Linear Models: Homework 3

Student: Khuong Quynh Long

December 09, 2021

Assignment

The one-way ANOVA factor effects model is given by

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \, ; i = 1, ..., n_j; j = 1, ..., t \tag{1}$$

Observation Y_{ij} refers to the i^{th} replicate from treatment j^{th}. With $\varepsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$
The zero-sum restriction

$$\sum_{j=1}^{t} \tau_j = 0$$

Question 1

The equivalent regression model formulation for the factor effects model (1) is

$$Y_i = \beta_0 + \sum_{j=1}^{t-1} \beta_j x_{ij} + \varepsilon_i \tag{2}$$

with $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$

Define the (1, 0, -1) dummies as ($k = 1, ..., t-1$)

x_{ik} = 1 if observation i belongs to treatment group k
x_{ik} = 0 if observation i does not belong to treatment group k
x_{ik} = -1 if observation i belongs to treatment group t

These dummy variables are also called *“deviation regressors”*

For example, dummy regressors coding with t = 3 are

	Dummy 1	Dummy 2
Treatment group 1	1	0
Treatment group 2	0	1
Treatment group 3	-1	-1

With these dummy definitions, the model by group becomes

Group 1: $Y_i = \mu + \tau_1 + \varepsilon_i = \beta_0 + \beta_1 + \varepsilon_i$

Group 2: $Y_i = \mu + \tau_2 + \varepsilon_i = \beta_0 + \beta_2 + \varepsilon_i$

...

Group t: $Y_i = \mu - (\tau_1 + \tau_2 +, ..., +\tau_{t-1}) + \varepsilon_i = \beta_0 - (\beta_1 + \beta_2 +, ..., +\beta_{t-1}) + \varepsilon_i$

with $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$

Therefore:

- The μ or β_0 presents the population grand mean
- The τ_j or β_j presents the distance of the group j from the grand mean. Thus, is the effect of being in group j as compared to the grand mean.

Question 2

Analyze the PWD data with outcome ADWG0050 and zero-sum restriction parameterisation.

```
library(tidyverse)
library(magrittr)
library(ggridges)
load("data/PWD.RData")
head(PWD) %>% knitr::kable(caption = "PWD data")
```

Table 2: PWD data

Pen	Treatment	Feeder	Sex	W0	P0	ADWG0021	ADWG2150	ADWG0050
1	A	1	1	110.0	16	166.6667	525.8621	375.0000
2	A	1	1	111.0	16	151.7857	471.9828	337.5000
3	C	2	1	108.5	16	130.9524	608.1178	407.7083
4	C	2	1	99.0	16	151.7857	568.9655	393.7500
5	E	3	1	103.0	16	133.9286	502.1552	347.5000
6	E	3	1	104.5	16	147.3214	500.0000	351.8750

```r
PWD %>% ggplot(aes(x = ADWG0050, y = Treatment,
                   fill = Treatment, color = Treatment)) +
    # Include lines for mean of each group
    stat_density_ridges(quantile_lines = T, quantiles = 2, quantile_fun = mean,
                        jittered_points = T, alpha = 0.5, show.legend = F) +
    theme_minimal() +
    coord_flip() +
    scale_fill_brewer(palette = "Set1") +
    scale_color_brewer(palette = "Set1")
```

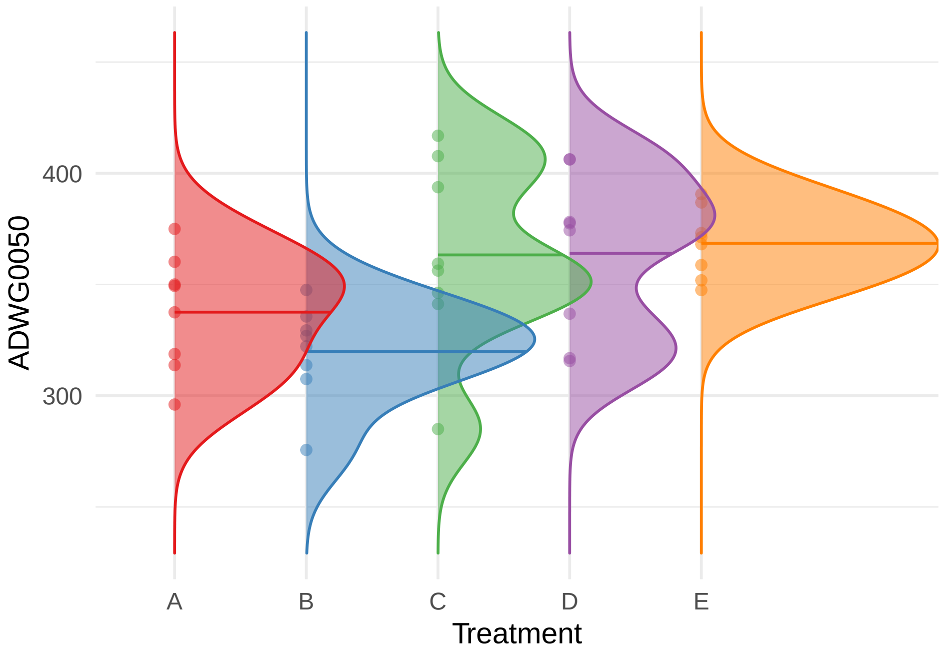

Figure 1: Distribution of ADWG0050 among groups

```r
# Create deviation regressors
PWD %<>% mutate(
    DummyA = ifelse(Treatment == "A", 1,
                    ifelse(Treatment == "E", -1, 0)),
    DummyB = ifelse(Treatment == "B", 1,
                    ifelse(Treatment == "E", -1, 0)),
    DummyC = ifelse(Treatment == "C", 1,
                    ifelse(Treatment == "E", -1, 0)),
    DummyD = ifelse(Treatment == "D", 1,
                    ifelse(Treatment == "E", -1, 0)),
)
```

```r
# Show dummy coding and descriptive statistics
PWD %>% group_by(Treatment) %>%
    summarise(A = mean(DummyA), B = mean(DummyB),
              C = mean(DummyC), D = mean(DummyD),
              `Sample size` = n(),
              `ADWG0050 (mean)` = mean(ADWG0050),
              `ADWG0050 (SD)` = sd(ADWG0050)) %>%
    knitr::kable(caption = "Dummy coding and descriptive statistics")
```

Table 3: Dummy coding and descriptive statistics

Treatment	A	B	C	D	Sample size	ADWG0050 (mean)	ADWG0050 (SD)
A	1	0	0	0	8	337.5781	26.33182
B	0	1	0	0	8	319.8073	21.73441
C	0	0	1	0	8	363.3073	42.66562
D	0	0	0	1	8	363.9844	36.55954
E	-1	-1	-1	-1	8	368.5156	15.43975

Each treatment group has 8 observations (8 pens of piglets), with mean of ADWG0050 (average daily weight gain between 0 and 50 days post-weaning (g/day)) are 337.58, 319.81, 363.31, 363.98, and 368.52 (g/day) for treatment group A, B, C, D, and E, respectively. (Table 3)

- The parameter estimates and 90% confidence intervals are

```r
m <- lm(ADWG0050 ~ DummyA + DummyB + DummyC + DummyD, data = PWD)
# Extracting coefficients and 90% Confidence interval
mCoef <- summary(m)$coefficients
mCI90 <- confint(m, level = 0.90)
cbind(mCoef, mCI90) %>%
    knitr::kable(caption = "Parameter estimates and 90% CI")
```

Table 4: Parameter estimates and 90% CI

	Estimate	Std. Error	t value	Pr(>|t|)	5 %	95 %
(Intercept)	350.63854	4.775505	73.424384	0.0000000	342.569979	358.707104
DummyA	-13.06042	9.551011	-1.367438	0.1802058	-29.197541	3.076708
DummyB	-30.83125	9.551011	-3.228061	0.0027073	-46.968375	-14.694125
DummyC	12.66875	9.551011	1.326430	0.1932878	-3.468375	28.805875
DummyD	13.34583	9.551011	1.397322	0.1711128	-2.791291	29.482958

- Interpret the parameter estimates
  - The $\beta_0 (intercept)$ presents the grand mean of the outcome, which is 350.64 (g/day).
  - The $\beta_1 = -13.06$, $\beta_2 = -30.83$, $\beta_3 = 12.67$, $\beta_4 = 13.35$ show that the mean ADWG0050 of piglets in treatment A and B are lower than the grand mean 13.06 and 30.83 (g/day), respectively (i.e., $\bar{Y_A} = 350.64 - 13.06 = 337.58$, $\bar{Y_B} = 350.64 - 30.83 = 319.81$ (g/day)). Whereas, the mean ADWG0050 of piglets in treatment C and D are higher than the grand mean 12.67 and 13.35 (g/day), respectively (i.e., $\bar{Y_C} = 350.64 + 12.67 = 363.31$, $\bar{Y_D} = 350.64 + 13.35 = 363.98$ (g/day)).
  - The $-(\beta_1 + \beta_2 + \beta_3 + \beta_4) = 17.87$ shows that the mean ADWG0050 of piglets in treatment E is higher than the grand mean 17.87 (g/day). (i.e., $\bar{Y_E} = 350.64 + 17.87 = 368.52$ (g/day)). (Table 4)
- Perform a test for testing $H_0 : \tau_C = \tau_D = \tau_E$ versus $H_1 :$ not H_0 at the 5% level of significance

Since the research of interest is to compare the treatment effects of methods C, D, and E. We use the original ANOVA factor effect model with (1, 0) dummy coding. It is noticed that the overall F-statistic and p-value will be the same if using the zero-sum restriction defined in Question 1.

```r
# Keep treatments C, D, and E
PWD2 <- PWD %>% filter(Treatment %in% c("C", "D", "E")) %>% droplevels()
m2 <- lm(ADWG0050 ~ Treatment, data = PWD2)
anova(m2) %>%
    knitr::kable(caption = "ANOVA table for effects of treatment C, D, and E")
```

Table 5: ANOVA table for effects of treatment C, D, and E

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	2	128.3131	64.15654	0.0566864	0.9450344
Residuals	21	23767.3882	1131.78039	NA	NA

F-statistic is 0.057 and p-value is 0.945. Thus, at the 5% level of signicance, we fail to reject the null hypothesis of equal ADWG0050 (average daily weight gain between 0 and 50 days post-weaning (g/day)) among the three treatment groups.

The assessment of model assumptions is shown in the Appendix

Appendix

Assessment of model assumptions

```
par(mfrow = c(2, 2))
# Normality
qqnorm(m$residuals)
qqline(m$residuals)
# Residual distribution across groups
boxplot(m$residuals ~ PWD$Treatment,
        xlab = "treatment", ylab= "residual",
        main = "Residual distribution across groups")
abline(h = 0, lty = 2, col = 4, size = 2)
# Residual vs fitted plot
boxplot(m$residuals ~ predict(m),
     xlab = "prediction", ylab = "residual",
     main = "Residual vs fitted plot")
abline(h = 0, lty = 2, col = 4, size = 2)
# Absolute residual distribution across groups
boxplot((abs(m$residuals)) ~ PWD$Treatment,
        xlab = "Treatment", ylab = "abs(residual)",
        main = "Absolute residual distribution across groups")
abline(h = mean(abs(m2$residuals)), lty = 2, col = 4, size = 2)
```

The normality assumption is satisfied since the normal QQ plot does not show a serious deviation from normality.

Regarding conditional mean, the boxplot of residual distribution across groups shows no serious deviation from 0 value.

In terms of constant variance, no serious deviations can be detected, except for group E shows some non-constancy. However, due to the small number of observations for each boxplot. This is reasonable.

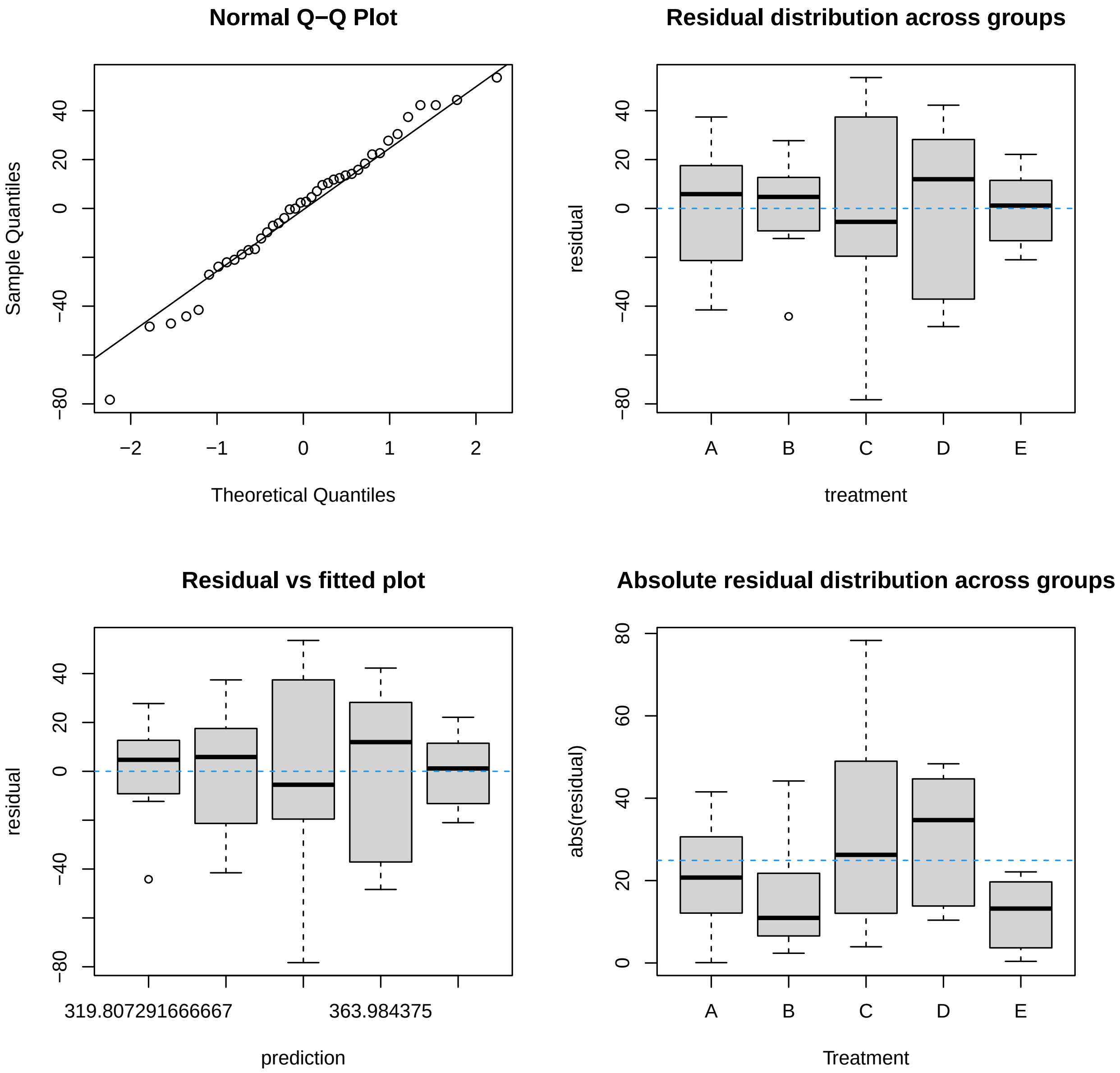

Figure 2: Assessment of model assumptions