Lab Report 3: Multiple Linear Regression

MA 575 Fall 2021 - C3 Team #2

Ali Taqi, Hsin-Chang Lin, Huiru Yang, Ryan Mahoney, Yulin Li

11/15/2021

Contents

1 Introduction

In this lab report, Multiple Linear Regression (MLR) is performed on three different responses variable and a subset of predictors chosen from the 2011 Bike Sharing dataset [1]. The response variables of our concern are:

1. the count of **casual** daily bike rentals
2. the count of **registered** daily bike rentals
3. the count of **total** daily bike rentals

The model should help answer the following question: what are the **daily** bike rentals under different conditions? Business owners may like to know the daily bike rentals in 2012 so that they could optimize the inventory to reduce costs, and they may also wonder whether it is worth leaving the bike-sharing system open on days with extreme weather conditions. This can be done by performing predictive modeling on the daily rental variable based on data given in 2011.

2 Preprocessing

2.1 Overview

Variable Interpretations (see [1])

Both hour.csv and day.csv have the following fields, except `hr` which is not available in bike-day.csv:

- `instant`: record index
- `dteday`: date
- `season`: season (1:spring, 2:summer, 3:fall, 4:winter)
- `yr`: year (0:2011, 1:2012)
- `mnth`: month (1 to 12)

- `hr`: hour (0 to 23)
- `holiday`: weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- `weekday`: day of the week
- `workingday`: if day is neither weekend nor holiday is 1, otherwise is 0.
- `weather-sit`:
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- `temp`: Normalized temperature in Celsius. The values are divided to 41 (max)
- `atemp`: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- `hum`: Normalized humidity. The values are divided to 100 (max)
- `windspeed`: Normalized wind speed. The values are divided to 67 (max)
- `casual`: count of casual users
- `registered`: count of registered users
- `cnt`: count of total rental bikes including both casual and registered

```
# A brief look at the data structure from day.csv
head(bikedata, 3)
```

```
##   instant     dteday season yr mnth holiday weekday workingday weathersit
## 1       1 2011-01-01      1  0    1       0       6          0          2
## 2       2 2011-01-02      1  0    1       0       0          0          2
## 3       3 2011-01-03      1  0    1       0       1          1          1
##       temp    atemp      hum windspeed casual registered  cnt
## 1 0.344167 0.363625 0.805833  0.160446    331        654  985
## 2 0.363478 0.353739 0.696087  0.248539    131        670  801
## 3 0.196364 0.189405 0.437273  0.248309    120       1229 1349
```

2.2 Data Type & Value Conversion

Typically, all variables whose numerical values are not attached to actual physical meanings are treated as **categorical** variables.

```
# Boolean variables (from int to logical type)
holiday <- as.logical(bikedata$holiday)          #0 or 1
workingday <- as.logical(bikedata$workingday)    #0 or 1

# Other categorical variables (from int to factor type)
season <- as.factor(bikedata$season)             #1 to 4
yr <- as.factor(bikedata$yr)                     #0 to 1
mnth <- as.factor(bikedata$mnth)                 #1 to 12
weekday <- as.factor(bikedata$weekday)           #0 to 6
weathersit <- as.factor(bikedata$weathersit)     #1 to 4
```

Note: Although `weathersit` (weather type) is a categorical variable, its numerical value (from 1 to 4) actually indicates a gradual change in the level of suitability for outdoor activities - the larger the number, the worse the weather condition (i.e., more fogs/rains/snows, see "Variable Interpretations" above).

Furthermore, the normalized weather condition measurements (see "Variable Interpretations" above) are also converted to their original values, so that the numerical values being used "make more sense" to us. This makes it easier for commonsense and real-life experience to be applied in later analysis.

```
# Re-scale the normalized measurements
temp <- bikedata$temp * 41
atemp <- bikedata$atemp * 50
hum <- bikedata$hum * 100
windspeed <- bikedata$windspeed * 67
```

3 Multiple Linear Regression (MLR) Modeling

4 Conclusion

Given the above diagnostic results, it is still hard to decide among the 3 models.

Model 2 is good in that it involves less higher order terms, which means more stability of the model. However, it is weak in that it does not fit as well under the lowest and highest temperatures.

Model 3_1 and 4_2 are good in that they fit the best among all, have all coefficients significant, and also have nice diagnostic performance. However, the existence of higher-order terms leads to a more unstable model, and the one at a medium level of polynomial degrees, Model 3_1, happens to have the bi-modal issue in its standardized residuals.

These problems could potentially be resolved by:

1. introducing a time-series model
2. including the weekday variable
3. breaking the model into high and low temperature parts and model them separately.

Further analysis is left for the next lab report.

5 References

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

[2] MA 575 Fall 2021 C3 Team #2, "Lab Report 2: Ordinary Least Squares".