STA610 Lab04

Yuren Zhou

2024-09-20

- Write down your answers in any blank sheet and submit your work in paper during the lab.
- Your work will not be graded. As long as you submit, you will get a full credit.
- For those who missed the lab today, you can submit it via email to me for half credit.

One-way ANOVA

Load the package.

```
library(lme4)
```

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

Generate data in case www2.stat.duke.edu is still down.

```
# Set seed for reproducibility
set.seed(0)

# Number of groups and observations per group
n_groups <- 4
n_obs_per_group <- 10

# Create a group factor
group <- factor(rep(1:n_groups, each = n_obs_per_group))

# Define group means
group_means <- c(5, 10, 15, 20)

# Generate the response variable (Y) with random noise
y <- rep(group_means, each = n_obs_per_group) + rnorm(n_groups * n_obs_per_group, mean=0, sd = 2)

# Combine into a data frame
data_anova <- data.frame(group, y)

# Take a look at the data
head(data_anova)
```

```
##   group        y
## 1     1 7.525909
```

```
## 2     1 4.347533
## 3     1 7.659599
## 4     1 7.544859
## 5     1 5.829283
## 6     1 1.920100
```

Fit a one-way ANOVA model using *lme4*.

```
# Fit the model using lme4
anova_model <- lmer(y ~ 1 + (1 | group), data = data_anova, REML = FALSE)

# Summarize the model
summary(anova_model)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: y ~ 1 + (1 | group)
##    Data: data_anova
##
##      AIC      BIC   logLik deviance df.resid
##      182      187      -88      176       37
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.23280 -0.66737 -0.04368  0.87337  2.32573
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  group    (Intercept) 30.769   5.547
##  Residual              2.995   1.731
## Number of obs: 40, groups:  group, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   12.610      2.787   4.525
```

Get confidence intervals using *confint*.

```
# Get confidence intervals for both fixed and random effects
confint(anova_model)
```

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

```
##                 2.5 %    97.5 %
## .sig01      3.124333 13.658199
## .sigma      1.396553  2.222698
## (Intercept) 5.531632 19.688103
```

```
# If you only want confidence intervals for the intercept
confint(anova_model, parm = "beta_")
```

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

```
##                 2.5 %  97.5 %
## (Intercept) 5.531632 19.6881
```

```
# If you only want confidence intervals for the standard deviations
confint(anova_model, parm = "theta_")
```

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

```
##           2.5 %    97.5 %
## .sig01 3.124333 13.658199
## .sigma 1.396553  2.222698
```

Extract random effect modes using *ranef*.

```
# Extract random effect modes
random_effects <- ranef(anova_model)

# View random effect modes for each group
random_effects$group
```

```
##   (Intercept)
## 1   -6.825581
## 2   -3.302684
## 3    2.504569
## 4    7.623696
```

What are the obtained estimates for the effect of each group (fixed effect intercept + random effect mode)?

How do they compare to the sample means of each group?

How do they compare to the population means of each group $(5, 10, 15, 20)$?

Linear Models with Interactions (Linear Fixed Effects Models)

Let $y = (y_1, \ldots, y_n)$ be the response variable, $x = (x_1, \ldots, x_n)$ be a continuous covariate, and $z = (z_1, \ldots, z_n)$ be a categorical covariate. Say each i corresponds to a student, x corresponds to household income, z corresponds to school, and y corresponds to test score. How to interpret the differences between the following linear models?

1. $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$;
2. $y_i = (\beta_0 + a_{0,z_i}) + \beta_1 x_i + \epsilon_i$;
3. $y_i = \beta_0 + (\beta_1 + a_{1,z_i}) x_i + \epsilon_i$;
4. $y_i = (\beta_0 + a_{0,z_i}) + (\beta_1 + a_{1,z_i}) x_i + \epsilon_i$.

We simulate some data in the following.

```r
# Set seed for reproducibility
set.seed(0)

# Define the number of levels for the categorical variable z
n_levels <- 3  # For example, three groups in z
n_obs_per_level <- 50  # Number of observations per level of z

# Create a categorical variable z with three levels
z <- factor(rep(1:n_levels, each = n_obs_per_level), labels = c("Group1", "Group2", "Group3"))

# Generate a continuous variable x (randomly sampled from a normal distribution)
x <- rnorm(n_levels * n_obs_per_level, mean = 50, sd = 10)

# Define interaction terms (different slopes for each group)
beta_0 <- c(5, 7, 9)  # Different intercepts for each group
beta_x <- c(1, 1, 1)  # Different slopes for x in each group

# Generate the response variable y with interaction effects
y <- numeric(length(x))
for (i in 1:n_levels) {
  group_idx <- which(z == levels(z)[i])
  y[group_idx] <- beta_0[i] + beta_x[i] * x[group_idx] + rnorm(length(group_idx), sd = 2)
}

# Combine into a data frame
data <- data.frame(x, z, y)

# Take a look at the data
head(data)
```

```
##          x      z        y
## 1 62.62954 Group1 70.57006
## 2 46.73767 Group1 49.11483
## 3 63.29799 Group1 68.10494
## 4 62.72429 Group1 72.46373
## 5 54.14641 Group1 60.92767
## 6 34.60050 Group1 39.09613
```

Fit all four linear models above to the simulated data. What is the R-square of each model?

Select the best model using information criteria (e.g. AIC/BIC) or any other approach you find appropriate (e.g. F-tests). Which model have you selected?