STAT 512 Homework 6

Kathleen Wendt

03/31/2020

Part 1: Irrigation data

A study was done to investigate the effectiveness of five methods for the irrigation of blueberry shrubs. Ten farms were included in the study. Each of the five treatments was evaluated at each of the ten farms (with irrigation treatments randomly assigned to plots). The response variable is weight of the harvested fruit. The data is available from Canvas as "Irrigation.csv". Note: Be sure to define Farm `as.factor`!

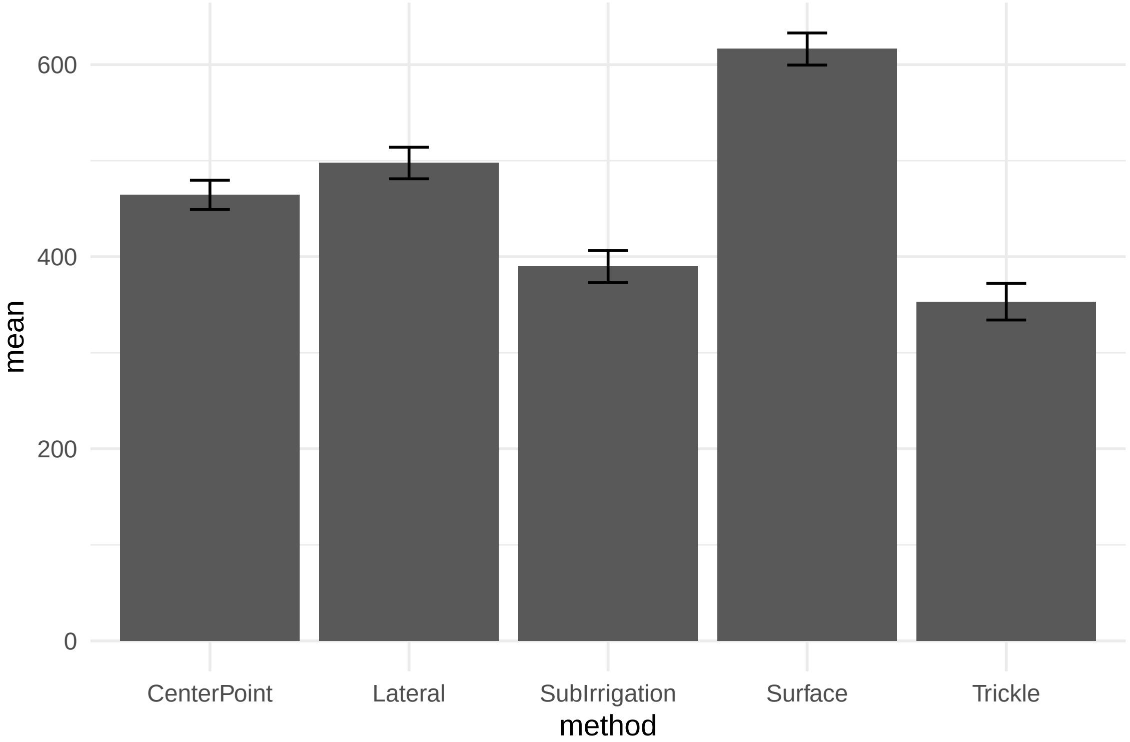
Question A: Summary table

Calculate the sample size, simple mean and SE for each method (averaging over farms). Include the resulting summary table in your assignment.

method	n	mean	sd	se
CenterPoint	10	464.4	48.23369	15.25283
Lateral	10	497.6	52.00684	16.44601
SubIrrigation	10	389.7	52.73635	16.67670
Surface	10	616.3	52.82266	16.70399
Trickle	10	353.2	60.32836	19.07750

Question B: Bar chart

Create a bar chart (with SE bars) to summarize the data. Include the resulting graph in your assignment.



Question C: RCB model and assumptions

Fit the RCB model. Inspect the diagnostic plots (Resids vs Fitted and Normal QQplot of Resids), and comment on what you see. Do the assumptions appear to be satisfied? Note: You do not have to include the diagnostic plot in your assignment, just comment on each graph. (4 pts)

Yes, assumptions appear to be satisfied.

- Residuals vs Fitted: Evenly distributed residuals. Supports assumption of linearity and constant variance.
- Normal Q-Q: Residuals tightly follow line. Supports assumption of normality.

Question D: Type 3 ANOVA for RCB

Continuing with the RCB model from the previous question, include the Type3 ANOVA table in your assignment.

term	sumsq	df	statistic	p.value
(Intercept)	577042.29	1	335.776676	0.0000000
farm	66311.52	9	4.287354	0.0007685
method	421212.52	4	61.275119	0.0000000
Residuals	61867.08	36	NA	NA

Question E: Differences by method

Can we conclude that there is a difference between the irrigation methods? Justify your response with a test statistic and p-value.

Yes, there is a statistically significant difference in blueberry weights by irrigation method, based on the test statistic (61.275), with a corresponding p-value $(1.4338943 \times 10^{-15}) < \alpha = 0.05$.

Question F: Effectiveness of blocking

Make a conclusion about the effectiveness of the blocking in this example. Justify your response with a test statistic and p-value.

Yes, blocking by farm was effective, based on the test statistic (4.287), with a corresponding p-value (0.001) $< \alpha = 0.05$.

Question G: Multiple comparisons

The investigators are interested in which irrigation methods are significantly different from each other. Use `emmeans()` function from the `emmeans` package to get Tukey-adjusted p-values for comparing treatments. Then use this information to create a "cld" display, where methods that are NOT significantly different from each other are given the same number grouping.

level1	level2	estimate	std.error	df	statistic	p.value
CenterPoint	Lateral	-33.2	18.53931	36	-1.790789	0.3944294
CenterPoint	SubIrrigation	74.7	18.53931	36	4.029276	0.0024208
CenterPoint	Surface	-151.9	18.53931	36	-8.193401	0.0000000
CenterPoint	Trickle	111.2	18.53931	36	5.998066	0.0000067
Lateral	SubIrrigation	107.9	18.53931	36	5.820066	0.0000115
Lateral	Surface	-118.7	18.53931	36	-6.402612	0.0000019
Lateral	Trickle	144.4	18.53931	36	7.788855	0.0000000
SubIrrigation	Surface	-226.6	18.53931	36	-12.222677	0.0000000
SubIrrigation	Trickle	36.5	18.53931	36	1.968790	0.3013898
Surface	Trickle	263.1	18.53931	36	14.191467	0.0000000

	method	emmean	SE	df	lower.CL	upper.CL	.group
5	Trickle	353.2	13.10927	36	326.6132	379.7868	1
3	SubIrrigation	389.7	13.10927	36	363.1132	416.2868	1
1	CenterPoint	464.4	13.10927	36	437.8132	490.9868	2
2	Lateral	497.6	13.10927	36	471.0132	524.1868	2
4	Surface	616.3	13.10927	36	589.7132	642.8868	3

Question H: Simple means and SEs

Are the simple means (part A) and emmeans (part G) the same for this analysis? What about the simple SEs (part A) versus SEs returned by emmeans (part G)?

The simple means (1A) and estimated marginal means (1G) are the same because this is a balanced design. The standard errors differ between 1A and 1G because, in balanced designs based on the model (1G), N is set, thus the standard errors are the same, reflecting equal confidence in parameter estimation by block and treatment.

Question I: One-way ANOVA

Run the analysis as a one-way ANOVA using just Method in the model. (In practice I would not do this, but try it here for illustration.) Include the ANOVA table in your assignment. How does dfResid compare to the RCB model? How does MSResid compare to the RCB model? (4 pts) Hint: Recall that MSResid = SSResid/dfResid.

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## method     4 421213  105303  36.969 1.096e-13 ***
## Residuals 45 128179    2848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `dfResid` for the one-way ANOVA is 45, which is higher, or more flexible, than the RCB model (`dfResid` = 36). The `MSResid` for the one-way ANOVA is 2848, which is higher, than the RCB model (`MSResid` = 1718.5277778).

Part 2: Fertilizer data

A fertilizer trial on a range grass (blue grama) was conducted in a randomized complete block design. Five fertilizer treatments were randomly assigned to the plots in each of five blocks, but two observations have missing values. The response variable (Y) represents phosphorous. The data is available from Canvas as "GrassMiss.csv". Note: Be sure to define Block `as.factor`!

Question A: Summary table

Calculate the simple mean for each trt (averaging over blocks). Include the resulting summary table in your assignment. Hint: Because of the NA values, it is easiest to use `aggregate()` here.

trt	y
Ctrl	2.0450
N100	1.8780
N100wP	2.3340
N50	2.0420
N50wP	2.4525

Question B: RCB model with Type 3 ANOVA

Fit the RCB model and include the Type 3 ANOVA table in your assignment.

term	sumsq	df	statistic	p.value
(Intercept)	10.1078455	1	1566.556767	0.0000000
block	0.0333232	4	1.291144	0.3204483
trt	0.9650632	4	37.392398	0.0000002
Residuals	0.0903318	14	NA	NA

Question C: Estimated marginal means

Calculate the emmeans and corresponding confidence intervals for each `trt` and include them in your assignment. Note that the SE is larger (and CIs are wider) for treatments that have missing values.

trt	estimate	std.error	df	conf.low	conf.high
Ctrl	2.053647	0.0411740	14	1.965337	2.141957
N100	1.878000	0.0359229	14	1.800953	1.955047
N100wP	2.334000	0.0359229	14	2.256953	2.411047
N50	2.042000	0.0359229	14	1.964953	2.119047
N50wP	2.447647	0.0411740	14	2.359338	2.535957

Question D: Means

Are the simple means (part A) and emmeans (part C) the same for this analysis?

The simple means (2A) and estimated marginal means (2C) are almost the same, except for the treatment categories (`Crtl` and `N50wP`) with missing `y` values.

Question E: Predict NA values

Use the coefficient estimates (from the `summary()` output) to compute predicted values for the two missing observations. Show your work for full credit. (Note that you can verify these using the `predict()` function.) (4 pts)

Obs 14 $\hat{y} = 2.02059 + 0.01365 + 0.39400 =$ **2.42824**

Obs 21 $\hat{y} = 2.02059 + 0.06765 =$ **2.08824**

```
##    block   trt  y     yhat
## 14     3 N50wP NA 2.428235
## 21     5  Ctrl NA 2.088235
```

Question F: N50wP average

Verify that the emmean for N50wP is the average of the five predicted values (one from each block) for N50wP. Show your work for full credit.

Yes, the estimated marginal mean for N50wP is the average of the five predicted values (one from each block) for N50wP.

By hand

predicted: (2.414588 + 2.410588 + 2.428235 + 2.502588 + 2.482235) / 5 = 2.447647

emmean: 2.45

Code

```
##   estimate
## 1 2.447647
## 2 2.447647
```

Appendix

```
# load packages
library(tidyverse)
library(janitor)
library(car)
library(broom)
library(kableExtra)
library(emmeans)
library(multcomp)
# set global options
knitr::opts_chunk$set(fig.width = 6,
                      fig.height = 4,
                      fig.path = "figs/",
                      echo = FALSE,
                      warning = FALSE,
                      message = FALSE)
# 1. load blueberry irrigation data
berry_data <- readr::read_csv("data/Irrigation.csv") %>%
  janitor::clean_names() %>%
  dplyr::mutate(method = as.factor(method),
                farm = as.factor(farm))
# 1a. create summary statistics table for blueberry data
berry_table <- berry_data %>%
  dplyr::group_by(method) %>%
  dplyr::summarize(
    n = n(),
    mean = mean(weight),
    sd = sd(weight),
    se = sd/sqrt(n)
    )
kableExtra::kable(berry_table)
# 1b. create summary bar chart
berry_plot <- berry_table %>%
  ggplot2::ggplot(aes(x = method, y = mean)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = mean - se,
                    ymax = mean + se),
                width = 0.2) +
  theme_minimal()
berry_plot
# 1c. fit rcb model for blueberry data
berry_rcb_lm <- lm(weight ~ farm + method, data = berry_data)
# 1c. rcb diagnostic plots
par(mfrow = c(2, 2))
plot(berry_rcb_lm)
# 1d. call and tidy anova type 3 on rcb berry
berry_type3 <- broom::tidy(car::Anova(berry_rcb_lm, type = 3))
# 1d. kable type 3 table
kableExtra::kable(berry_type3)
# 1g. emmeans for berry data by irrigation method, avg by farm block
berry_em <- emmeans::emmeans(berry_rcb_lm, pairwise ~ method)
# 1g. tidy table for emmeans with tukey-adjusted p-values
```

```r
kableExtra::kable(broom::tidy(berry_em$contrasts))
# 1g. kable cld for emmeans
kableExtra::kable(multcomp::cld(berry_em$emmeans))
# 1i. construct one-way anova with just method
berry_lm <- lm(weight ~ method, data = berry_data)
# call anova table for berry-method lm
anova(berry_lm) # fine to use `anova()` b/c only one predictor in model
# 2. load fertilizer data
fertilizer_data <- readr::read_csv("data/GrassMiss.csv") %>%
  janitor::clean_names() %>%
  dplyr::mutate(block = as.factor(block),
                trt = as.factor(trt))
# 2a. create summary statistics table for fertilizer data
fert_table <- aggregate(y ~ trt,
                        FUN = mean,
                        data = fertilizer_data)
kableExtra::kable(fert_table)
# 2b. build fertilizer rcb model
fert_rcb_lm <- lm(y ~ block + trt, data = fertilizer_data)
# 2b. call and tidy anova type 3 on rcb fertilizer
fert_type3 <- broom::tidy(car::Anova(fert_rcb_lm, type = 3))
# 2b. kable type 3 table
kableExtra::kable(fert_type3)
# 2c. calculate emmeans for fertilizer rcb model by trt, avg by block
fert_em <- emmeans::emmeans(fert_rcb_lm, pairwise ~ trt)
# 2c. tidy table for emmeans with tukey-adjusted p-values
kableExtra::kable(broom::tidy(fert_em$emmeans))
# 2e. review fertilizer rcb model summary for coeff estimates
summary(fert_rcb_lm)
# 2e. show predicted missing value for control and N50wP
fert_est_data <- data.frame(fertilizer_data,
                            yhat = predict(fert_rcb_lm,
                                           newdata = fertilizer_data))
# 2e. subset predicted observations to cross-check hand calculations
fert_est_data[c(14, 21), ]
# 2f. calculate average five predicted y values of N50wP
fert_pred_n50wp <- fert_est_data %>%
  dplyr::filter(trt == "N50wP") %>%
  dplyr::summarize(estimate = mean(yhat))
# 2f. cross-check with corresponding emmean
fert_n50wp <- broom::tidy(fert_em$emmeans) %>%
  dplyr::filter(trt == "N50wP") %>%
  dplyr::select(estimate)
# 2f. compare predicted vs estimated
dplyr::full_join(fert_pred_n50wp, fert_n50wp, by = "estimate")
```