Seoul Bike Sharing Demand Analysis

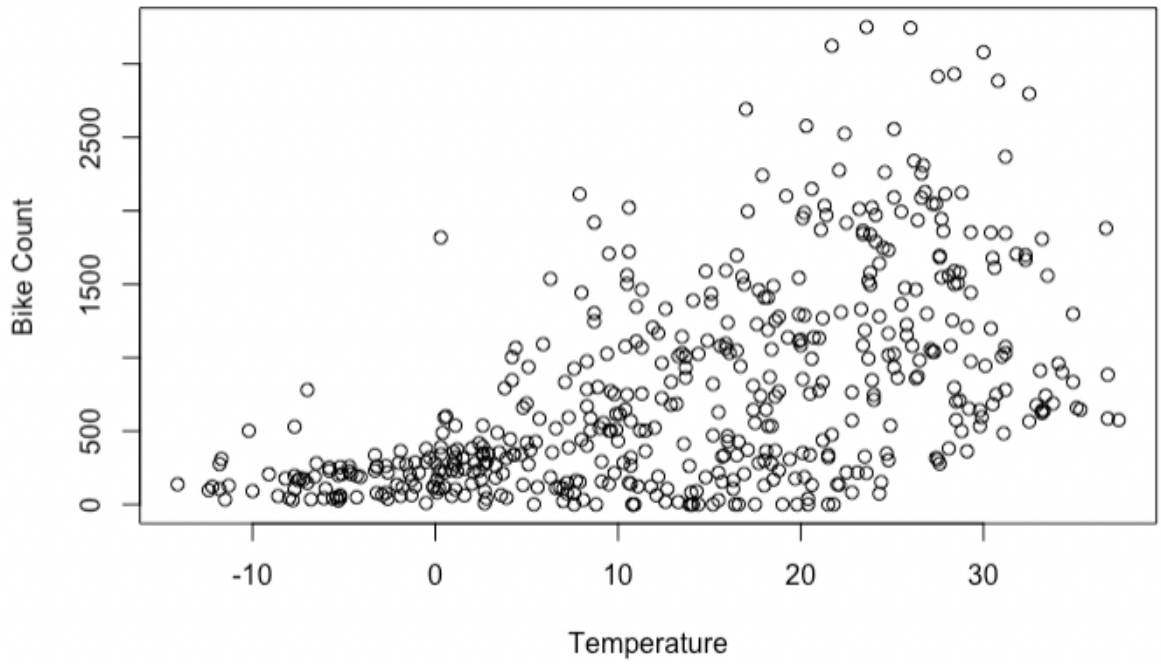
STA 9700 - Spring 2022
Professor Rongning Wu

Authors:
Andrea Escalona, Sierra Levy, Huong Thao Nguyen,
Balakumaran Ramaswamy Kannan, Zijing Yang

PART 1

This report will explore and analyze the Seoul Bike Sharing dataset from the Machine Learning Repository at UCI. The dataset contains the count of public bikes rented at each hour, with other weather and holiday information to help support the analysis. We have chosen Bike Count (which is the number of bikes rented) as our response variable since our goal is to understand demand and how weather and external factors affect bike count. For our predictor variable, we have chosen Temperature given that extreme weather conditions can potentially influence people's ability or willingness to use cycling as a form of transportation.

It should be noticed that the original data set contains more than 3000 observations, so we first subsampled the data and randomly chose 500 observations before the analysis. Below is a scatter plot of Bike Count and Temperature. The plot indicates that there is a positive relationship between the two variables. As temperature increases, the number of bikes rented per hour seems to increase as well.

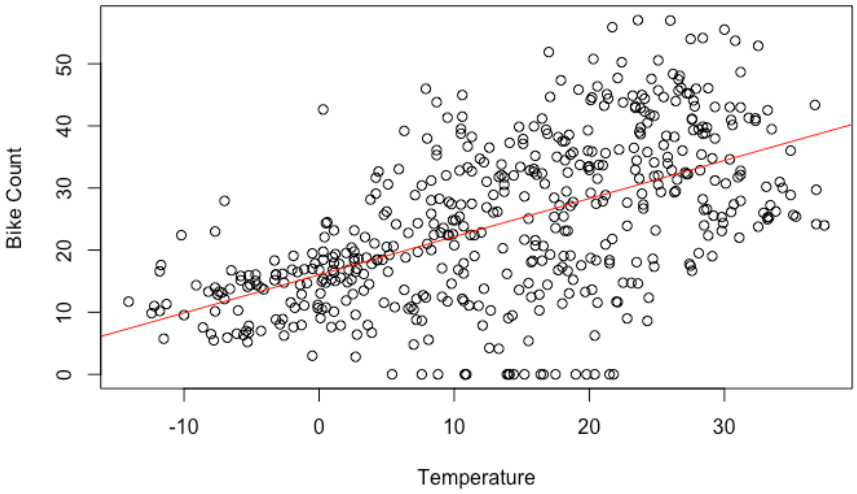


However, given the curvature of the data points, a square root transformation on Y is applied. As shown in the scatterplot of transformed Y against X below, such a transformation of

Y would likely fit the data better. After transforming the data, the estimated regression is as follows:

$$\widehat{Y}' = 16.03 + 0.61X$$

```
Call:
lm(formula = new.y ~ X, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-29.400  -7.256   0.639   6.843  26.545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.02543    0.73173   21.90   <2e-16 ***
X            0.61350    0.04026   15.24   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.73 on 498 degrees of freedom
Multiple R-squared:  0.318,     Adjusted R-squared:  0.3167
F-statistic: 232.2 on 1 and 498 DF,  p-value: < 2.2e-16
```

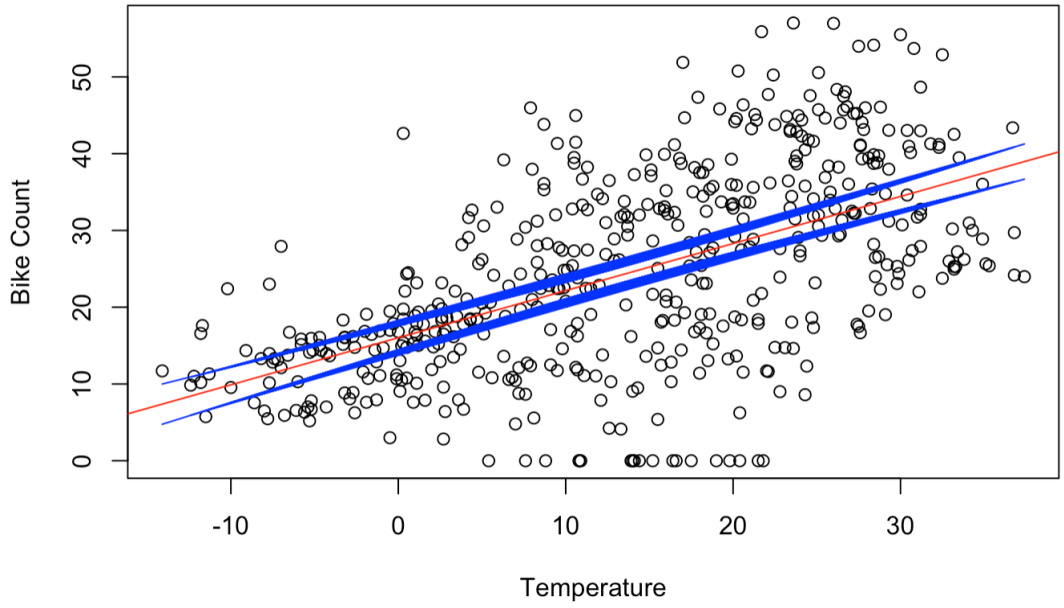


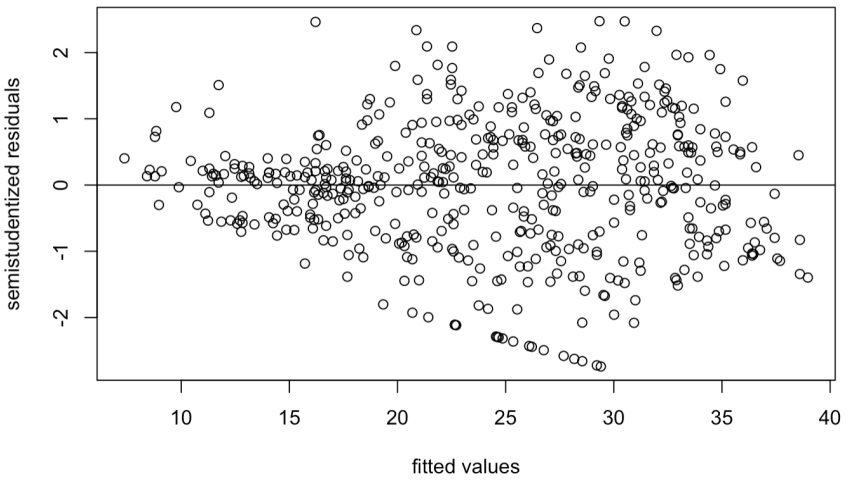
After fitting the model, our MSE is 115.19, and our R^2 value is .318041. MSE measures the average of the squared errors, that is, the average squared difference between the estimated values and the actual value. It is an unbiased estimator of the error variance σ^2 for the regression model and can also be used to estimate the standard deviation of the error terms. Our R^2 value indicates that 30.53% of the variation in bike count (Y) is accounted for by introducing temperature (X) into the regression model.

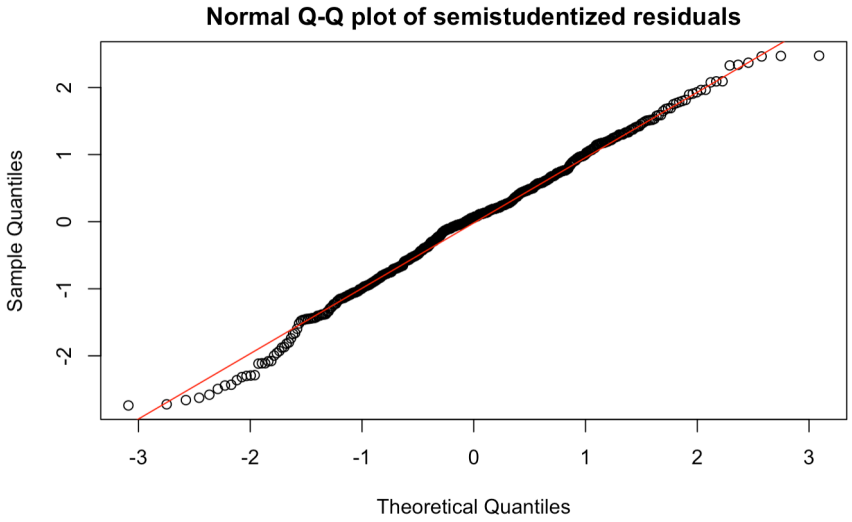
```
Analysis of Variance Table

Response: new.y
           Df Sum Sq Mean Sq F value    Pr(>F)
X           1  26753 26752.8  232.25 < 2.2e-16 ***
Residuals 498  57365   115.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To better understand if there is a linear relationship between Temperature and Bike Count, we're going to perform a t-test to help draw a conclusion. The following hypothesis is as follows: H_o: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$. Our decision rule is if t* ≤ t-critical, conclude H_o, otherwise, conclude H_a. From the regression analysis, we obtained a t* value of 15.24, a critical t value of 1.965 and a p-value of 2e-16. Since the t* value (15.24) is larger than the critical value (1.965), we reject the null hypothesis and conclude H_a, which tells us that β_1 is different from zero and there is a linear relationship between Temperature and Bike Count.

Below is a scatterplot of our 90% confidence band. This shows us the uncertainty in our data.







The above scatterplot of the semi-studentized residuals against the predicted values and the normal probability plot of the semi-studentized residuals help us in running a model diagnostic. The semi-studentized residuals plot against the fitted values suggests an apparent violation of the constant error variance assumption. From the plot, the larger the fitted value is, the more spread out the residuals are. This means that the error variance is larger for larger values of the predicted value than for smaller values of the predicted value. From the normal probability plot, most of the points lie on the theoretical straight line against the expected residuals although in bottom left and top right, a few data points move away from the line. The normal probability plot is suggesting weak departure of the residuals from normality.

PART 2

Model 1

After exploring the data, we decided that we would like to introduce these three variables into the model: Hour (X_1), Temperature (X_2), and Solar Radiation (X_3).

	bike.count	hour	temp	humidity	wind.speed	visibility	dew.point.temp	solar.radiation	rainfall	snowfall
bike.count	1.00	0.44	0.55	-0.22	0.12	0.19	0.40	0.27	-0.14	-0.13
hour	0.44	1.00	0.14	-0.28	0.24	0.10	0.01	0.14	-0.04	-0.01
temp	0.55	0.14	1.00	0.10	-0.08	0.07	0.91	0.33	0.02	-0.18
humidity	-0.22	-0.28	0.10	1.00	-0.39	-0.52	0.49	-0.50	0.28	0.17
wind.speed	0.12	0.24	-0.08	-0.39	1.00	0.18	-0.22	0.36	-0.08	-0.04
visibility	0.19	0.10	0.07	-0.52	0.18	1.00	-0.13	0.14	-0.22	-0.17
dew.point.temp	0.40	0.01	0.91	0.49	-0.22	-0.13	1.00	0.08	0.11	-0.10
solar.radiation	0.27	0.14	0.33	-0.50	0.36	0.14	0.08	1.00	-0.09	-0.07
rainfall	-0.14	-0.04	0.02	0.28	-0.08	-0.22	0.11	-0.09	1.00	0.12
snowfall	-0.13	-0.01	-0.18	0.17	-0.04	-0.17	-0.10	-0.07	0.12	1.00

```
Call:
lm(formula = df$bike.count ~ df$hour + df$temp + df$solar.radiation,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1337.43  -322.51   -41.92   240.80  1955.60

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -70.868     51.151  -1.385    0.167
df$hour             36.180      3.443  10.509   <2e-16 ***
df$temp             28.089      2.095  13.406   <2e-16 ***
df$solar.radiation  42.381     28.572   1.483    0.139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 523.9 on 496 degrees of freedom
Multiple R-squared:  0.4366,    Adjusted R-squared:  0.4332
F-statistic: 128.1 on 3 and 496 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: df$bike.count
                   Df    Sum Sq  Mean Sq  F value Pr(>F)
df$hour             1  45964865 45964865 167.4978 <2e-16 ***
df$temp             1  58896117 58896117 214.6198 <2e-16 ***
df$solar.radiation  1    603774   603774   2.2002 0.1386
Residuals         496 136112685   274421
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

After running the model, we get the following SST, SSR, and SSE values:

SSR : 105,464,756 - d.o.f : 3
SSE : 136,112,685 - d.o.f : 496
SST : 241,577,441 - d.o.f : 499

To test the overall linear relationship of the model, we conducted an F test to see if the group of variables are jointly significant. Our hypothesis is as follows: H_o: $\beta_1 = \beta_2 = \beta_3 = 0$, H_a: At least one $\beta_k \neq 0$ (k = 1,2,3). F* follows F(1-α; p-1, n-p) distribution under H_o. We will conclude Ha if our F- statistic is greater than the F-critical value, or if the p-value is less than 0.05. Otherwise, we will conclude H_Θ. After conducting an ANOVA test, we obtained an F-statistic of

128.1 (F*=MSR/MSE=(105,464,756/3)/274,421=128.1), an F-critical value of 2.622879 and a p-value of 0. Since our F-statistic is greater than the F-critical value and our p-value is less than the alpha of 0.05, we reject the null hypothesis and conclude that at least one of the slopes in the model does not equal 0 and there is a linear relationship.

After rejecting the null hypothesis and concluding that there is an overall linear relationship, we'll compute the Extra Sum of Squares(SSR ($X_3|X_1, X_2$)) and Coefficient of partial determination. Our SSR ($X_3|X_1X_2$) value is 603,774, and the coefficient of partial determination is .004416246. For interpretation, SSR ($X_3|X_1X_2$) is the extra sum of squares of solar radiation, given that hour and temperature are in the model, and R^2Y3l12 is the amount of variation in average bike count per hour that is reduced by introducing solar radiation into the model, given that hour and temperature are already included.

```
Analysis of Variance Table

Response: df$bike.count
                  Df    Sum Sq  Mean Sq  F value Pr(>F)
df$hour            1  45964865 45964865 167.4978 <2e-16 ***
df$temp            1  58896117 58896117 214.6198 <2e-16 ***
df$solar.radiation 1    603774   603774   2.2002 0.1386
Residuals        496 136112685   274421
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
$partial.rsq
[1] 0.004416246
```

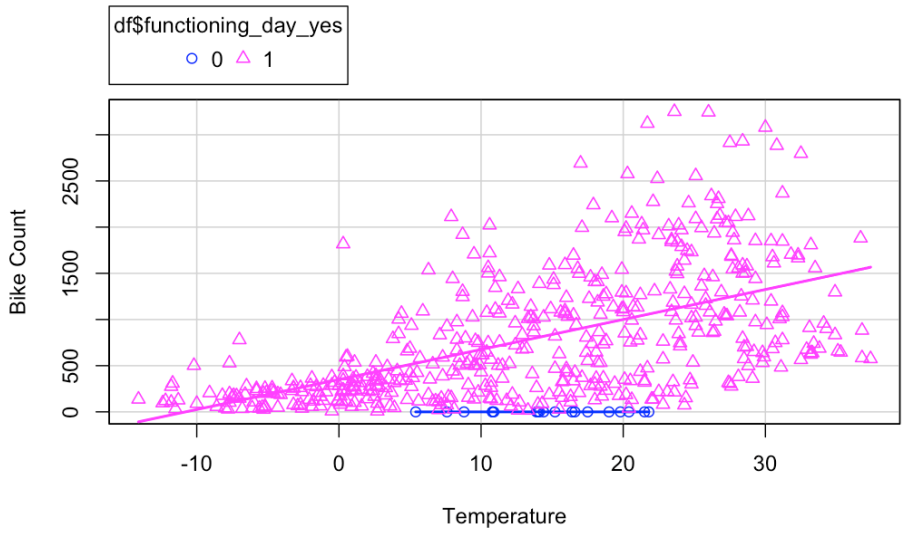
After analyzing the model and introducing new variables, we'll conduct an F test to determine whether X_3 is statistically helpful (given X_1 and X_2 are in the model). Our hypothesis is as follows: H_o: $\beta_3 = 0$, H_a: $\beta_3 \neq 0$. We will conclude H_o if our F- statistic is less than or equal to the F-critical value, or if the p-value is greater than 0.05. Otherwise, we will conclude H_a. After conducting a partial F-test, we obtained an F-statistic of 2.2002, an F-critical value of 3.86, and a

p-value of 0.1386. Since the F-statistic (2.2002) is less than the F-critical (3.86), and our p-value is greater than the alpha of 0.05, we fail to reject the null hypothesis and conclude that solar radiation does not significantly improve our model, given that temperature and hour of the day are already included. Therefore, we can exclude that predictor from the model.

Model 2

Our next model will introduce a categorical variable into the mix. We have chosen Functioning Day as our categorical variable which states whether an observation was made during functional or nonfunctional hours. The indicator variables are defined to take the value one if there are functional hours and zero if there are non functional hours. The second model will also contain Temperature as our numerical predictor. After fitting the model, we find that the coefficient of the categorical variable Functioning Day is 351.0, and the coefficient of the interaction term between Functioning Day and Temperature is 32.52. For interpretation, the average count of bikes rented per hour on Functioning days is 351 more than that of non-Functional days, when the temperature is 0 degrees. On the other hand, the effect of temperature on the average count of bikes rented per hour is 32.52 higher on Functional days than on non-Functional days.

```
Call:
lm(formula = df$bike.count ~ df$temp * df$functioning_day_yes,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1067.2  -352.8   -41.8   252.3  2132.6

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -5.806e-12  4.272e+02   0.000    1.000
df$temp                        2.839e-13  2.776e+01   0.000    1.000
df$functioning_day_yes         3.510e+02  4.289e+02   0.818    0.414
df$temp:df$functioning_day_yes 3.252e+01  2.784e+01   1.168    0.243

Residual standard error: 559.5 on 496 degrees of freedom
Multiple R-squared:  0.3572,    Adjusted R-squared:  0.3533
F-statistic: 91.87 on 3 and 496 DF,  p-value: < 2.2e-16
```



In order to test whether or not we should drop the interaction term, we'll run a t* test. Our hypothesis is as follows: H_o: $\beta_3 = 0$, H_a: $\beta_3 \neq 0$. We will conclude H_o if t* ≤ 1.965, and conclude H_a if t* > 1.965. After running the test, we obtained a t* value of 1.168 and a t-critical value of 1.965. Since 1.168 < 1.965, we conclude H_o which means we should drop the interaction term. In our case, this means that while the effect of temperature on Functional and non-Functional days is different, that difference is not statistically significant.

Model 3

Our last model is going to be built with the AIC Criterion and "Backwards Elimination" procedure to automatically obtain the best fit model. We started with 12 predictor variables, and after completing Backwards Elimination, we find that the best fit model includes the 7 following variables: Solar Radiation, Rainfall, Season, Dew Point Temperature, Humidity,

Functioning Day, and Hour. The AIC value is 6,162.43. Our adjusted R^2 value improved from Model 1 to model 3, increasing from .4332 to .5439.

```
Step:  AIC=6162.43
df$bike.count ~ hour + humidity + dew.point.temp + solar.radiation +
    rainfall + season + functioning.day

                  Df Sum of Sq       RSS    AIC
<none>                         108198177 6162.4
- solar.radiation  1    688816 108886993 6163.6
- rainfall         1   2136540 110334717 6170.2
- season           3   5278838 113477015 6180.3
- dew.point.temp   1   8807574 117005751 6199.6
- humidity         1  12496141 120694318 6215.1
- functioning.day  1  14243277 122441454 6222.3
- hour             1  21850623 130048800 6252.4
```

```
Call:
lm(formula = df$bike.count ~ hour + humidity + dew.point.temp +
    solar.radiation + rainfall + season + functioning.day, data = df[,
    3:14])

Residuals:
    Min      1Q  Median      3Q     Max
-966.66 -325.22  -59.86  242.41 1812.82

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        343.888    164.142   2.095 0.036678 *
hour                32.529      3.270   9.948  < 2e-16 ***
humidity           -14.456      1.922  -7.523 2.58e-13 ***
dew.point.temp      27.367      4.333   6.316 6.04e-10 ***
solar.radiation    -55.057     31.173  -1.766 0.077985 .
rainfall          -121.607     39.094  -3.111 0.001976 **
seasonSpring       -13.336     61.331  -0.217 0.827957
seasonSummer       -82.072     78.619  -1.044 0.297040
seasonWinter      -338.199     91.854  -3.682 0.000257 ***
functioning.dayYes  926.039    115.302   8.031 7.22e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 469.9 on 490 degrees of freedom
Multiple R-squared:  0.5521,    Adjusted R-squared:  0.5439
F-statistic: 67.12 on 9 and 490 DF,  p-value: < 2.2e-16
```