STAT4520 HW5

Anton Yang

2024-11-02

Problem 1a

```r
set.seed(2)
sige = 3
siga = 4
a = 4
n = 20
alpha = rnorm(a, 0, siga)
mu = 2.5
x = runif(a*n, 0, 10)
b1 = 1.3
z = matrix(0, n*a, a)
for(i in 1:a){
  z[((i-1)*n+1):(i*n), i] = 1
}
y = mu + x * b1 + z %*% alpha + rnorm(n*a, 0, sige)
dat = data.frame(y = y,
                 group = rep(1:a, each = n) |> as.factor(),
                 x = x)
```

```r
library(ggplot2)
library(lme4)
```

```
## Loading required package: Matrix
```

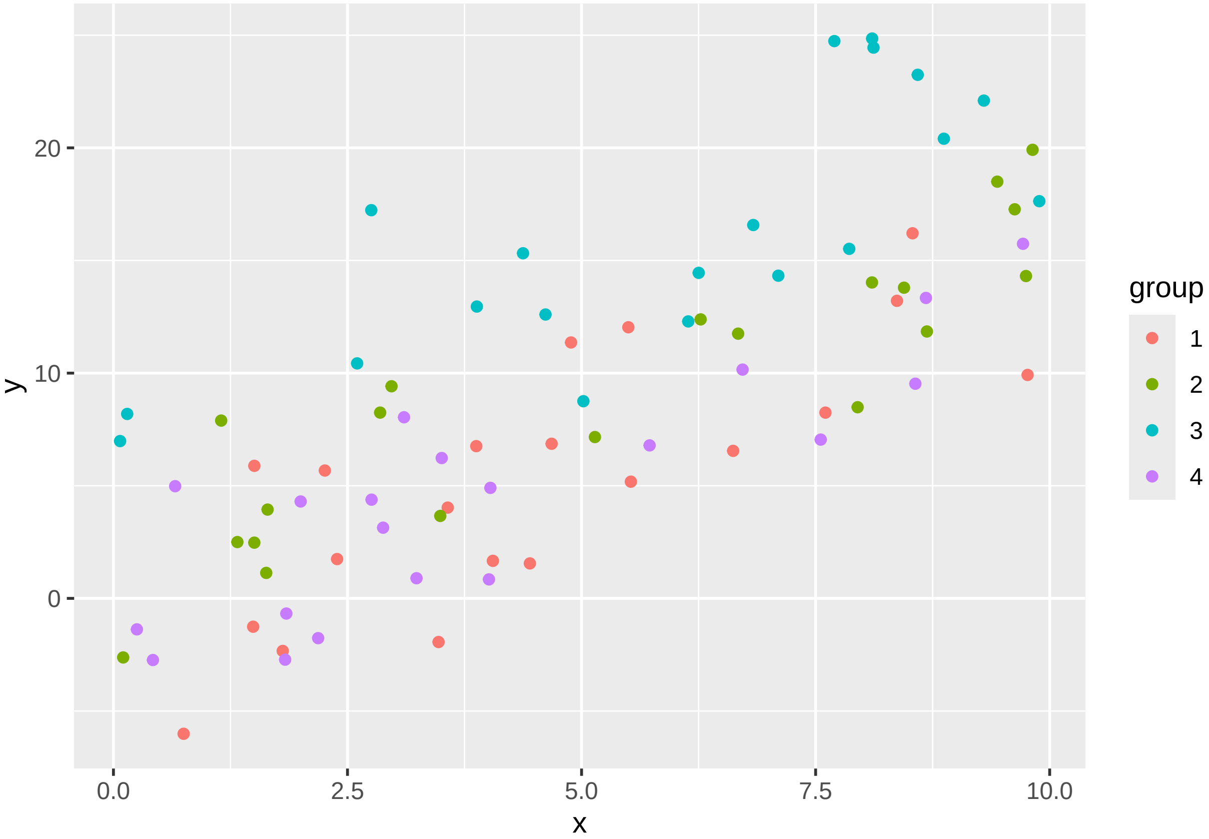
```r
library(lmerTest)
```

```
##
## Attaching package: 'lmerTest'

## The following object is masked from 'package:lme4':
##
##     lmer

## The following object is masked from 'package:stats':
##
##     step
```

```
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(faraway)
library(RLRsim)

ggplot(data = dat, aes(x = x, y = y, color = group))+
  geom_point()
```



From the plot, we can see that there's a clear distinction between each group. We can see that the group 3 tend to have the highest y value. Group 1 and 4 look pretty similar in term of the y value.

Problem 1b

```
options(contrasts = c("contr.sum", "contr.poly"))

model <- aov(y ~ group + x, dat)
summary(model)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        3 1696.7   565.6   53.66 <2e-16 ***
## x            1 1682.0  1682.0  159.57 <2e-16 ***
## Residuals   75  790.5    10.5
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

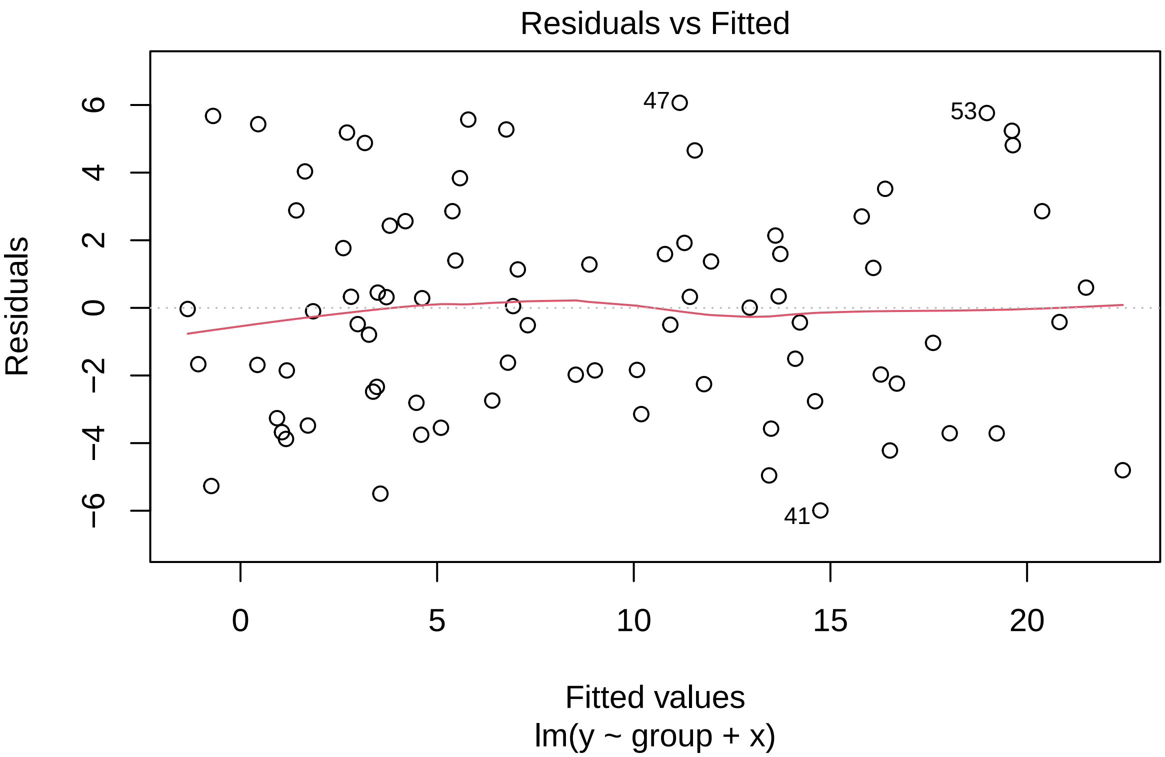
```
lm_model<-lm(y ~ group + x, dat)
summary(lm_model)
```

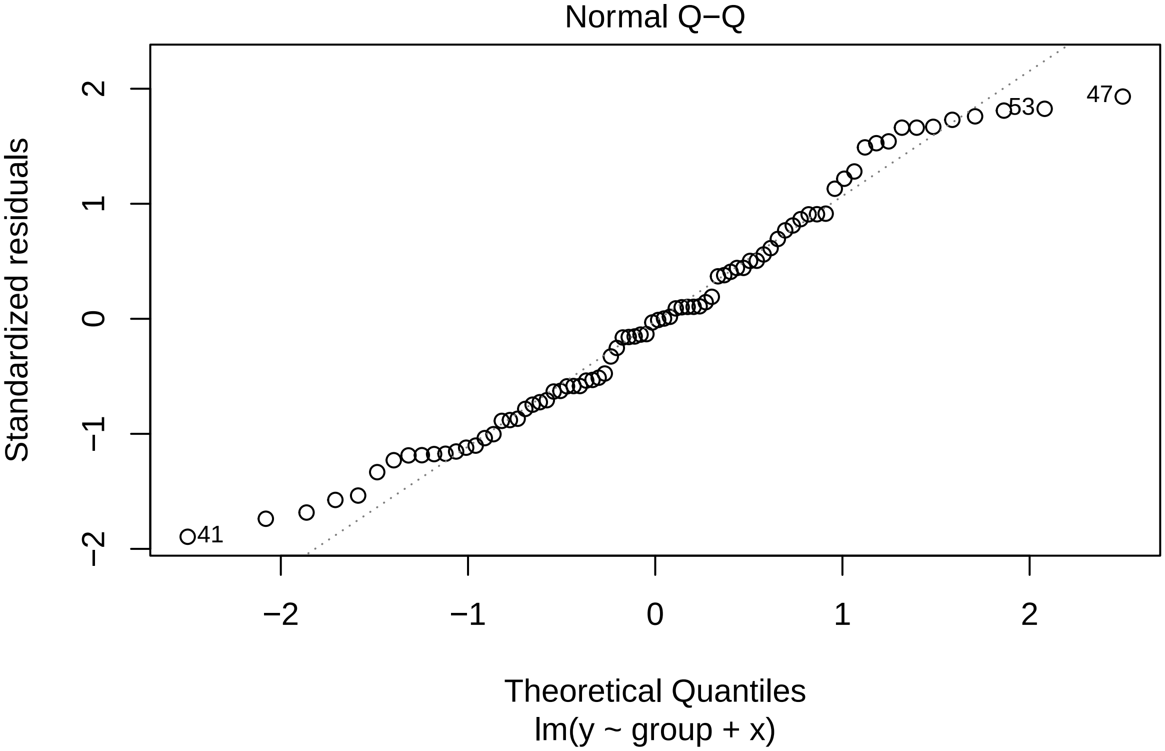
```
##
## Call:
## lm(formula = y ~ group + x, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.992 -2.372 -0.067  2.212  6.065
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0114     0.7168   1.411    0.162
## group1       -2.9372     0.6306  -4.658 1.35e-05 ***
## group2       -0.1208     0.6305  -0.192    0.849
## group3        5.8077     0.6402   9.072 1.09e-13 ***
## x             1.5790     0.1250  12.632  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.247 on 75 degrees of freedom
## Multiple R-squared:  0.8104, Adjusted R-squared:  0.8003
## F-statistic: 80.14 on 4 and 75 DF,  p-value: < 2.2e-16
```

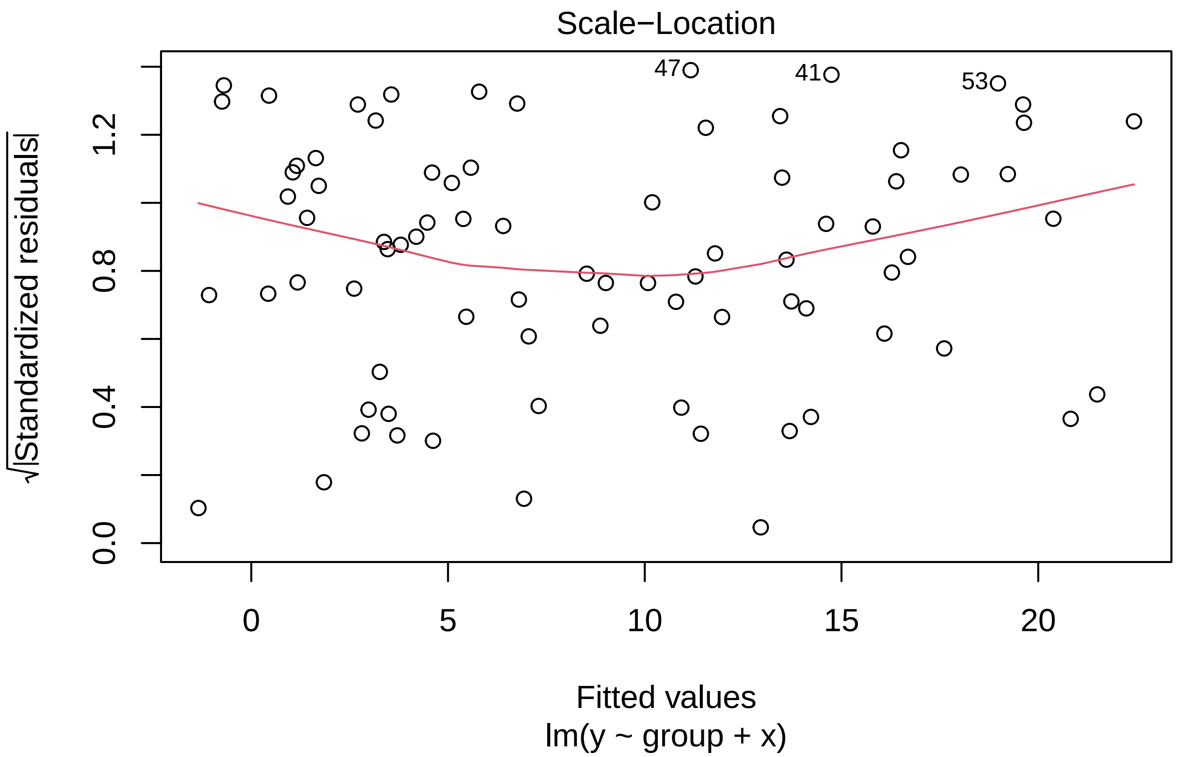
From the summary of the model, we can see that both x and groups are significant. Thus, we’ll reject the null hypothesis $H_0 : \alpha_i = 0, \forall i$, both group and x are significant. The MSE provides the estimate of $\hat{\sigma} = 10.5$. The estimated overall mean is 1.0113543.

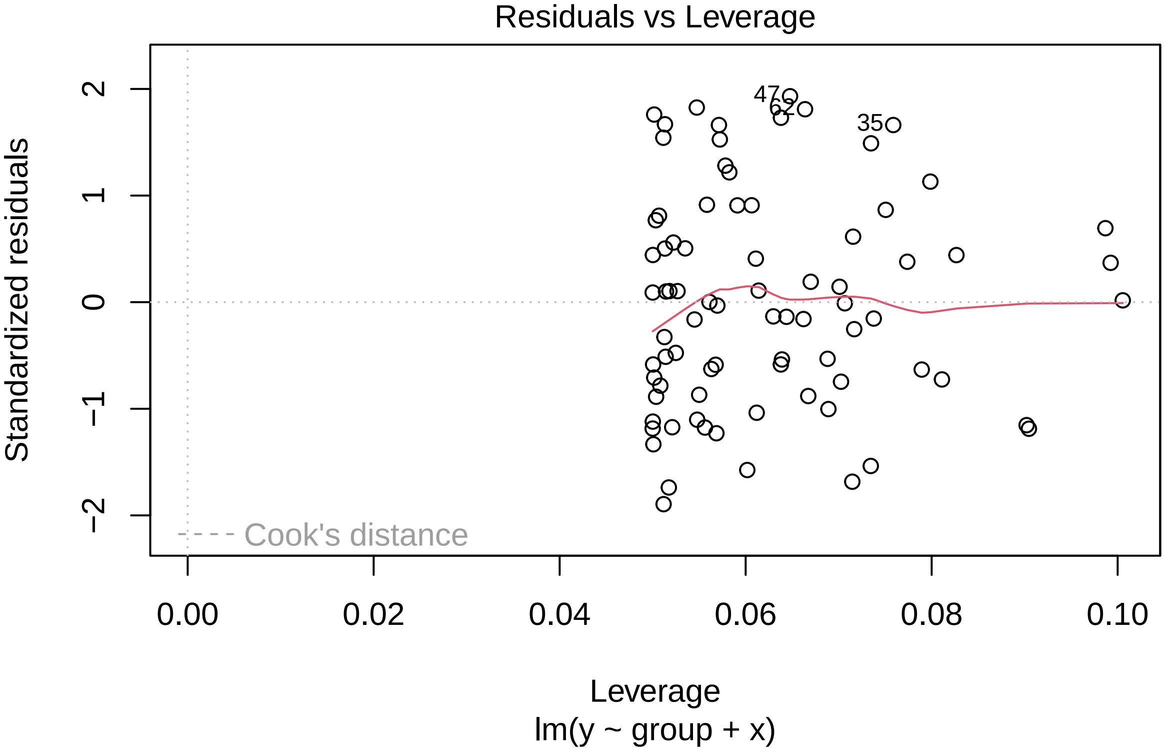
From the linear model, we can see that all variables are significant except for the group2.

```
plot(lm_model)
```


Residuals vs Fitted
Residuals
6
4
2
0
−2
−4
−6
47
53
41
0
5
10
15
20
Fitted values
lm(y ~ group + x)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
47
53
41
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ group + x)





According to the residual plot, we can see that the residuals are random and uniform, which means the model are efficient in capturing the information in the data. From the Q-Q plot, we can see that the observations are not quite normal where it is off on both sides from theoretical quantile line.

Problem 1c

```
model2<-lmer(y ~ 1 + (1|group) + x, data = dat)
summary(model2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: y ~ 1 + (1 | group) + x
##    Data: dat
##
## REML criterion at convergence: 426.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7864 -0.7475 -0.0404  0.6432  1.9334
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  group    (Intercept) 16.07    4.009
##  Residual             10.54    3.247
## Number of obs: 80, groups:  group, 4
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   0.9679     2.1287  3.5623   0.455    0.676
## x             1.5877     0.1249 75.2918  12.715   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##   (Intr)
## x -0.290
```

```
coef(model2)$group
```

```
##   (Intercept)        x
## 1  -1.8727001 1.587736
## 2   0.8476946 1.587736
## 3   6.5830360 1.587736
## 4  -1.6863289 1.587736
```

```
print(unique(z %*% alpha))
```

```
##            [,1]
## [1,] -3.5876582
## [2,]  0.7393967
## [3,]  6.3513813
## [4,] -4.5215027
```

We see that this gives identical estimates to the ANOVA method: $\hat{\sigma}^2 = 10.54, \hat{\sigma}^2_\alpha = 16.07$, and $\hat{\mu} = 0.9679$.

We can see that the coefficient of x for the random effect model is similar to the original linear model. We know the truth for group 1 is -3.5877, group 2 is 0.7394, group 3 is 6.3514, and group 4 is -4.5215. Thus, we can see that from the intercept, we can see that group 2 and group 3 is pretty close to the truth, but group 1 and 4 are off from the truth.

Problem 1d & 1e

```
anova_results_fixed<-anova(model2, ddf = "Kenward-Roger")
anova_results_random <- rand(model2)

print(anova_results_fixed)
```

```
## Type III Analysis of Variance Table with Kenward-Roger's method
##   Sum Sq Mean Sq NumDF  DenDF F value    Pr(>F)
## x 1699.4  1699.4     1 75.297  161.22 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(anova_results_random)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## y ~ x + (1 | group)
##             npar  logLik    AIC    LRT Df Pr(>Chisq)
## <none>         4 -213.16 434.31
## (1 | group)    3 -237.26 480.52 48.205  1  3.838e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the summary of ANOVA (Kenward-Roger), we can see that the fixed effect has a p-value lower than 0.05, and this means that we will reject the null hypothesis, which fixed effect is not significant. Therefore, the fixed effect is significant according to the ANOVA. In addition, we can see that the result for the random effect is also significant with a p-value lower than 0.05. This means that the random effect is also significant. Hence, the results show us that we should keep both the fixed effect and the random effect

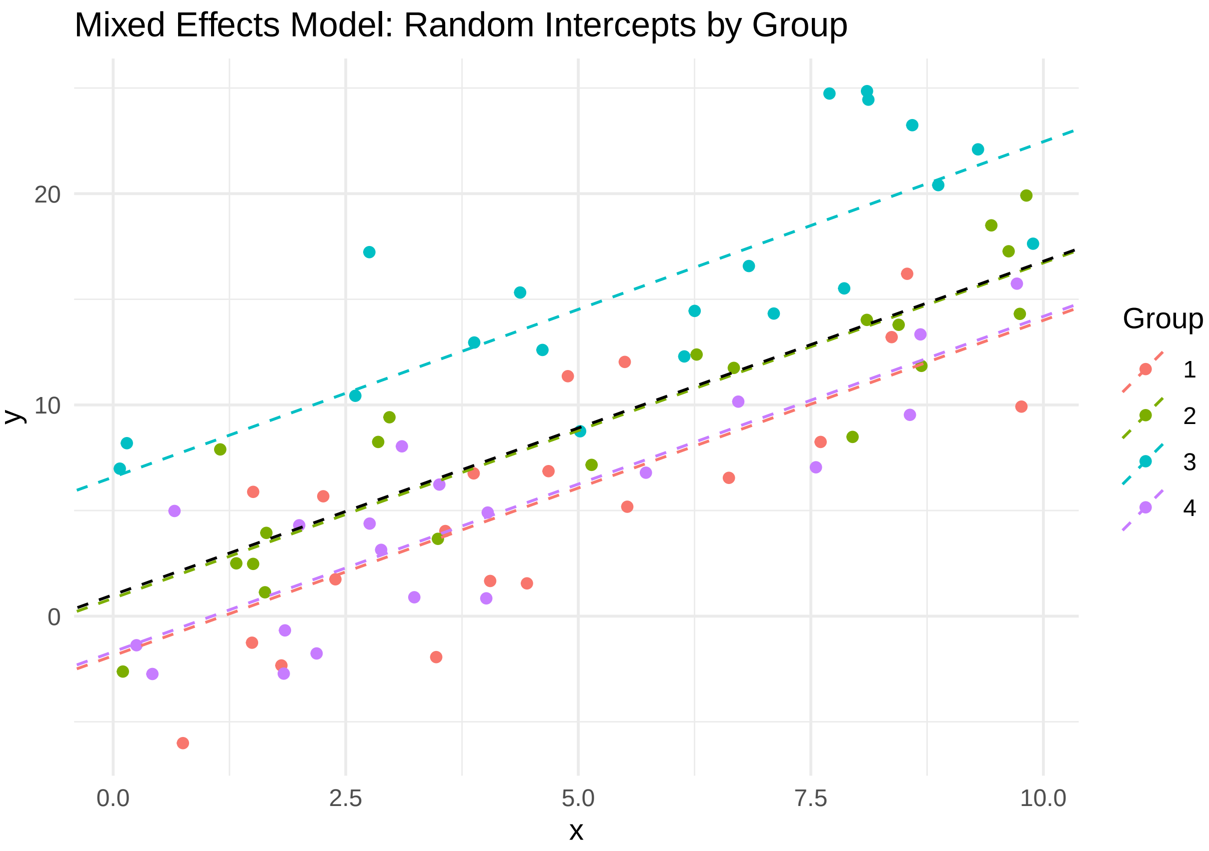
Problem 1f

```
ran_intercepts <- ranef(model2)$group %>%
  as.data.frame() %>%
  rename(intercept = `(Intercept)`) %>%
  mutate(group = rownames(ranef(model2)$group))

ran_intercepts$intercept <- ran_intercepts$intercept + fixef(model2)[1]
ran_intercepts$slope <- fixef(model2)["x"]

lm_intercept <- coef(lm_model)[1]
lm_slope <- coef(lm_model)[5]
```

```
ggplot(dat, aes(x = x, y = y, color = as.factor(group))) +
  geom_point() +
  geom_abline(data = ran_intercepts, aes(intercept = intercept, slope = slope, color = as.factor(group)
  geom_abline(intercept = lm_intercept, slope = lm_slope, color = "black", linetype = "dashed", aes(lin
  labs(title = "Mixed Effects Model: Random Intercepts by Group",
       x = "x",
       y = "y",
       color = "Group") +
  theme_minimal() +
  scale_color_discrete(name = "Group") +
  scale_linetype_manual(values = c("Linear Model" = "dashed")) +
  scale_linetype_manual(name = "Model Type", values = c("dashed", "solid"), labels = c("Mixed Effects (
```

```
## Warning: ‘geom_abline()‘: Ignoring ‘mapping‘ because ‘slope‘ and/or ‘intercept‘ were
## provided.
```

```
## Scale for linetype is already present.
## Adding another scale for linetype, which will replace the existing scale.
```



We can see that the random effect model fits well with the observation with different slope between each groups. We can see that the linear model (black, dashed line) is in the middle of the observations. However, we can see that there's clear distinction between the group, so this might mean that the Random Effect model is preferred.

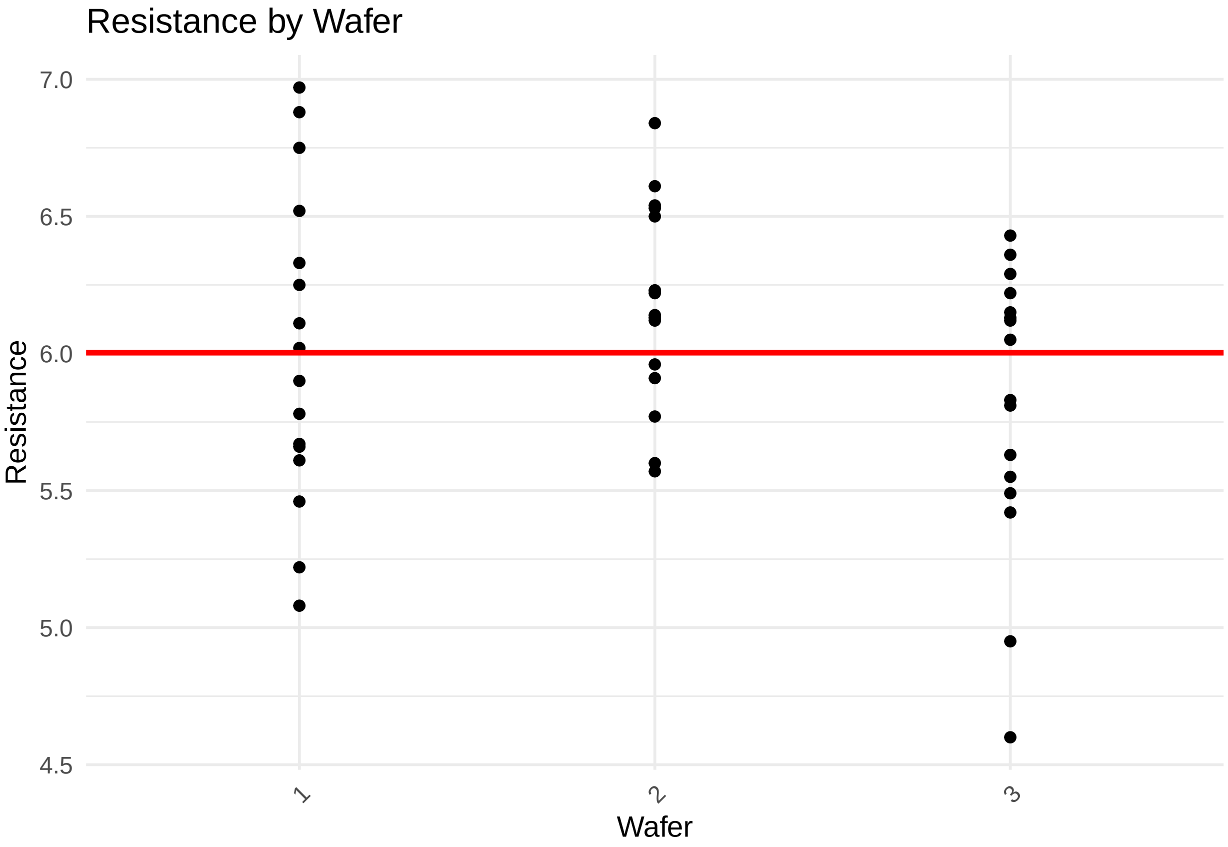
Problem 2a

```r
data<-semicond

overall_mean <- mean(data$resistance, na.rm = TRUE)
print(overall_mean)
```

```
## [1] 6.002917
```

```r
ggplot(data, aes(x = factor(Wafer), y = resistance)) +
  geom_point() +
  geom_hline(yintercept = overall_mean, color = "red", linetype = "solid", size = 1) +
  labs(title = "Resistance by Wafer",
       x = "Wafer",
       y = "Resistance") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



We can see that the Wafer 2 has the highest mean resistance and Wafer 3 has the lowest.

Problem 2b

```
contrasts(data$ET) <- contr.sum
contrasts(data$position) <- contr.sum

model<-lm(resistance ~ ET * position, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = resistance ~ ET * position, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01333 -0.25750  0.04333  0.28333  0.74667
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.002917   0.067228  89.292   <2e-16 ***
## ET1           -0.377083   0.116442  -3.238   0.0028 **
## ET2           -0.037083   0.116442  -0.318   0.7522
## ET3            0.084583   0.116442   0.726   0.4729
## position1      0.017917   0.116442   0.154   0.8787
## position2      0.131250   0.116442   1.127   0.2681
## position3     -0.255417   0.116442  -2.194   0.0357 *
## ET1:position1 -0.030417   0.201683  -0.151   0.8811
## ET2:position1  0.009583   0.201683   0.048   0.9624
## ET3:position1  0.031250   0.201683   0.155   0.8778
## ET1:position2 -0.307083   0.201683  -1.523   0.1377
## ET2:position2  0.089583   0.201683   0.444   0.6599
## ET3:position2  0.127917   0.201683   0.634   0.5304
## ET1:position3  0.182917   0.201683   0.907   0.3712
## ET2:position3  0.056250   0.201683   0.279   0.7821
## ET3:position3 -0.058750   0.201683  -0.291   0.7727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4658 on 32 degrees of freedom
## Multiple R-squared:  0.4211, Adjusted R-squared:  0.1498
## F-statistic: 1.552 on 15 and 32 DF,  p-value: 0.1449
```

We can see that only ET1 and position3 are significant. We can also see that the intercept and the overall mean is the same.

Problem 2c

```
random_model<-lmer(resistance ~ 1 + ET * position + (1|ET:Wafer), data = data)
summary(random_model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: resistance ~ 1 + ET * position + (1 | ET:Wafer)
##    Data: data
##
## REML criterion at convergence: 72.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.91111 -0.45920  0.01029  0.46868  1.31146
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ET:Wafer (Intercept) 0.1058   0.3253
##  Residual             0.1111   0.3334
## Number of obs: 48, groups:  ET:Wafer, 12
##
## Fixed effects:
##               Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)    6.002917   0.105506  8.000000  56.897 1.01e-11 ***
## ET1           -0.377083   0.182741  8.000000  -2.063  0.07296 .
## ET2           -0.037083   0.182741  8.000000  -0.203  0.84426
## ET3            0.084583   0.182741  8.000000   0.463  0.65580
## position1      0.017917   0.083348 24.000000   0.215  0.83161
## position2      0.131250   0.083348 24.000000   1.575  0.12841
## position3     -0.255417   0.083348 24.000000  -3.064  0.00532 **
## ET1:position1 -0.030417   0.144362 24.000000  -0.211  0.83490
## ET2:position1  0.009583   0.144362 24.000000   0.066  0.94762
## ET3:position1  0.031250   0.144362 24.000000   0.216  0.83045
## ET1:position2 -0.307083   0.144362 24.000000  -2.127  0.04388 *
## ET2:position2  0.089583   0.144362 24.000000   0.621  0.54075
## ET3:position2  0.127917   0.144362 24.000000   0.886  0.38437
## ET1:position3  0.182917   0.144362 24.000000   1.267  0.21729
## ET2:position3  0.056250   0.144362 24.000000   0.390  0.70024
## ET3:position3 -0.058750   0.144362 24.000000  -0.407  0.68764
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

We can see that $\hat{\sigma}^2 = 0.1111, \hat{\sigma}^2_\alpha = 0.1058$, and $\hat{\mu} = 6.002917$ for the model with random effect ET:Wafer.

Problem 2d

```
anova_fixed<-anova(random_model, ddf = "Kenward-Roger")

print(anova_fixed)
```

```
## Type III Analysis of Variance Table with Kenward-Roger’s method
##              Sum Sq Mean Sq NumDF DenDF F value  Pr(>F)
## ET          0.64740 0.21580     3     8  1.9415 0.20150
## position    1.12889 0.37630     3    24  3.3855 0.03451 *
## ET:position 0.80948 0.08994     9    24  0.8092 0.61253
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From the summary of ANOVA table (Kenward-Roger’s method), we can see that only the variable position has a p-value lower than 0.05, which means that the variable position is significant. Therefore, we’ll only keep the variable position as the fixed effect in our final model.

```
final_model<-lmer(resistance ~ 1 + position + (1|ET:Wafer), data = data)
anova_final<-anova(final_model, ddf = "Kenward-Roger")
print(anova_final)
```

```
## Type III Analysis of Variance Table with Kenward-Roger’s method
##          Sum Sq Mean Sq NumDF DenDF F value  Pr(>F)
## position 1.1289  0.3763     3    33  3.5714 0.02427 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From the test of final model, we can see that the position variable is significant, so we can conclude that the variable position is signficant to the model.

Problem 2e

```
anova_random<-rand(final_model)
print(anova_random)
```

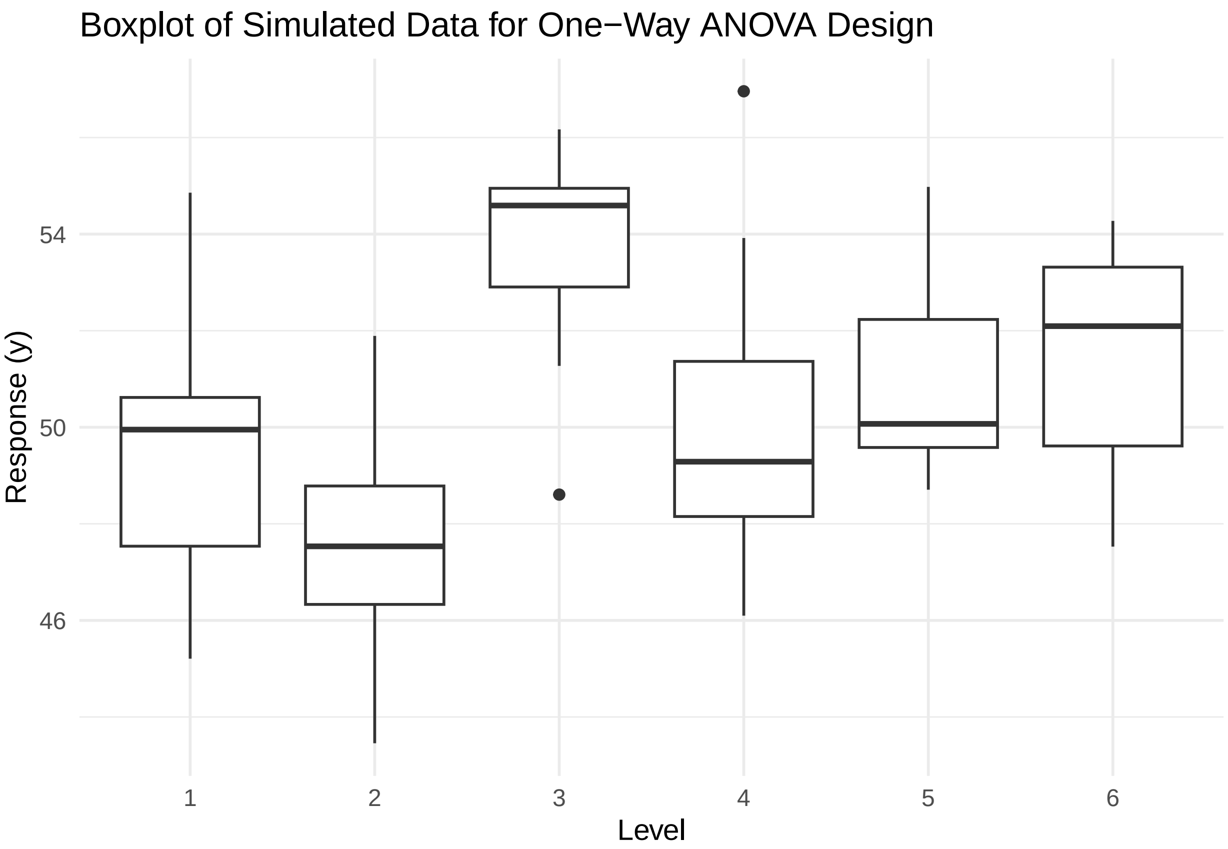
```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## resistance ~ position + (1 | ET:Wafer)
##                 npar  logLik    AIC    LRT Df Pr(>Chisq)
## <none>             6 -29.469 70.938
## (1 | ET:Wafer)     5 -38.017 86.033 17.096  1  3.555e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From the ANOVA for the random effect, we can see that the random effect (1|ET:Wafer) has a p-value lower than 0.05, which means that we will reject the null hypothesis. This means that the random effect (1|ET:Wafer) is significant. Therefore, we shall keep the random effect in our model.

Problem 3

```
set.seed(123)
mu <- 50
sigma2_alpha <- 2
```

```r
sigma2 <- 5 * sigma2_alpha

num_levels <- 6
obs_per_level <- 10

alpha <- rnorm(num_levels, mean = 0, sd = sqrt(sigma2_alpha))

# Simulate observations
data <- data.frame(
  level = rep(1:num_levels, each = obs_per_level),
  y = mu + rep(alpha, each = obs_per_level) + rnorm(num_levels * obs_per_level, mean = 0, sd = sqrt(sig
)

ggplot(data, aes(x = factor(level), y = y)) +
  geom_boxplot() +
  labs(title = "Boxplot of Simulated Data for One-Way ANOVA Design",
       x = "Level",
       y = "Response (y)") +
  theme_minimal()
```



In this simulation, we chose the $\mu = 3, \hat{\sigma}^2 = 1, \hat{\sigma}_\alpha^2 = 5$, and $\hat{\mu} = 50$. According to the box plot, we can see that the third level has the highest median response y and level 2 has the lowest median response y. From the simulation, we can clearly see that there's a difference between each leven so the random effect model is suggested.

```r
lmer_model<-lmer(y ~ (1|level), data = data)
summary(lmer_model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite’s method [
## lmerModLmerTest]
## Formula: y ~ (1 | level)
##    Data: data
##
## REML criterion at convergence: 295.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.75095 -0.70831 -0.09417  0.65777  2.56298
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  level    (Intercept) 3.392    1.842
##  Residual             7.049    2.655
## Number of obs: 60, groups:  level, 6
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  50.5803     0.8264  5.0000   61.21  2.2e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova_model <- aov(y ~ factor(level), data = data)
summary(anova_model)
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## factor(level)  5  204.9   40.97   5.813 0.00023 ***
## Residuals     54  380.6    7.05
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova_summary<-summary(anova_model)

MSA <- anova_summary[[1]]["factor(level)", "Mean Sq"]
MSE <- anova_summary[[1]]["Residuals", "Mean Sq"]

estimated_sigma2_a<-(MSA - MSE) / (obs_per_level)
estimated_sigma2 <- MSE

cat("Estimated sigma^2_alpha:", estimated_sigma2_a)
```

```
## Estimated sigma^2_alpha: 3.392458
```

```
cat("Estimated sigma^2:", estimated_sigma2)
```

```
## Estimated sigma^2: 7.048796
```

Thus, our maximum likelihood gives us $\sigma_a^2 lpha = 3.392458$ and $\sigma^2 = 7.048796$.

```r
mle<-VarCorr(lmer_model)
print(mle)
```

```
##  Groups   Name        Std.Dev.
##  level    (Intercept) 1.8419
##  Residual             2.6550
```

We can see that the ANOVA estimators for σ_α^2 and σ^2 are similar to the the maximum likelihood estimator with $\sigma_\alpha^2 = 1.8419^2 = 3.39259561$ and $\sigma^2 = 2.655^2 = 7.102225$. However, it's off from our true σ_α^2 and σ^2.