STAR 513: HW 7

Yvette Uwineza

Total points: 30
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 - Q4 (Bike Share)

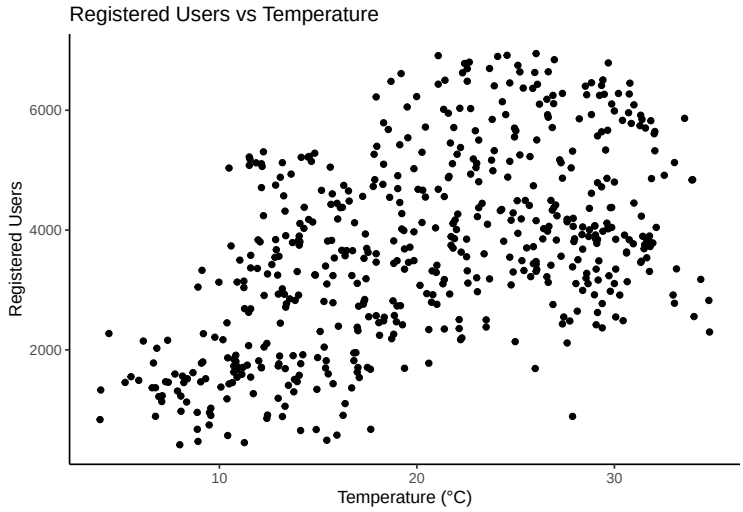
For this group of questions, we will model the number of users for a bike sharing program. The data contains daily observations for n = 551 days.

For this assignment, we use the following variables:
registered_users (y)
temp (x): average ambient temperature in degrees Celsius

This data is available from Canvas as `bike_sharing.csv`.

Q1

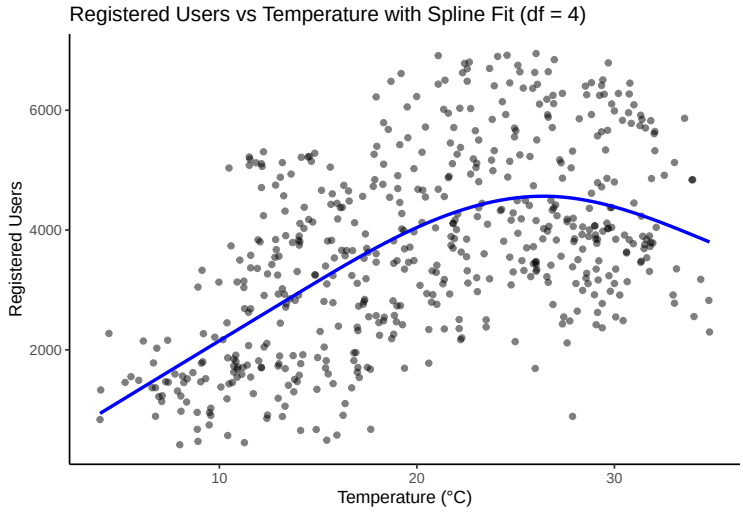
Create a scatterplot of registered users (y) vs temperature (x).



Q2 (4 pts)

Use the ns() function from the splines package to model the relationship between registered users vs temperature. Show the scatterplot with fitted curve overlaid. Experiment with different values for the df = function to see how the spline fit changes with this value. Choose a value that you feel is appropriate for the data and show only the plot of this fit. (You will get full credit for any reasonable value.)

```
##
## Call:
## lm(formula = registered_users ~ ns(temp, df = 4), data = bike_sharing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3643.0  -987.5  -174.5  1010.6  2786.0
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            938.8      368.9   2.545   0.0112 *
## ns(temp, df = 4)1     3301.7      358.7   9.205  < 2e-16 ***
## ns(temp, df = 4)2     3318.0      336.4   9.863  < 2e-16 ***
## ns(temp, df = 4)3     4763.6      847.4   5.621 3.03e-08 ***
## ns(temp, df = 4)4     1872.2      366.3   5.111 4.44e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1283 on 546 degrees of freedom
## Multiple R-squared:  0.3378, Adjusted R-squared:  0.3329
## F-statistic: 69.62 on 4 and 546 DF,  p-value: < 2.2e-16
```

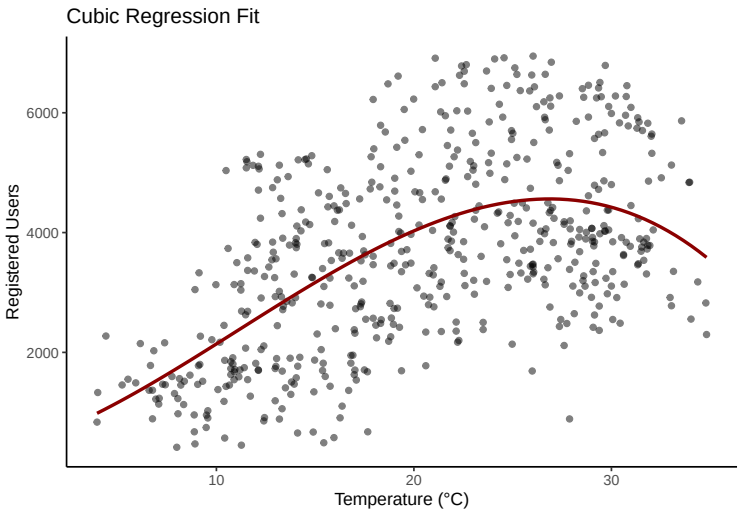


Q3 (4 pts)

Fit a cubic regression model. Show the coefficients table and a scatterplot with the fitted curve overlaid. For consistency, please use `poly(,3)` to fit this model.

```
##
```

```
## Call:
## lm(formula = registered_users ~ poly(temp, 3), data = bike_sharing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3658.7  -997.6  -152.6  1020.9  2792.7
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3667.52      54.61  67.159  < 2e-16 ***
## poly(temp, 3)1 19302.09    1281.87  15.058  < 2e-16 ***
## poly(temp, 3)2 -8851.88    1281.87  -6.905 1.39e-11 ***
## poly(temp, 3)3 -2731.85    1281.87  -2.131   0.0335 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1282 on 547 degrees of freedom
## Multiple R-squared:  0.3377, Adjusted R-squared:  0.3341
## F-statistic: 92.99 on 3 and 547 DF,  p-value: < 2.2e-16
```



Q4

Consider the cubic regression from the previous question. Do you see any reason to prefer a quadratic model? Do you see any reason to prefer (or consider) a more complex model (spline or quartic)? Briefly discuss.

Looking at the quadratic model above, I could see that the association was not linear, so there was an need for a quadratic model to cover the curvature. On the other hand though, I don't see a reason for the spline if the quadratic model can cover the same relationship. But in different cases, we might need a non parametric model if it is hard to capture the relationship.

Q5 - Q9 (Biodiversity)

We continue with the Biodiversity data from Ott & Longnecker Problem 16.23 with the 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.

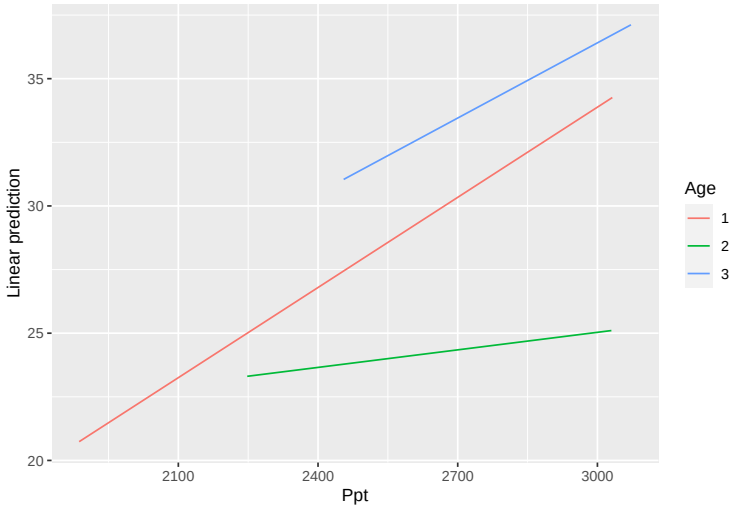
Q5

Fit an appropriate model Age (factor) and Ppt predictors with interaction. 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, etc.

```
##
## Call:
## lm(formula = FisherAlpha ~ Age * Ppt, data = ex16_23_1_)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.859  -4.899   1.168   2.769  20.879
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.548882  10.183315  -0.152  0.87989
## Age2        19.688559  18.287184   1.077  0.28826
## Age3         8.415358  30.434859   0.277  0.78362
## Ppt          0.011810   0.004230   2.792  0.00807 **
## Age2:Ppt    -0.009512   0.007164  -1.328  0.19200
## Age3:Ppt    -0.001963   0.011792  -0.166  0.86867
## ---
## 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:  0.3401, Adjusted R-squared:  0.2555
## F-statistic:  4.02 on 5 and 39 DF,  p-value: 0.004883
```

Q6

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



Q7 (6 pts)

For each Age category, provide the estimated intercept and slope (corresponding to Ppt). Using the interaction model from Q5, provide numeric values for each estimate.

Age 1:
Estimated Intercept = -1.549
Estimated Slope = 0.012

Age 2:
Estimated Intercept = -1.548882 + 19.688559 = 18.139677
Estimated Slope = 0.011810 + (-0.009512) = 0.002298

Age 3:
Estimated Intercept = -1.548882 + 8.415358 = 6.866476
Estimated Slope = 0.011810 + (-0.001963) = 0.009847

Q8

Use emmeans to estimate and compare mean diversity at different values of Ppt. Use the default Tukey adjustment. For these questions, just show the output. No need to discuss.

Q8A

Provide emmeans and contrasts at Ppt = 2000.

Q8B

Provide emmeans and contrasts at Ppt = 3000.

Q9

Run a test for Age:Ppt **interaction**.

Q9A

Provide an appropriate test statistic and p-value.

Response

Q9B

Is the interaction (Age*Ppt) or additive (Age + Ppt) model preferred? Briefly justify your response using the test result from the previous question.

Response

Appendix

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


# uploading the datasets

# the bikesharing dataset
bike_sharing <- read_csv("/data/yuwineza/star_513/Homework - 7/bike_sharing.csv")

#glimpse(bike_sharing)

# the ex something too

ex16_23_1_ <- read_csv("/data/yuwineza/star_513/Homework - 7/ex16-23 (1).csv")

#glimpse(ex16_23_1_)

#Q1

ggplot(bike_sharing, aes(x = temp, y = registered_users)) +
  geom_point() +
  labs(
    title = "Registered Users vs Temperature",
    x = "Temperature (°C)",
    y = "Registered Users"
  ) +
  theme_classic()


#Q2

model_1<- lm(registered_users ~ ns(temp, df = 4), data = bike_sharing)
summary(model_1)


g<- ggplot(bike_sharing, aes(x = temp, y = registered_users)) +
  geom_point(alpha = 0.5)  +
  labs(
    title = "Registered Users vs Temperature with Spline Fit (df = 4)",
    x = "Temperature (°C)",
    y = "Registered Users"
  ) +
  theme_classic()

g + geom_smooth(method = "lm",
              formula = y ~ ns(x, df = 4),
              se = FALSE,
```

```
              color = "blue",
              linewidth = 1)


#Q3

model_2 <- lm(registered_users ~ poly(temp, 3), data = bike_sharing)

# Show coefficients table
summary(model_2)

g<-ggplot(bike_sharing, aes(x = temp, y = registered_users)) +
  geom_point(alpha = 0.5) +
  labs(
    title = "Cubic Regression Fit",
    x = "Temperature (°C)",
    y = "Registered Users"
  ) +
  theme_classic()

 g + geom_smooth(method = "lm",
              formula = y ~ poly(x, 3),
              se = FALSE,
              color = "darkred",
              linewidth = 1)

#Q5

ex16_23_1_ <- ex16_23_1_ %>%
  mutate(Age = as_factor(Age))


Model_3 <- lm(FisherAlpha ~ Age * Ppt, data = ex16_23_1_)
summary(Model_3)



#Q6
emmip(Model_3, Age ~ Ppt, cov.reduce = FALSE)

#Q7

#Q8


#Q9
```