Data Analytics Assessment #5

Beiming Zhang

Conceptual Part

Question 1

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}{X_1}^2$$

When one-unit increase in X_1:

$$\widehat{Y'} = \widehat{\beta_0} + \widehat{\beta_1}(X_1 + 1) + \widehat{\beta_2}(X_1 + 1)^2 = \widehat{\beta_0} + \widehat{\beta_1}(X_1 + 1) + \widehat{\beta_2}({X_1}^2 + 2X_1 + 1)$$

The average value of $\hat{Y}$:

$$\widehat{Y'} - \hat{Y} = \widehat{\beta_1} + \widehat{\beta_2}(2X_1 + 1)$$

Programming Part

Question 1

a)

Output:

```
Call:
lm(formula = Rentals ~ Temperature, data = Bikeshare)

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 ***
Temperature   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
```

- The temperature has a statistically significant effect on the number of rentals. It can be noted that the p-value corresponding to Temperature is less than 0.001, therefore we reject the null hypothesis (H0).
- Based on the output, the regression results are easily obtained as follows:

$$\hat{Y} = 1214.6 + 6640.7X$$

- When one degree (Celsius) change in temperature on the rentals, the normalized temperature change:

$$t_n = \frac{t_d - t_{min}}{t_{max} - t_{min}} = \frac{t_d}{47} + \frac{8}{47}$$

$$\frac{dt_n}{dt_d} = \left(\frac{t_d}{47}\right)' + \left(\frac{8}{47}\right)' = \frac{1}{47}$$

- Differentiating the linear regression formula yields:

$$\frac{d\hat{Y}}{dX} = (1214.6)' + (6640.7X)' = 6640.7$$

- So, when one degree (Celsius) change in temperature on the rentals, the rentals change equals 6640.7*1/47, which is 141.29.

b)

Output:

```
Call:
lm(formula = Rentals ~ Humidity, data = Bikeshare)

Residuals:
     Min        1Q    Median       3Q       Max
-4741.0 -1386.9      50.3    1439.3   4036.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     5364.0         322.7    16.623    < 2e-16 ***
Humidity         -1369.1          501.2   -2.732    0.00645 **
---
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1929 on 729 degrees of freedom
Multiple R-squared:   0.01013,     Adjusted R-squared:   0.008774
F-statistic: 7.462 on 1 and 729 DF,   p-value: 0.006454
```

- The Humidity has a statistically significant effect on the number of rentals. It can be noted that the p-value corresponding to Humidity is less than 0.01, therefore we reject the null hypothesis (H0).
- Based on the output, the regression results are easily obtained as follows:

$$\hat{Y} = 5364.0 - 1369.1X$$

- When one change in Humidity on the rentals, the normalized Humidity change:

$$h_n = \frac{h_d}{100}$$

$$\frac{dh_n}{dh_d} = \left(\frac{h_d}{100}\right)' = \frac{1}{100}$$

- Differentiating the linear regression formula yields:

$$\frac{d\hat{Y}}{dX} = (5364.0)' - (1369.1X)' = -1369.1$$

- So, when one change in Humidity on the rentals, the rentals change equals -1369.1*1/100, which is -13.69.

c)

Output:

```
Call:
lm(formula = Rentals ~ Windspeed, data = Bikeshare)

Residuals:
    Min      1Q  Median      3Q     Max
-4522.7 -1374.7   -74.6  1461.8  4544.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5621.2      185.1  30.374  < 2e-16 ***
Windspeed    -5862.9      900.0  -6.514 1.36e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1884 on 729 degrees of freedom
Multiple R-squared:  0.05501,	Adjusted R-squared:  0.05372
F-statistic: 42.44 on 1 and 729 DF,  p-value: 1.36e-10
```

- The Windspeed has a statistically significant effect on the number of rentals. It can be noted that the p-value corresponding to Windspeed is less than 0.001, therefore we reject the null hypothesis (H0).
- Based on the output, the regression results are easily obtained as follows:

$$\hat{Y} = 5621.2 - 5862.9X$$

- When one change in Windspeed on the rentals, the normalized Windspeed change:

$$w_n = \frac{w_d}{w_{max}} = \frac{w_d}{67}$$

$$\frac{dw_n}{dw_d} = \left(\frac{w_d}{67}\right)' = \frac{1}{67}$$

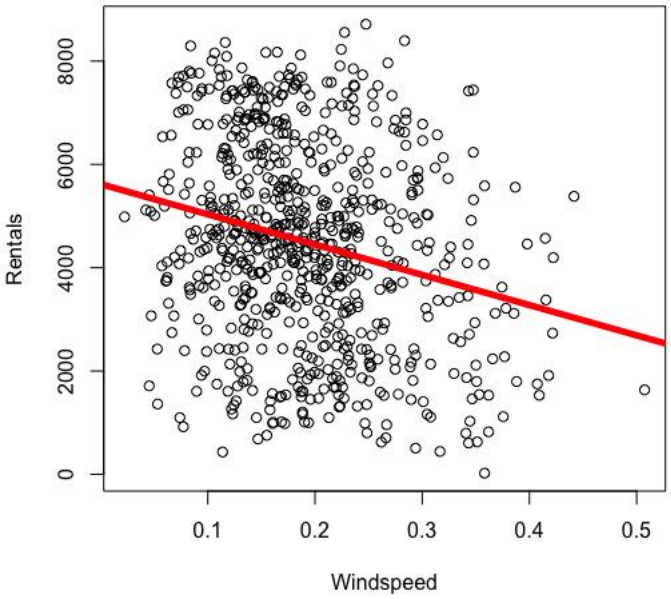
- Differentiating the linear regression formula yields:

$$\frac{d\hat{Y}}{dX} = (5621.2)' - (5862.9X)' = -5862.9$$

- So, when one change in Windspeed on the rentals, the rentals change equals -5862.9*1/67, which is -87.51.

d)

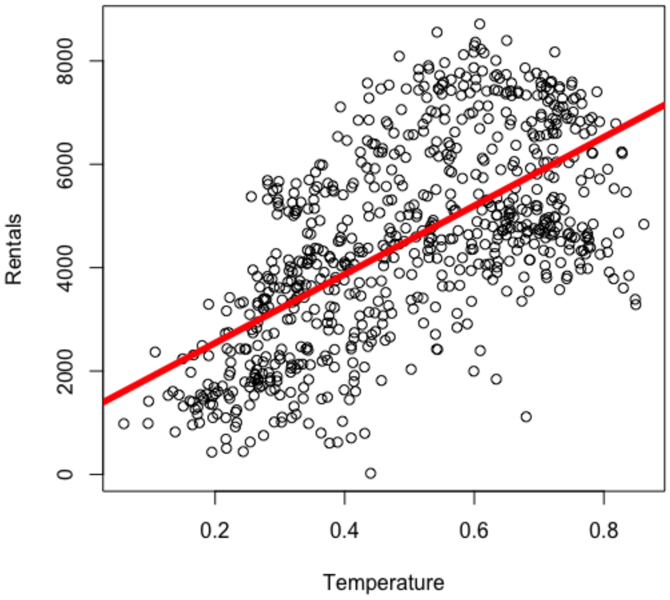
Output:



- The R² value in part c) is 0.05501. A lower R² value indicates a worse proportion of variance in the dependent variable being explained by the independent variables. This means that the model has a poor ability to explain the data.
- From the graph, we can see that the distribution of data points is quite scattered. We should consider adjusting the number of variables in the model for optimization.
- However, considering the small p-value corresponding to Windspeed, there still exists a significant linear relationship between them; it's just that this relationship is not suitable to be represented with a linear regression model.

e)

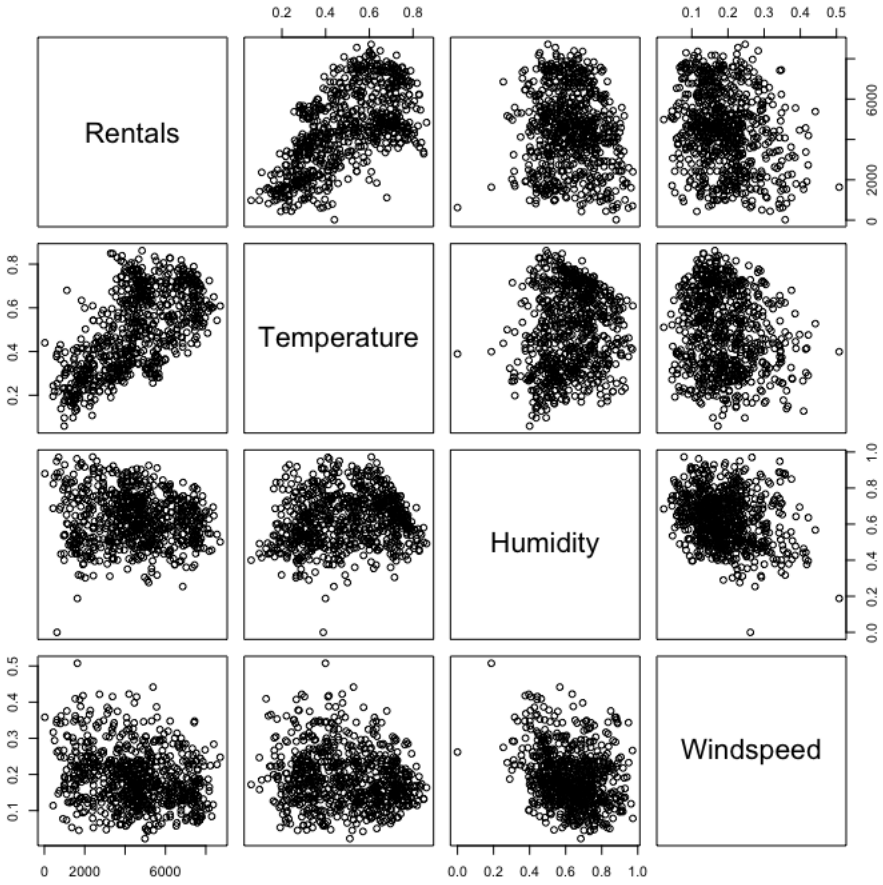
Output:



- The scatter plot and regression line for 'Rentals' versus 'Temperature' are shown above. The two show a certain degree of linear relationship.

f)

Output:



- ♦ The input variable "Temperature" appears to have an effect on "Rentals". Their graph can be seen as extending from the bottom left to the top right. Clearly, Temperature and Rentals show an obvious positive correlation.
- ♦ Although Humidity and Windspeed may also show a certain linear relationship, it is not significant. I believe this image does not convincingly demonstrate their relationship with Rentals.

g)

Output:

```
Call:
lm(formula = Rentals ~ . - Day, data = Bikeshare)

Residuals:
     Min        1Q   Median        3Q       Max
-4780.5 -1082.6      -62.2   1056.5   3653.5

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     4084.4         337.9    12.089    < 2e-16 ***
```

```
Temperature     6625.5          293.1    22.606    < 2e-16 ***
Humidity        -3100.1          384.0    -8.073 2.83e-15 ***
Windspeed       -4806.9          708.9    -6.781 2.48e-11 ***
---
Signif. codes:   0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1425 on 727 degrees of freedom
Multiple R-squared:    0.4609, Adjusted R-squared:    0.4587
F-statistic: 207.2 on 3 and 727 DF,    p-value: < 2.2e-16
```

- First, observe the p-values. The p-values corresponding to Temperature, Humidity, and Windspeed are all less than 0.001, therefore these three variables are quite significant.
- When we observe the R^2 value, we can see that it is higher compared to the models involved in the previous questions, therefore this model provides a better explanation of the data.
- Therefore, Temperature, Humidity and Windspeed all have a statistically significant effect on 'Rentals'.

h)

- Based on the results of question g):

$$\hat{Y} = 4084.4 + 6625.5X_t - 3100.1X_h - 4806.9X_w$$

- Normalized all parameters:

$$X_t = \frac{t_d}{47} + \frac{8}{47} = \frac{15}{47} + \frac{8}{47} = \frac{23}{47}$$

$$X_h = \frac{h_d}{100} = \frac{50}{100} = \frac{1}{2}$$

$$X_h = \frac{w_d}{67} = \frac{5}{67}$$

- So, the predicted number of rentals is:

$$\hat{Y} = 4084.4 + 6625.5 \times \frac{23}{47} - 3100.1 \times \frac{1}{2} - 4806.9 \times \frac{5}{67} = 5417.89$$

Question 2

a)

Output:

```
Call:
lm(formula = Rentals ~ Temperature + Humidity + Windspeed + Weekend,
     data = BikeshareUpdated)

Residuals:
     Min          1Q    Median          3Q          Max
-4796.4 -1085.1     -64.6    1045.3    3688.4
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4101.65     340.55  12.044  < 2e-16 ***
Temperature  6620.59     293.49  22.558  < 2e-16 ***
Humidity    -3102.00     384.24  -8.073 2.84e-15 ***
Windspeed   -4804.77     709.32  -6.774 2.59e-11 ***
Weekend       -48.96     116.69  -0.420    0.675
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1426 on 726 degrees of freedom
Multiple R-squared:  0.461,  Adjusted R-squared:  0.4581
F-statistic: 155.3 on 4 and 726 DF,  p-value: < 2.2e-16
```

- The p-values for Temperature, Humidity, and Windspeed are less than 0.001, indicating high significance. However, the p-value for Weekend is greater than 0.1, indicating low significance.
- Therefore, Temperature, Humidity, and Windspeed have a statistically significant effect on the number of rentals.

b)

Output:

```
Call:
lm(formula = Registered ~ Temperature + Humidity + Windspeed +
    Weekend, data = BikeshareUpdated)

Residuals:
    Min      1Q  Median      3Q     Max
-4018.3  -922.9   -52.7   910.8  2919.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3805.79     286.74  13.273  < 2e-16 ***
Temperature  4490.44     247.11  18.172  < 2e-16 ***
Humidity    -2277.38     323.52  -7.039 4.48e-12 ***
Windspeed   -3657.16     597.24  -6.123 1.50e-09 ***
Weekend      -861.59      98.25  -8.769  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1201 on 726 degrees of freedom
Multiple R-squared:  0.411,  Adjusted R-squared:  0.4077
F-statistic: 126.6 on 4 and 726 DF,  p-value: < 2.2e-16
```

- The p-values for Temperature, Humidity, Windspeed and Weekend are less than 0.001, indicating high significance.
- Therefore, Temperature, Humidity, Windspeed and Weekend have a statistically significant effect on the number of rentals by registered users.

c)

Output:

```
Call:
lm(formula = Casual ~ Temperature + Humidity + Windspeed + Weekend,
    data = BikeshareUpdated)

Residuals:
     Min       1Q    Median      3Q      Max
-1367.44  -215.35   -18.26   165.46  1840.78

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   295.85     101.63   2.911  0.00371 **
Temperature  2130.15      87.58  24.321  < 2e-16 ***
Humidity     -824.61     114.67  -7.191 1.60e-12 ***
Windspeed   -1147.61     211.68  -5.421 8.06e-08 ***
Weekend       812.63      34.82  23.335  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 425.6 on 726 degrees of freedom
Multiple R-squared:  0.6179, Adjusted R-squared:  0.6158
F-statistic: 293.5 on 4 and 726 DF,  p-value: < 2.2e-16
```

- The p-values for Temperature, Humidity, Windspeed and Weekend are less than 0.001, indicating high significance.
- Therefore, Temperature, Humidity, Windspeed and Weekend have a statistically significant effect on the number of rentals by casual users.

d)

- Weekend is significantly negatively correlated with rentals by registered users and significantly positively correlated with rentals by casual users. It is not significantly correlated with rentals overall.
- Therefore, weekends decrease the rental demand for registered users, increase the rental demand for casual users, and have a non-significant impact on the overall rental demand.