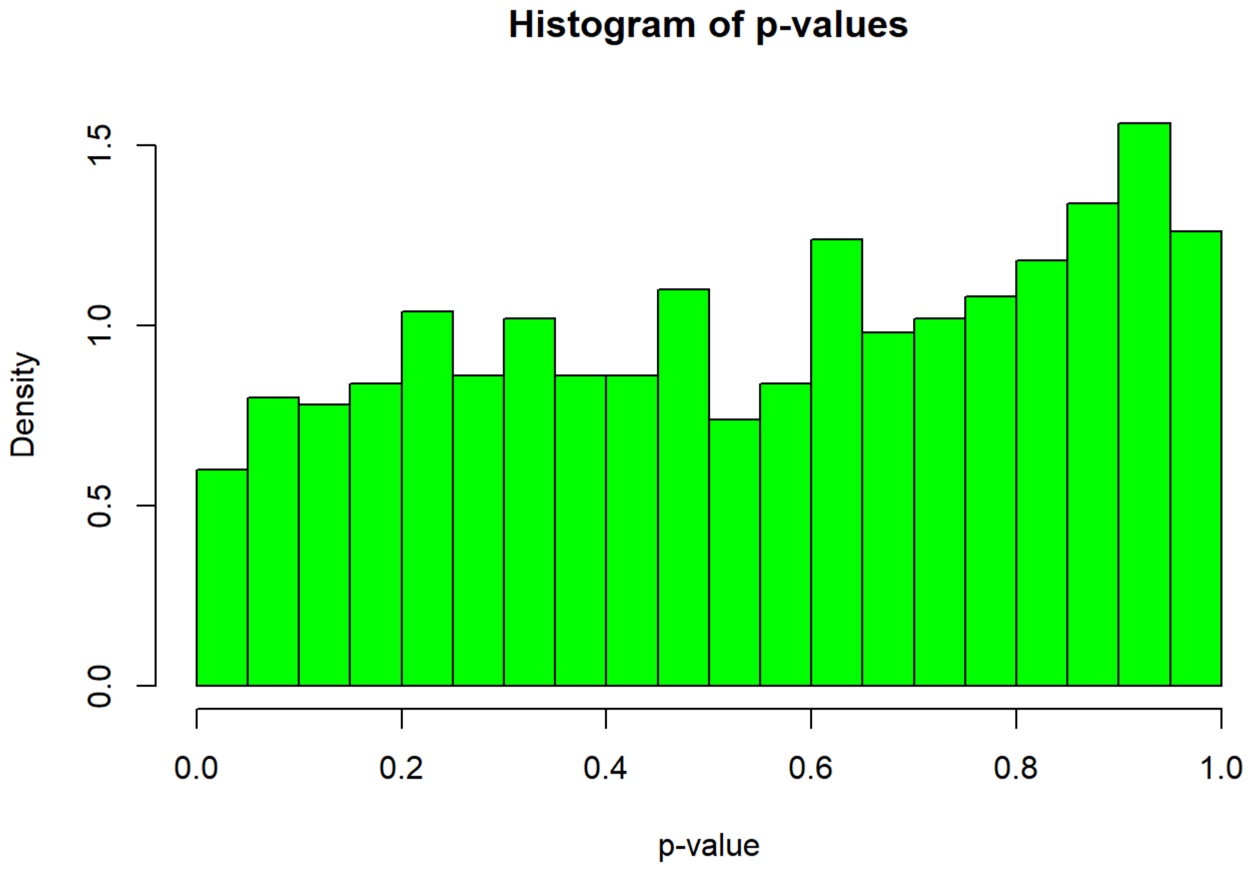
We will first try with the model parameters:

```
m <- 20
n <- 20
num_simulations <- 1000

beta0 <- 3
beta1 <- 6
beta2 <- 9
sigma2 <- 1
psi0_sq <- 2
psi1_sq <- 0
psi01 <- 0

LRT_stats <- numeric(num_simulations)
p_values <- numeric(num_simulations)

for (s in 1:num_simulations) {

  w_j <- rnorm(m)
  a0_j <- rnorm(m, mean = 0, sd = sqrt(psi0_sq))
  x_ij <- rnorm(m * n)
  epsilon_ij <- rnorm(m * n, mean = 0, sd = sqrt(sigma2))

  group <- rep(1:m, each = n)
  y <- beta0 + beta1 * x_ij + beta2 * rep(w_j, each = n) + rep(a0_j, each = n) + epsilon_ij
  data_sim <- data.frame(y = y, x = x_ij, w = rep(w_j, each = n), group = factor(group))

  full_model <- lmer(y ~ x + w + (x | group), data = data_sim, REML = FALSE)
  restricted_model <- lmer(y ~ x + w + (1 | group), data = data_sim, REML = FALSE)

  lambda <- 2 * (logLik(full_model) - logLik(restricted_model))
  LRT_stats[s] <- as.numeric(lambda)

  p_values[s] <- 0.5 * (1 - pchisq(lambda, df = 1)) + 0.5 * (1 - pchisq(lambda, df = 2))
}
```

```
mpar()
hist(LRT_stats, col="lightblue", prob=TRUE, xlab="lambda", nclass=15)
xs <- seq(0, 10, length=100)
lines(xs, 0.5 * dchisq(xs, df=1) + 0.5 * dchisq(xs, df=2), col="red")
```



```
# Plot histogram of p-values
hist(p_values, breaks = 20, probability = TRUE, col = "green",
main = "Histogram of p-values",
xlab = "p-value")
```

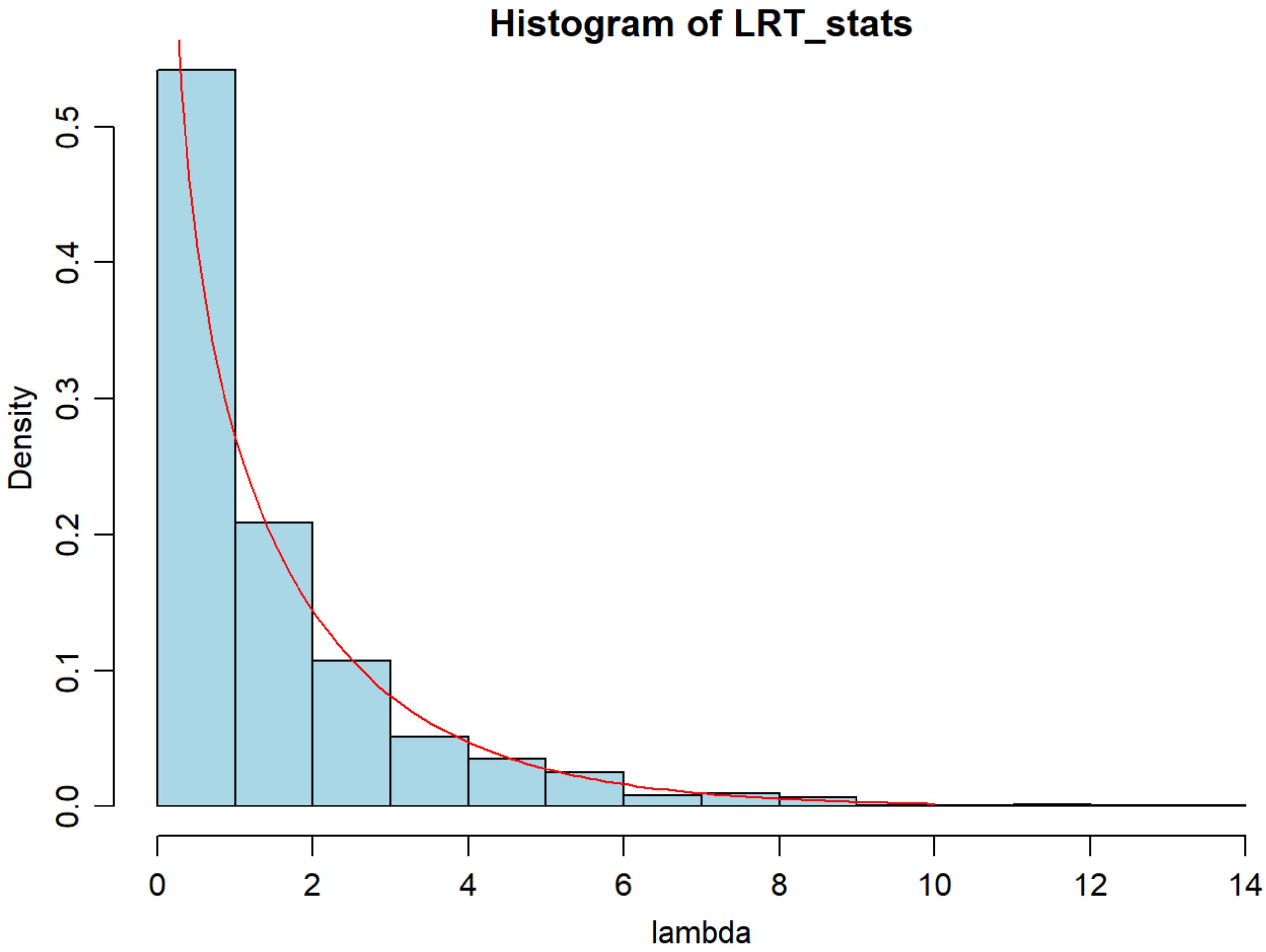


With different model parameters, the result appears to be closely matching with what we got from part (d), where the LRT statistics follows a $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$ distribution, and the histogram of the p-value appears to be uniformly distributed that is slightly centered at larger p-values.

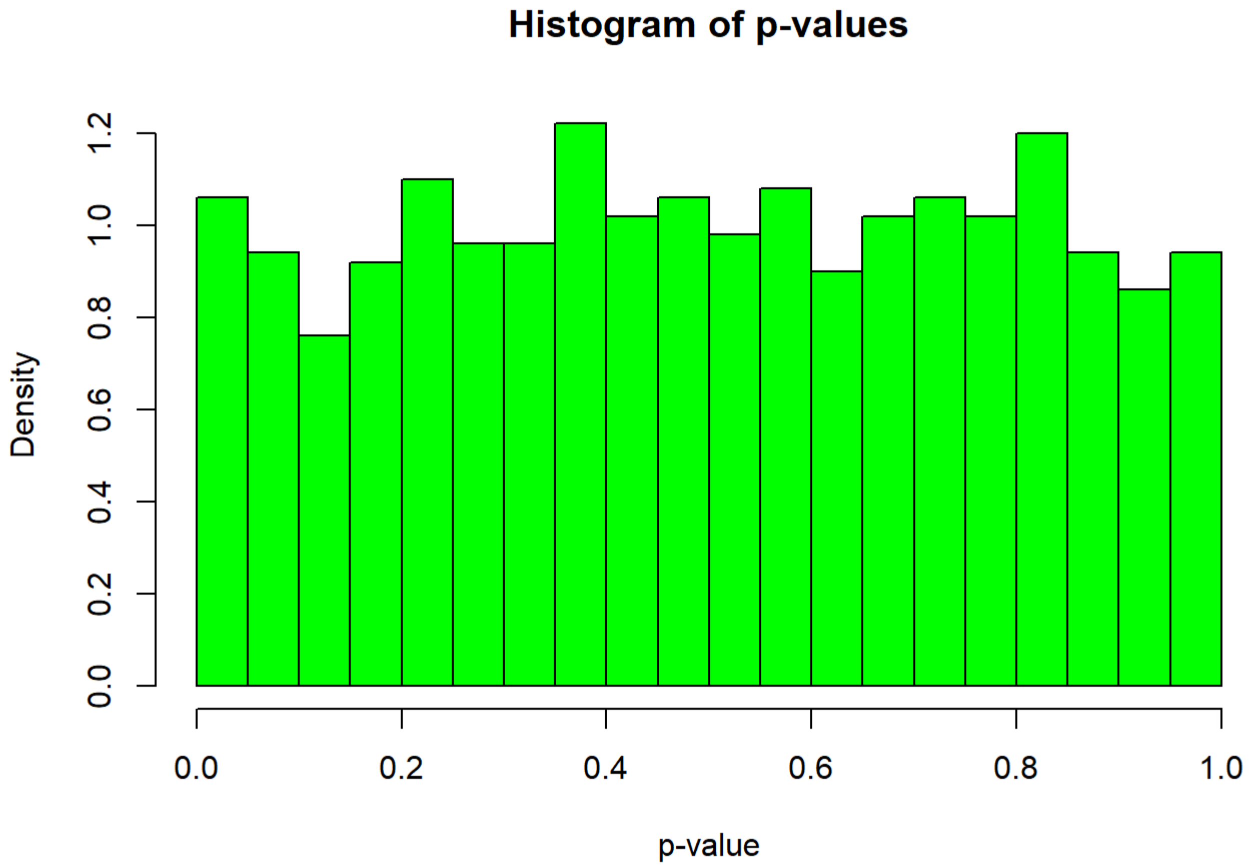
We then only change the number of groups and see if the results remains similar.

```
m <- 500
n <- 20
num_simulations <- 1000

beta0 <- 1
beta1 <- 1
beta2 <- 1
sigma2 <- 1
psi0_sq <- 1
psi1_sq <- 0
psi01 <- 0

LRT_stats <- numeric(num_simulations)
p_values <- numeric(num_simulations)

for (s in 1:num_simulations) {

  w_j <- rnorm(m)
  a0_j <- rnorm(m, mean = 0, sd = sqrt(psi0_sq))
  x_ij <- rnorm(m * n)
  epsilon_ij <- rnorm(m * n, mean = 0, sd = sqrt(sigma2))

  group <- rep(1:m, each = n)
  y <- beta0 + beta1 * x_ij + beta2 * rep(w_j, each = n) + rep(a0_j, each = n) + epsilon_ij
  data_sim <- data.frame(y = y, x = x_ij, w = rep(w_j, each = n), group = factor(group))

  full_model <- lmer(y ~ x + w + (x | group), data = data_sim, REML = FALSE)
  restricted_model <- lmer(y ~ x + w + (1 | group), data = data_sim, REML = FALSE)

  lambda <- 2 * (logLik(full_model) - logLik(restricted_model))
  LRT_stats[s] <- as.numeric(lambda)

  p_values[s] <- 0.5 * (1 - pchisq(lambda, df = 1)) + 0.5 * (1 - pchisq(lambda, df = 2))
}
```

```
mpar()
hist(LRT_stats, col="lightblue", prob=TRUE, xlab="lambda", nclass=15)
xs <- seq(0, 10, length=100)
lines(xs, 0.5 * dchisq(xs, df=1) + 0.5 * dchisq(xs, df=2), col="red")
```



```
# Plot histogram of p-values
hist(p_values, breaks = 20, probability = TRUE, col = "green",
main = "Histogram of p-values",
xlab = "p-value")
```



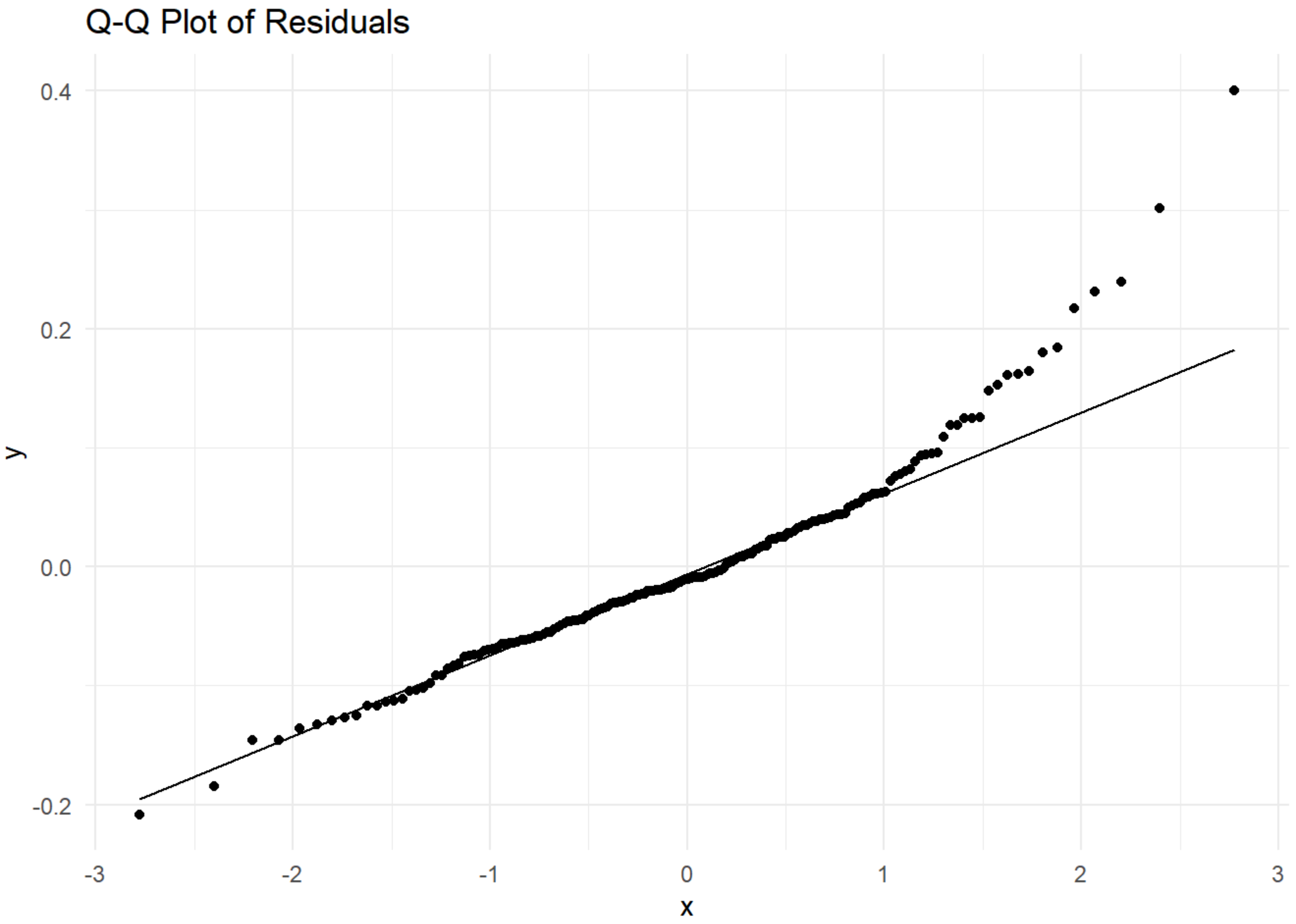
When we increase the number of groups, we see that the p-value distribution appears to be more uniformly distributed. This is expected. As the number of groups increases, the random effects are estimated with greater precision. The likelihood ratio test becomes more sensitive, and the p-values start to behave more like what you would expect under the null hypothesis, which is uniformly distributed between 0 and 1. With more groups, there’s less variability in the test statistic under the null, so the p-values become more evenly spread across their range.

Q2 (a)

```
data = dget("C:/Users/LEGION/OneDrive - Duke University/610/Data/Earthquake.txt")
hlm_model <- lmer(accel ~ Richter + distance + soil + (distance + soil | Quake), REML = F, data
= data)
summary(hlm_model)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: accel ~ Richter + distance + soil + (distance + soil | Quake)
##    Data: data
##
##      AIC      BIC   logLik deviance df.resid
##   -293.5   -258.3    157.8   -315.5      171
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3123 -0.5807 -0.1150  0.4370  4.4474
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  Quake    (Intercept) 8.928e-03 0.094491
##           distance    4.058e-06 0.002014 -0.92
##           soil1       9.657e-06 0.003108 -0.85  0.57
##  Residual             8.113e-03 0.090073
## Number of obs: 182, groups:  Quake, 23
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept) -0.0505713  0.1145529  -0.441
## Richter      0.0443509  0.0188922   2.348
## distance    -0.0030945  0.0006114  -5.062
## soil1        0.0080566  0.0206672   0.390
##
## Correlation of Fixed Effects:
##          (Intr) Richtr distnc
## Richter  -0.966
## distance  0.089 -0.272
## soil1    -0.247  0.107 -0.040
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.0167446 (tol = 0.002, component 1)
```

```
residuals = residuals(hlm_model)
fitted_values = fitted(hlm_model)
qqplot <- ggplot(data.frame(residuals = residuals), aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("Q-Q Plot of Residuals") +
  theme_minimal()
print(qqplot)
```



We observe the following results from the model summary:

- Random effects:
    - The random intercept for each earthquake (quake) group has a variance of 8.928e-03, indicating variability in acceleration across different earthquakes.
    - The random slopes for distance and soil type have variances of 4.058e-06 and 9.657e-06, respectively.

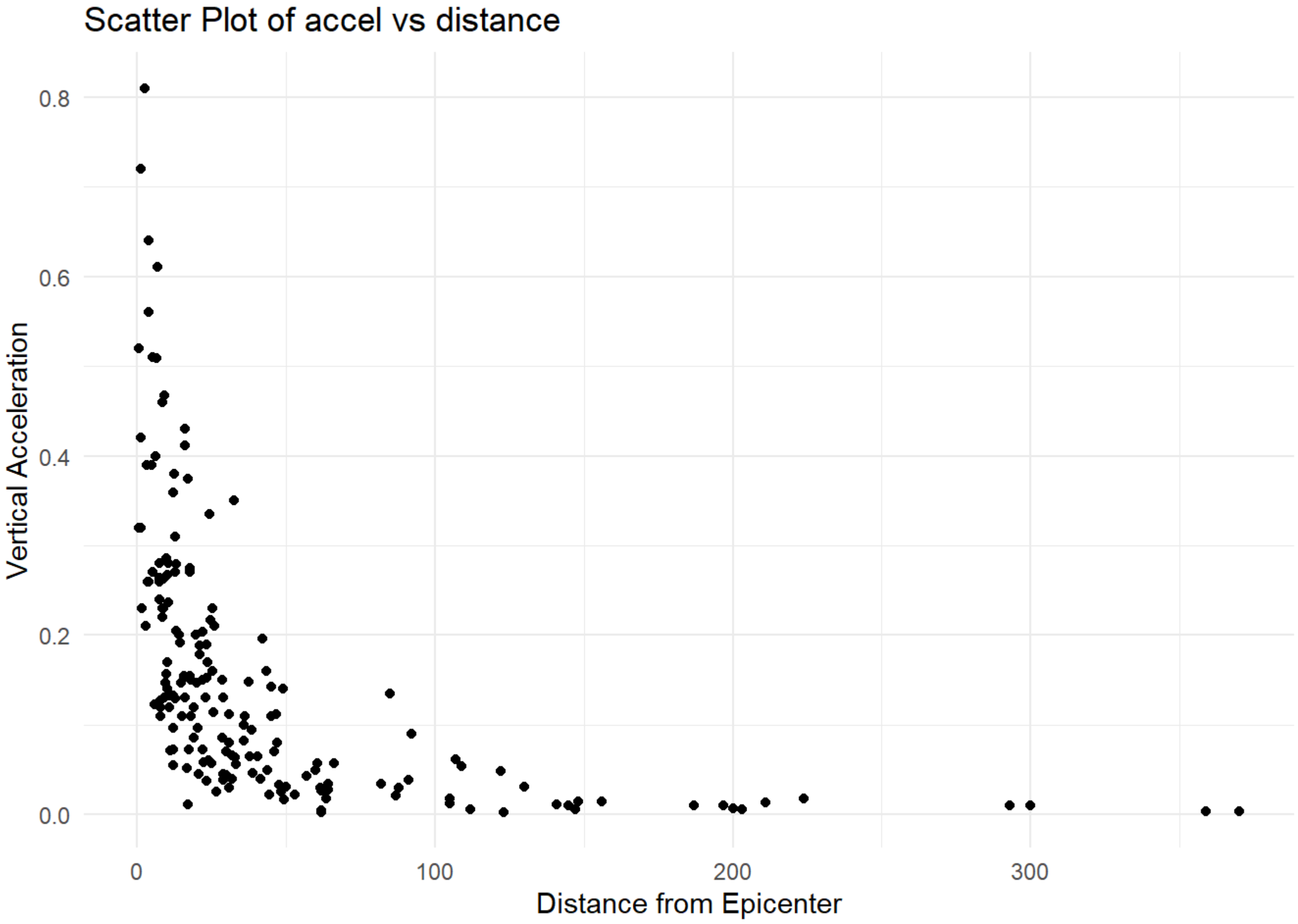
These small variances suggest that the random effects associated with distance and soil type are relatively minor compared to the overall variation. There is a high negative correlation (-0.92) between the random effects for distance and soil type, suggesting a strong inverse relationship in how these covariates influence the acceleration data at the random effects level.

- Fixed effects:
    - The fixed effect for the Richter scale is positive and statistically significant $(|t| = 2.348)$, indicating that higher earthquake magnitudes lead to higher vertical accelerations.
    - The fixed effect for distance is negative and statistically significant $(|t| = 5.062)$, showing that acceleration decreases as the distance from the epicenter increases.
    - The fixed effect for soil type is negative but not statistically significant $(|t| = 0.39)$.

Regarding the observation for normality from the Q-Q plot, the points deviate noticeably from the diagonal line, especially in the tails. This suggest that the residuals are not normally distributed. The Q-Q plot indicates that the assumptions of normally distributed residuals does not hold for this model.

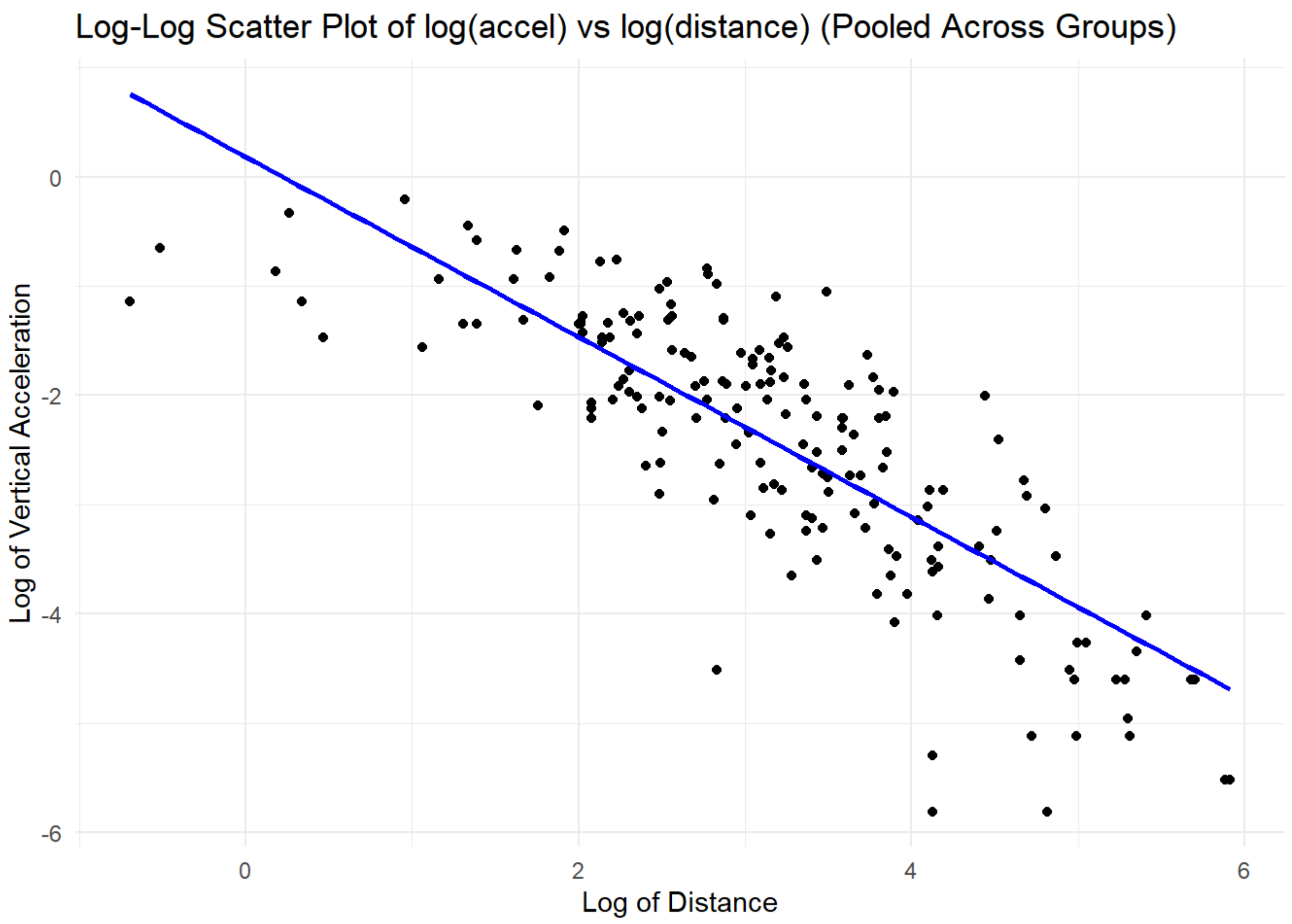
Q2 (b)

```
library(ggplot2)
ggplot(data, aes(x = distance, y = accel)) +
  geom_point() +
  ggtitle("Scatter Plot of accel vs distance") +
  xlab("Distance from Epicenter") +
  ylab("Vertical Acceleration") +
  theme_minimal()
```



No, the relationship does not look linear.

```
# Log-log transformation on both variables pooled across groups
ggplot(data, aes(x = log(distance), y = log(accel))) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  ggtitle("Log-Log Scatter Plot of log(accel) vs log(distance) (Pooled Across Groups)") +
  xlab("Log of Distance") +
  ylab("Log of Vertical Acceleration") +
  theme_minimal()
```



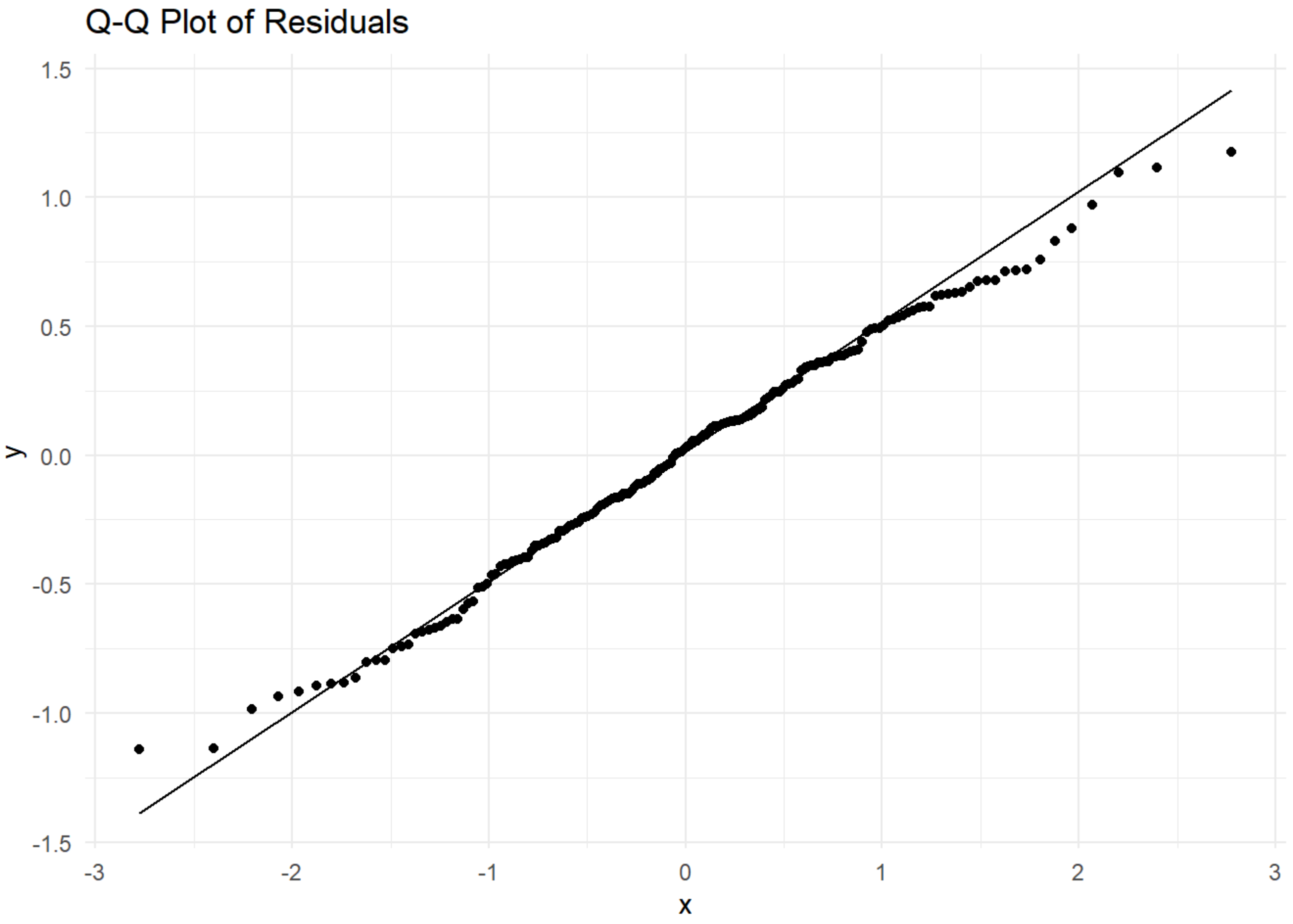
The log-log scatter plot shows a much more linear relationship compared to the original plot.

Q2 (c)

```
hlm_log_log <- lmer(log(accel) ~ log(distance) + Richter + soil + (log(distance) + soil | Quak
e), data = data)
summary(hlm_log_log)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(accel) ~ log(distance) + Richter + soil + (log(distance) +
##     soil | Quake)
##    Data: data
##
## REML criterion at convergence: 332.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.22114 -0.63347  0.05285  0.69244  2.28937
##
## Random effects:
##  Groups   Name          Variance Std.Dev. Corr
##  Quake    (Intercept)   2.2877   1.5125
##           log(distance) 0.2015   0.4489   -0.93
##           soil1         0.1711   0.4136   -0.63  0.30
##  Residual               0.2636   0.5134
## Number of obs: 182, groups:  Quake, 23
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)    -2.6528     0.6916  -3.836
## log(distance)  -1.1446     0.1205  -9.498
## Richter         0.6104     0.1067   5.721
## soil1           0.2114     0.1764   1.198
##
## Correlation of Fixed Effects:
##             (Intr) lg(ds) Richtr
## log(distnc) -0.387
## Richter     -0.801 -0.172
## soil1       -0.246  0.145 -0.096
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```
residuals = residuals(hlm_log_log)
fitted_values = fitted(hlm_log_log)
qqplot <- ggplot(data.frame(residuals = residuals), aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("Q-Q Plot of Residuals") +
  theme_minimal()
print(qqplot)
```



The residuals are much more closely aligned with the diagonal line, especially in the middle range of the distribution. This suggests that the assumption of normality for the residuals is better satisfied in the log-log model.

Q2 (d)

We will first find out the null distribution. For testing of the random slope of soil, since there are three random effects under the full model and 2 random effects under the full model, the null distribution is

$$\lambda \sim \frac{1}{2}\chi_2^2 + \frac{1}{2}\chi_3^2.$$

This is the same with the log-log transformed model.

For the fixed effect test, there is exactly one fixed effect being zero. Thus, the LRT statistic follows a chi-squared distribution with 1 df. Namely, $\lambda \sim \chi_1^2$.

```
library(lme4)
data$log_accel <- log(data$accel)
data$log_distance <- log(data$distance)
fit_full <- lmer(log_accel ~ log_distance + soil + Richter +
                 (log_distance + soil | Quake), data = data, REML = FALSE)

# Random Slope of Soil
fit_no_soil_slope <- lmer(log_accel ~ log_distance + soil + Richter +
                          (log_distance | Quake), data = data, REML = FALSE)
LRT_soil <- 2 * (logLik(fit_full) - logLik(fit_no_soil_slope))
LRT_soil_value <- as.numeric(LRT_soil)
p_value_soil <- 0.5 * (1 - pchisq(LRT_soil_value, df = 2)) +
                0.5 * (1 - pchisq(LRT_soil_value, df = 3))

# Random Slope of log_distance
fit_no_distance_slope <- lmer(log_accel ~ log_distance + soil + Richter +
                              (soil | Quake), data = data, REML = FALSE)
LRT_distance <- 2 * (logLik(fit_full) - logLik(fit_no_distance_slope))
LRT_distance_value <- as.numeric(LRT_distance)
p_value_distance <- 0.5 * (1 - pchisq(LRT_distance_value, df = 2)) +
                    0.5 * (1 - pchisq(LRT_distance_value, df = 3))

# Fixed Effect of Soil
fit_no_soil_fixed <- lmer(log_accel ~ log_distance + Richter +
                          (log_distance + soil | Quake), data = data, REML = FALSE)

LRT_soil_fixed <- 2 * (logLik(fit_full) - logLik(fit_no_soil_fixed))
LRT_soil_fixed_value <- as.numeric(LRT_soil_fixed)

p_value_soil_fixed <- 1 - pchisq(LRT_soil_fixed_value, df = 1)

# Fixed Effect of log_distance
fit_no_distance_fixed <- lmer(log_accel ~ soil + Richter +
                              (log_distance + soil | Quake), data = data, REML = FALSE)

LRT_distance_fixed <- 2 * (logLik(fit_full) - logLik(fit_no_distance_fixed))
LRT_distance_fixed_value <- as.numeric(LRT_distance_fixed)

p_value_distance_fixed <- 1 - pchisq(LRT_distance_fixed_value, df = 1)

# Fixed Effect of Richter
fit_no_richter_fixed <- lmer(log_accel ~ log_distance + soil +
                             (log_distance + soil | Quake), data = data, REML = FALSE)

LRT_richter_fixed <- 2 * (logLik(fit_full) - logLik(fit_no_richter_fixed))
LRT_richter_fixed_value <- as.numeric(LRT_richter_fixed)

p_value_richter_fixed <- 1 - pchisq(LRT_richter_fixed_value, df = 1)

test_results <- data.frame(
  Test = c("Random Slope of Soil", "Random Slope of log_distance",
           "Fixed Effect of Soil", "Fixed Effect of log_distance",
           "Fixed Effect of Richter"),
  LRT_Value = c(LRT_soil_value, LRT_distance_value, LRT_soil_fixed_value,
                LRT_distance_fixed_value, LRT_richter_fixed_value),
  P_value = c(p_value_soil, p_value_distance, p_value_soil_fixed,
              p_value_distance_fixed, p_value_richter_fixed)
)
print(test_results)
```

```
##                            Test  LRT_Value      P_value
## 1         Random Slope of Soil  5.7646557 8.982154e-02
## 2 Random Slope of log_distance 50.6963268 3.329181e-11
## 3         Fixed Effect of Soil  0.9563451 3.281094e-01
## 4 Fixed Effect of log_distance 34.8308103 3.596368e-09
## 5      Fixed Effect of Richter 22.3260840 2.300604e-06
```

- Random slope of soil The LRT statistic for the random slope of soil is 5.765. The p-value is 0.0898, which is above the 0.05 threshold. This suggests that there is no significant across-quake heterogeneity in the slope of soil.
- Random slope of log-transformed distance The LRT statistic for the random slope of log_distance is 50.696. The p-value is 3.33e-11, which is extremely small. This provides very strong evidence against the null hypothesis, indicating significant heterogeneity in the slope of log_distance across quakes.
- Fixed effect of soil The LRT statistic for the fixed effect of soil is 0.956. The p-value is 0.328, which is above 0.05. This suggests that soil does not have a statistically significant fixed effect.
- Fixed effect of log-transformed distance The LRT statistic for the fixed effect of log_distance is 34.83. The LRT statistic for the fixed effect of log_distance is 34.83.
- Fixed effect of Richter The LRT statistic for the fixed effect of Richter is 22.33. The p-value is 2.30e-06, which is very small, indicating a highly significant fixed effect for Richter scale magnitude.

In conclusion, there is significant across-quake heterogeneity in the slope of log_distance but no significant heterogeneity in the slope of soil. The fixed effects for both log_distance and Richter scale are statistically significant, while the fixed effect of soil is not significant.

Q2 (e)

Based on the findings from the LRT and the hierarchical linear model, we can draw several conclusions regarding the relationship between vertical acceleration and the explanatory variables.

First, the fixed effect of log-transformed distance is statistically significant and negative. This indicates that as the distance from the earthquake's epicenter increases (after log transformation), the vertical acceleration decreases. This result is expected, as greater distances from the epicenter typically result in reduced ground motion and therefore lower acceleration. The p-value for this effect is extremely small (3.59e-09), providing strong evidence that the distance from the epicenter is a key determinant of earthquake acceleration.

The Richter scale magnitude also has a highly significant positive fixed effect. This suggests that as the magnitude of the earthquake increases, the vertical acceleration at a given location also increases. Given that larger earthquakes release more energy, this positive relationship is consistent with physical expectations, and the very small p-value (2.30e-06) highlights the importance of this variable.

On the other hand, the fixed effect of soil type is not statistically significant (p-value = 0.328). This suggests that, in this model, soil type does not have a strong or consistent effect on the vertical acceleration when controlling for other factors like distance and earthquake magnitude. However, the random slope of soil type across different quakes is somewhat marginal, with a p-value of 0.0898. This could indicate mild heterogeneity in how soil type impacts acceleration across different quakes, although it is not strongly significant.

Finally, there is significant across-quake heterogeneity in the slope of log-transformed distance, as evidenced by the very small p-value (3.33e-11). This means that the effect of distance on vertical acceleration varies significantly across different earthquakes. This heterogeneity may arise due to varying geological or environmental factors that alter how distance affects acceleration in different quakes.

In summary, both distance and Richter scale magnitude are important and significant predictors of vertical acceleration. While soil type does not have a significant fixed effect, there may be some variability in how it influences acceleration across different earthquakes.