# STAR 513: HW 3

Yvette Uwineza

Total points: 40
Questions are worth **2 pts** each, except where noted.
See Canvas calendar for due date.

Homework should be submitted as a pdf, doc or docx file via Canvas.
Use of R markdown HW template is strongly encouraged.
Add or delete code chunks as needed.
Knit frequently to avoid last minute problems!
Your submitted assignment should be neatly formatted and organized.

**Ott & Longnecker Example 8.7:** It is conjectured that if fields are overgrazed by cattle there will be soil compaction (which could lead to reduced grass). A horticulturist at the agriculture experiment station designed a study to evaluate the conjecture. Three grazing regimens are considered:
- Continuous: continuous grazing
- Rest1week: three-week grazing then one-week no grazing and
- Rest2weeks: two-week grazing then two-weeks no grazing.

A total of 21 similar plots of land are selected for the study. Each of the three grazing regimens are randomly assigned to 7 plots per regimen. After the plots are subject to the grazing regimens for 4 mounts, the the soil density (g/cm3) is measured for each plot.

The data GrazeData.csv is available from Canvas.

Prior to starting the statistical analysis, you will first need to transpose the data to long format using code something like the following. Modify the code as needed. For consistency, the levels of graze should match the bullet list above. Be sure to check the modified data!

## Q1 (4 pts)

What is the predictor variable (x)? Is this categorical or numeric?
What is the response variable (y)? Is this categorical or numeric?

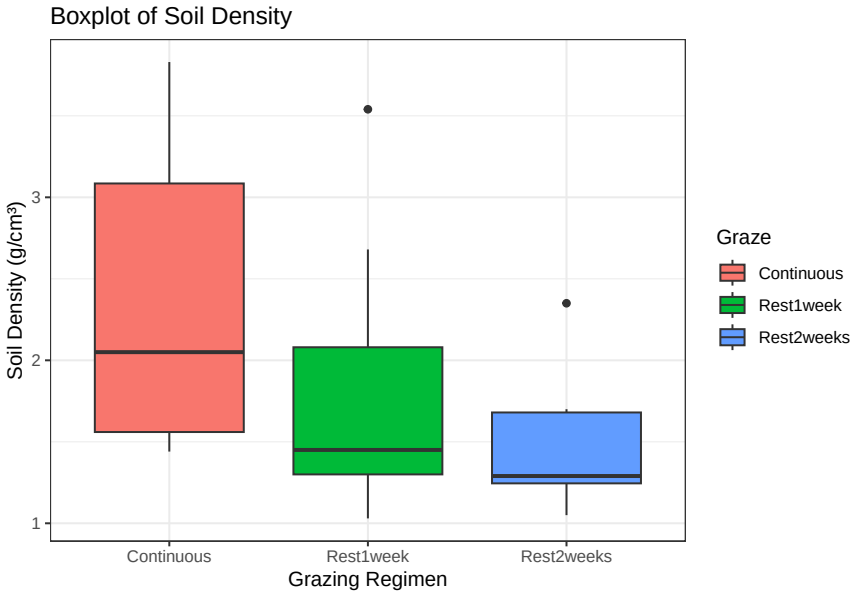
---

The predictor variable (x) is Graze, and it is a categorical variable
The response variable (y) is Soil density, and it is a numerical variable.

---

## Q2 (4 pts)

Create a boxplot of the data. Your plot should include axis labels that include the units where appropriate. Briefly comment on at least one thing you learn from this plot.




---

Continuous grazing has higher soil density compared to Rest1week and Rest2weeks median which is what we would expect when the soil is overgrazed.

---

## Q3 (4 pts)

Create a table of summary statistics of SoilDensity by Graze including n, mean and standard deviation. Hint: Use tidyverse group_by() and summarise().

---

```
## # A tibble: 3 x 4
##   Graze          n  mean std_dev
##   <fct>      <int> <dbl>   <dbl>
## 1 Continuous     7  2.37   0.952
## 2 Rest1week      7  1.83   0.925
## 3 Rest2weeks     7  1.51   0.441
```

---

## Q4

Fit an appropriate model and include the detailed "coefficients table" in your assignment. This table includes estimates, standard errors, test statistics and p-values. This can be done using tidy() or summary().

---

```
## # A tibble: 3 x 5
##   term            estimate std.error statistic     p.value
##   <chr>              <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)        2.37      0.305      7.77 0.000000367
## 2 GrazeRest1week    -0.547     0.432     -1.27 0.221
## 3 GrazeRest2weeks   -0.867     0.432     -2.01 0.0598
```

---

## Q5 (3 pts)

Calculate the estimated average soil density for each of grazing regimens, using the coefficient (or parameter) estimates from the previous question.
Notes:
(1) You must show your work to get full credit for this question.
(2) Use echo = TRUE to show your work for this question.
(3) Check your own work using simple means from Q3.

---

```r
#Q5

coefficients_values <- coef(model)

#GrazeContinuous
value_1 <- coefficients_values["(Intercept)"]
value_1
```

```
## (Intercept)
##    2.372857
```

```r
#Rest1week
value_2 <- coefficients_values["(Intercept)"] + coefficients_values["GrazeRest1week"]
value_2
```

```
## (Intercept)
##    1.825714
```

```r
#Rest2weeks
value_3 <- coefficients_values["(Intercept)"] + coefficients_values["GrazeRest2weeks"]
value_3
```

```
## (Intercept)
##    1.505714
```

---

## Q6 (3 pts)

The estimate labeled "Intercept" is $\hat{\beta}_0$. Provide a detailed one-sentence interpretation of what is being estimated in context of this study.

---

The intercept shows the estimated average soil density for plots under the continuous grazing regimen.

---

## Q7 (4 pts)

The estimate labeled "GrazeRest1week" is $\hat{\beta}_1$. Provide a detailed one-sentence interpretation of what is being estimated in context of this study.

---

This is the estimated difference in mean soil density between Rest1week regimen and the continuous grazing regimen

---

## Q8

Use model.matrix() to examine the design or model matrix (but you do not need to include it in your assignment). Provide a brief description of the second column labeled "GrazeRest1week".

---

Response

It is a dummy variable that shows the rest1week grazing regimen. This allows us to compared the soil density to the reference group which is continuous grazing.

---

## Q9

Which grazing regimen does R treat as the reference group? Why is this the reference group?

---

The continuous grazing regimen is treated as the reference group. R set it as the reference group because it is the first category that we have in our dataset.

---

## Q10

Provide an ANOVA table corresponding to the model.

---

```
## Analysis of Variance Table
##
## Response: SoilDensity
##           Df Sum Sq Mean Sq F value Pr(>F)
## Graze      2  2.692  1.3460  2.0641 0.1559
## Residuals 18 11.738  0.6521
```

---

## Q11

The ANOVA F-test from the previous question corresponds to a null hypothesis of $H_0 : \mu_1 = \mu_2 = \mu_3$. Provide a brief, statistical conclusion using $\alpha = 0.05$ (ex: Reject H0 or Fail to Reject H0.)

---

The p value is $0.1559 > 0.05$, so we fail to reject the null hypothesis.

At 0.05 significance level, there is not much evidence that the mean soil density is different from the three grazing regimens.

---

## Q12

Now provide a conclusion in context of the research study, limiting statistical jargon. Remember we don't accept H0!

---

At 0.05 significance level, there is not much evidence that the mean soil density is different from the three grazing regimens. So, this means that the grazing regimens don't significantly affects soil density. Switching up to any of the grazing regimen may not necessarily affect soil density differently.

---

## Q13

Use emmeans() to provide the emmeans (estimated marginal means).

---

```
##  Graze      emmean    SE df lower.CL upper.CL
##  Continuous   2.37 0.305 18    1.732     3.01
##  Rest1week    1.83 0.305 18    1.184     2.47
##  Rest2weeks   1.51 0.305 18    0.864     2.15
##
## Confidence level used: 0.95
```

---

### Q14

Regardless of any previous results, use pairs(, adjust = "none") to provide the unadjusted pairwise comparisons. Note: I ask for unadjusted pairwise comparisons here for learning purposes. In most cases, we prefer Tukey adjusted pairwise comparisons.

---

```
##  contrast                estimate    SE df t.ratio p.value
##  Continuous - Rest1week     0.547 0.432 18   1.268  0.2211
##  Continuous - Rest2weeks    0.867 0.432 18   2.009  0.0598
##  Rest1week - Rest2weeks     0.320 0.432 18   0.741  0.4680
```

---

### Q15

Which of the pairwise comparisons from Q14 already appeared in the default output (from Q4)? Why do we get one additional comparison in Q14?

---

The comparisons of Continuous - Rest1week and Continuous - Rest2weeks already appeared in the default ouput on from question 4. The addition comparison in Q14 is Rest1week - Rest2weeks and we have this because the function provides all the pairwise comparisons which is different from the regression that only compares it to the reference group

---

## Appendix

```r
#Retain this code chunk!!!
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
library(knitr)
library(tinytex)
library(tidyverse)
library(broom)
library(emmeans)
```

```r
 #tinytex::install_tinytex()

#Import and transpose

GrazeData <- read.csv("GrazeData.csv")
GrazeData <- GrazeData %>%
  pivot_longer(cols = everything(), names_to = "Graze", values_to = "SoilDensity") %>%
  mutate(Graze = as_factor(Graze)) %>%
  arrange(Graze)


#glimpse(GrazeData)


#Q2

ggplot(GrazeData, aes(x = Graze, y = SoilDensity, fill=Graze)) +
  geom_boxplot() +
  labs(title = "Boxplot of Soil Density ",
       x = "Grazing Regimen",
       y = "Soil Density (g/cm³)") +
  theme_bw()

#Q3


GrazeData %>%
  group_by(Graze) %>%
  summarise(
    n = n(),
    mean = mean(SoilDensity),
    std_dev = sd(SoilDensity)
  )




#Q4


model <- lm (SoilDensity ~ Graze, data = GrazeData)

tidy(model)


#Q5

coefficients_values <- coef(model)

#GrazeContinuous
value_1 <- coefficients_values["(Intercept)"]
```

```r
value_1
#Rest1week
value_2 <- coefficients_values["(Intercept)"] + coefficients_values["GrazeRest1week"]
value_2
#Rest2weeks
value_3 <- coefficients_values["(Intercept)"] + coefficients_values["GrazeRest2weeks"]
value_3

#Q8

the_matrix <- model.matrix(model)



#Q10


Anova_table<- anova(model)

Anova_table

#Q13

emmeans_values<- emmeans(model, ~ Graze)

emmeans_values

#Q14

unadjusted_pairwise<-pairs(emmeans(model, ~Graze), adjust = "none")

unadjusted_pairwise
```