STAT 512 Homework 4

Kathleen Wendt

02/25/2020

Part A: Panama Canal

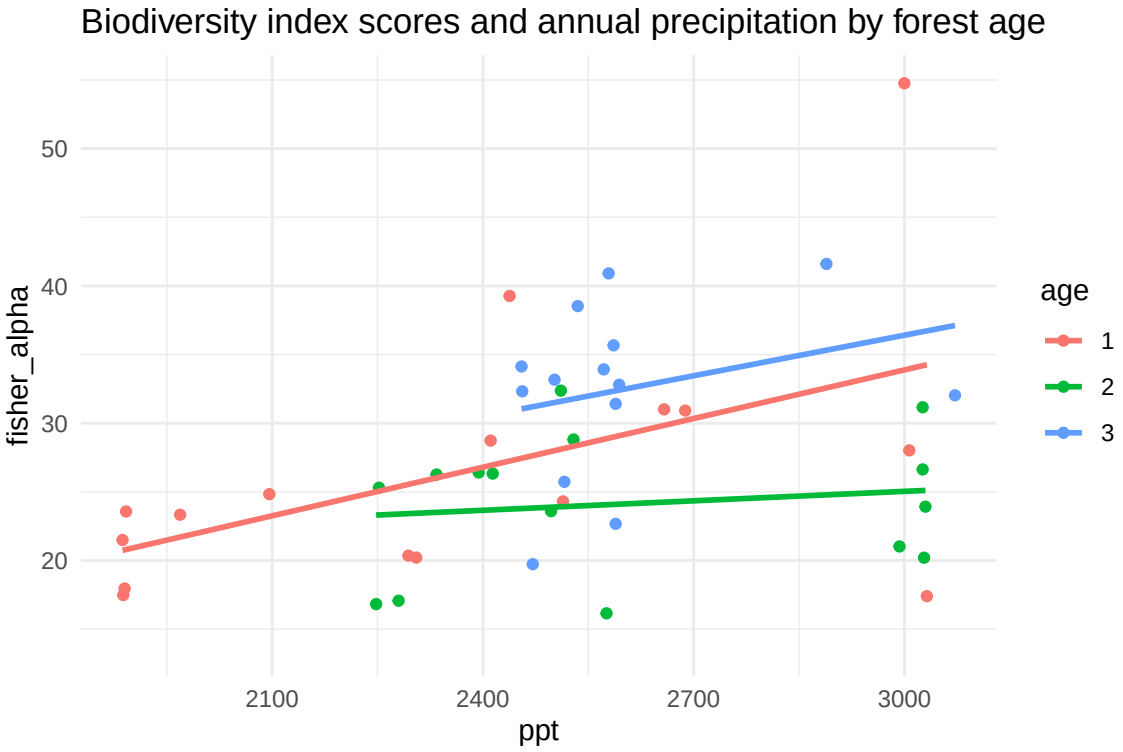
Ott & Longnecker Problem 16.23 describes a study original published in Pyke (2001). Researchers studied the floristic composition of lowland tropical forest in the watershed of the Panama Canal. For this group of question we will work on fitting a model to explain `FisherAlpha` (Y) using `Age` and `Ppt` as predictors. The following characteristics were measured on 45 plots:

- `FisherAlpha`: a biodiversity index
- `Age`: 1 = secondary forest, 2 = mature secondary, 3 = old growth, primary forest
- `Ppt`: annual precipitation (mm)

Note that `Age` should be defined as.factor in R.

Question 1: Plot

Construct a scatterplot of FisherAlpha (Y) vs Ppt (X) for all Age groups on the same plot. Overlay a fitted regression line for each Age group. (2 pts)



Question 2: ANCOVA + interaction

Fit the ANCOVA model WITH interaction. Include the Type 3 ANOVA table in your assignment. What can we conclude about differences between the slopes for the Age groups? Briefly justify your response.

term	sumsq	df	statistic	p.value
(Intercept)	1.085508	1	0.0231344	0.8798922
ppt	365.737866	1	7.7946229	0.0080743
age	54.668811	2	0.5825522	0.5632566
ppt:age	83.360922	2	0.8882960	0.4195143
Residuals	1829.950837	39	NA	NA

Based on the analysis of covariance with an interaction term (precipitation by forest age), we can conclude there is no statistically significant difference between the slopes for the forest age groups, $p = 0.4195143 > \alpha = 0.05$.

Note: Continue using the ANCOVA WITH interaction model for questions 3-6.

Question 3: Diagnostic plots

Consider the diagnostic plots (Resids vs Fitted and QQplot of residuals). You do not need to include these plots in your assignment, but briefly discuss your findings.

The plot of residuals vs. fitted values indicated equal scatter and no concerning patterns, which support assumptions of linearity and constant variance.

The Q-Q plot of residuals showed some evidence for heavy tails, indicating possible outliers, but, considering the small sample size, this distribution of residuals is sufficiently approximately normal.

Question 4: Forest age differences

For each Age group, provide the estimated intercept, slope (corresponding to Ppt) and p-value corresponding to a test of the slope. (6 pts)

```
##
## Call:
## lm(formula = fisher_alpha ~ -1 + age + age:ppt, data = bio_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.859  -4.899   1.168   2.769  20.879
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## age1     -1.548882  10.183315  -0.152  0.87989
## age2     18.139678  15.189510   1.194  0.23960
## age3      6.866477  28.680668   0.239  0.81204
## age1:ppt  0.011810   0.004230   2.792  0.00807 **
## age2:ppt  0.002298   0.005782   0.397  0.69319
## age3:ppt  0.009847   0.011007   0.895  0.37647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.85 on 39 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9429
```

```
## F-statistic: 124.8 on 6 and 39 DF,  p-value: < 2.2e-16
```

As indicated above, the slope for precipitation in secondary forest (`age` 1) differs from 0. No other slope differs from 0.

Question 5: Pairwise comparisons

Calculate Tukey adjusted pairwise comparisons of the slopes. What can we conclude about differences between the slopes for the Age groups? Briefly justify your response.

level1	level2	estimate	std.error	df	statistic	p.value
1	2	0.0095118	0.0071641	39	1.3277040	0.3886548
1	3	0.0019627	0.0117919	39	0.1664450	0.9848489
2	3	-0.0075491	0.0124332	39	-0.6071723	0.8170641

Based on the Tukey-adjusted pairwise comparisons, there are no differences in the slopes by forest age and precipitation level.

Question 6: Estimated marginal means

Calculate emmeans for the Age groups at (A) Ppt = 2500 and (B) Ppt = 3000.

Question 6A: 2500 `ppt`

age	estimate	std.error	df	conf.low	conf.high
1	27.97600	1.794697	39	24.34588	31.60612
2	23.88504	1.877969	39	20.08649	27.68359
3	31.48461	2.138170	39	27.15976	35.80947

Question 6B: 3000 `ppt`

age	estimate	std.error	df	conf.low	conf.high
1	33.88098	3.156859	39	27.49563	40.26633
2	25.03412	2.869460	39	19.23009	30.83815
3	36.40824	4.764627	39	26.77087	46.04561

Questions 7 and 8 (FisherAlpha continued): Use the ANCOVA WITH interaction model above as the "full" model. But our goal is to choose a model that predicts FisherAlpha.

Question 7: Backward elimination

Based on a backwards elimination approach, which model is preferred? Briefly justify your response. Use $\alpha = 0.05$.

1. Starting with the full model (ANCOVA with interaction term), the interaction between precipitation and forest age is not significant, $p = 0.4195143$, indicating that the interaction term should be removed.

term	sumsq	df	statistic	p.value
(Intercept)	1.085508	1	0.0231344	0.8798922
ppt	365.737866	1	7.7946229	0.0080743
age	54.668811	2	0.5825522	0.5632566
ppt:age	83.360922	2	0.8882960	0.4195143
Residuals	1829.950837	39	NA	NA

2. The interaction term was removed from the full model and re-tested. Because both main effects (precipitation and forest age) are significant at $\alpha = 0.05$, the ANCOVA without an interaction term is the preferred model based on backward elimination.

term	sumsq	df	statistic	p.value
(Intercept)	27.24322	1	0.5837898	0.4492057
ppt	327.34463	1	7.0146069	0.0114261
age	513.21406	2	5.4987841	0.0076633
Residuals	1913.31176	41	NA	NA

Question 8: AIC

Based on AIC, which model is preferred? Briefly justify your response. Hint: Use `dredge()` from `MuMIn`.

```
## Global model call: lm(formula = fisher_alpha ~ ppt * age, data = bio_data)
## ---
## Model selection table
##    (Int) age      ppt age:ppt df   logLik   AIC delta weight
## 4  6.039   + 0.008613          5 -148.226 306.5  0.00  0.673
## 8 -1.549   + 0.011810       +  7 -147.223 308.4  2.00  0.248
## 2 26.480   +                   4 -151.779 311.6  5.11  0.052
## 3  6.491     0.008353          3 -153.572 313.1  6.69  0.024
## 1 27.560                       2 -156.577 317.2 10.70  0.003
## Models ranked by AIC(x)
```

Based on AIC, Model 4 with forest age and precipitation as predictors (no interaction) and 5 degrees of freedom is preferred. This model has the lowest AIC value.

Part B: Body Fat

Return to the Body Fat data from HW2. The data is available from Canvas as "BodyFat.csv". With 3 predictors, there are 8 possible models. Which model would you choose? To identify the model, just state which predictors are included.

Question 9: Backward elimination

Choose a model using "backwards elimination" (hypothesis testing) approach. Use $\alpha = 0.05$. No need to discuss, just state your final model.

Based on backward elimination, the preferred model is an ANCOVA (no interaction) with tricep and midarm measurements as predictors.

Question 10: Forward selection

Choose a model using "forward selection" (hypothesis testing) approach. Use $\alpha = 0.05$. No need to discuss, just state your final model.

Based on forward selection, the preferred model is an ANOVA with thigh measurement as the primary predictor of body fat.

Question 11: AICc

Choose a model using AICc. Hint: Use `dredge()` from `MuMIn`. No need to discuss, just state your final model.

Just like Q10, an ANOVA with `thigh` as the sole predictor has the lowest AICc value.

Appendix

```
# load packages
library(tidyverse)
library(janitor)
library(car)
library(emmeans)
library(broom)
library(kableExtra)
library(MuMIn)
# set global options
knitr::opts_chunk$set(fig.width = 6,
                      fig.height = 4,
                      fig.path = "figs/",
                      echo = FALSE,
                      warning = FALSE,
                      message = FALSE)
# read panama data
bio_data <- readr::read_csv("data/ex16-23.txt") %>%
  janitor::clean_names() %>%
  dplyr::select(fisher_alpha, age, ppt) %>%
  dplyr::mutate(age = as.factor(age))
# 1. plot fisher_alpha and ppt by age group
bio_data %>%
  dplyr::group_by(age) %>%
  ggplot2::ggplot(aes(x = ppt, y = fisher_alpha, color = age)) +
  geom_point() +
  geom_smooth(formula = "y ~ x", method = "lm", fill = NA) +
  ggtitle("Biodiversity index scores and annual precipitation by forest age") +
  theme_minimal()
# 2. fit ancova with interaction term
bio_ancova_int <- lm(fisher_alpha ~ ppt*age, data = bio_data)
# 2. tidy model
bio_ancova_tidy <- broom::tidy(car::Anova(bio_ancova_int, type = 3))
kableExtra::kable(bio_ancova_tidy)
# 3. check diagnostic plots
plot(bio_ancova_int)
# 4. build alternate parameterization; remove common intercept and main effects
bio_ancova_alt <- lm(fisher_alpha ~ - 1 + age + age:ppt, data = bio_data)
summary(bio_ancova_alt)
# 5. tukey pairwise comparisons of slopes for age
bio_slope_em <- emmeans::emtrends(model = bio_ancova_int,
                                  specs = "age",
                                  var = "ppt")
kableExtra::kable(broom::tidy(pairs(bio_slope_em)))
# 6A. emmeans for age at specific ppt levels
bio_ppt_em1 <- emmeans::emmeans(bio_ancova_int,
                                pairwise ~ age,
                                at = list(ppt = 2500))
kableExtra::kable(broom::tidy(bio_ppt_em1$emmeans))
# 6B. emmeans for age at specific ppt levels
bio_ppt_em2 <- emmeans::emmeans(bio_ancova_int,
                                pairwise ~ age,
```

```
                                at = list(ppt = 3000))
kableExtra::kable(broom::tidy(bio_ppt_em2$emmeans))
# 7. call full model - ancova with interaction
kableExtra::kable(bio_ancova_tidy)
# 7. int term NS - build ancova without interaction
bio_ancova <- lm(fisher_alpha ~ ppt + age, data = bio_data)
kableExtra::kable(broom::tidy(car::Anova(bio_ancova, type = 3)))
# 8. fisher_alpha model selection based on AIC
options(na.action = "na.fail")
MuMIn::dredge(bio_ancova_int, rank = "AIC")
fat_data <- readr::read_csv("data/BodyFat.csv") %>% janitor::clean_names()
# 9. build full fat ancova model with interactions
fat_full <- lm(body_fat ~ triceps + thigh + midarm + triceps*thigh*midarm,
               data = fat_data)
# 9. build fat ancova without interaction terms
fat_ancova <- lm(body_fat ~ triceps + thigh + midarm, data = fat_data)
# 9. drop thigh term (highest p-value) from ancova and review
fat_ancova <- update(fat_ancova, ~ . -thigh)
drop1(fat_ancova, test = "F")
summary(fat_ancova)
# 10. create and test null model
fat_null <- lm(body_fat ~ 1, data = fat_data)
add1(fat_null, scope = fat_full, test = "F")
# 10. add thigh as predictor and test
fat_forward_1 <- update(fat_null, ~ . + thigh)
add1(fat_forward_1, scope = fat_full, test = "F")
# 11. compare fat models using AICc
options(na.action = "na.fail")
MuMIn::dredge(fat_full, rank = "AICc")
```