Problem set2

Group member:
Shiming Zhao(MISM6202-05)
Haozhe Liu(MISM6202-02)

Regression Analysis.

Question 1

<1> Regression equation

The p-value of daily bookings are small enough so that we have 95% confidence to say there is a regression relationship between daily completed rides and daily bookings.

Daily completed rides = 145.34331 +0.90088*daily bookings

<2> R_2 value

R-squared: 0.9739 Adjusted R-squared: 0.9734

<3> Interpretation

Regression equation: The coefficient of daily bookings is 0.90088, suggesting a positive influence of daily bookings on daily completed. To be more specific, if the number of daily bookings increases by 1 unit, then, the predicted number of daily completed rides is expected to increase by 0.90088, nearly one, holding all other conditions constant.

R_2 value: The coefficient of determination R2 for this model is 0.9739, which means that 97.39% of the sample variation in daily completed rides is explained by the regression model.

Question2:

<1> Execute three different multiple regressions and pick the best one.

	Model1	Model2	Model3
Intercept	86.78463	81.303832	94.945744
Starting a session	0.23139	0.239466	0.213144
Tapping on the sidebar	-0.66689	-0.623606	NA
Tapping on a stop	0.02856(insignificant)	NA	NA
Viewing van ETAs	-0.13938	-0.158358	-0.205511
Se	140	**139.2**	170.6
R-suqared	0.9625	0.9623	0.9424
Adjusted R-squared	0.96	**0.9604**	0.9405
F-test (P-value)	378.8(0.0000)	510.3(0.0000)	499.4(0.0000)

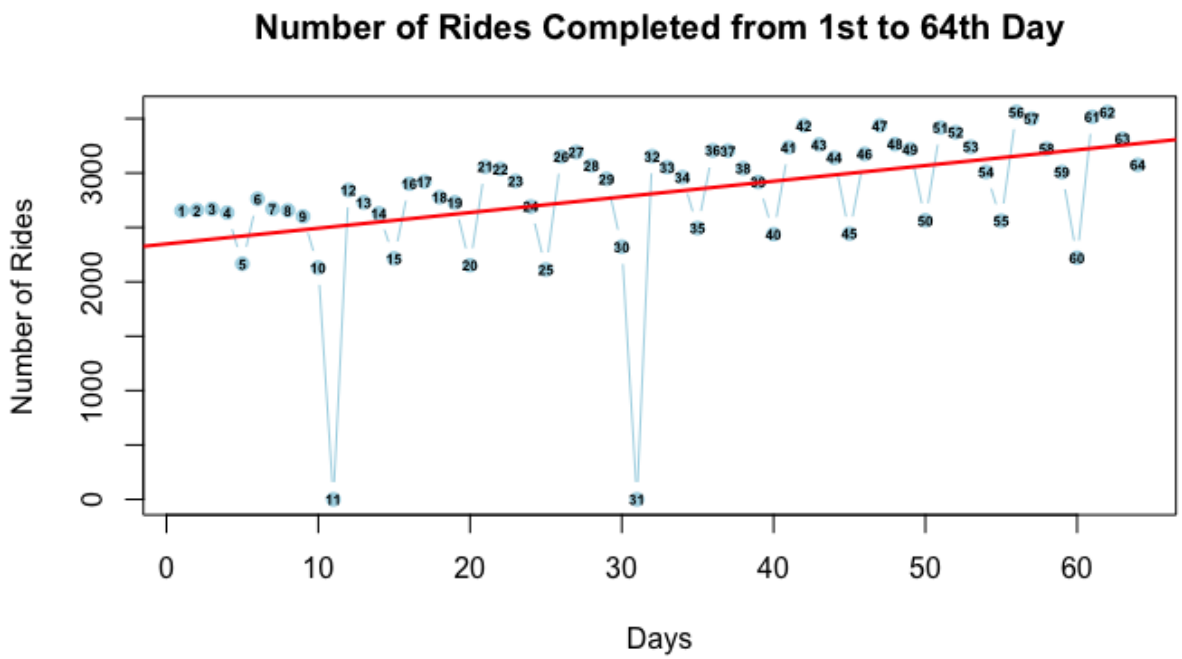
In order to estimate which multiple regression models are more superior in predicting ride bookings, we estimate three multiple regression models.
The first model contains all four possible predictor variables, which are starting a session, tapping on a stop, tapping on the sidebar, and viewing van ETAs.
The second model excludes the predictor variable, tapping on a stop, because it is insignificant at the 5% confidence interval in the first multiple regression model.
The third model excludes both tapping on a stop and tapping on the sidebar, because customers care more about ETA(estimated time of arrival) to make decisions compared with viewing sidebar or stop.
According to goodness-of-fit measure, **the model 2** containing starting a session, tapping on the sidebar, and viewing van ETAs **is best fit** because of its lowest standard error of estimate(SEE), 139.2, and highest adjusted R^2 value, 0.9604.

<2> The estimated regression equation
Daily bookings= 81.303832 + 0.239466*(starts.session) - 0.623606*(tapped.sidebar) - 0.158358*(viewed.eta)
Individual tests of significance showed that all three variables were significant at the 5% level.
Besides, F-statistic was 510.3, and p-value is so small that the predictor variables are jointly significant in explaining daily bookings.
There are negative relationships between tapping on the sidebar and daily bookings, and between viewing van ETAs and daily bookings. There is a positive relationship between starting a session and daily bookings. Moreover, holding all other predictor variables constant, one customer tapping on the sidebar will decrease 0.62 customers to book tickets.

<3> R2 value
The coefficient of determination, R2, was 0.9623, which revealed that approximately 96.23% of the sample variability in daily bookings was explained by the model, which further suggested that this model is pretty good.

***Forecasting*.**

Question3:



This scatter plot is about daily completed rides, with dots connected, and every weekday is labeled from 1 to 64. This graph describes a linear regression model for trends and seasonality.

First, there is a persistent upward movement in daily rides completed, except the number 11 and 31, two days off.

Second, the seasonality is shown in the graph where every 5 weekdays repeat themselves week after week. For example, the daily rides completed are lower in the fifth weekday as compared to the other four weekdays.

Question4:
<1> We choose 5 as k-period because seasonality patterns are repeated every 5 days.

<2> When calculating MSE, MAD, MAPE, the first thing we have done is to replace zero value of rides with nearby values. For instance, we use 12th ride, 2844, to replace 11th rides, 0 and 32nd ride, 3156, to replace 31st ride,0. This is because both 11th and 31st are Monday, which always have similar ride numbers with Tuesday's by going through all dataset.

By calculating by R, we get MSE = 111581.38, MAD = 270.55, and MAPE = 9.80%.

Question5:

<1> Based on new dataframe we have created in Q4, we got the equation:

Rides = 2540.048 + 11.401*t. This equation means every weekday, the number of rides completed will increase by about 11.491.

<2> R-squared: 0.3026, adjusted R-squared: 0.2914.

This R2 suggests that about 30.26% of the sample variations in rides are explained by the sample trend line.

Question6:

<1> First, we used “Friday” as a reference category, and created d1, d2, d3, d4 dummy variables, where b1, b2, b3, b3 are coefficients on Monday, Tuesday, Wednesday, and Thursday. The estimated regression equation is:

Rides = 1938.6418 + 825.6473*d1 + 841.4978*d2 + 693.8099*d3 + 553.9681*d4 +11.6879*t

The coefficients for dummy variables indicate that, relative to Friday, the numbers of rides completed are about 826, 841, 694, and 554 higher on Monday, Tuesday, Wednesday, and Thursday. The estimated coefficient for the trend variable suggests that the rides’ number increases by about 12 every weekday.

<2> Multiple R-squared: 0.929

This R2 suggests that about 92.9% of the sample variations in rides are explained by the linear trend model with day-of-week dummy variables.

<3> Adjusted R-squared: 0.9229

In question5, the adjusted R-squared is 0.2914, which is way smaller than 0.9229 of question6’s. The big difference between two adjusted R2 indicates that the linear trend with day-of-week dummy variables is more preferable compared with a single linear trend model.

Question7-(a):

<1> Forecasts based on the question6's estimated model are made as follows:

Monday rides = 1938.6418 + 825.6473*d1 + 11.6879*t = 2764.289 + 11.6879*t
Tuesday rides = 1938.6418 + 841.4978*d2 + 11.6879*t = 2780.14 + 11.6879*t
Wednesday rides = 1938.6418 + 693.8099*d3 + 11.6879*t = 2632.452 +11.6879*t
Thursday rides = 1938.6418 + 553.9681*d4 +11.6879*t = 2492.61 + 11.6879*t
Friday rides = 1938.6418 +11.6879*t

Based on the equation above, we calculate the estimated value of rides from day 1 to day 64, and **results are shown in R**.

<2> Calculate MSE, MAD, and MAPE for this forecast

	MSE	MAD	MAPE
Question4	111581.38	270.55	9.80%
Question6	10409.92	77.30	2.76%

By comparing to question4's MSE, MAD, and MAPE, the new regression model has lower MSE, MAD, and MAPE. Thus, the new regression model from question6 performs better.

Question7-(b):

Time	Date	Weekday	Dummy Variable				Prediction of rides
65	01/04/2016	Friday	0	0	0	0	2698.3553
66	02/04/2016	Monday	1	0	0	0	3535.6904
67	03/04/2016	Tuesday	0	1	0	0	3563.2293
68	04/04/2016	Wednesday	0	0	1	0	3427.2292
69	05/04/2016	Thursday	0	0	0	1	3299.0751
70	06/04/2016	Friday	0	0	0	0	2756.7948
71	07/04/2016	Monday	1	0	0	0	3594.1299
72	08/04/2016	Tuesday	0	1	0	0	3621.6688
73	09/04/2016	Wednesday	0	0	1	0	3485.6687
74	10/04/2016	Thursday	0	0	0	1	3357.5146
75	11/04/2016	Friday	0	0	0	0	2815.2343
76	12/04/2016	Monday	1	0	0	0	3652.5694
77	13/04/2016	Tuesday	0	1	0	0	3680.1083
78	14/04/2016	Wednesday	0	0	1	0	3544.1082
79	15/04/2016	Thursday	0	0	0	1	3415.9541
80	16/04/2016	Friday	0	0	0	0	2873.6738
81	17/04/2016	Monday	1	0	0	0	3711.0089
82	18/04/2016	Tuesday	0	1	0	0	3738.5478
83	19/04/2016	Wednesday	0	0	1	0	3602.5477
84	20/04/2016	Thursday	0	0	0	1	3474.3936
85	21/04/2016	Friday	0	0	0	0	2932.1133
86	22/04/2016	Monday	1	0	0	0	3769.4484
87	23/04/2016	Tuesday	0	1	0	0	3796.9873
88	24/04/2016	Wednesday	0	0	1	0	3660.9872
89	25/04/2016	Thursday	0	0	0	1	3532.8331
90	26/04/2016	Friday	0	0	0	0	2990.5528
91	27/04/2016	Monday	1	0	0	0	3827.8879
92	28/04/2016	Tuesday	0	1	0	0	3855.4268
93	29/04/2016	Wednesday	0	0	1	0	3719.4267

We put the estimated regression equation (Monday to Friday) from the 7(a) to Excel, then got the full April forecasts data.

Question8

In general, by drawing the scatterplot between the number of completed rides and time, we could see that the overall trend is upward, but it was not a pure linear trend model because of seasonality where 5 weekdays repeated themself with four days going up and the last day going down. Then we respectively build two models: a linear trend model for the ‘rides’ variable and a linear trend model with day-of-week dummy variables for the ‘rides’ variable. By calculating their MAD, MAPE, MSE, we found the linear trend model with seasonality has smaller ratios than the linear trend model, which means that the second model has smaller forecast errors. Besides, the adjusted R-squared of the second model tends to be higher than the first one, which means that the second model is able to explain more of the sample variation in rides variable. In addition, its R-squared is high up to 92.9%. In the end, combining the above analysis, we could draw the conclusion that the linear trend model with day-of-week dummy variables for the ‘rides’ variable is well suited in prediction.

Thus, we highly recommend CVE managers to utilize the second model with seasonality to forecast the number of rides in the short period of time due to its reliable characteristics above. When we do the forecast for April, we find that the predicted rides are corresponding with our expectations. More importantly, due to the seasonality of ridership data, CVE managers cannot directly use simple linear regression models to predict other variables, such as revenue. They, instead, need to consider the 5 days cycling pattern into other predictive models so as to improve accuracy of prediction.