STAT 512 Homework 3

Kathleen Wendt

02/18/2020

Part A: Grain

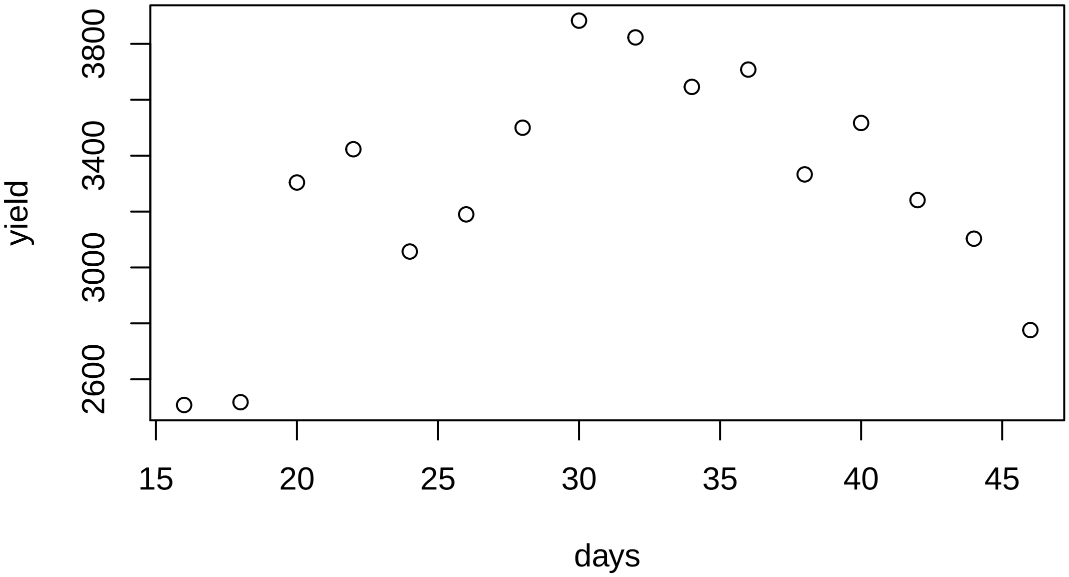
Questions 1 through 6 (Grain Yield): Data relating grain Yield (Y) to the number of Days (X) after flowering that harvesting took place was examined in "Determination of Biological Maturity and Effect of Harvesting and Drying Conditions on Milling Quality of Paddy" (J of Ag Engr. Research (1975):353-361.) The data is available from Canvas as "Grain.csv".

Notes:

- For consistency, please use the `I()` or `poly( , raw = TRUE)` functions for fitting the quadratic and cubic models.
- For questions 2-4, you do NOT need to include the diagnostic plots in your assignment. Just discuss your findings.

Question 1: Scatterplot

Create a scatterplot of Yield vs Days. Include this plot in your assignment.



Question 2: Simple linear regression

Fit a linear regression model of Yield on Days. Include the parameter estimate information ("Coefficients" table) in your assignment. Examine a plot of the residuals versus predicted values. What does the residual plot suggest? (4 pts)

```
##
## Call:
## lm(formula = yield ~ days, data = grain_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -691.07 -217.65   45.85  271.77  612.14
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2902.96     364.67   7.961 1.45e-06 ***
## days           12.26      11.28   1.088    0.295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 415.8 on 14 degrees of freedom
## Multiple R-squared:  0.07791,    Adjusted R-squared:  0.01205
## F-statistic: 1.183 on 1 and 14 DF,  p-value: 0.2951
```

The plot of the residuals vs. predicted values for the simple linear regression of yield (Y) on days (X) revealed curvature, indicating a violation of the assumption of linearity.

Question 3: Quadratic regression model

Fit a quadratic regression model (including both linear and quadratic terms). Include the parameter estimate information ("Coefficients" table) in your assignment. Examine a plot of the residuals versus predicted values and comment. (4 pts)

```
##
## Call:
## lm(formula = yield ~ days + I(days^2), data = grain_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -303.96 -118.11   13.86  115.67  319.06
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1070.3977   617.2527  -1.734    0.107
## days          293.4829    42.1776   6.958 9.94e-06 ***
## I(days^2)      -4.5358     0.6744  -6.726 1.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.9 on 13 degrees of freedom
## Multiple R-squared:  0.7942, Adjusted R-squared:  0.7625
## F-statistic: 25.08 on 2 and 13 DF,  p-value: 3.452e-05
```

The plot of the residuals vs. predicted values for the quadratic regression of yield (Y) on days (X) with both linear (X) and quadratic terms of days (X^2) showed an improvement in scatter, although non-constant variance might be a slight issue.

Question 4: Cubic regression model

Fit a cubic regression model (including linear, quadratic, and cubic terms). Include the parameter estimate information ("Coefficients" table) in your assignment. Again examine a plot of the residuals versus predicted values and comment. (4 pts)

```
##
## Call:
## lm(formula = yield ~ days + I(days^2) + I(days^3), data = grain_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -281.97 -113.21   -6.11   97.75  330.92
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -203.60852 2285.13020  -0.089    0.930
## days         199.07674  242.92513   0.819    0.428
## I(days^2)     -1.32071    8.16843  -0.162    0.874
## I(days^3)     -0.03457    0.08751  -0.395    0.700
##
## Residual standard error: 210.8 on 12 degrees of freedom
## Multiple R-squared:  0.7968, Adjusted R-squared:  0.746
## F-statistic: 15.68 on 3 and 12 DF,  p-value: 0.0001876
```

The plot of the residuals vs. predicted values for the cubic regression of yield (Y) on days (X) with linear, quadratic, and cubic terms for days (X) is similar to the corresponding plot for the quadratic regression model.

Question 5: Hypothesis test for cubic model

In the cubic model (#4), test the hypothesis that the linear, quadratic and cubic regression coefficients are all simultaneously zero. Give the F-statistic and p-value and make a conclusion about the test.

Based on the linear hypothesis test, we reject the null hypothesis. At least one of the partial regression coefficients (linear, quadratic, cubic of `days`) in the cubic model for grain yield is non-zero, $F = 15.684$, $p = 1.9 \times 10^{-4} < \alpha = 0.05$.

Question 6: Model selection

Which model would you choose: linear, quadratic or cubic? Justify your choice. Hint: Think about the simplest model that satisfies assumptions.

I would select the *quadratic* model because it is the simplest model that satisfies assumptions reasonably well and seems to fit the data well ($R^2 = 0.79$).

Part B: Drugs

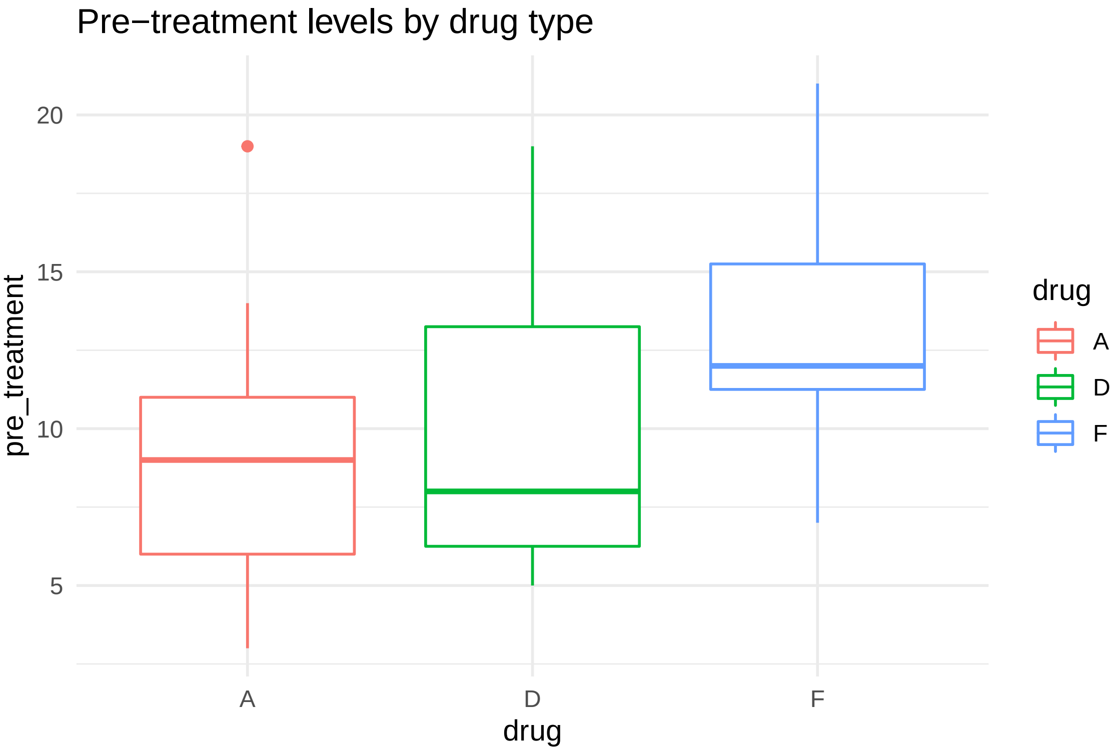
Data was collected to compare 3 drug treatments for leprosy. Three variables are included in the data set:

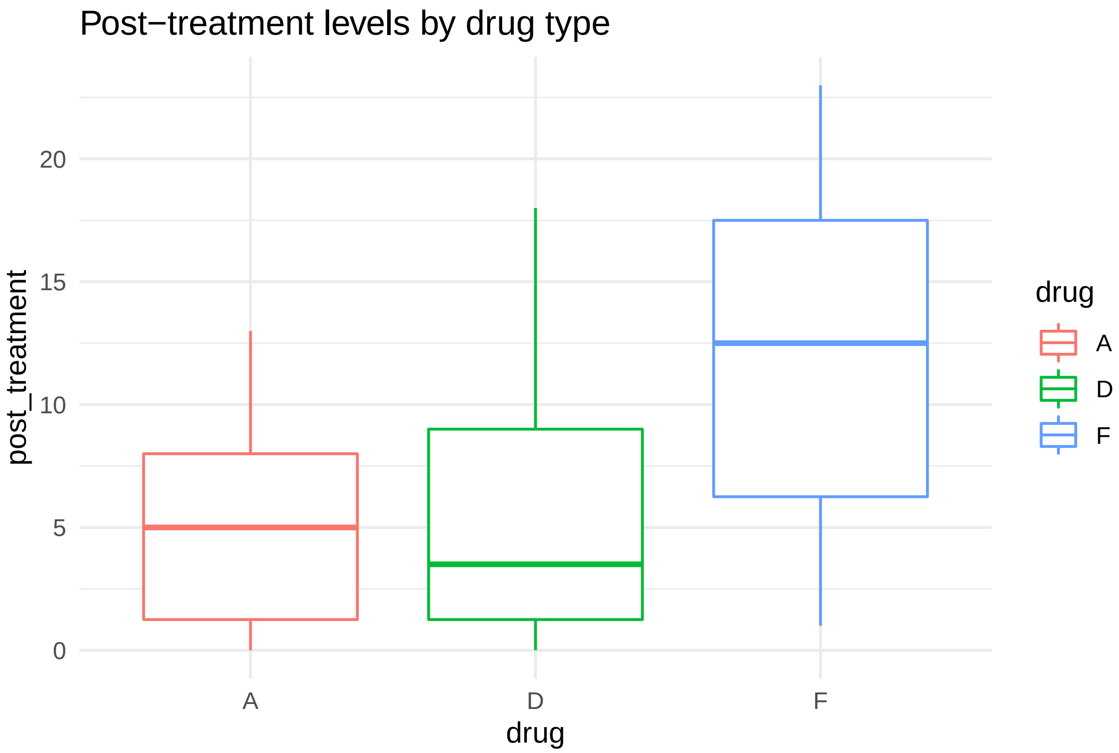
- `Drug`: drug treatment (A, D or F)
- `PreTreatment`: a pretreatment score of leprosy bacilli
- `PostTreatment`: a posttreatment score of leprosy bacilli

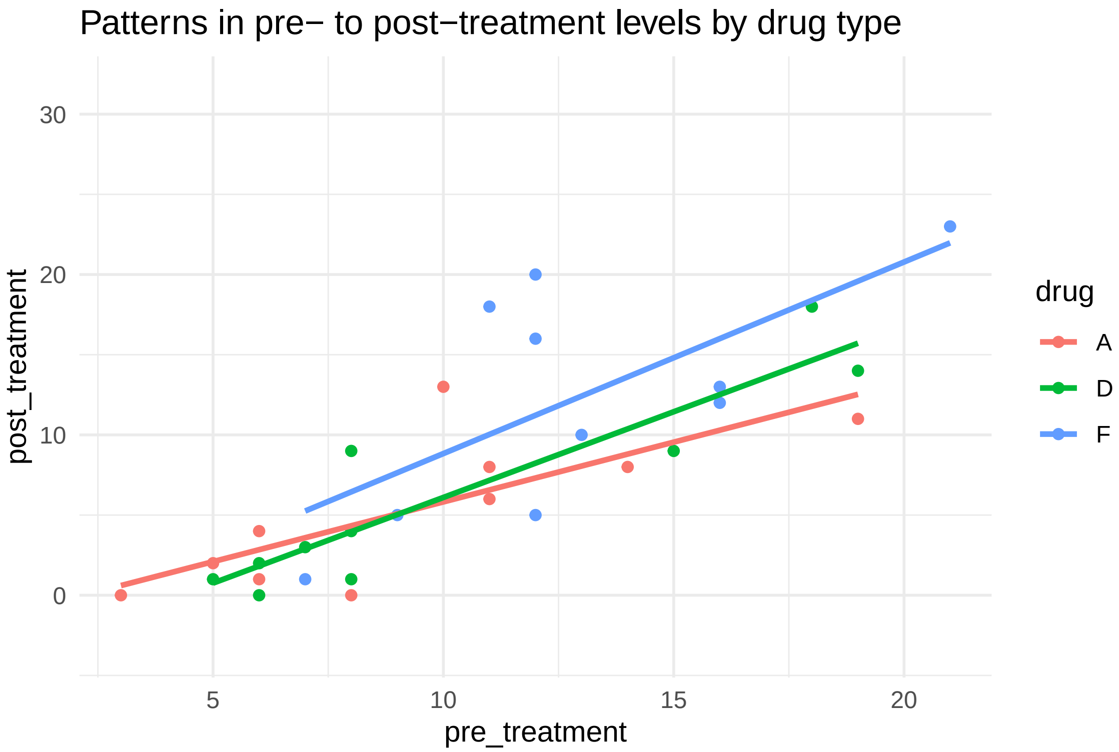
Ten patients were randomly assigned to each Drug treatment (for a total of n = 30 subjects). The goal of the study is to compare mean PostTreatment values for the 3 drugs. Hence, PostTreatment is the response variable and Drug is the predictor variable. But the researchers would like to consider including the PreTreatment value as a covariate. The data is available from Canvas as "DrugTest.csv".

Question 7: Plots

Construct side-by-side boxplots of (1) PreTreatment vs Drug and (2) PostTreatment vs Drug. Also construct (3) a scatterplot of PostTreatment vs PreTreatment for all Drugs on the same plot. Overlay a fitted regression line for each Drug. Include these plots in your assignment. (4 pts)




Post−treatment levels by drug type


Patterns in pre− to post−treatment levels by drug type

Question 8: One-way ANOVA

Fit the one-way ANOVA model (using Drug as the only predictor). Include the ANOVA table and Tukey adjusted pairwise comparisons in your assignment. What can we conclude about differences between the Drugs? (4 pts)

```
## Analysis of Variance Table
##
## Response: post_treatment
##           Df Sum Sq Mean Sq F value  Pr(>F)
## drug       2  293.6 146.800  3.9831 0.03049 *
## Residuals 27  995.1  36.856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

contrast	estimate	SE	df	t.ratio	p.value
A - D	-0.8	2.714979	27	-0.2946616	0.9533479
A - F	-7.0	2.714979	27	-2.5782888	0.0403261
D - F	-6.2	2.714979	27	-2.2836272	0.0754125

Based on an analyis of variance with one predictor (`drug`), there is a difference in post-treatment levels by drug type. Specifically, Drugs A and F differ in post-treatment levels at α of 0.05.

Question 9

Now fit the ANCOVA model with NO Interaction (using Drug and PreTreatment as the predictors). Include the ANOVA table and Tukey adjusted pairwise comparisons in your assignment. What can we conclude about differences between the Drugs? (4 pts)

```
## Analysis of Variance Table
##
## Response: post_treatment
##               Df Sum Sq Mean Sq F value    Pr(>F)
## drug           2  293.6  146.80  9.1486 0.0009812 ***
## pre_treatment  1  577.9  577.90 36.0145 2.454e-06 ***
## Residuals     26  417.2   16.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

contrast	estimate	SE	df	t.ratio	p.value
A - D	-0.1089713	1.795135	26	-0.0607037	0.9979706
A - F	-3.4461383	1.886781	26	-1.8264647	0.1808765
D - F	-3.3371669	1.853866	26	-1.8001119	0.1893423

Based on the analysis of covariance without an interaction term, there are no differences in post-treatment levels by drug type and pre-treatment level, using $\alpha = 0.05$.

Question 10

Comparing your conclusions from #8 (one-way ANOVA) vs #9 (ANCOVA) you should have found different conclusions regarding significant differences between the Drug treatments. Give a brief explanation of why the conclusions change when we include PreTreatment as a covariate. Hint: Consider the boxplots from #7.

The possible difference between Drugs A and F is less obvious when considering their pre-treatment levels. The pre-treatment level “washes out” any difference by drug type in post-treatment level.

Question 11

An alternative approach to the ANCOVA above is to calculate the difference (Diff = PostTreatment − PreTreatment) and use this Diff as the response in a one-way ANOVA model. Do this and include the

ANOVA table and Tukey adjusted pairwise comparisons in your assignment. What can we conclude about differences between the Drugs? (4 pts)

```
## Analysis of Variance Table
##
## Response: diff
##           Df Sum Sq Mean Sq F value Pr(>F)
## drug       2  74.87  37.433   2.422 0.1078
## Residuals 27 417.30  15.456
```

contrast	estimate	SE	df	t.ratio	p.value
A - D	-0.1	1.758156	27	-0.0568778	0.9982181
A - F	-3.4	1.758156	27	-1.9338448	0.1486429
D - F	-3.3	1.758156	27	-1.8769670	0.1647452

Based on the analysis of variance with a difference score between pre- and post-treatment as the response and drug as the predictor, there are still no differences between drugs in measured levels, using $\alpha = 0.05$.

Appendix

```r
# load packages
library(tidyverse)
library(janitor)
library(broom)
library(car)
library(emmeans)
library(kableExtra)
# set global options
knitr::opts_chunk$set(fig.width = 6,
                      fig.height = 4,
                      fig.path = "figs/",
                      echo = FALSE,
                      warning = FALSE,
                      message = FALSE)
# read grain data
grain_data <- readr::read_csv("data/Grain.csv") %>% janitor::clean_names()
# 1. pairwise plot (base) for grain data
plot(grain_data)
# 2. grain simple lin reg
grain_lm <- lm(yield ~ days, data = grain_data)
summary(grain_lm)
# 2. res vs. fitted plot of grain simple lin reg model
plot(grain_lm, which = 1)
# 3. grain quad reg model
grain_quadlm <- lm(yield ~ days + I(days^2), data = grain_data)
summary(grain_quadlm)
# 3. res vs. fitted plot of grain quadratic model
plot(grain_quadlm, which = 1)
# 4. grain cubic reg model
grain_cubiclm <- lm(yield ~ days + I(days^2) + I(days^3), data = grain_data)
summary(grain_cubiclm)
# 4. res vs. fitted plot of grain quadratic model
plot(grain_cubiclm, which = 1)
# 5. test if linear, quadratic, and cubic reg coeffs are 0
# 5. create matrix pattern
grain_matrix_q5 <- matrix(c(0, 1, 0, 0,
                            0, 0, 1, 0,
                            0, 0, 0, 1),
                          nrow = 3,
                          ncol = 4,
                          byrow = TRUE)
# 5. complete hypothesis test for coeffs
grain_betas <- broom::tidy(car::lht(grain_cubiclm,
                                    grain_matrix_q5,
                                    rhs = c(0, 0, 0)))
# read drug data
drug_data <- readr::read_csv("data/DrugTest.csv") %>%
  janitor::clean_names() %>%
  dplyr::mutate(drug = as.factor(drug))
# 7. side-by-side boxplot of pre by drug
drug_data %>%
  dplyr::group_by(drug) %>%
  ggplot2::ggplot(aes(x = drug, y = pre_treatment, color = drug)) +
  geom_boxplot() +
  ggtitle("Pre-treatment levels by drug type") +
  theme_minimal()
# 7. side-by-side boxplot of post by drug
drug_data %>%
  dplyr::group_by(drug) %>%
  ggplot2::ggplot(aes(x = drug, y = post_treatment, color = drug)) +
  geom_boxplot() +
  ggtitle("Post-treatment levels by drug type") +
  theme_minimal()
# 7. scatter plot of pre vs. post by drug
drug_data %>%
  dplyr::group_by(drug) %>%
  ggplot2::ggplot(aes(x = pre_treatment, y = post_treatment, color = drug)) +
  geom_point() +
  geom_smooth(formula = "y ~ x", method = "lm", fill = NA) +
  ggtitle("Patterns in pre- to post-treatment levels by drug type") +
  theme_minimal()
# 8. drug one-way anova
drug_anova <- lm(post_treatment ~ drug, data = drug_data)
anova(drug_anova)
# 8. Tukey-adjusted pairwise comparisons
drug_anova_em <- emmeans::emmeans(drug_anova, pairwise ~ drug)
kableExtra::kable(drug_anova_em$contrasts)
# 9. drug ancova with no interaction
drug_ancova <- lm(post_treatment ~ drug + pre_treatment, data = drug_data)
anova(drug_ancova)
# 9. Tukey-adjusted pairwise comparisons
drug_ancova_em <- emmeans::emmeans(drug_ancova, pairwise ~ drug)
kableExtra::kable(drug_ancova_em$contrasts)
# 11. create pre-post difference score
drug_data <- drug_data %>%
  dplyr::mutate(diff = post_treatment - pre_treatment)
# 11. drug diff anova
drug_diff_anova <- lm(diff ~ drug, data = drug_data)
anova(drug_diff_anova)
# 11. Tukey-adjusted pairwise comparisons
drug_diff_em <- emmeans::emmeans(drug_diff_anova, pairwise ~ drug)
kableExtra::kable(drug_diff_em$contrasts)
```