BIOSTAT 200B Project 1 Report by Xinyang Li

- **Introduction**

The main purpose of this project is to analyze the relationship between day of week, mode of transportation, departure time, presence of rain, and the average speed during a commute for students from home to UCLA campus. By setting up simple and multiple linear regression models, I found that there are some underlying connections between those factors and average speed (miles/hour).

- **Methods**

The observations are collected from 60 UCLA students in both 2019 and 2020 by completing online google sheets. The sample data includes students ID, date (MM/DD/YYYY), minutes from home to school, distance, mode of transportation, departure time (HH:MM), day of week, and presence of rain (Y/N). From the data, the average speed (miles/hour) is easy to be calculated and the relationship can be analyzed. In convenience, I labeled different methods of transportation by numbers 0 – 6, which represent by walk (0), by car (1), by bus (2), ride share (3), by bike (4), Uber or Lyft (5), and other (6), respectively. Also, I numbered day of week by 1 – 5 through Monday to Friday. Then, I divided the departure time into intervals and labeled by number 0 – 7. (0: Depart before 7am, 1: between 7am and 8am, 2: between 8am and 9am, 3: between 9am and 10am, 4: between 10am and 11am, 5: between 11am and 12pm, 6: between 12pm and 1pm, 7: after 1pm).

Before setting up regression models and do analysis, I first plotted univariate distributions of variables to get critical statistical data (e.g., mean and percentile) from the variables. Since we want to find variables affecting the value of average speed, the distribution of average speed (Figure1.1) is necessary. From the histogram I got, most students had travel speed from 1 mile/hour to 15 miles/hour. And it is obvious to see the existence of an extreme point of 45 miles/hour. From the distribution of day of week (Figure 1.2), we can see the sample observations are kind of equally distributed on Monday to Friday. From the distribution of presence of rain (Figure1.3), I observed that nearly all students contributed data with the absence of rain. This might cause the analysis result to be inaccurate because the sample size is too small. Then, look at the distribution of different departure time intervals (Figure1.4), we observed that no students departed between 11am and 12pm, and most students departed between 10am and 11am. Finally, the distribution of different methods of transportation (Figure1.5) showed that most students went to campus by walk, and many of the other were by car or by bus.

After basically getting to know these variables, I tried to scatterplot and fit loess curves of average speed versus the four predictors, separately. Since the four predictors are all categorical variables, the loess curves connected the mean differences of average speed on different categories (Figure3.1 – 3.4). Further, I produced the correlation matrix tables of between average speed (Y) and dummy variables for each predictor. From the correlation matrices, the level of correlation between a dummy variable (one category in a predictor) and the average speed is not critical, since the effect of a single category on average speed (Y) cannot illustrate the effect of the entire predictor.

In the next step, I modeled four simple linear regressions of average speed (miles/hour) on day of the week, mode of transportation, departure time, and presence of rain, individually. In the regression model of average speed on day of the week, I selected Monday as the reference group. And w1 = 1 for Tuesday, w2 = 1 for Wednesday, w3 = 1 for Thursday, and w4 = 1 for Friday. From the result (Figure4.1.1), we can see that the F-value = 1.42, which is very small. Also the p-value of β's are greater than 0.05, and the model assumptions are not satisfied (no constant variance and not normal). Therefore I considered to do model transformation. Since the regression contains only dummy variables as predictors, I have no way to do transformation on x. Therefore I transformed y to both y^2 and log2(y) (Figure4.1.2). After running regression model, I found log2(y) achieved better model assumptions but the ANOVA table and parameter estimates are still not good. In conclusion, I picked the regression of log2(average speed) on day of week. Followed the same procedure, in the regression model of average speed on mode of transportation, I chose walk as the reference group and finally picked the regression with average speed transformed by taking log base 2. From my observation, in the fit diagnostic plots (Figure4.2.2), transforming to log2(average speed) can make the model satisfied the normality and linearity assumptions better, and the p-values of the regression coefficients are still significant. Then, in the model regressing on departure time, I chose

the time interval before 7am as the reference and considered the original regression model (see Figure4.3.1) as the best one because this model decently satisfied the model assumptions and had significant F-value and p-values. In the fourth model, regressing average speed on rain. This model (Figure4.4.1) barely satisfied the assumptions, and no critical values found, no matter being transformed or not. Finally, I multiply regressed average speed on all the four categorical variables. From the fit diagnostic graphs (Figure5.1), I could see that the model assumptions are basically satisfied. From the ANOVA table and parameter estimates table (Figure5.2), the F-value = 13.34, which is critical; and some of the regression coefficients contributed a lot to explain variation the linearity of average speed. By taking log on average speed and doing regression, I did not see very big improvement on both model assumptions and p-values (see Figure5.3). So I chose not to transform.

- **Results**

In the simple regression model on day of week, the ANOVA table (Figure4.1.1) showed F = 1.42, which is very small. Therefore the null hypothesis is accepted, showing that all the regression coefficients are equal to zero. Also, all p-values in the parameter estimates are > 0.05, which means there are no mean difference in average speed between Monday (reference group) and other days of week. Next, Figure4.1.3 showed the result of hypothesis testing if all regression coefficients are equal. Since p-value = 0.1470, it indicated different days of week will not affect the average speed.

In the simple regression model of log2(average speed) on transportation, the ANOVA table (Figure4.2.2) gave the p-value < 0.0001, which means the null hypothesis is rejected, and mean difference of average speed is detected. From the parameter estimates, all p-values are less than 0.05 except for t6. Since the reference group is by walk, it showed commuting by car, by bus, by riding share, by bike, by Uber or Lyft, are all have major difference compared to commuting by feet. It fitted our intuition because walking should be the slowest way in all categories. Also, t6 has no critical difference in mean average speed as walking, since this group contains commuting by "scooter" and "ride share and bus". In addition, I tested the mean difference among all regression coefficients (Figure4.2.3). Since p-value = 0.0049, the null hypothesis is rejected, showing the detection of mean difference of average speed among mode of transportation.

In the simple regression model on departure time, after testing for comparing equality for all pairs of coefficients, I found that there is mean speed difference between 11am-12pm and after 1pm; between 10am-11am and after 1pm; between 9am-10am and after 1am; between 8am-9am and 11-12pm; between 8am-9am and 10am-11am; between 8am-9am and 9am-10am (Figure4.3.3-4.3.9).

In the simple regression model on rain, based on the ANOVA table, F = 0.13, which is super small, and the parameter estimates table also showed no relationship between whether rain or not and travel speed.

In the multiple regression model, the ANOVA table (Figure5.2) showed that at least four of these predictors will affect the average commute speed to campus. By comparing the p-values in the parameter estimates table, it is clearly shown that mode of transportation and departure time will significantly affect students' commute speed.

- **Conclusion**

From those regression models, I concluded that departure time and mode of transportation are the two factors that will significantly affect the average travelling speed, while the day of week and presence of rain have slightly influence on the fluctuation of average speed, or we don't have enough evidence to show the connections between these two factors and average commute speed for now.

- **Limitations**

The result can to some extent show the potential factors of average speed from the door of house to UCLA campus, but many underlying relationships may be ignored because of the limitation of sample size. For example, when I regressed average speed on the presence of rain, there are only two samples of "Rain", and the remaining contains 57 of "No Rain" and one missing value. Since the sample size of "Rain" is too small, the model may not be able to represent the true conditions. Also, from the departure time data, we can see there is only one student who departed before 7AM, which cause the average speed to be 45miles/hr., and dramatically affected the loess curve and the regression model on departure time. Furthermore, students who filled out this survey may be limited in the public health department, so their destination might be the same. Hence, similar travel routes may undermine many possible factors.

- **Tables And Graphs**
- **Useful Univariate Distributions**

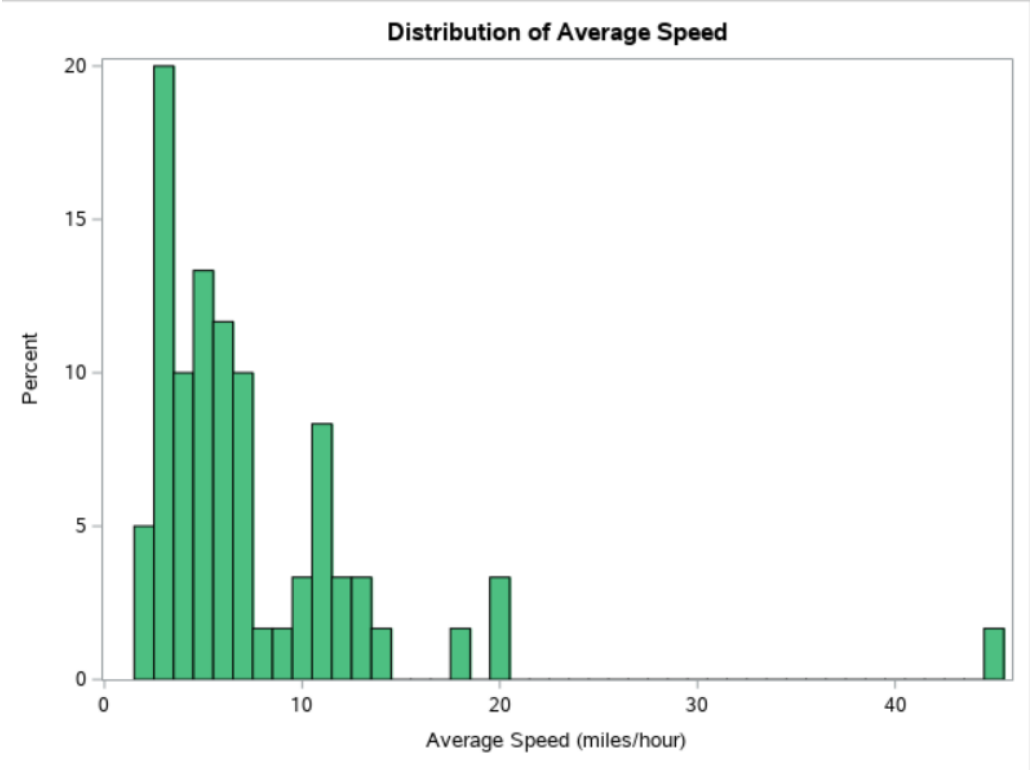


Figure1.1 Distribution of Average Speed

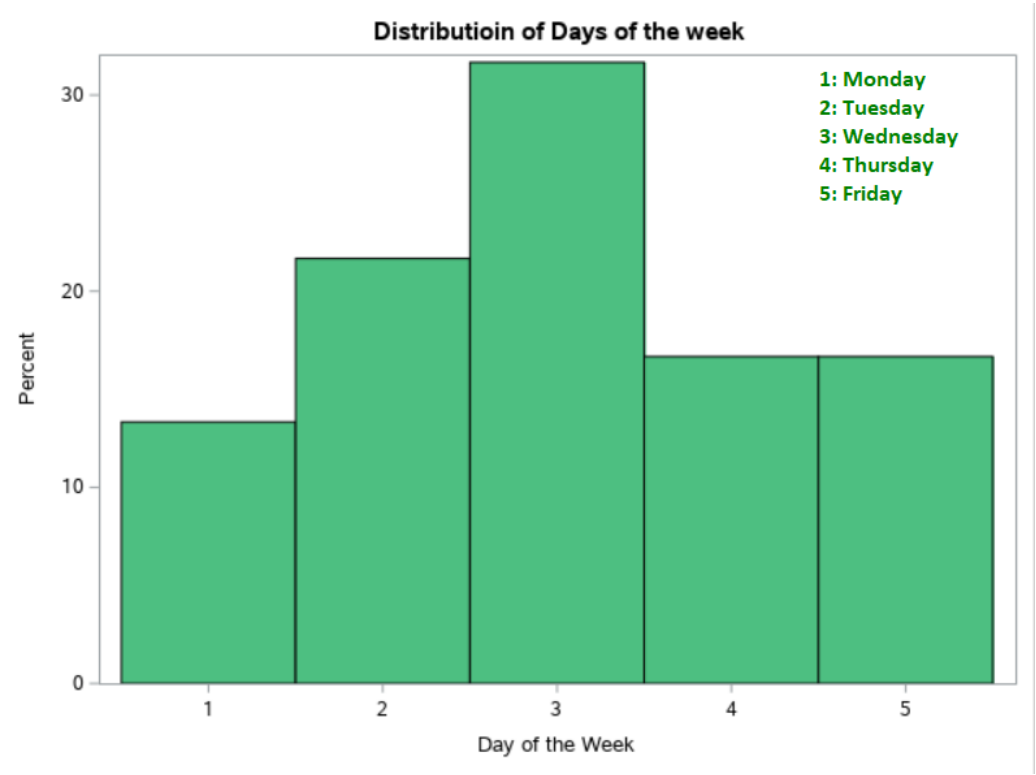


Figure1.2 Distribution of Days of the Week

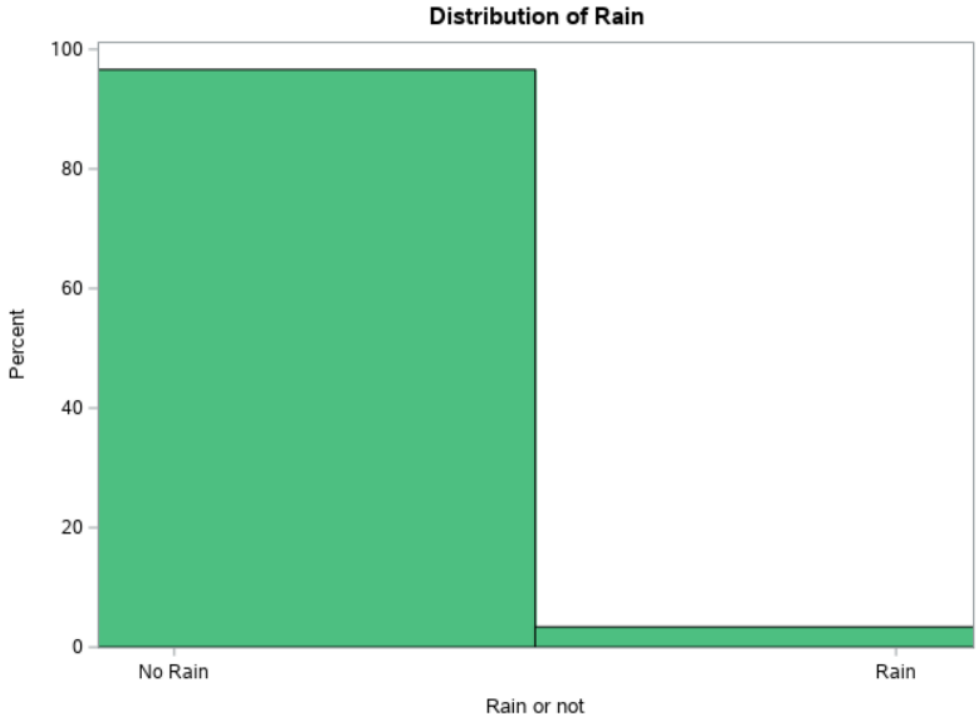


Figure1.3 Distribution of Rain

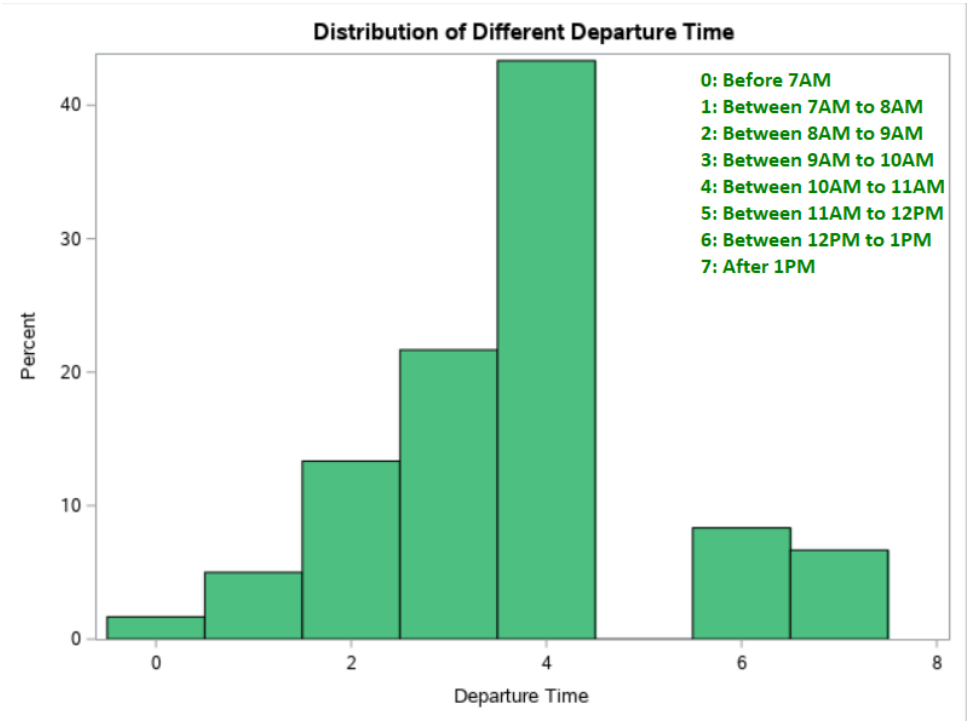


Figure1.4 Distribution of Different Departure Time

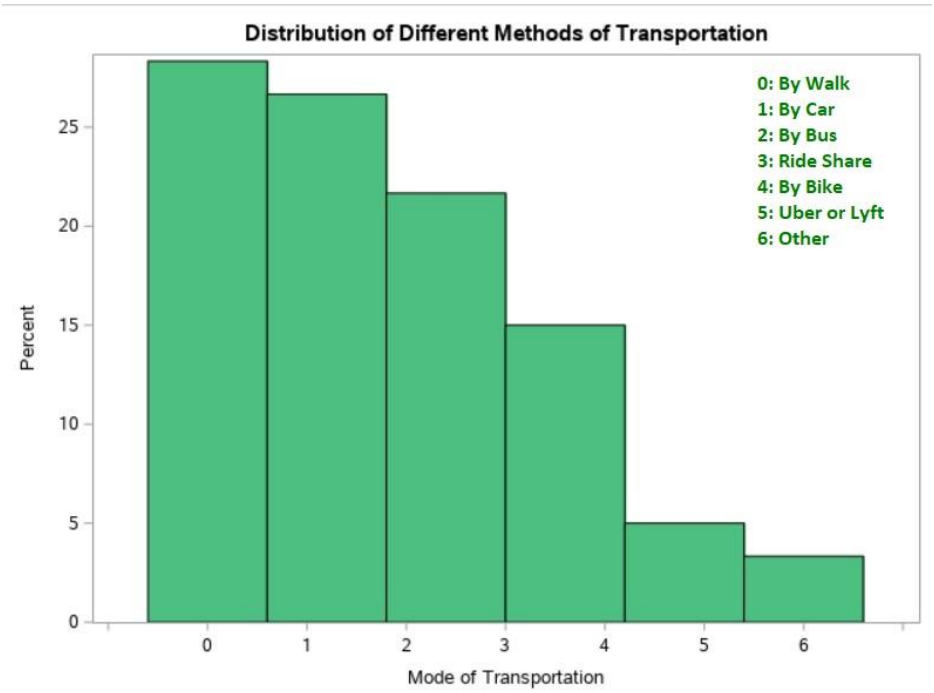


Figure1.5 Distribution of Different Methods of Transportation

---

- **Bivariate Distributions**

Pearson Correlation Coefficients, N = 60
Prob > |r| under H0: Rho=0

	Ave_Speed	w1	w2	w3	w4
Ave_Speed	1.00000	0.29451 0.0224	-0.06785 0.6065	-0.14825 0.2583	-0.05794 0.6601
w1	0.29451 0.0224	1.00000	-0.35802 0.0050	-0.23520 0.0705	-0.23520 0.0705
w2	-0.06785 0.6065	-0.35802 0.0050	1.00000	-0.30444 0.0180	-0.30444 0.0180
w3	-0.14825 0.2583	-0.23520 0.0705	-0.30444 0.0180	1.00000	-0.20000 0.1255
w4	-0.05794 0.6601	-0.23520 0.0705	-0.30444 0.0180	-0.20000 0.1255	1.00000

Pearson Correlation Coefficients, N = 60
Prob > |r| under H0: Rho=0

	Ave_Speed	d2	d3	d4	d5	d7	d8
Ave_Speed	1.00000	0.12847 0.3280	-0.13126 0.3175	-0.17180 0.1893	-0.12348 0.3472	0.23070 0.0762	-0.04516 0.7319
d2	0.12847 0.3280	1.00000	-0.08998 0.4941	-0.12066 0.3585	-0.20062 0.1243	-0.06917 0.5995	-0.06131 0.6417
d3	-0.13126 0.3175	-0.08998 0.4941	1.00000	-0.20628 0.1138	-0.34300 0.0073	-0.11826 0.3681	-0.10483 0.4254
d4	-0.17180 0.1893	-0.12066 0.3585	-0.20628 0.1138	1.00000	-0.45991 0.0002	-0.15857 0.2262	-0.14056 0.2841
d5	-0.12348 0.3472	-0.20062 0.1243	-0.34300 0.0073	-0.45991 0.0002	1.00000	-0.26366 0.0418	-0.23371 0.0723
d7	0.23070 0.0762	-0.06917 0.5995	-0.11826 0.3681	-0.15857 0.2262	-0.26366 0.0418	1.00000	-0.08058 0.5405
d8	-0.04516 0.7319	-0.06131 0.6417	-0.10483 0.4254	-0.14056 0.2841	-0.23371 0.0723	-0.08058 0.5405	1.00000

Pearson Correlation Coefficients, N = 60
Prob > |r| under H0: Rho=0

	Ave_Speed	t1	t2	t3	t4	t5	t6
Ave_Speed	1.00000	0.47682 0.0001	-0.13294 0.3113	0.04886 0.7109	-0.00567 0.9657	0.16718 0.2017	-0.02643 0.8412
t1	0.47682 0.0001	1.00000	-0.31714 0.0135	-0.18182 0.1644	-0.16116 0.2186	-0.13834 0.2918	-0.11198 0.3943
t2	-0.13294 0.3113	-0.31714 0.0135	1.00000	-0.15857 0.2262	-0.14056 0.2841	-0.12066 0.3585	-0.09766 0.4579
t3	0.04886 0.7109	-0.18182 0.1644	-0.15857 0.2262	1.00000	-0.08058 0.5405	-0.06917 0.5995	-0.05599 0.6709
t4	-0.00567 0.9657	-0.16116 0.2186	-0.14056 0.2841	-0.08058 0.5405	1.00000	-0.06131 0.6417	-0.04963 0.7065
t5	0.16718 0.2017	-0.13834 0.2918	-0.12066 0.3585	-0.06917 0.5995	-0.06131 0.6417	1.00000	-0.04260 0.7465
t6	-0.02643 0.8412	-0.11198 0.3943	-0.09766 0.4579	-0.05599 0.6709	-0.04963 0.7065	-0.04260 0.7465	1.00000

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Ave_Speed	r
Ave_Speed	1.00000 60	0.04749 0.7210 59
r	0.04749 0.7210 59	1.00000 59

Table 2.1 Correlation Matrices Between Day of Week and Average Speed, Departure Time and Average Speed, Transportation and Average Speed, Rain and Average Speed.

- **Loess Curves Fitting**

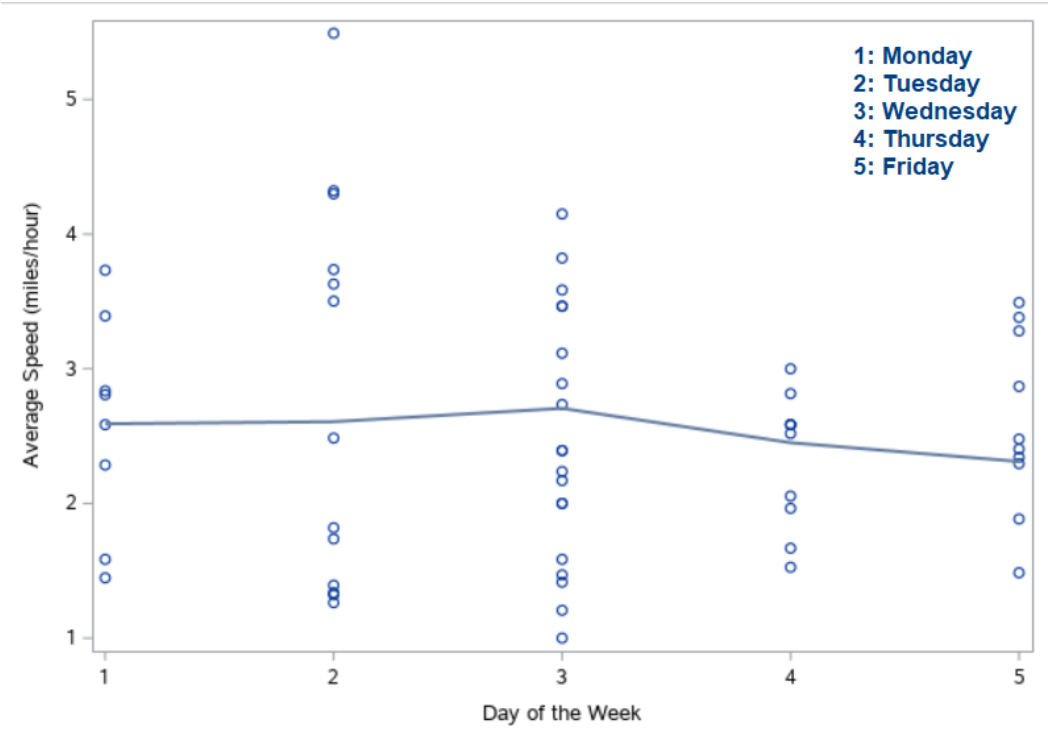


Figure3.1 Loess Curve of Average Speed and Day of the week

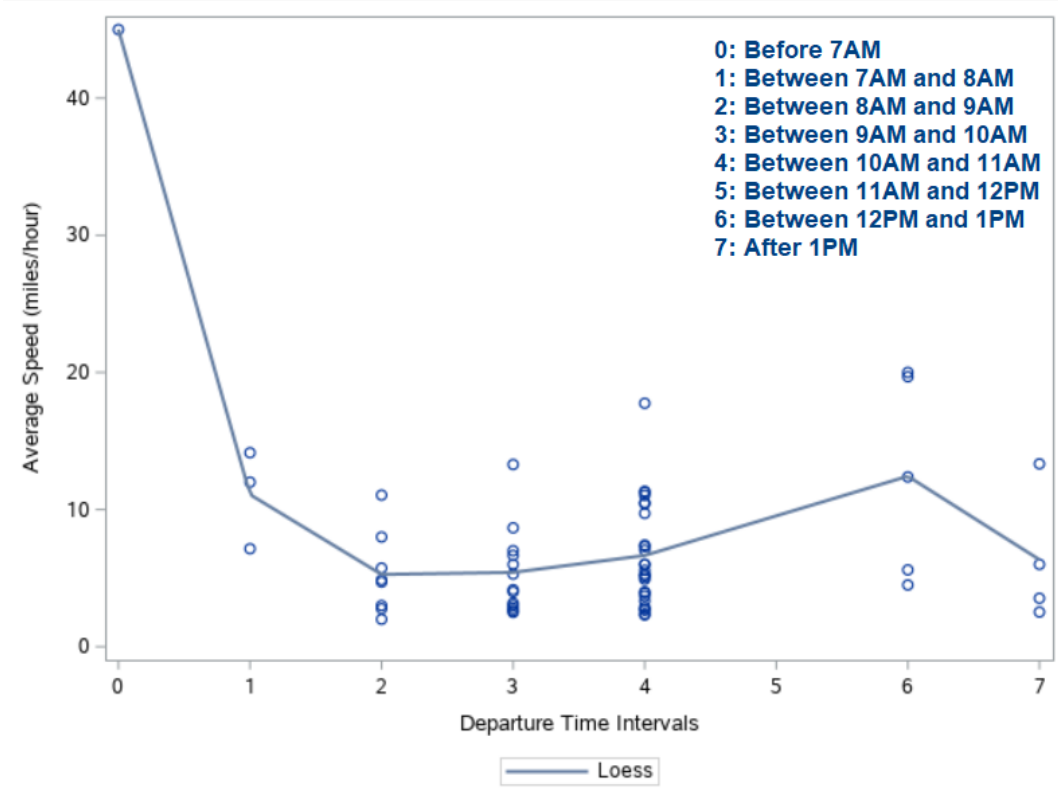


Figure 3.2 Loess Curve of Average Speed and Departure Time

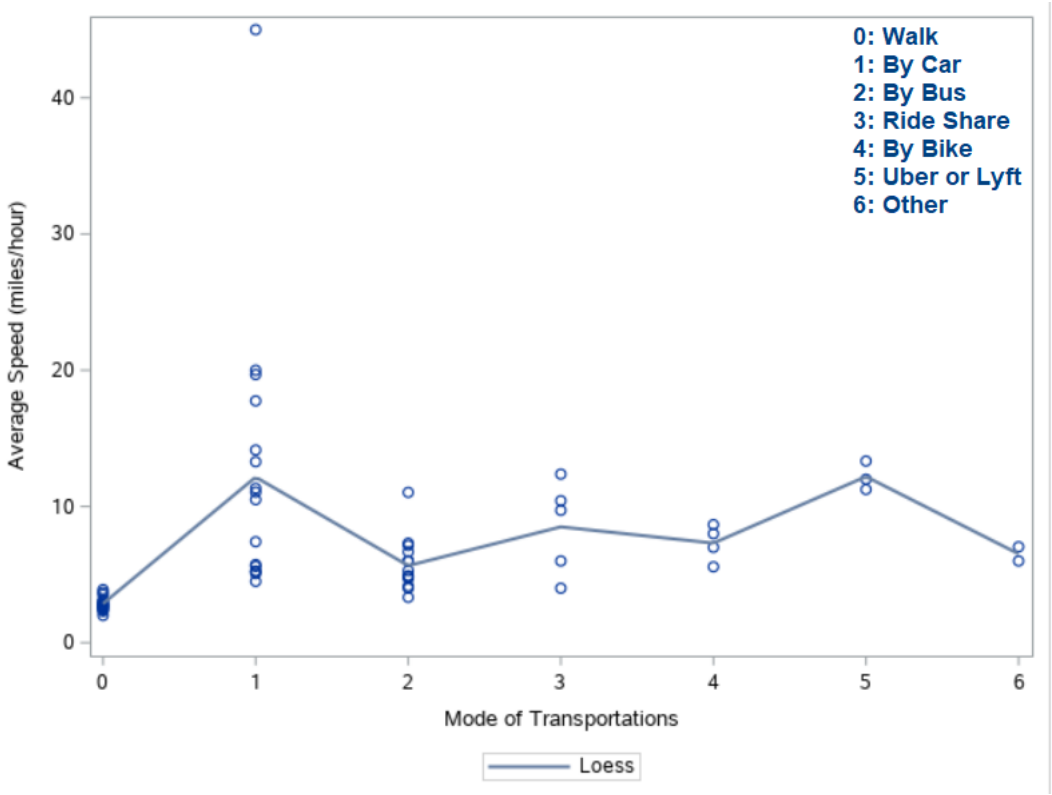


Figure 3.3 Loess Curve of Average Speed and Transportation

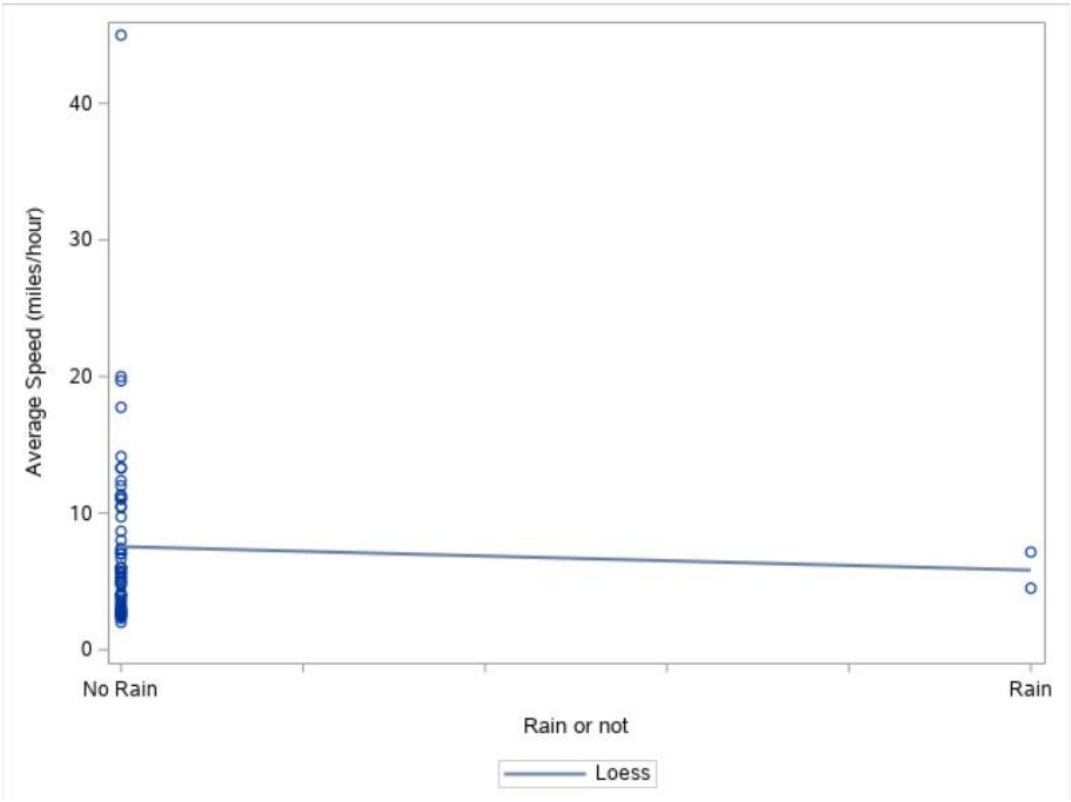
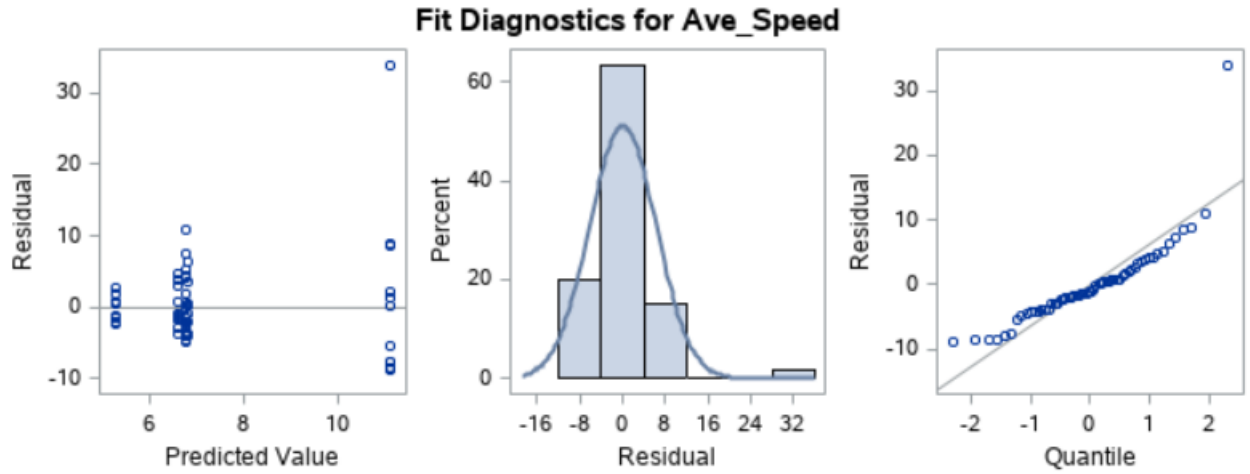


Figure 3.4 Loess Curve of Average Speed and Rain

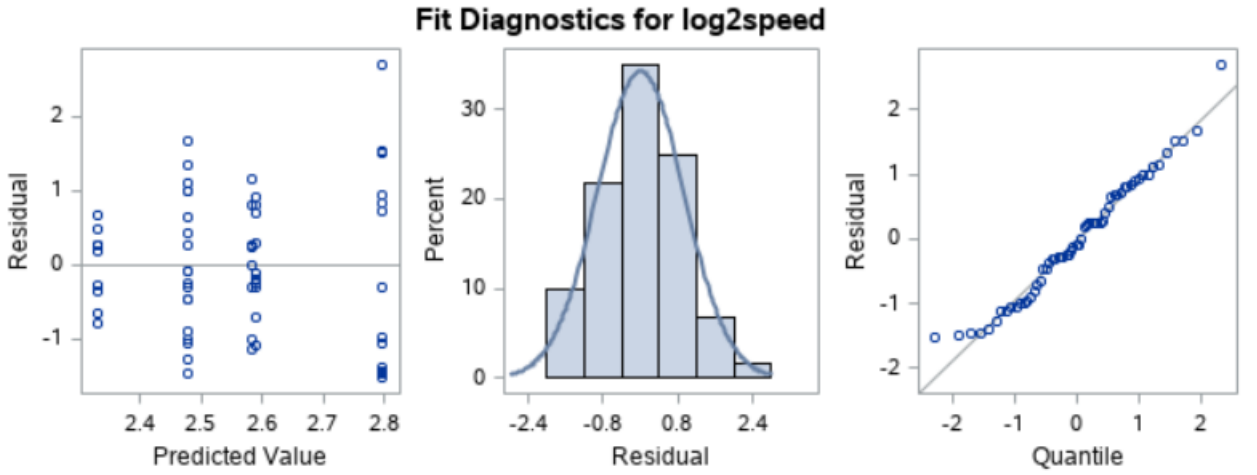
- **Simple Linear Regressions**
  1. **Regression of Average Speed on Day of the Week:**



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	237.89734	59.47434	1.42	0.2404
Error	55	2307.66844	41.95761		
Corrected Total	59	2545.56578			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.81636	2.29013	2.98	0.0043
w1	1	4.27893	2.91071	1.47	0.1472
w2	1	-0.01780	2.73002	-0.01	0.9948
w3	1	-1.52781	3.07253	-0.50	0.6210
w4	1	-0.21243	3.07253	-0.07	0.9451

Figure4.1.1



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.38624	0.34656	0.37	0.8269
Error	55	51.09632	0.92902		
Corrected Total	59	52.48255			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.58385	0.34078	7.58	<.0001
w1	1	0.21136	0.43312	0.49	0.6275
w2	1	-0.10518	0.40623	-0.26	0.7967
w3	1	-0.25368	0.45720	-0.55	0.5812
w4	1	0.00763	0.45720	0.02	0.9867

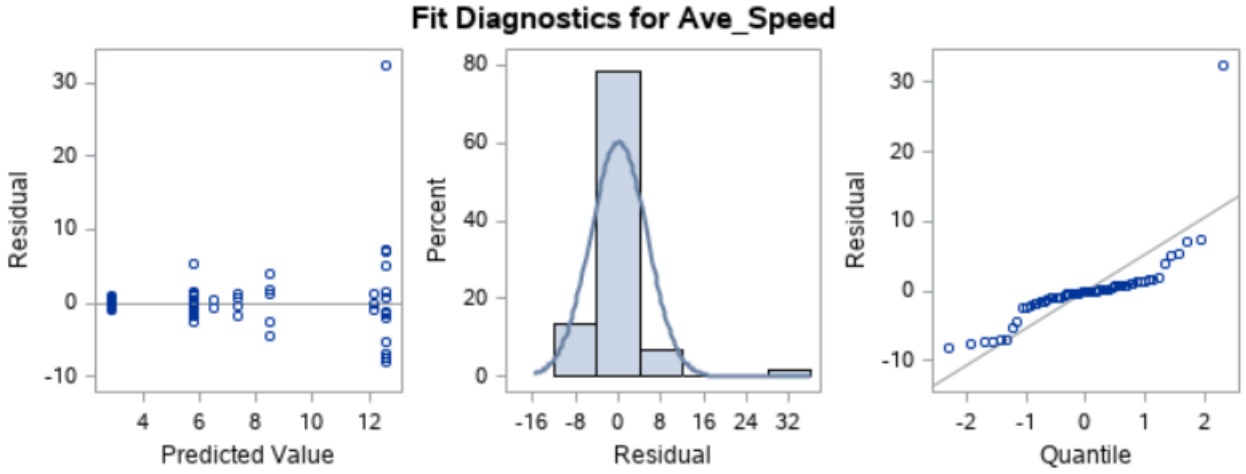
Figure4.1.2

Test 1 Results for Dependent Variable Ave_Speed				
Source	DF	Mean Square	F Value	Pr > F
Numerator	3	78.07236	1.86	0.1470
Denominator	55	41.95761		

Figure4.1.3 Hypothesis Test Result

---

2. Regression of Average Speed on Mode of Transportation:

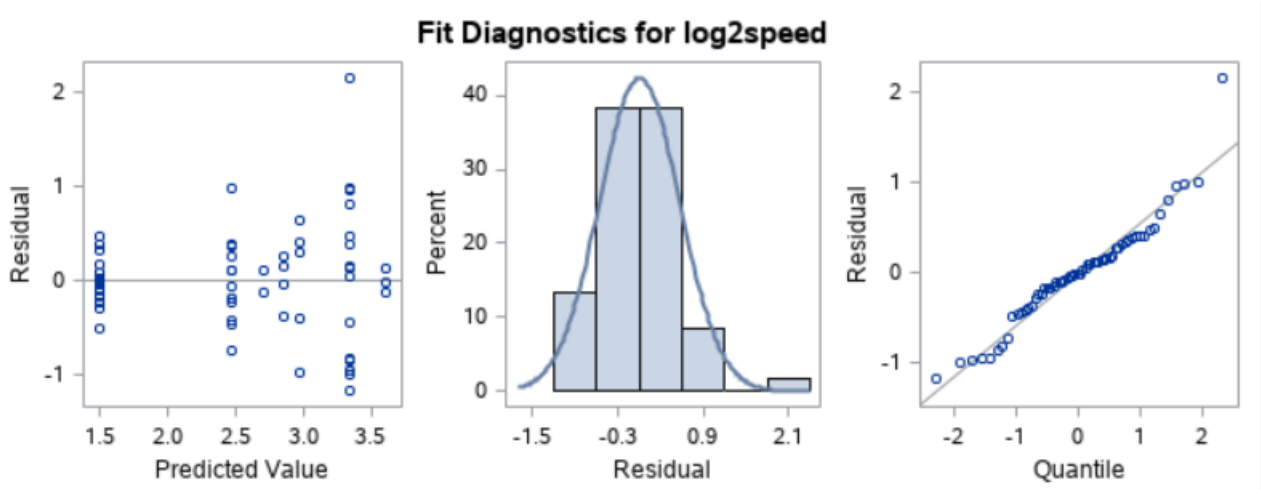


Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	893.51740	148.91957	4.78	0.0006
Error	53	1652.04838	31.17072		
Corrected Total	59	2545.56578			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.85296	1.35409	2.11	0.0399
t1	1	9.74513	1.94467	5.01	<.0001
t2	1	2.94840	2.05702	1.43	0.1576
t3	1	5.65025	2.84037	1.99	0.0518
t4	1	4.45656	3.10262	1.44	0.1568
t5	1	9.34148	3.49626	2.67	0.0100
t6	1	3.66792	4.17360	0.88	0.3835

Figure4.2.1

Taking log base 2 on Average Speed:



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	33.62660	5.60443	15.75	<.0001
Error	53	18.85596	0.35577		
Corrected Total	59	52.48255			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.49230	0.14466	10.32	<.0001
t1	1	1.84336	0.20776	8.87	<.0001
t2	1	0.97599	0.21976	4.44	<.0001
t3	1	1.48305	0.30345	4.89	<.0001
t4	1	1.35792	0.33147	4.10	0.0001
t5	1	2.11230	0.37352	5.66	<.0001
t6	1	1.20815	0.44589	2.71	0.0091

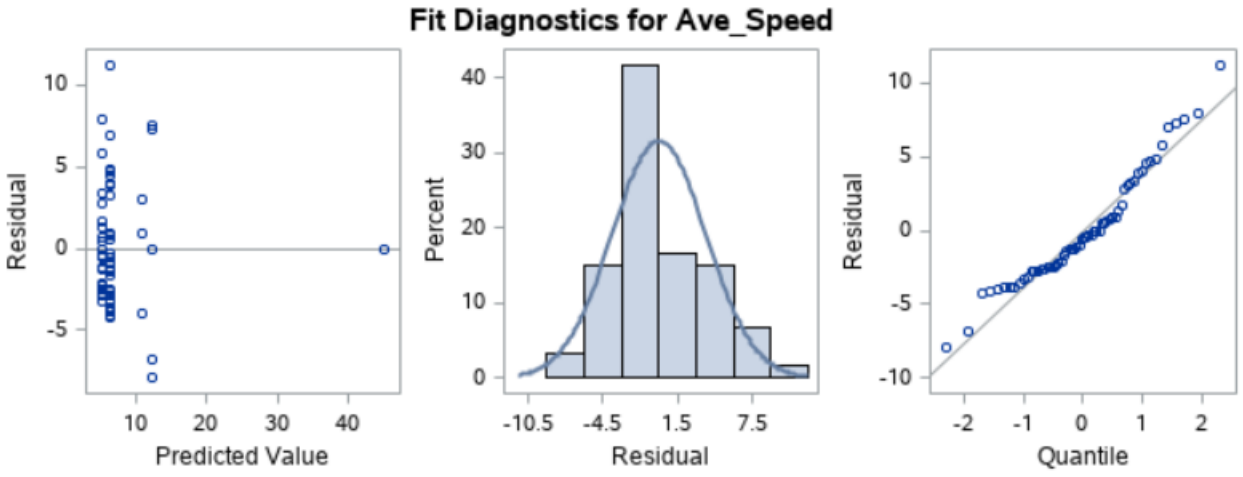
Figure4.2.2 (Including the model assumption tables in previous page)

Test 1 Results for Dependent Variable log2speed				
Source	DF	Mean Square	F Value	Pr > F
Numerator	5	1.36424	3.83	0.0049
Denominator	53	0.35577		

Figure4.2.3 Hypothesis Test Result

---

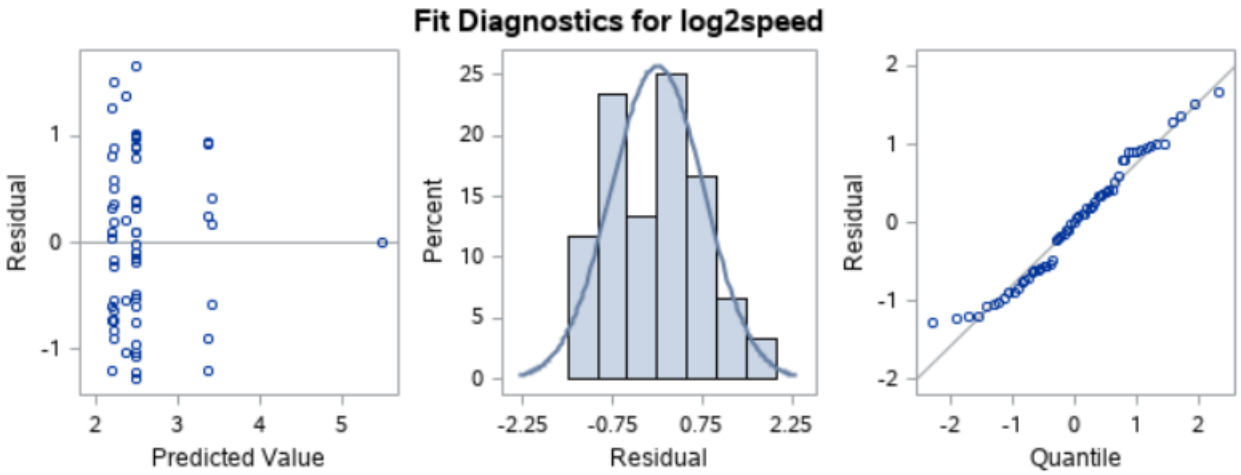
3. Regression of Average Speed on Departure Time:



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1697.97555	282.99593	17.70	<.0001
Error	53	847.59023	15.99227		
Corrected Total	59	2545.56578			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	45.00000	3.99903	11.25	<.0001
d2	1	-33.90476	4.61769	-7.34	<.0001
d3	1	-39.73194	4.24162	-9.37	<.0001
d4	1	-39.67997	4.14999	-9.56	<.0001
d5	1	-38.47195	4.07521	-9.44	<.0001
d7	1	-32.56841	4.38072	-7.43	<.0001
d8	1	-38.65273	4.47106	-8.65	<.0001

Figure4.3.1



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	16.97597	2.82933	4.22	0.0015
Error	53	35.50658	0.66994		
Corrected Total	59	52.48255			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.49185	0.81850	6.71	<.0001
d2	1	-2.07736	0.94512	-2.20	0.0323
d3	1	-3.29653	0.86815	-3.80	0.0004
d4	1	-3.26988	0.84939	-3.85	0.0003
d5	1	-3.00139	0.83409	-3.60	0.0007
d7	1	-2.11075	0.89662	-2.35	0.0223
d8	1	-3.12226	0.91511	-3.41	0.0012

x

Figure4.3.2 Taking log on average speed

Test 1 Results for Dependent Variable log2speed				
Source	DF	Mean Square	F Value	Pr > F
Numerator	5	1.64132	2.45	0.0453
Denominator	53	0.66994		

Figure4.3.3 Test all equals

Test 14 Results for Dependent Variable log2speed				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	3.32649	4.97	0.0301
Denominator	53	0.66994		

Figure4.3.4 test beta5 = beta7

Test 12 Results for Dependent Variable log2speed				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	4.85178	7.24	0.0095
Denominator	53	0.66994		

Figure4.3.5 Test beta4 = beta 7

Test 9 Results for Dependent Variable log2speed				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	4.32638	6.46	0.0140
Denominator	53	0.66994		

Figure4.3.6 Test beta3 = beta7

Test 4 Results for Dependent Variable log2speed				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	2.29650	3.43	0.0697
Denominator	53	0.66994		

Figure4.3.7 Test beta2 = beta5

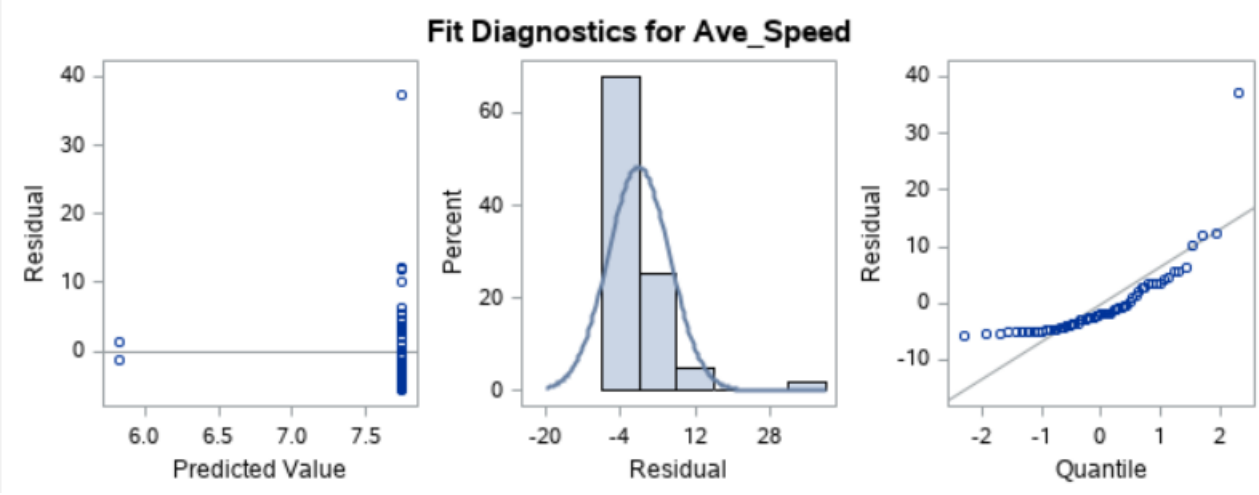
Test 3 Results for Dependent Variable log2speed				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	3.46633	5.17	0.0270
Denominator	53	0.66994		

Figure4.3.8 Test beta2 = beta4

Test 3 Results for Dependent Variable log2speed				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	3.24298	4.84	0.0322
Denominator	53	0.66994		

Figure4.3.9 (Left) Test beta2 = beta3

4. Regression of Average Speed on Rain:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.72935	5.72935	0.13	0.7210
Error	57	2534.92432	44.47236		
Corrected Total	58	2540.65368			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.82143	4.71553	1.23	0.2221
r	1	1.72197	4.79754	0.36	0.7210

Figure4.4.1

- **Multiple Linear Regression**

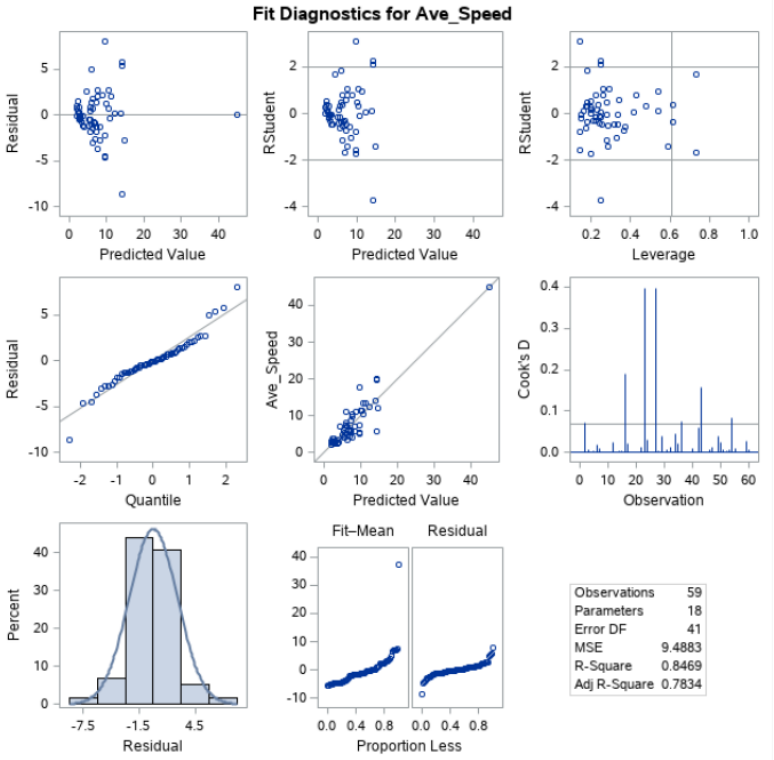
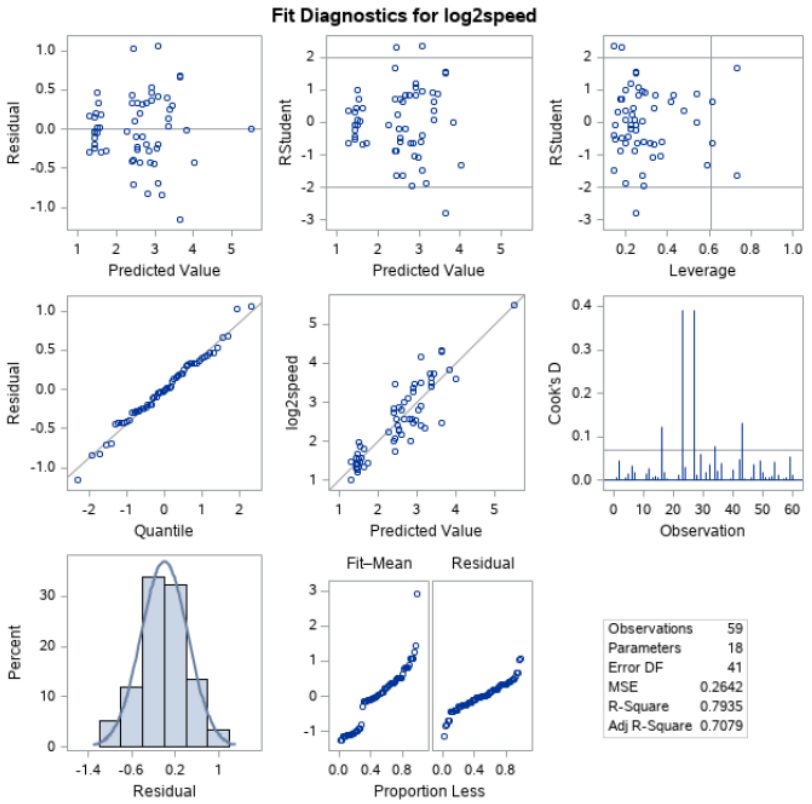


Figure5.1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	2151.63245	126.56661	13.34	<.0001
Error	41	389.02122	9.48832		
Corrected Total	58	2540.65368			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	32.10000	4.35530	7.37	<.0001
t1	1	6.83481	1.27972	5.34	<.0001
t2	1	3.14353	1.19025	2.64	0.0116
t3	1	4.82828	1.71926	2.81	0.0076
t4	1	4.60727	1.77631	2.59	0.0131
t5	1	7.65576	2.18023	3.51	0.0011
t6	1	3.86773	2.74006	1.41	0.1656
d2	1	-30.70047	4.00447	-7.67	<.0001
d3	1	-35.79455	3.71759	-9.63	<.0001
d4	1	-35.18381	3.61275	-9.74	<.0001
d5	1	-34.96473	3.48086	-10.04	<.0001
d7	1	-30.73384	3.43927	-8.94	<.0001
d8	1	-34.50443	3.70821	-9.30	<.0001
w1	1	-0.76250	1.84516	-0.41	0.6816
w2	1	-1.10117	1.35205	-0.81	0.4201
w3	1	-1.53996	1.60503	-0.96	0.3430
w4	1	-1.13372	1.62592	-0.70	0.4896
r	1	6.82769	2.79681	2.44	0.0190

Figure5.2 (Include two tables)



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.09134	0.72674	4.25	0.0001
t1	1	1.62947	0.21354	7.63	<.0001
t2	1	0.97761	0.19861	4.92	<.0001
t3	1	1.35924	0.28688	4.74	<.0001
t4	1	1.34429	0.29640	4.54	<.0001
t5	1	1.81128	0.36380	4.98	<.0001
t6	1	1.11719	0.45722	2.44	0.0189
d2	1	-1.66366	0.66820	-2.49	0.0169
d3	1	-2.56464	0.62033	-4.13	0.0002
d4	1	-2.43176	0.60284	-4.03	0.0002
d5	1	-2.40269	0.58083	-4.14	0.0002
d7	1	-1.84820	0.57389	-3.22	0.0025
d8	1	-2.23481	0.61877	-3.61	0.0008
w1	1	-0.27147	0.30789	-0.88	0.3831
w2	1	-0.27148	0.22561	-1.20	0.2358
w3	1	-0.23185	0.26782	-0.87	0.3917
w4	1	-0.17750	0.27131	-0.65	0.5166
r	1	1.04251	0.46669	2.23	0.0310

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	41.62372	2.44845	9.27	<.0001
Error	41	10.83181	0.26419		
Corrected Total	58	52.45553			

Figure5.3 (Include three tables)