Lab Report 2: Ordinary Least Squares

MA 575 Fall 2021 - C3 Team #2

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

10/4/2021

Contents

1 Introduction

In this lab report, Ordinary Least Squares (OLS) modeling is performed on one response variable and a single predictor variable chosen from the 2011-2012 Bike Sharing dataset [1]. The dataset contains two main response variables of our concerns:

1. the total count of **daily** bike rentals
2. the total count of **hourly** bike rentals.

For simplicity, the former is chosen as the response variable to be studied in this lab. The model thus helps answer the following question as mentioned in Lab Report 1:

- What are the **daily** bike rentals under different conditions? (Business owners may like to know the daily bike rentals in 2013 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 and 2012.)

The studies in this report should serve as a starting point for later attempts of more sophisticated modeling approaches engaging more predictors.

2 Preprocessing

2.1 Overview

```
# A brief look at the data structure
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
```

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:springer, 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

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
```

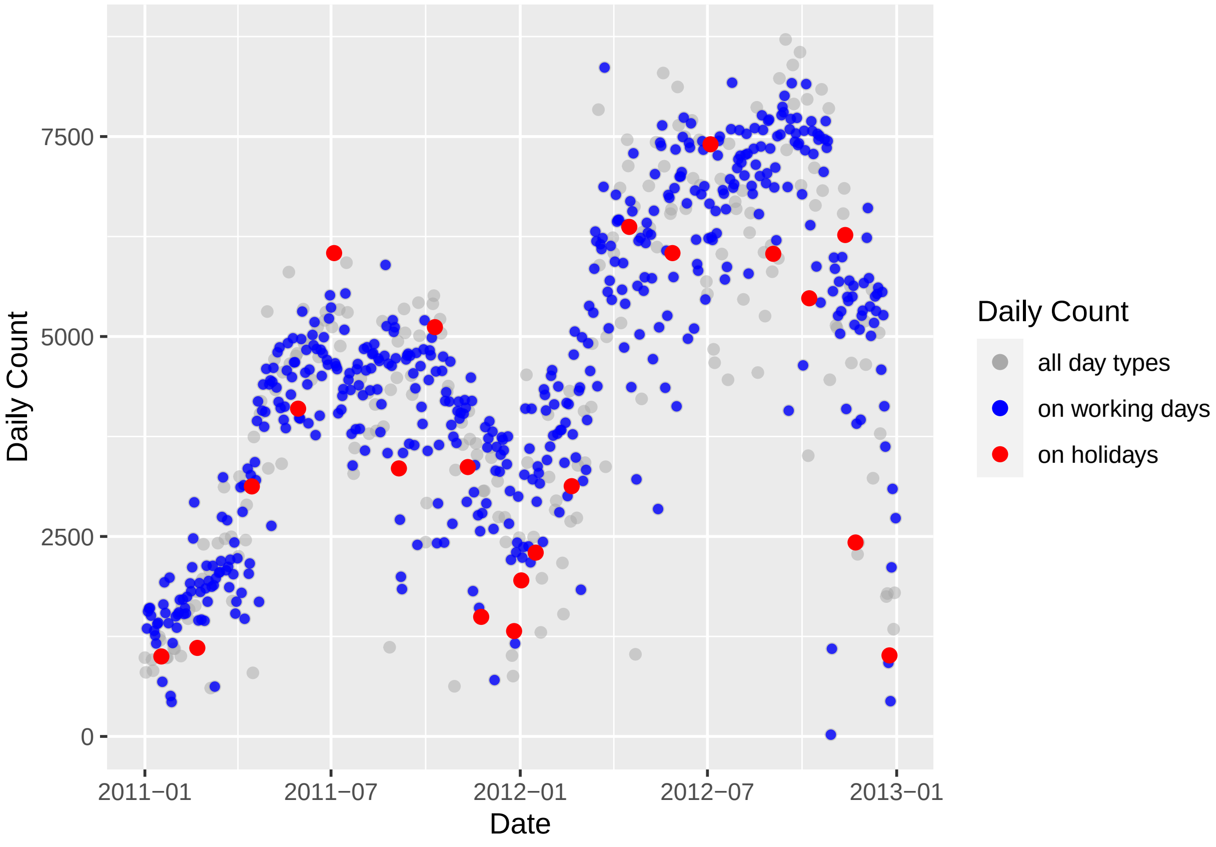
The normalized weather condition measurements (see "Variable Interpretations" above) are also converted to their original values, so that the numerical values being used "make more sense" to us. This makes it easier for commonsense and real-life experience to be applied in later analysis.

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

2.3 Visualization

Response variable `cnt`(count) by time (some groups labeled):

Fig.1: Total Count of Daily Bike Rentals by Date



Compare the response variable `cnt`(count) to a group of temporal and environmental variables that might be influential:

Fig.2: Pairs Plot

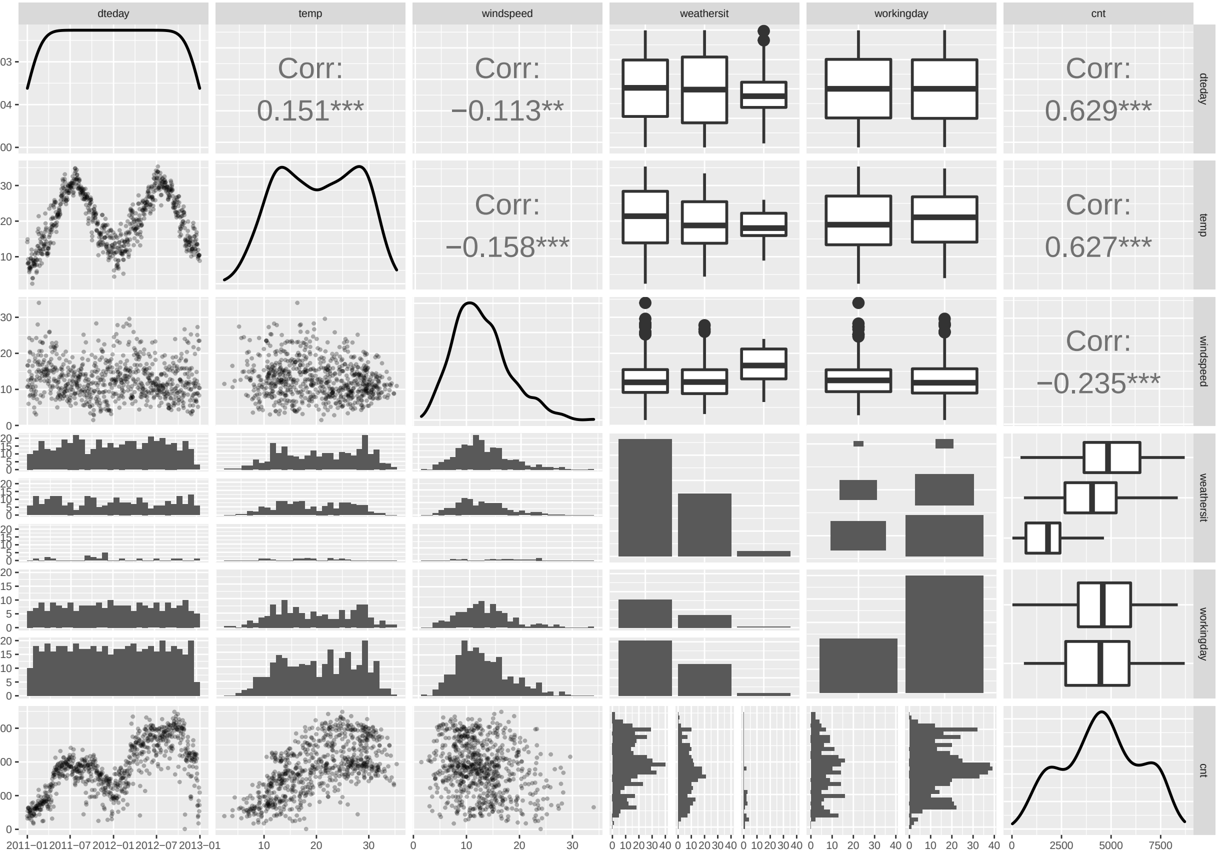


Fig.2 Column names in order: `dteday`, `temp`, `windspeed`, `weathersit`, `workingday`, `cnt`

Observations

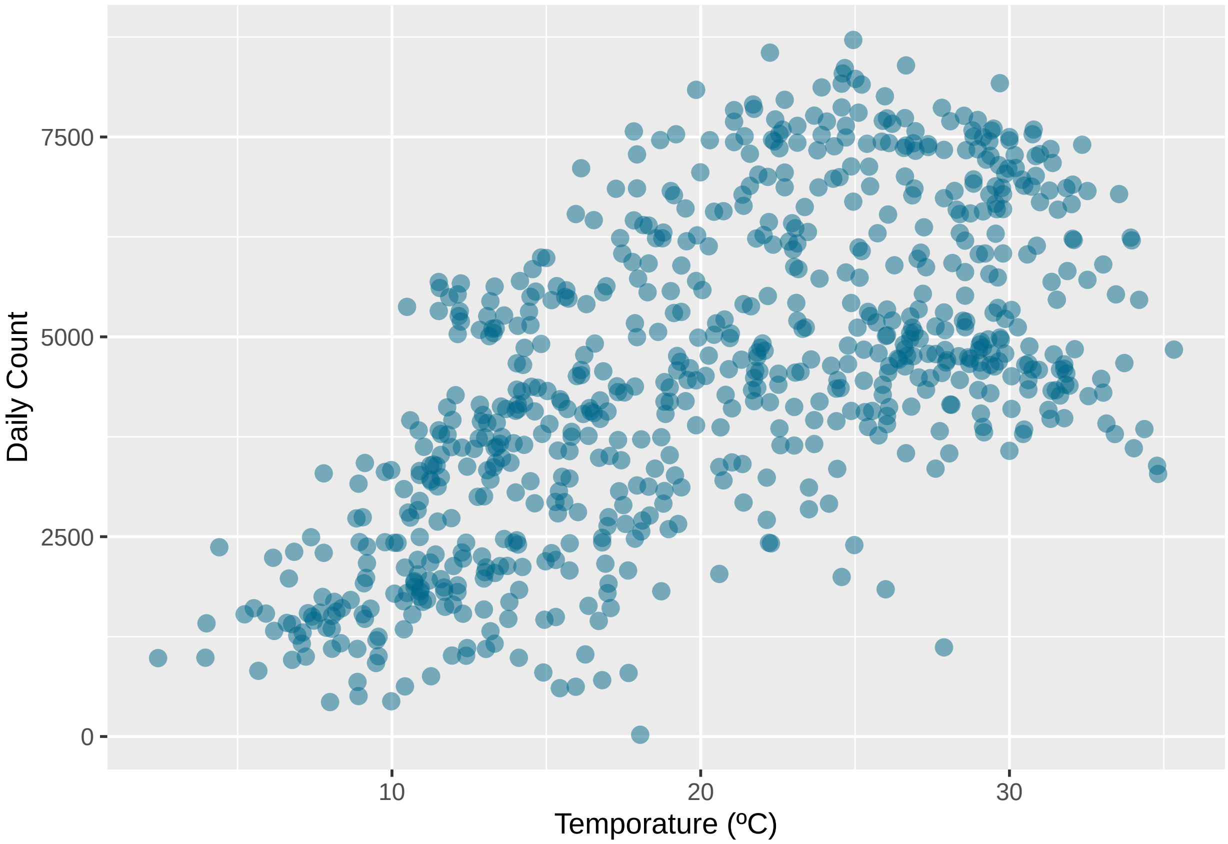
- In the last row of the pairs plot (Fig.2) where `cnt` is on the y-axis, the clearest trend shows up at column 1 & 2 where `dteday` and `temp` are on the x-axis, respectively. The corresponding correlations are also the highest in value (row 1 & 2 of the last column), indicating that there might be some linear relationship to be explored.
- In both Fig.1 and Fig.2, the differences that the categorical variables make (e.g., `weathersit`, `workingday`) are still not immediately clear.

Given that the impact of weather conditions is of great concerns and that the use of numerical variables are more common for a starting model, it is then decided to chose the weather condition measurement `temp`(temperature) as the predictor in this lab.

3 Model: Single-Predictor OLS

3.1 Methods

Fig.3: Daily Bike Usage Count by Recorded Temperature (2011-2012)



The curvy trend in the scatter plot suggests the need of (at least) a quadratic term in the predictor for the linear model.

Using the measured temperature values as our predictor (t), we attempt OLS models of the daily bike usage count (c) that take the following forms:

1. Linear:
$$c \sim \beta_0 + \beta_1 t$$
2. Quadratic:
$$c \sim \beta_0 + \beta_1 t + \beta_2 t^2$$
3. Cubic:
$$c \sim \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$$
4. Quartic:
$$c \sim \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4$$

3.2 Results

- **Model 1: Linear**

```r
# Ordinary LS: Linear
m.ols.1 <- lm(cnt ~ temp, data = bikedata)
summary(m.ols.1)
```

```
##
## Call:
## lm(formula = cnt ~ temp, data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4615.3 -1134.9  -104.4  1044.3  3737.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1214.6      161.2   7.537 1.43e-13 ***
## temp          6640.7      305.2  21.759  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1509 on 729 degrees of freedom
## Multiple R-squared:  0.3937, Adjusted R-squared:  0.3929
## F-statistic: 473.5 on 1 and 729 DF,  p-value: < 2.2e-16
```

- **Model 2: Quadratic**

```r
# Ordinary LS: Quadratic
m.ols.2 <- lm(cnt ~ I(temp^2) + temp, data = bikedata)
summary(m.ols.2)
```

```
##
## Call:
## lm(formula = cnt ~ I(temp^2) + temp, data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4580.4 -1043.6   -79.1  1150.7  3274.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1902.0      382.4  -4.974 8.19e-07 ***
## I(temp^2)   -15055.0     1692.5  -8.895  < 2e-16 ***
## temp         21406.9     1685.2  12.703  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1434 on 728 degrees of freedom
## Multiple R-squared:  0.4532, Adjusted R-squared:  0.4517
## F-statistic: 301.7 on 2 and 728 DF,  p-value: < 2.2e-16
```

- **Model 3: Cubic**

```
# Ordinary LS: Cubic
m.ols.3 <- lm(cnt ~ I(temp^3) + I(temp^2) + temp, data = bikedata)
summary(m.ols.3)
```

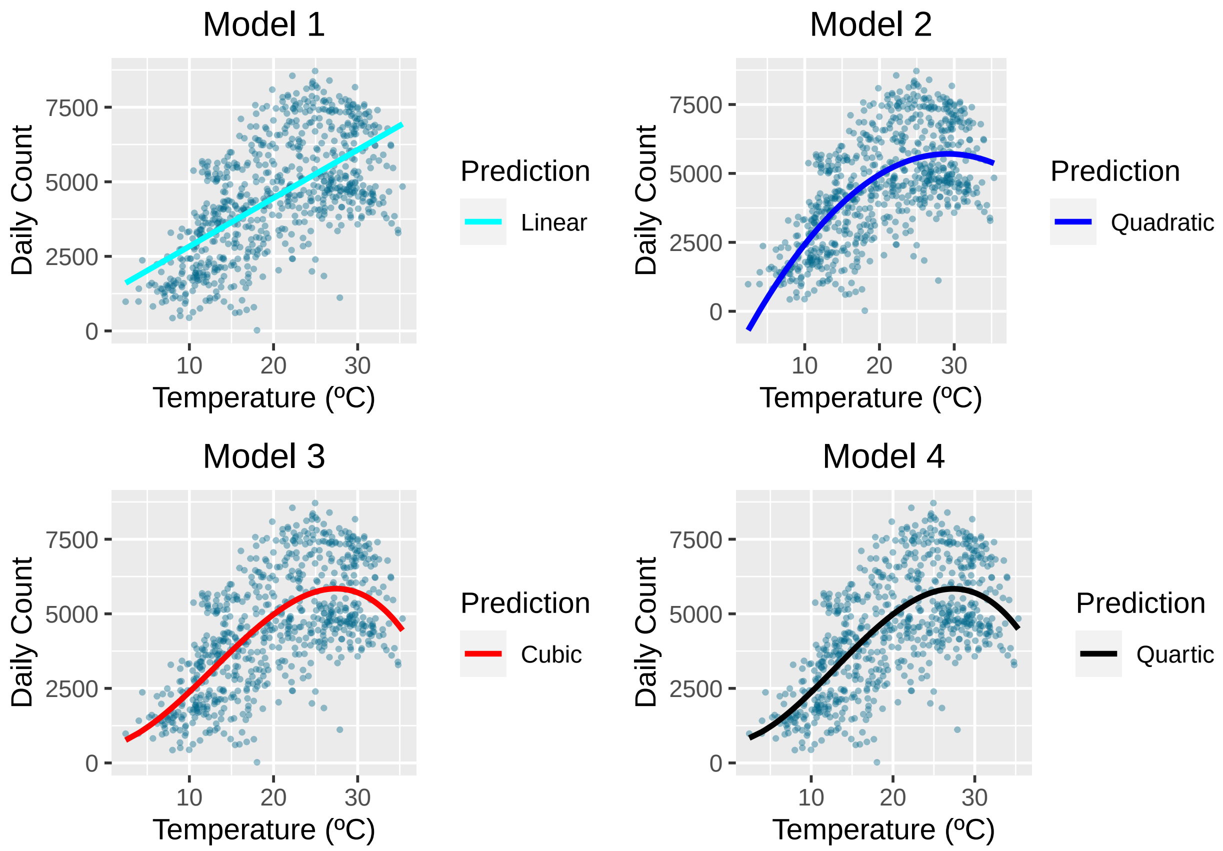
```
##
## Call:
## lm(formula = cnt ~ I(temp^3) + I(temp^2) + temp, data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4724.0 -1034.4   -99.6  1130.1  3160.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    519.0      775.3   0.669 0.503472
## I(temp^3)   -29799.9     8323.7  -3.580 0.000366 ***
## I(temp^2)    27962.0    12132.3   2.305 0.021461 *
## temp          2588.8     5515.7   0.469 0.638964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1423 on 727 degrees of freedom
## Multiple R-squared:  0.4627, Adjusted R-squared:  0.4604
## F-statistic: 208.6 on 3 and 727 DF,  p-value: < 2.2e-16
```

- **Model 4: Quartic**

```
# Ordinary LS: Quartic
m.ols.4 <- lm(cnt ~ I(temp^4) + I(temp^3) + I(temp^2) + temp, data = bikedata)
summary(m.ols.4)
```

```
##
## Call:
## lm(formula = cnt ~ I(temp^4) + I(temp^3) + I(temp^2) + temp,
##     data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4717.6 -1033.3  -103.1  1136.5  3155.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    676.6     1396.4   0.485    0.628
## I(temp^4)     5437.6    40063.4   0.136    0.892
## I(temp^3)   -40152.6    76730.2  -0.523    0.601
## I(temp^2)    34788.2    51738.5   0.672    0.502
## temp           783.2    14402.7   0.054    0.957
##
## Residual standard error: 1424 on 726 degrees of freedom
## Multiple R-squared:  0.4627, Adjusted R-squared:  0.4597
## F-statistic: 156.3 on 4 and 726 DF,  p-value: < 2.2e-16
```

Fig.4: Daily Bike Usage Count by Recorded Temperature, Predictions Overlayed Seperately



3.3 Discussion

Compare the four models:

- **Model 1: Linear**

This model is considered bad, even though all coefficients are indicated as “statistically significant” by the low p-values. This is because the model fails to capture the quadratic variation of data that is visually clear in the scatter plot; as a result, the residuals are clearly not centered around the predicted curve, which indicates violation of the “zero mean” assumption for noises in this model.

- **Model 2: Quadratic**

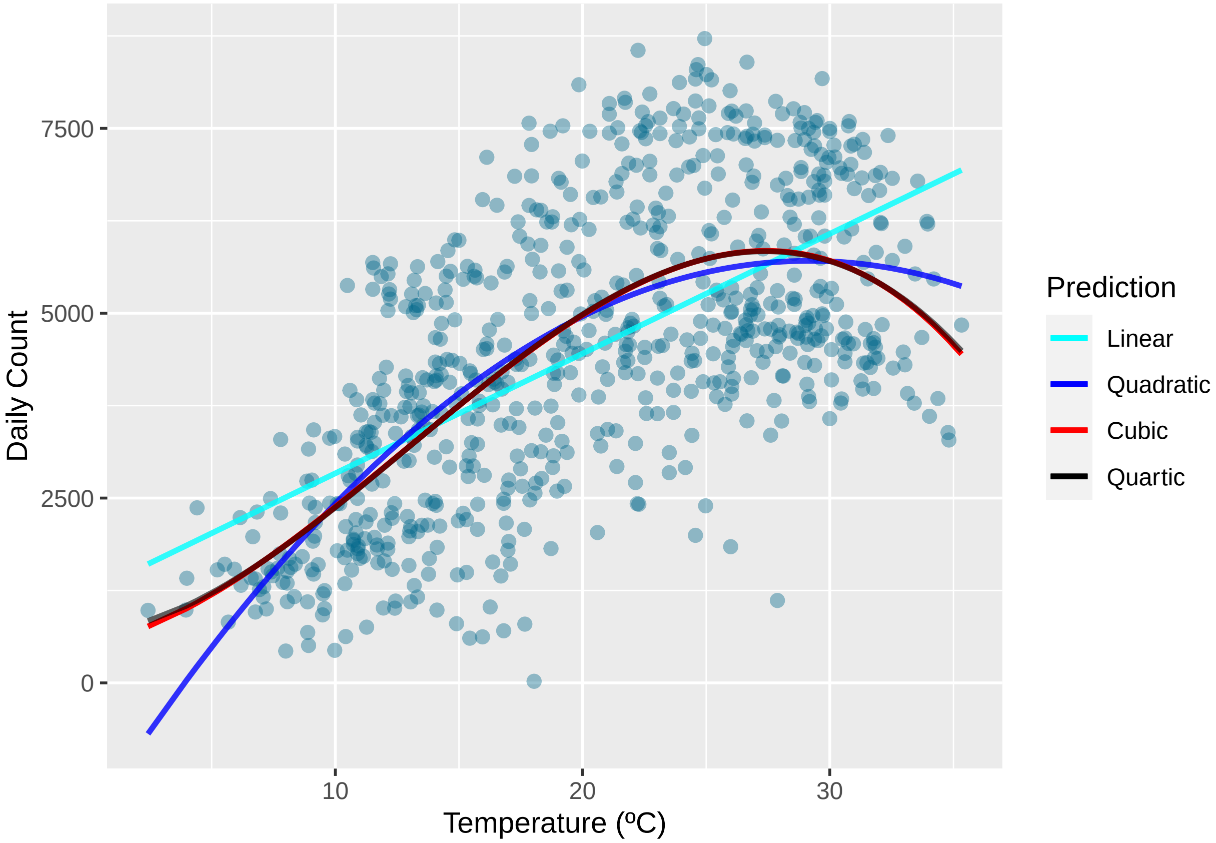
This model is fair enough in that:

1. It captures the non-linear variations.
2. All coefficients are indicated as “statistically significant”.
3. The variance of noises seems to be roughly constant over different temperatures.

However, it still has the following problems:

1. In the higher temperature part, the residuals are no longer centered around the predicted values.
2. In the lower temperature part, the predicted curve even crosses over the x-axis, which is highly impossible since there is no negative rental counts.

Fig.5: Daily Bike Usage Count by Recorded Temperature, Predictions Overlayed



- **Model 3: 3rd-degree polynomial**

Interestingly, in this model, the coefficients of the lower-order terms which used to be "significant" in Model 1 & 2 are no longer as significant. Instead, only the cubic term exhibits the same level of significance in p-value, which suggests that the cubic term might be better at explaining the variations than the lower-order terms.

Now, this model improves on the problems of Model 2 while maintaining a relatively strong predictive power.

- **Model 4: 4th-degree polynomial**

Interestingly, the prediction generated by this model almost completely overlaps with that of the 3rd-degree model, so increasing model complexity by adding higher-order terms at this point no longer leads to more modeling power. Further, the coefficients of this model are no longer indicated as statistically significant, which suggest that this model should probably be rejected. Overall, this would not be a better model than Model 3.

3.4 Conclusion

From the above, the 3rd-degree model with the predictor variable `temp` is considered the current best that can be reach with a single-predictor OLS model. Lastly, we verify that this model makes sense by our commonsense:

When the temperature approaches the comfortable temperature for outdoor activities, which should be warm but neither too high nor too low, the bike rental count is supposed to increase. That temperature should typically be between 20ºC and 30ºC, as is in the early summer or early autumn, and that corresponds to the peak of predicted value as well as the scatter point trends in Fig.5. From this perspective, the 3rd-degree polynomial model is a relatively convincing model.

Notably, in the scatter plot (Fig.3), the data points seem to be roughly divided into three layers, which could possibly be captured separately by three quadratic curves. This indicates that a categorical variable is probably needed to capture this characteristics. This is left for our further studies.

4 Reference

[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.