STAR 513: HW 6

Yvette Uwineza

Total points: 48
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.

Q1 - Q5 (Brain Study)

We continue with the BrainData.csv from the Siddarth (2018) article. For this group of questions, we will use the multiple regression model using MTL as the response and including 3 predictors: Age, Sitting, Activity. Start by fitting this model.

```
##
## Call:
## lm(formula = MTL ~ Age + Sitting + Activity, data = BrainData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30500 -0.13119 -0.02225  0.13292  0.45199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.417e+00  2.706e-01   8.933 4.42e-10 ***
## Age          4.391e-03  3.945e-03   1.113   0.2743
## Sitting     -2.090e-02  9.549e-03  -2.189   0.0362 *
## Activity     1.992e-06  2.566e-05   0.078   0.9386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1812 on 31 degrees of freedom
## Multiple R-squared:  0.1892, Adjusted R-squared:  0.1108
## F-statistic: 2.412 on 3 and 31 DF,  p-value: 0.0856
```

Q1

Use the `confint()` to construct 95% confidence intervals for all 4 of the regression coefficients (βs).

```
##                     2.5 %        97.5 %
## (Intercept)  1.865245e+00  2.969005e+00
## Age         -3.655503e-03  1.243808e-02
## Sitting     -4.037823e-02 -1.429366e-03
## Activity    -5.033088e-05  5.431456e-05
```

Q2

Choose ONE of the predictors (not the intercept) for this question. Use the corresponding confidence interval to briefly discuss whether we have evidence of linear association between mean MTL and that predictor (based on this model, controlling for the other predictors). Use alpha = 0.05 for this question. Remember: We do not accept H0!

Predictor: Sitting

The 95% confidence interval is (-0.0404, -0.00143). This means that we have evidence of a negative linear association between (average) MTL and Sitting. This is mainly because zero is not included in the confidence interval. A one hour increase Sitting is associated with an estimated/predicted decrease of 0.02 mm in average MTL thickness, controlling for Age and Activity

Q3

We will make predictions for two scenarios. (A) A 60 year old individual who sits for 7 hours per day and has activity level of 1500 MET minutes per week. (B) A 20 year old individual who sits for 2 hours per day and has activity level of 8000 MET minutes per week.

Q3A (4 pts)

Calculate predicted values ($\hat{y}$) and corresponding 95% **prediction** intervals for each of these scenarios using `predict()`. Hint: Start by creating a small data.frame containing the values for the two scenarios. Your column names need to exactly match the original data.

```
##        fit      lwr      upr
## 1 2.537263 2.162491 2.912036
## 2 2.479078 1.889915 3.068240
```

Q3B

If we had constructed 95% **confidence** intervals for mean responses would the confidence intervals be wider or more narrow that your prediction intervals from the previous question?

The confidence intervals for the mean responses would be narrower than the prediction intervals from the previous questions.

Q3C

The width of the prediction interval is driven by the SE. Use code like the following to get the SE for each of the predicted values. Replace "ScenarioData" with whatever you called your data.frame in Q3A.

```
## $fit
##        1        2
## 2.537263 2.479078
##
## $se.fit
##          1          2
## 0.03074603 0.22500533
##
## $df
## [1] 31
##
## $residual.scale
## [1] 0.1811652
```

Q3D (4 pts)

Briefly discuss why scenario B has a noticeably larger SE as compared to scenario A. To motivate your discussion, provide the summary statistics (mean, sd, min, median, max) for each variable. Suggested code is provided below; modify as needed.

Response

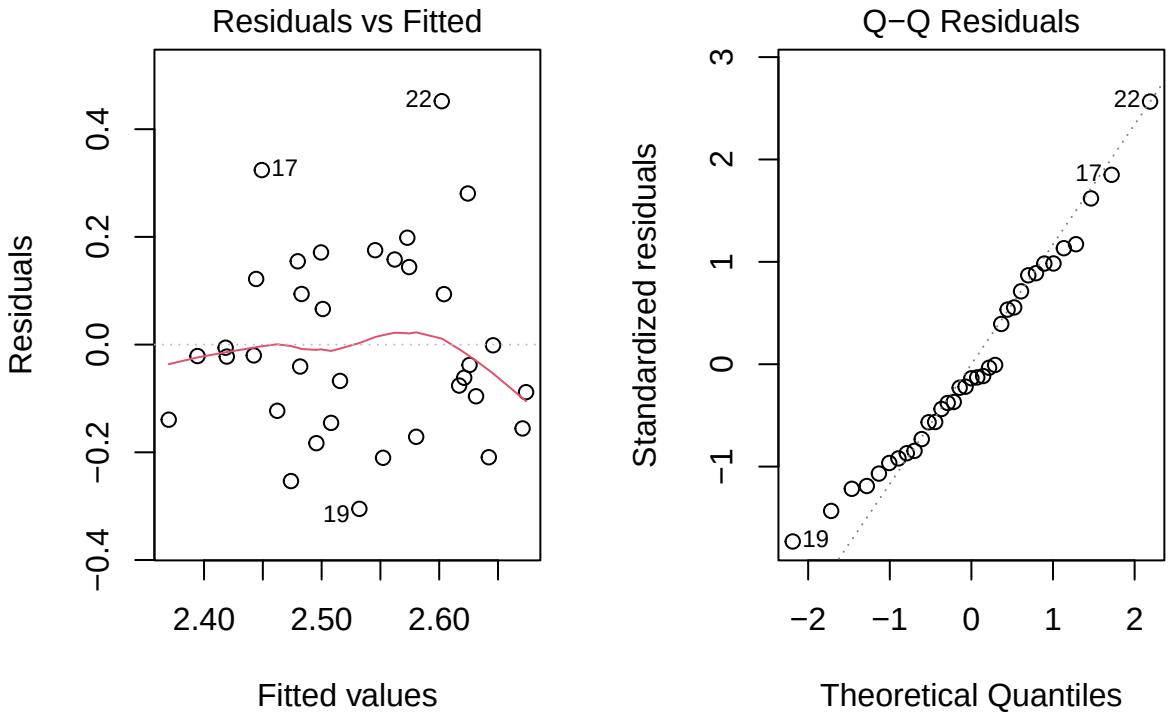
variable	n	min	median	max	mean	sd
Activity	35	99.00	1039.80	5112.00	1521.26	1225.68
Age	35	46.00	63.00	75.00	60.37	8.08
MTL	35	2.22	2.54	3.05	2.53	0.19
Sitting	35	2.00	7.00	15.00	7.20	3.32

Looking a the summary statistics table, there are a couple things that we observe that could tell us the reason scenario b has a larger standard error.

The standard error is larger for scenario B because the values of scenario b are far from the dataset's mean predictor values. And looking at the range values of our variables, they are outside of the observed range of the dataset.This is true for both age and activity. For sitting, the entry value is at the minimum value of the observed data range where errors are less stable.

Q4

Show residual diagnostic plots (A) Residuals vs Fitted values (B) QQplot of residuals Please show *just* these two plots.



Q5 (8 pts)

State the four assumptions of linear regression (in words) and briefly discuss whether each assumption is satisfied. Where applicable, your discussion should mention evidence from a specific plot (graph) from the previous question. A clear discussion of specific evidence is more important than a firm conclusion!

Assumption 1: Linearity, looking at the residuals vs fitted plot, we can see some curvature in the red line, meaning that the linearity assumption is slightly but not entirely violated

Assumption 2: Independence, given that the data comes from individual subjects, the assumption is sort of satisfied

Assumption 3: Homoscedasticity, looking at the residuals vs fitted plot, we can see that our residuals are spread across fitted values and there is specific pattern that they are following. So, we can say that the assumption is satisfied

Assumption 4: Normality of Residuals, looking at the qqplot, we can see that the points follow the line, but we have outliers such as 220, but they don't seem to be a major concern given that they are below 3 standardized residuals. so the normality, is somehow satisfied as well.

Q6 - Q10 (Biodiversity)

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. Characteristics were measured on n = 45 plots.

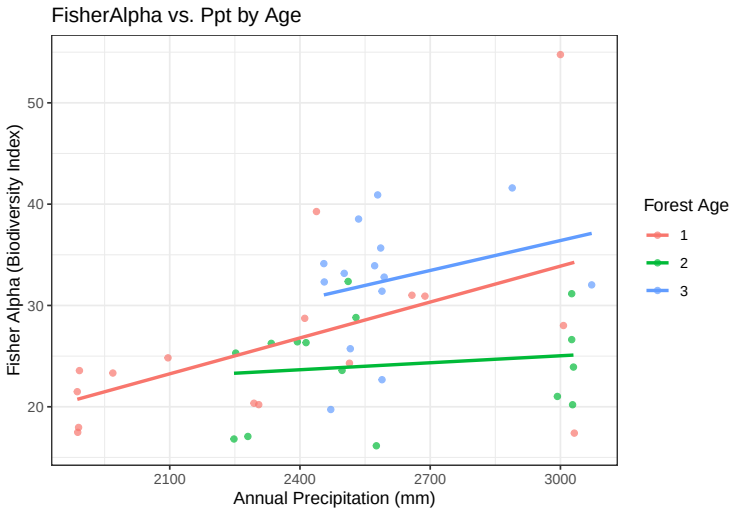
For this assignment, we use the following variables:
FisherAlpha: a biodiversity index
Age: 1 = secondary forest, 2 = mature secondary, 3 = old growth, primary forest
Ppt: annual precipitation (mm)

We will use FisherAlpha as the response (Y) and Age (factor) and Ppt as predictors. This data is available from Canvas as `ex16-23.csv`.

Important Note: Specify Age as.factor in R before model fitting!
Using base R, use code like `AlphaData$Age <- as.factor(AlphaData$Age)`
Using tidyverse, use the as_factor() function.

Q6 (4 pts)

Create a scatterplot of FisherAlpha (Y) vs Ppt (X) color coded by Age. Overlay separate regression lines for each Age.



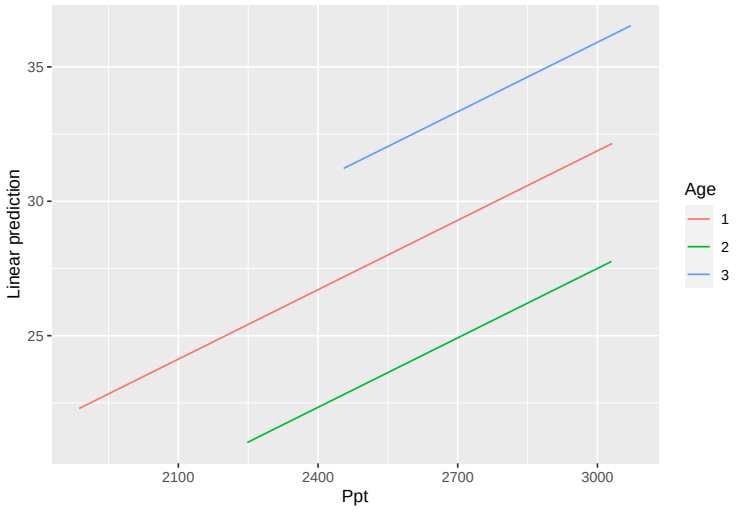
Q7

Fit an appropriate model Age (factor) and Ppt predictors. This should be an additive model, no interaction. Show the coefficients table. For consistency, please use Age as the first predictor and Ppt as the second predictor. Such that Age corresponds to β_1, β_2 and Ppt corresponds to β_3.

```
##
## Call:
## lm(formula = FisherAlpha ~ Age + Ppt, data = ex16_23)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7520  -4.3565   0.4203   3.4343  22.8836
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.038791   7.903542   0.764   0.4492
## Age2        -4.374617   2.572421  -1.701   0.0966 .
## Age3         4.038399   2.606959   1.549   0.1290
## Ppt          0.008613   0.003252   2.649   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.831 on 41 degrees of freedom
## Multiple R-squared:  0.3101, Adjusted R-squared:  0.2596
## F-statistic: 6.142 on 3 and 41 DF,  p-value: 0.001508
```

Q8

Use emmip (from the emmeans package) to create a visual summary of the model. Suggested code is provided; modify as needed.



Q9 (6 pts)

For each Age category, provide the estimated intercept and slope (corresponding to Ppt). Using the model from Q7, provide numeric values for each estimate.

Age 1:
Estimated Intercept = 6.0388
Estimated Slope = 0.0086

Age 2:
Estimated Intercept = 1.6642
Estimated Slope = 0.0086

Age 3:
Estimated Intercept = 10.0772
Estimated Slope = 0.0086

Q10

Now we will test for a difference between the mean FisherAlpha comparing between levels of Age, controlling for Ppt. In other words, we are testing for an Age effect using this model.

Q10A

Write the null hypothesis using standard greek letter notation. (For consistency, please use Age as the first predictor and Ppt as the second predictor. Such that Age corresponds to β_1, β_2 and Ppt corresponds to β_3.)

H0: $\beta_1 = \beta_2 = 0$

Q10B

Use `anova()` to perform a partial F-test comparing an appropriate reduced vs full model. Show your result.

```
## Analysis of Variance Table
##
## Model 1: FisherAlpha ~ Ppt
## Model 2: FisherAlpha ~ Age + Ppt
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     43 2426.5
## 2     41 1913.3  2    513.21 5.4988 0.007663 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q10C

Now check your previous answer using code `Anova(FullModel ,type = 3)`. Recall that `Anova()` from the car package. The F-test and p-value corresponding to "Age" should match your previous result.

```
## Anova Table (Type III tests)
##
## Response: FisherAlpha
##              Sum Sq Df F value   Pr(>F)
## (Intercept)   27.24  1  0.5838 0.449206
## Age          513.21  2  5.4988 0.007663 **
## Ppt          327.34  1  7.0146 0.011426 *
## Residuals   1913.31 41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q10D

Based on your result from the previous questions, do we have evidence that (population) mean FisherAlpha differs between Age categories, when controlling for Ppt? You can just answer yes or no! Use alpha = 0.05 for this question.

Yes! we have evidence that population mean fisher alpha differs between age categories when controlling for ppt at alpha equal to 0.05. The p value is 0.007663<0.05

Appendix

```
#Retain this code chunk!!!
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
library(knitr)
library(knitr)
library(tinytex)
library(broom)
library(emmeans)
library(tidyverse)
library(kableExtra)
library(ggplot2)
library(car)

# loading all the datasets

ex16_23 <- read_csv("/data/yuwineza/star_513/Homework-6/ex16-23.csv")
BrainData <- read_csv("/data/yuwineza/star_513/Homework-6/BrainData.csv")

glimpse(ex16_23)
glimpse(BrainData)


#Q1-Q5 Brain Model

Model_1 <- lm(MTL ~ Age + Sitting + Activity, data = BrainData)
summary(Model_1)

#Q1

confint(Model_1)

#Q3A

Scenarios <- data.frame(Age = c(60, 20),
                        Sitting = c(7, 2),
                        Activity = c(1500, 8000))


predictions <- predict(Model_1, newdata = Scenarios, interval = "prediction")
predictions


#Q3C

predict(Model_1, newdata = Scenarios, se.fit = TRUE)

#Q3D

#BrainData <- read.csv("BrainData.csv")
SumStats <- BrainData %>%
    select(Subject, MTL, Age, Sitting, Activity) %>%
```

```
    pivot_longer(!Subject, names_to = "variable", values_to = "value" ) %>%
    group_by(variable) %>%
    summarize(n = n(),
              min = min(value),
              median = median(value),
              max = max(value),
              mean = mean(value),
              sd = sd(value))
SumStats %>%
  kable(digits = 2)

#Q4


par(mfrow = c(1,2))
plot(Model_1, which = c(1,2))



#Q6

ex16_23$Age <- as.factor(ex16_23$Age)


# Create the scatterplot
ggplot(ex16_23, aes(x = Ppt, y = FisherAlpha, color = Age)) +
  geom_point(size = 1.5, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "FisherAlpha vs. Ppt by Age",
       x = "Annual Precipitation (mm)",
       y = "Fisher Alpha (Biodiversity Index)",
       color = "Forest Age") +
  theme_bw()


#Q7

Model_2 <- lm(FisherAlpha ~ Age + Ppt, data = ex16_23)

summary(Model_2)


#Q8
library(emmeans)
emmip(Model_2, Age ~ Ppt, cov.reduce = FALSE)

#Q10B

ReducedModel <- lm(FisherAlpha ~ Ppt, data = ex16_23)

FullModel <- lm(FisherAlpha ~ Age + Ppt, data = ex16_23)
```

```r
anova(ReducedModel, FullModel)


#Q10C
#install.packages("car")


Anova(FullModel ,type = 3)
```