STAR 513: HW 9

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.

We continue with the Bike Share data from the previous assignment. The data contains daily observations for n = 551 days. This data is available from Canvas as `bike_sharing.csv`.

Q1 - Q6 (Polynomial and Splines)

For this group of questions use:

- registered_users (y)
- temp (x): average ambient temperature in degrees Celsius

Q1

Use the poly() function to model a quadratic, cubic, and quartic (**degree = 2, 3, 4**, respectively) relationship between temperature and the number of registered users.

Please NOT show the output, just save the results for later use.

Q2 (4 pts)

Using your polynomial models from Q1, choose a model based on (manual) "backwards elimination" based on p-values. We will follow the "Principle of Hierarchy".

What is the full model? From that full model, give the relevant p-value to decide whether to stick with the full model or simplify.

Depending on your answer to the question above, consider further simplifications to the model.

Report the degree of "final" selected model.

Full Model: Degree = 4
From this full model, we consider the test of the 4th-order term.
Since p = 0.7895, we drop this term and fit a cubic model (degree=3)

Additional steps as needed. we fit the cubic model and the higest order ter has a p-value = 0.0335, which is statistically significant, so we keep this model.

Selected Model: Degree = 3

```
##
## 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
```

Q3

Using your polynomial models from Q1, choose a model based on AIC. We will follow the "Principle of Hierarchy".
Give the AIC values for each of the models.
Report the selected model.

Selected Model: Degree = 3 (it has the lowest AIC)

```
## [1] 9458.205
```

```
## [1] 9455.649
```

```
## [1] 9457.577
```

Q4

Use the ns() function from the splines package to model the relationship between temperature and the number of registered users for **df = 3, 4, 5**.

Please do NOT show the output, just save the results for later use.

Q5

Using your spline models from Q4, choose a model based on AIC.
Give the AIC values for each of the models.
Report the selected model.

Selected Model: df = 5 (it has the lowest AIC)

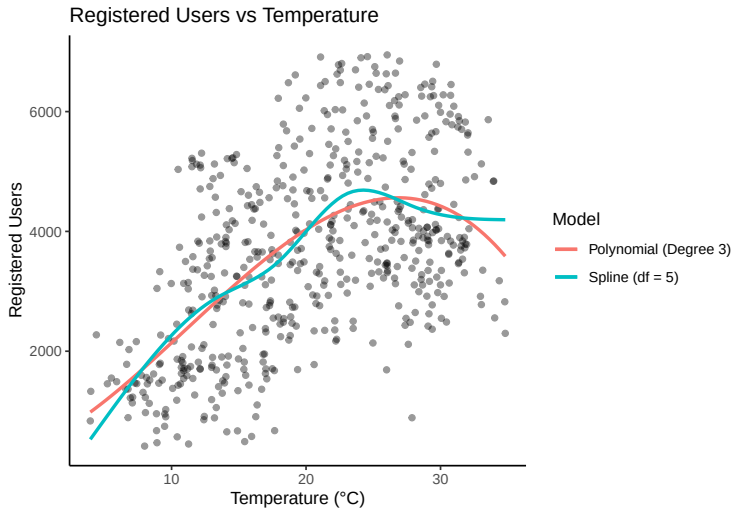
```
## [1] 9454.823
```

```
## [1] 9457.619
```

```
## [1] 9453.265
```

Q6 (4 pts)

Create a scatterplot of registered users (y) vs temperature (x). Overlay the selected polynomial model (from Q3) and the selected spline model (from Q5) on a single plot using different color lines. Include an informative legend.



Q7 - Q12 (AIC)

We continue using the bike data with registered users as the response (y). But now we will consider additional predictors:

- **as.factor(year)**: 2011 or 2012
- season: winter, spring, summer, or fall
- weather: Clear, Light Precip or Mist
- **poly(temp,3)**
- humidity: percent humidity (0-100)
- windspeed: peak windspeed in kilometers per hour.

Q7 (4 pts)

Use MuMIn::dredge() to perform all subsets selection using AIC criteria. Consider **additive models** only (no interactions). For the selected model, report the (multiple) R2 value and the selected predictors.

Note: By default, dredge() will rank models by AICc. Since this question specifically asks for AIC, specify rank = "AIC".

R2 = 0.8025
Predictors: as.factor (year), poly (temp,3), humidity, season, weather, and windspeed.

```
##
## Call:
## lm(formula = registered_users ~ as.factor(year) + humidity +
##     poly(temp, 3) + season + weather + windspeed + 1, data = bike_sharing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3526.9  -411.2   117.3   497.1  1313.6
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4861.173    225.395  21.567  < 2e-16 ***
## as.factor(year)2012   1750.014     61.584  28.417  < 2e-16 ***
## humidity               -14.302      2.912  -4.912 1.20e-06 ***
## poly(temp, 3)1       14807.384   1315.850  11.253  < 2e-16 ***
## poly(temp, 3)2       -6357.512    864.483  -7.354 7.19e-13 ***
## poly(temp, 3)3       -4789.476    727.195  -6.586 1.07e-10 ***
## seasonspring          -734.693     97.888  -7.505 2.55e-13 ***
## seasonsummer          -435.104    131.625  -3.306 0.001011 **
## seasonwinter         -1185.071    103.363 -11.465  < 2e-16 ***
## weatherLight Precip -1274.858    217.937  -5.850 8.55e-09 ***
## weatherMist          -265.828     80.282  -3.311 0.000991 ***
## windspeed             -33.854      6.222  -5.441 8.06e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 705.3 on 539 degrees of freedom
## Multiple R-squared:  0.8025, Adjusted R-squared:  0.7984
## F-statistic: 199.1 on 11 and 539 DF,  p-value: < 2.2e-16
```

Q8

Consider the diagnostic plots: Resids vs Fitted and QQplot of Residuals. You do NOT need to show the graphs, just discuss any concerns you might have.

Resids vs Fitted graphs, I woiuld look for strong curvature or fan shape to determined non linearity or poor model fit. For the QQplot, if the residuals goes far from the diagonal, it would indicate non-normality. With our regression, the R2 is pretty high so we might not have marjor concerns.

Q9

A colleague suggests that you consider the variable `tempfeel` as a potential predictor. Explain why we do NOT include `tempfeel` as a potential predictor. Reference specific evidence.

We don't include tempfeel as a potential predictor because it would be redundant, given that temp is already include in the model. temp and tempfeel are correlated as shown by the correlation test below. We would have perfect multicollinearity.

```
##
##  Pearson's product-moment correlation
##
## data:  bike_sharing$temp and bike_sharing$tempfeel
## t = Inf, df = 549, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  1 1
## sample estimates:
## cor
##   1
```

Q10

Looking at the MuMIn::dregde() results, a colleague asks why the + sign appears in the output. Briefly explain.

The + sign appears in the output to indicate the categorical predictor that was included in the model

---

Q11

A colleague suggests that you try stepwise AIC model selection. Do you expect this would change the selected model? Briefly discuss.

---

we wouldn't expect stepwise AIC to select the exact same model as dredge() given that the two approaches examines different set of models.

Q12

Propose at least one thing we could try to improve the model fit. You don't need to actually do this, just propose an idea.

One thing that we haven't done yet is to maybe include some interactions. We could interact season and temperature. This allows different slopes across groups and it might capture complex relationships in the data.

Appendix

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


# loading the data

bike_sharing <- read_csv("/data/yuwineza/star_513/Homework 9/bike_sharing.csv")

#Q1 - Q6: Polynomial and Splines
#Q1

# Polynomial models
model_deg_2 <- lm(registered_users ~ poly(temp, degree = 2), data = bike_sharing)
model_deg_3 <- lm(registered_users ~ poly(temp, degree = 3), data = bike_sharing)
```

```r
model_deg_4 <- lm(registered_users ~ poly(temp, degree = 4), data = bike_sharing)

#Q2

model_deg_3 <- lm(registered_users ~ poly(temp, 3), data = bike_sharing)
summary(model_deg_3)



#Q3

AIC(model_deg_2)
AIC(model_deg_3)
AIC(model_deg_4)

#Q4

spline_df_3 <- lm(registered_users ~ ns(temp, df = 3), data = bike_sharing)
spline_df_4 <- lm(registered_users ~ ns(temp, df = 4), data = bike_sharing)
spline_df_5 <- lm(registered_users ~ ns(temp, df = 5), data = bike_sharing)
#Q5
AIC(spline_df_3)
AIC(spline_df_4)
AIC(spline_df_5)
#Q6

ggplot(bike_sharing, aes(x = temp, y = registered_users)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 3), se = FALSE, aes(color = "Polynomial (Degree 3)")) +
  geom_smooth(method = "lm", formula = y ~ ns(x, 5), se = FALSE, aes(color = "Spline (df = 5)")) +

  labs(title = "Registered Users vs Temperature",
       x = "Temperature (°C)",
       y = "Registered Users",
       color = "Model") +
  theme_classic()


#Q7 - Q12: AIC
#Q7
full_model <- lm(registered_users ~ as.factor(year) + season + weather +
                   poly(temp, 3) + humidity + windspeed, data = bike_sharing)


options(na.action = "na.fail")  # required for dredge to work
model_set <- dredge(full_model, rank = "AIC")


# View top model
top_model <- get.models(model_set, 1)[[1]]
summary(top_model)
```

```r
#Q9

cor.test(bike_sharing$temp, bike_sharing$tempfeel)
# intentionally empty to show code later'
```