Travel Demand Analysis
Exercise 1

310702051 葉家榮

2022-03-05

This assignment is also uploaded in the following website, for detailed information and code.
https://chiajung-yeh.github.io/Travel-Demand-Analysis/linear-regression.html

Problem

Based on the provided survey data (南臺區域運輸規劃-高雄市), answer the following questions.
1. Summarize the trip distance variable using figures or/and tables. Briefly explain your result.
2. Linear regression (LR)
   - Develop LR models for identifying the effect of age on trip distance.
   - Develop LR models for predicting the average trip distance for each district of the Kaohsiung City.

Summarize the trip distance variable

The histogram and density plot of trip distance is shown in Figure 1. Note that the bin width is set at 1, which means that the frequency is counted by 1 kilometer interval of trip distance. And the density is calculated under Gaussian kernel density function in the figure. From the result, we can find out that most of the respondents have a trip distance in 1~2 kilometer(s) long. Also, the figure illustrates that the distribution of trip distance is right skewed, which indicates that the mean of trip distance (5.03 km) is higher than the median (2.73 km). This phenomenon is very common in most of the cities, that is, most people travel in a short distance, while there indeed occurs some extremely long travel distance due to the specific economic activities or personal needs.

A practical issue we should address is that over half of the travelers do not travel over 3 kilometers in a sub-trip, and it implies that the major trips in Kaohsiung is composed of short distance trip. In addition, to clearly understand the trip distance of each transport mode, the calculation is shown in Table 1.

Table 1: Trip Distance Summary (All and by mode)

	n	mean	sd	median	min	max	skew	kurtosis
所有運具	4156	5030.74	6731.46	2732.0	4	74795	3.54	18.72
步行	306	770.58	1738.79	397.5	4	23681	8.96	103.58
自行車	114	1454.87	1622.56	1058.5	94	12632	4.12	22.04
機車	3199	4274.67	4590.22	2706.0	42	49586	2.56	10.91
汽車	537	12721.53	12145.67	9343.0	313	74795	1.88	4.01

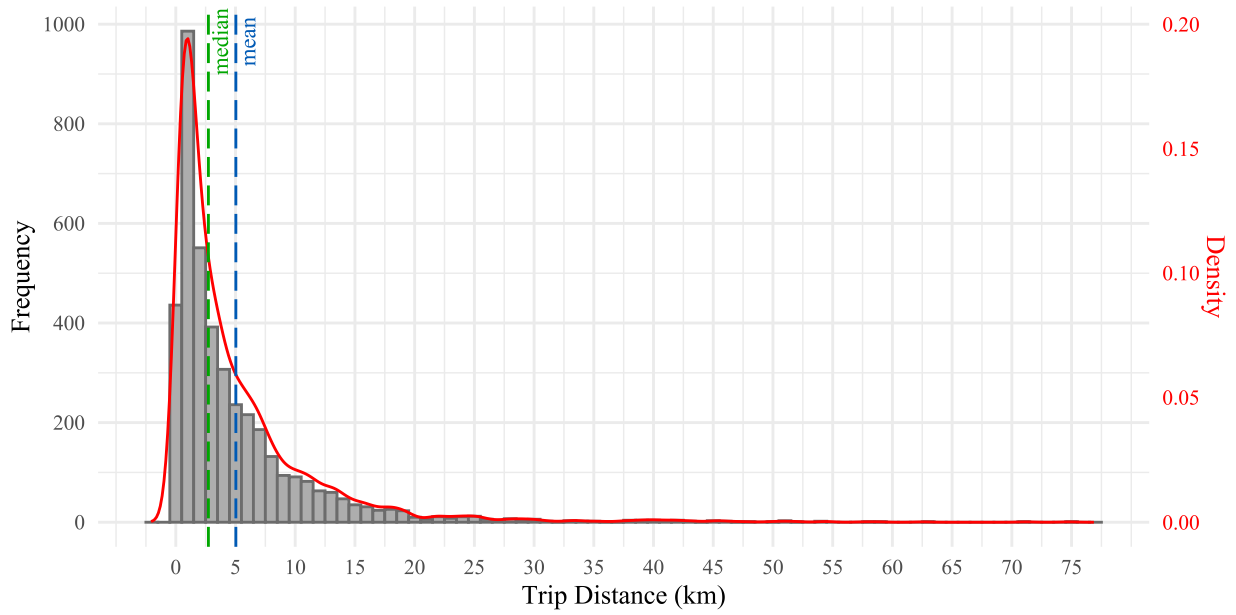

Figure 1: Histogram of trip distance

Here we can figure out that there are totally 4156 samples in the survey, and the major transport mode is scooter, which accounts for approximately 77%. The average travel distance of all modes is 5.03 kilometers, while the maximum distance can be up to 74.8 kilometers. As for different type of mode, we can find that the average distance of personal car is the highest among the four modes, which is 12.7 kilometers. And the second place is scooter, only 4.27 kilometers, much lower than the car is. The absolute value of skew and kurtosis of personal car is the lowest, indicating that the travel distance distribution of this mode is more likely to have no peaks and the value of median as well as mean is closer compared to other modes. In terms of walking, the standard deviation is low, which means that the travel distance of walking does not differ a lot. The skew and kurtosis of walking are the highest among all the modes, it again implies that most people walk only for a small range of distance.

The violin plot which is used to visualize the distribution of numerical data is shown in Figure 2. It can prove what we summarize above. The trip distance of the highest density for each mode is very short, especially for walking, bike and scooter. It again says that most people have a short trip for these modes in Kaohsiung.

Linear regression (the effect of age on trip distance)

Naive model

The naive regression model to evaluate the effect of age on trip distance is let the dependent variable be trip distance, while let the independent variable be ages. The model can be formulated as below.

$$TripDistance = \beta_{age} * x_{age}$$

The result of naive model and residual plots are shown in Table 2 and Figure 3 respectively. The coefficient of age is -23.025 (significant), which means the trip distance would approximately decrease 23 meters as the age increases by 1. From the figures, the mean of residuals are less than 0 for any given fitted values, indicating that the model may be biased. Also, the residual is apparently not under normal distribution.

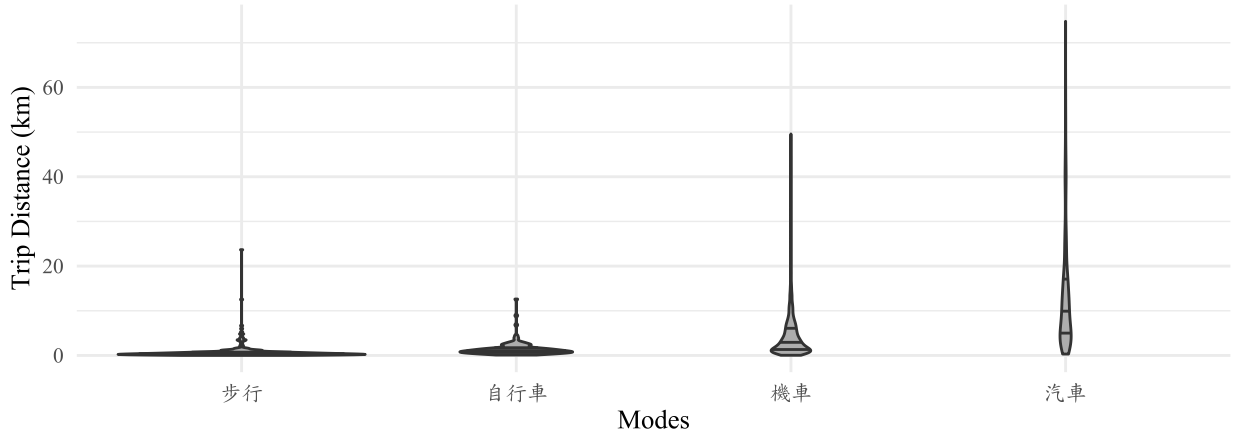

Figure 2: Violin Plot of Trip Distance

The adjusted R square of the model is only 0.0037. And hence, it is definitely not a good model. Since we found that the residuals have an exponential growth, we can then try to test the log-linear model.

Table 2: Naive Model of Age and Trip Distance

	Dependent variable:
	trip_dis
age	-23.025^{***}
	(5.681)
Constant	$6{,}004.324^{***}$
	(261.837)
Observations	4,156
R^2	0.004
Adjusted R^2	0.004
Residual Std. Error	6,719.000 (df = 4154)
F Statistic	16.428^{***} (df = 1; 4154)
Note:	*p<0.1; **p<0.05; ***p<0.01

Log-linear model

The log-linear regression model uses natural log values for dependent variable (trip distance) and keep the independent variables (age) in original scale. The model can be formulated as below.

$$ln(TripDistance) = \beta_{age} * x_{age}$$

The result of log-linear model and residual plots are shown in Table 3 and Figure 4 respectively. The coefficient of age is -0.0113, which means the trip distance would approximately decrease $|exp(-0.0113)-1| = 1.1\%$ (significant) as the age increases by 1. From the figures, the residuals are equally spread around the horizontal line, indicating that the model is not biased. The residual is likely to be normally distributed.

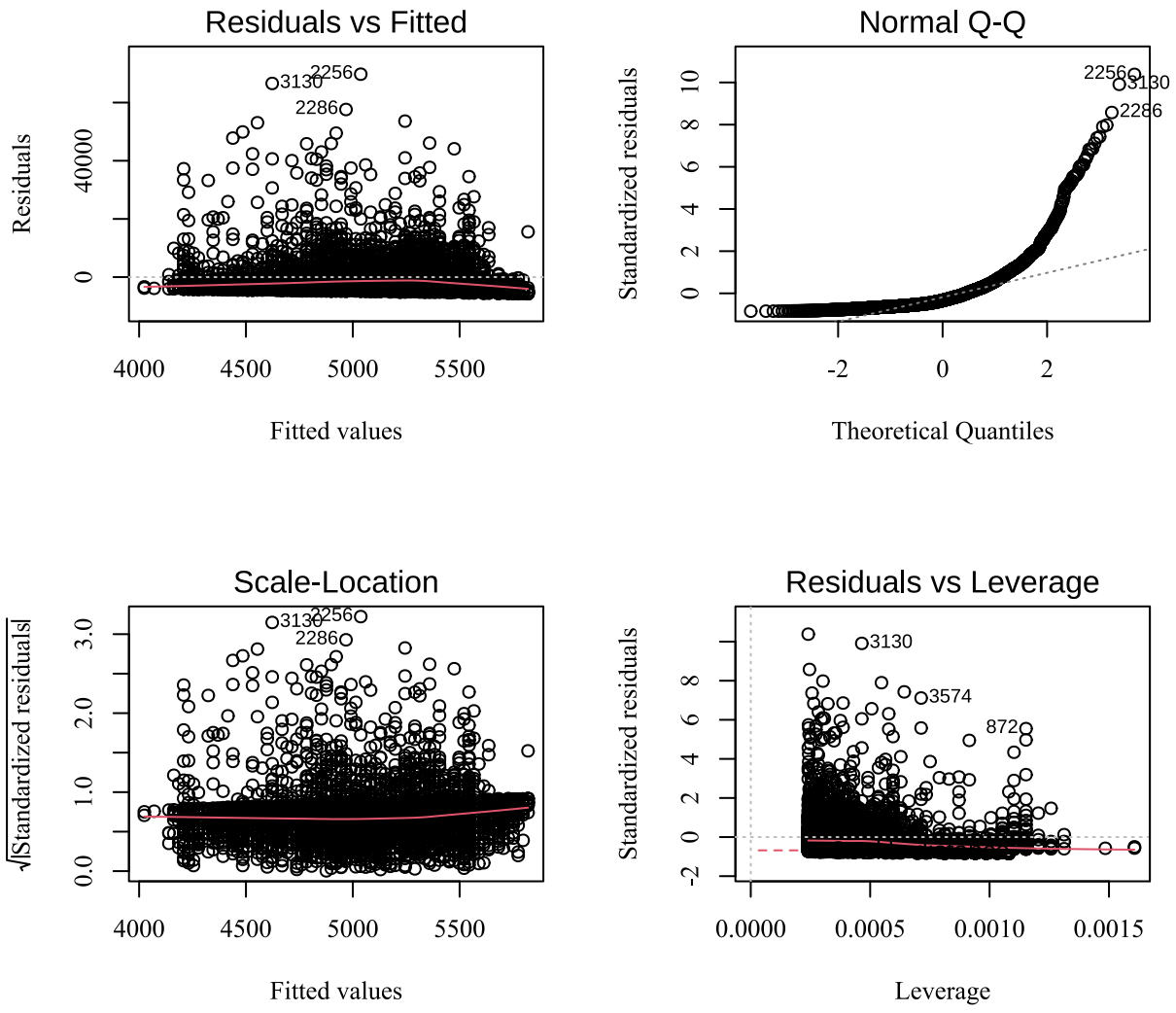

Figure 3: Residual Plot of Naive Model

Also, from the scale-location plot, we can find that the variance of residuals are equal for any fitted values. Thus, the results show that the linear regression assumption is met. Though the adjusted R square of the model is only 0.0256, it is much better than the previous model.

Table 3: Log-linear Model of Age and Trip Distance

	Dependent variable:
	log(trip_dis)
age	-0.011^{***}
	(0.001)
Constant	8.298^{***}
	(0.050)
Observations	4,156
R^2	0.026
Adjusted R^2	0.026
Residual Std. Error	1.276 (df = 4154)
F Statistic	110.254^{***} (df = 1; 4154)
Note:	*p<0.1; **p<0.05; ***p<0.01

Polynomial model

Based on the prior knowledge, the trip distance would be long for the youngsters and youth, while the kids and aged population may have a shorter trip distance. Thus, the relationship between trip distance and age might not be linear. Here, the polynomial model is introduced to verify our observations. The model resembles the log linear model in the former, while adding a quadratic term of age. The model is formulated as follows.

$$ln(TripDistance) = \beta_{age} * x_{age} + \beta_{age^2} * x_{age^2}$$

The result of polynomial model and residual plots are shown in Table 4 and Figure 5 respectively. The coefficient of age is 0.0815 (significant), and the coefficient of quadratic term is -1.064e-03 (significant). The result indicates that the function of trip distance is concave, and thus, the trip distance would increase in the initial stage, while reach a peak, and decrease after a specific age. From the figures, the residuals bounce randomly around the horizontal line, indicating that the model is not biased. Also, the variance of residuals are equal for any fitted values. The adjusted R square of the model is improved to be 0.1129, better than the two former models.

Dummy Regression

The relationship between age and trip distance might not simply be linear form. The trip distance may be different under the specific age interval. To classify the age group, we should first dig out the boundary of each group in the original data. As Table 5 shows, there are a huge gap between 18 to 19 years old, probably because the group turned out to be adults, and thus have a higher mobility. And, there is also a large gap between 64 to 65 years old, mainly because the adults are retired. Based on these observations, we can separate all the ages into 3 groups, namely, "<=18", "19-64", and ">=65". The model can be formulated as the followings, note that the variable used is dummy.

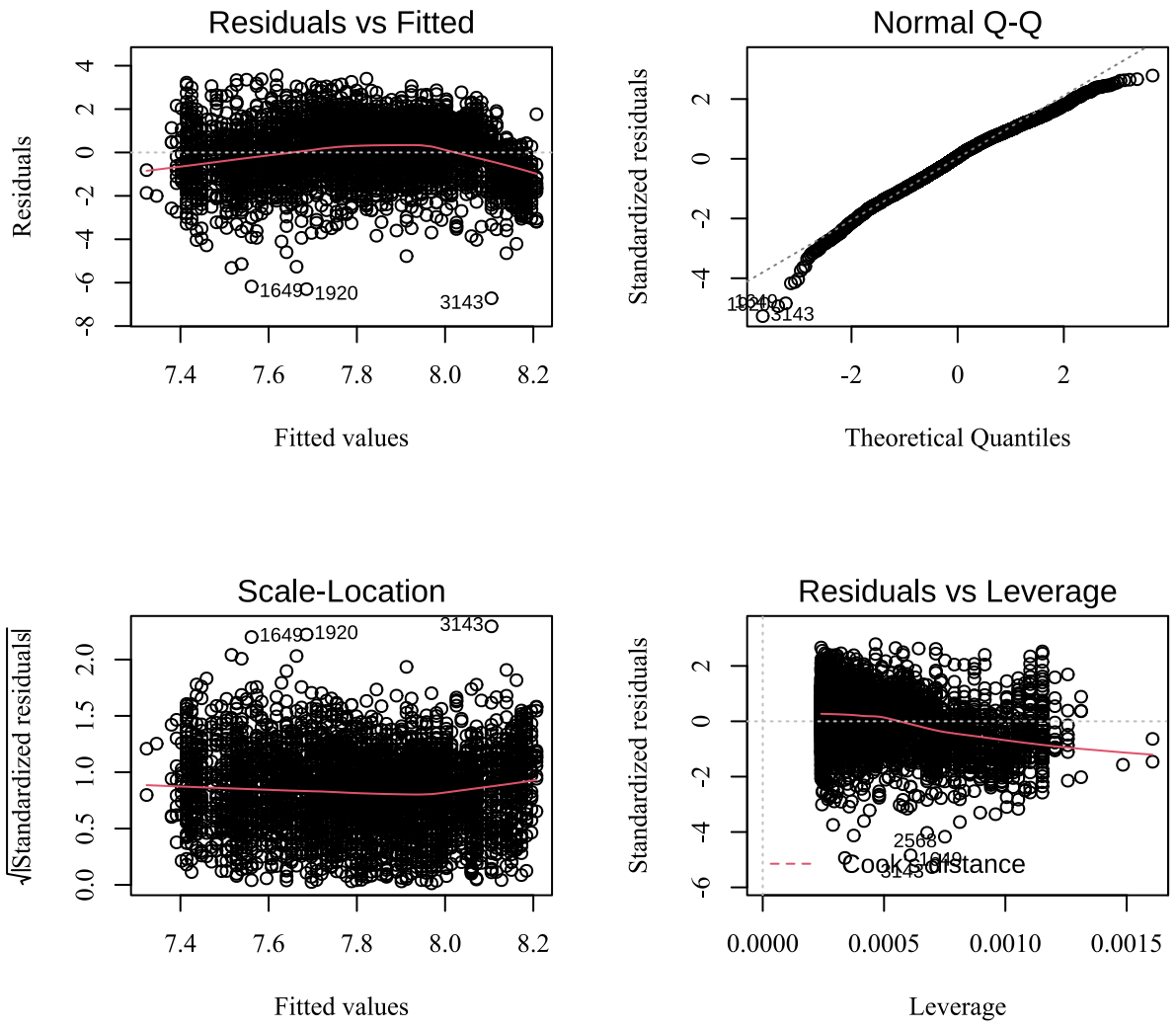

Figure 4: Residual Plot of Log-linear Model

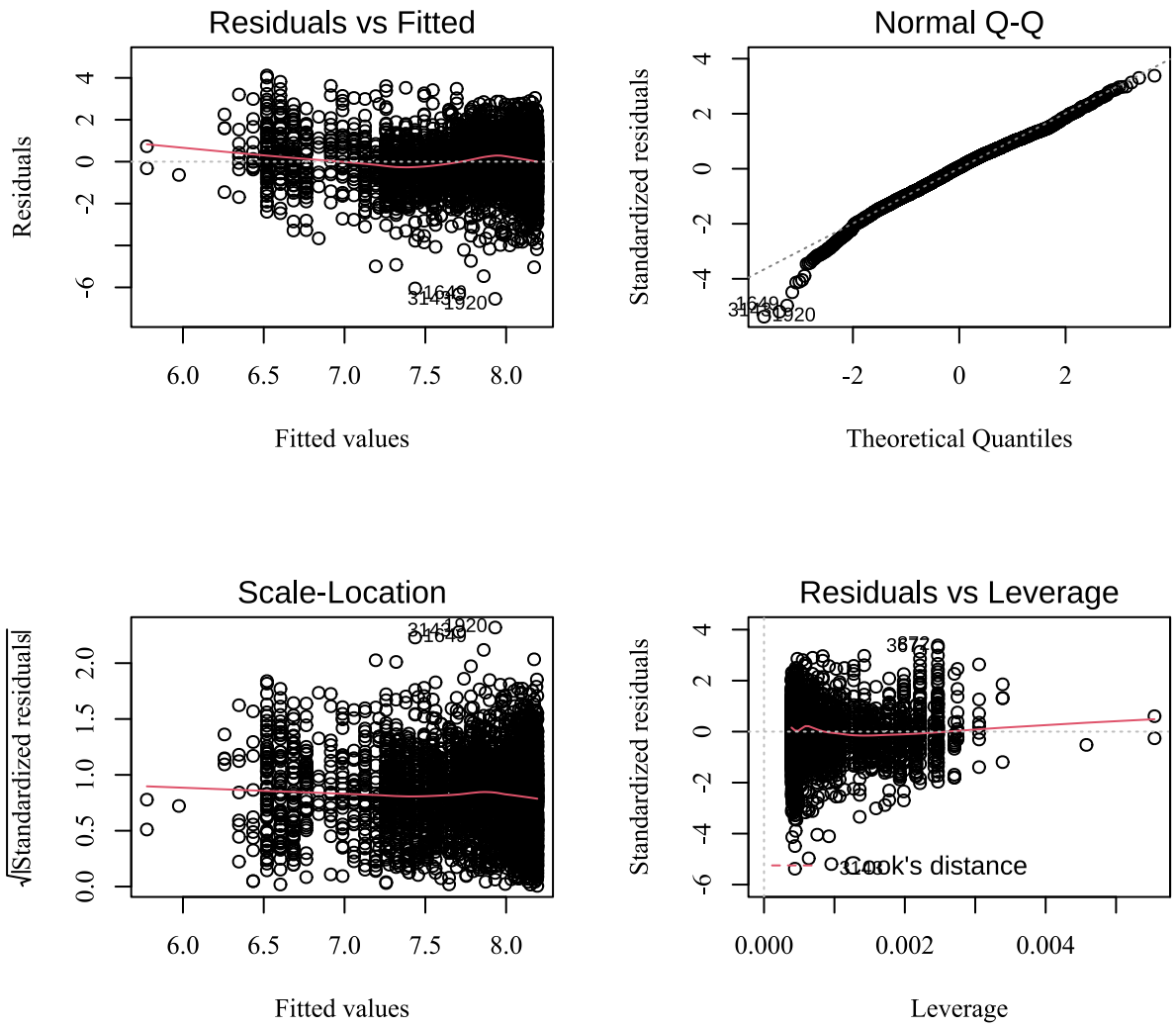

Figure 5: Residual Plot of Polynomial Model

Table 4: Polynomial Model of Age and Trip Distance

	Dependent variable:
	log(trip_dis)
age	0.082^{***}
	(0.005)
I(age^2)	-0.001^{***}
	(0.0001)
Constant	6.630^{***}
	(0.095)
Observations	4,156
R^2	0.113
Adjusted R^2	0.113
Residual Std. Error	1.218 (df = 4153)
F Statistic	265.317^{***} (df = 2; 4153)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5: Trip Distance for Each Ages

Age	Trip Distance	Age	Trip Distance
14	1694.59	61	4287.81
15	1634.60	62	2587.29
16	3437.27	63	4096.12
17	3146.88	64	5238.85
18	4949.53	65	1963.47
19	7770.55	66	3118.58
20	8412.22	67	1459.15
21	6623.52	68	3170.08
22	7647.37	69	2264.76

$$ln(TripDistance) = \beta_{AgeGroup} * x_{AgeGroup}$$
$$AgeGroup \in [\leq 18, 19-64, \geq 65]$$

The regression results by using dummy variables is shown in Table 6 and the residual plot is shown in Figure 6. Note that the base dummy is age "<=18" in the result. The coefficient of age 19-64 dummy variable is 1.026, and the t-test tells to reject the null hypothesis. We can conclude that age between 19-64 has a significant longer trip distance than age under 18 by $|exp(1.026) - 1| = 179\%$. The coefficient of age over 65 cannot reject to be 0, and hence we can say that the trip distance of age over 65 and under 18 have no significant difference. From the residual plot, we can find that it indeed randomly distributed around the horizontal line, and it is under normal distribution. They indicate that the model is not biased and the variance is equal. The adjusted R square of the model is 0.1239, the highest among four regression models.

The four models above are summarized in Table 7 and the prediction of trip distance for each model is shown in Figure 7.

Table 6: Dummy Model of Age and Trip Distance

	Dependent variable:
	log(trip_dis)
age_group19-64	1.026^{***}
	(0.062)
age_group>=65	-0.029
	(0.075)
Constant	7.060^{***}
	(0.058)
Observations	4,156
R^2	0.124
Adjusted R^2	0.124
Residual Std. Error	1.210 (df = 4153)
F Statistic	294.698^{***} (df = 2; 4153)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: Summary of All Models

	Dependent variable:			
	trip_dis		log(trip_dis)	
	Naive	Log-linear	Polynomial	Dummy
	(1)	(2)	(3)	(4)
age	-23.025^{***}	-0.011^{***}	0.082^{***}	
	(5.681)	(0.001)	(0.005)	
I(ageˆ2)			-0.001^{***}	
			(0.0001)	
age_group19-64				1.026^{***}
				(0.062)
age_group>=65				-0.029
				(0.075)
Constant	$6{,}004.324^{***}$	8.298^{***}	6.630^{***}	7.060^{***}
	(261.837)	(0.050)	(0.095)	(0.058)
Observations	4,156	4,156	4,156	4,156
R^2	0.004	0.026	0.113	0.124
Adjusted R^2	0.004	0.026	0.113	0.124
Residual Std. Error	6,719.000 (df = 4154)	1.276 (df = 4154)	1.218 (df = 4153)	1.210 (df = 4153)
F Statistic	16.428^{***} (df = 1; 4154)	110.254^{***} (df = 1; 4154)	265.317^{***} (df = 2; 4153)	294.698^{***} (df = 2; 4153)
Note:				*p<0.1; **p<0.05; ***p<0.01

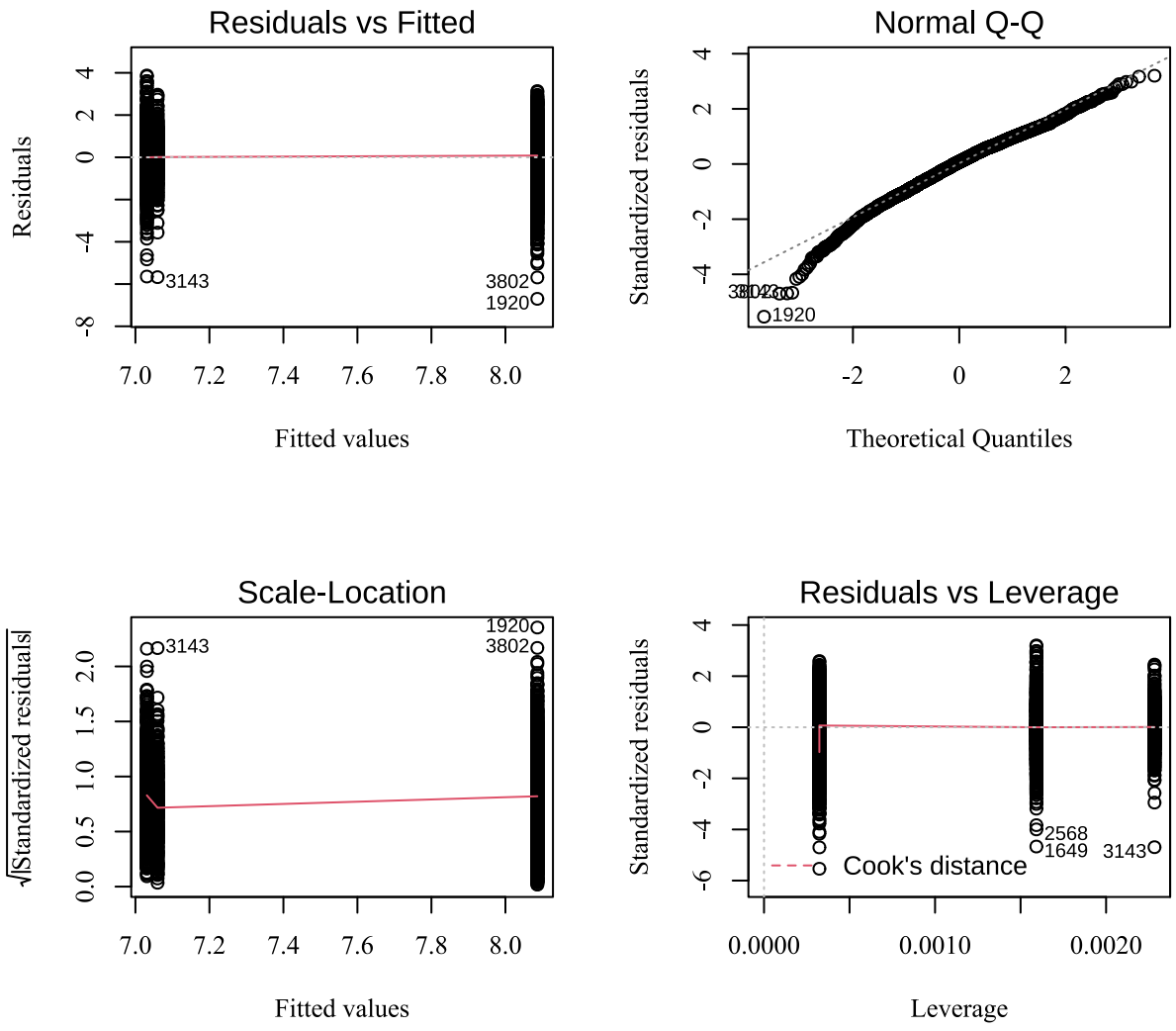

Figure 6: Residual Plot of Dummy Model

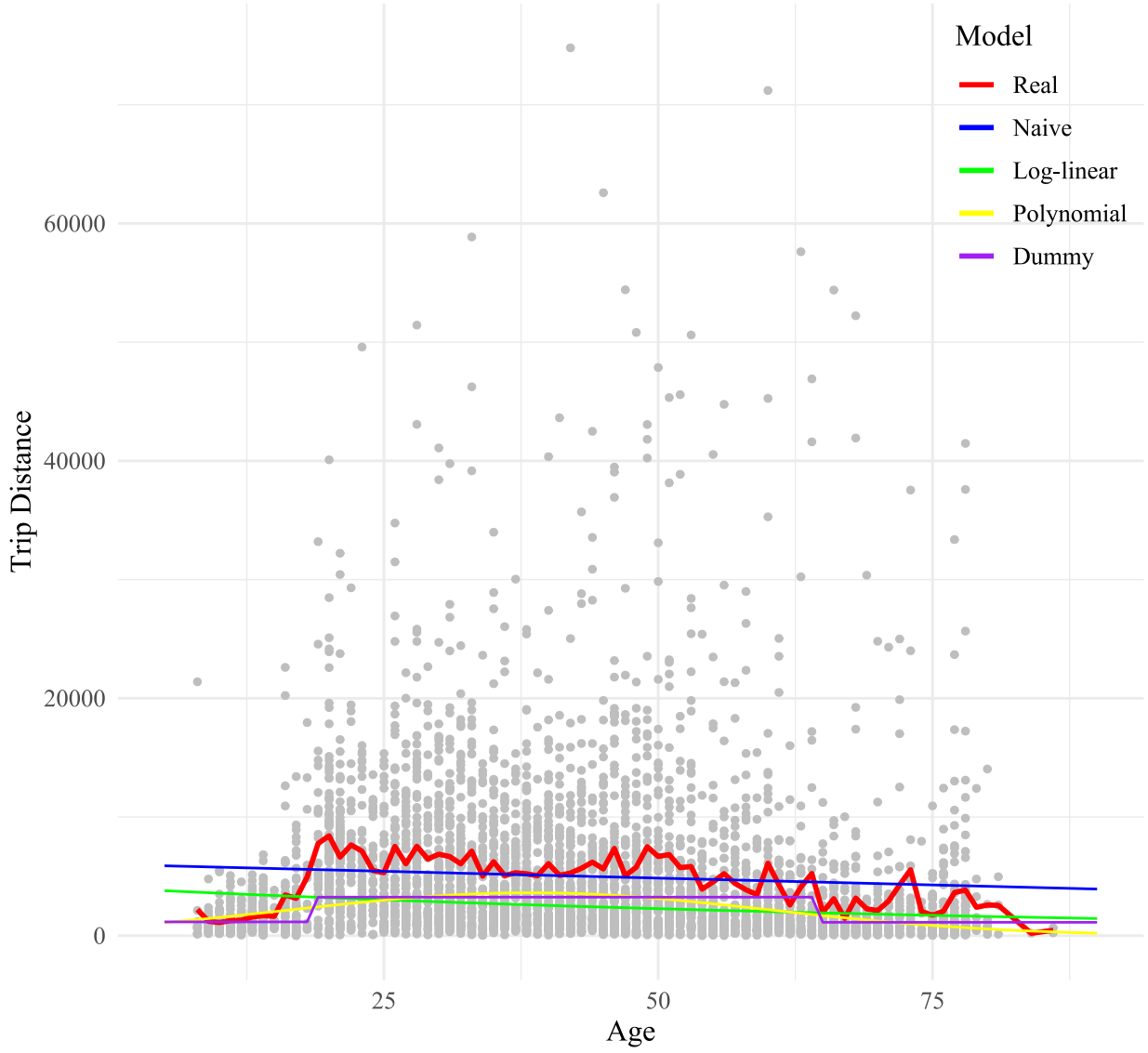

Figure 7: Prediction Plot of Dummy Model

Linear regression (the average trip distance for each district)

To first observe the mean trip distance of each district in Kaohsiung, it is essential to plot the visualization map, shown in Figure 8. The top three highest mean trip distance is distributed in 田寮區, 六龜區 and 內門區, which are located in the mountainous part of Kaohsiung, and far from the city center. And hence, people should ride for a longer distance. In the contrary, the center of the city such as 三民區, 前金區, 苓雅區 and so forth, which are located in the "舊高雄市", have the relatively lower travel distance, mainly because the accessibility of vital POIs regarding essential lives are much higher. But surprisingly, 甲仙區, 茂林區, 那瑪夏區, 桃源區, etc., where are also located in the remote area have the lowest trip distance. This may result from the socioeconomic features of those districts. The fact that people are not willing to go outside might because the road network is not operated well, or since the population is aged, and has no frequent need to travel for a long distance.

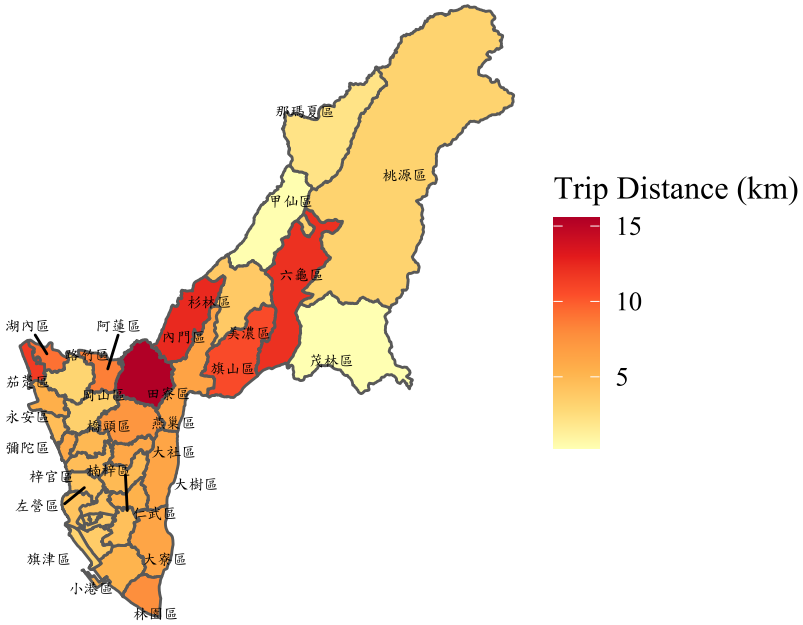

Figure 8: Average Trip Distance of Each District in Kaohsiung

To predict the average trip distance for each district of the Kaohsiung City by regression model, it can be formulated as below.

$$TripDistance = \beta_{district} * x_{district}$$

Note that the district variable is dummy. Since there are 38 districts in Kaohsiung, the model would use 37 dummies. The result is shown in Table 8, and here we use 旗津區 as the base. The adjusted R-squared of the model is 0.0798. In addition, the residual plot is shown in Figure 9.

The result shows that the intercept is 3172.25, it is exactly the average trip distance of the base district (旗津區). And coefficient of each dummy variables represent the difference between that district and 旗津區. For instance, the trip distance of 林園區 is significantly higher than that of 旗津區 by 4603.34 meters. From the residuals and fitted values plot, we can find that the regression is severely biased, and the heterogeneity exists, which indicates that there are other more vital variables we do not consider in the model. Also, the Q-Q plot tells that the residual is not normally distributed, violating the fundamental assumption of linear

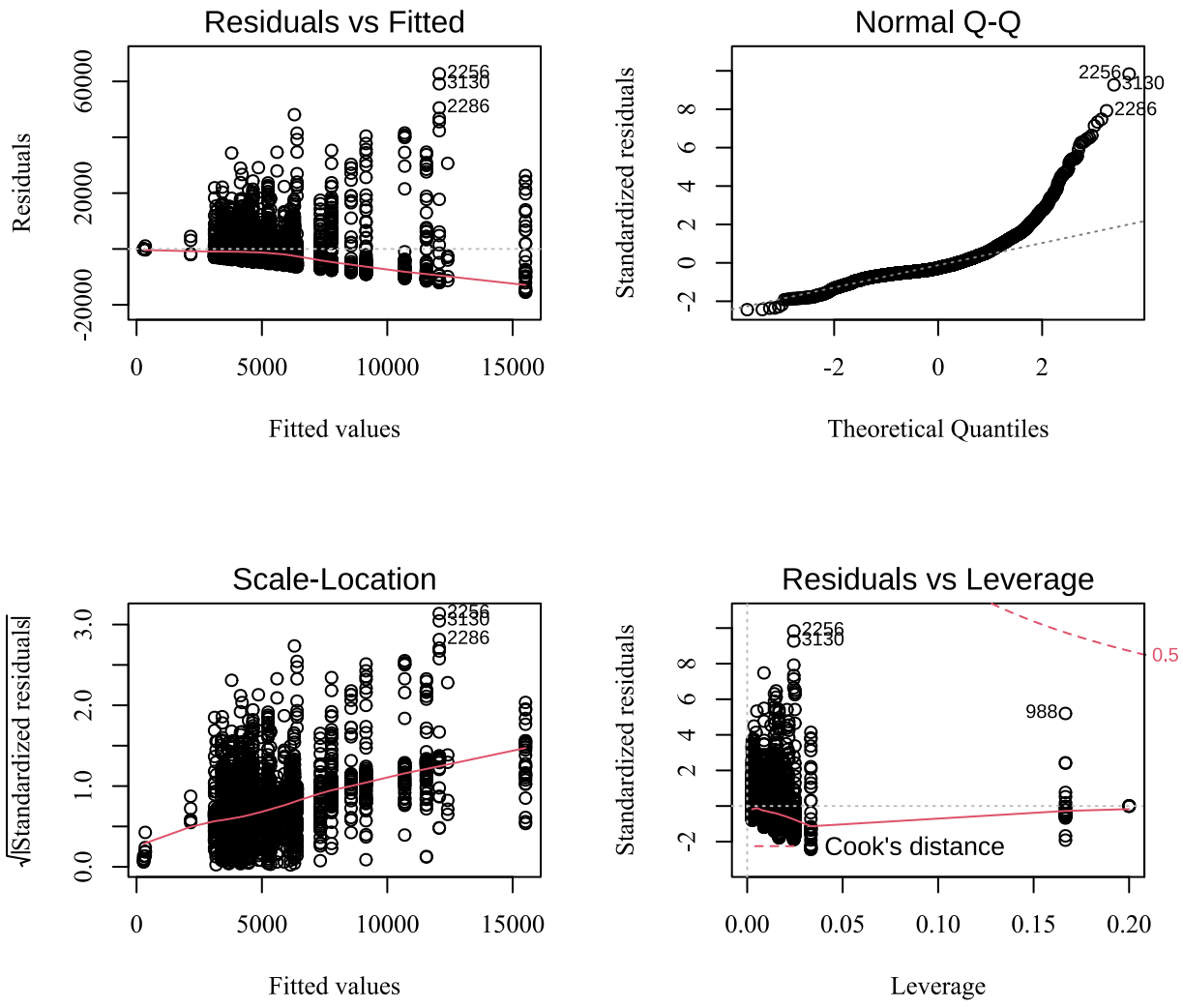

Figure 9: Average Trip Distance of Each District in Kaohsiung

Table 8: Regression Model of District and Trip Distance

District	Coefficients	p_value	District	Coefficients	p_value
(Intercept)	3172.25	0.00	茄萣區	8387.47	0.00
小港區	2071.97	0.05	永安區	2430.46	0.08
鳳山區	1297.23	0.23	彌陀區	1458.41	0.29
鹽埕區	332.05	0.77	梓官區	2718.57	0.05
鼓山區	1040.88	0.35	大樹區	3109.72	0.01
三民區	1035.16	0.35	大社區	2967.26	0.02
新興區	213.72	0.85	仁武區	1687.69	0.18
前金區	620.43	0.58	燕巢區	4155.64	0.00
苓雅區	515.30	0.64	阿蓮區	5383.55	0.00
前鎮區	502.14	0.65	田寮區	12339.92	0.00
左營區	973.38	0.37	旗山區	3228.93	0.01
楠梓區	1470.81	0.18	美濃區	7521.55	0.00
林園區	4603.34	0.00	六龜區	8900.24	0.00
大寮區	3123.67	0.01	甲仙區	-2825.08	0.32
鳥松區	2167.11	0.07	杉林區	667.58	0.81
岡山區	247.55	0.85	內門區	9236.42	0.00
橋頭區	1796.97	0.17	茂林區	-2890.25	0.35
路竹區	-58.54	0.96	桃源區	44.08	0.99
湖內區	5979.09	0.00	那瑪夏區	-1013.92	0.72

regression. Though the model performs not well, it indeed explain the difference of average trip distance for each district.

However, the regression above is not about "prediction", it is only used to prove that when the dependent variable is trip distance, and the independent variable is district, the intercept would be the "average trip distance" of the district which is the base of dummy variable, and the slope would be the difference between the base and each district.

To predict the average trip distance of each district by using regression, the first task is to construct a regression model, whose independent variable is the socioeconomic factors from the provided data. In order to formulate the model easily, some variables should be cleaned or condensed into fewer attributes. The age is categorized into three groups as the former problem, namely, "<=18", "19-24" and ">=65". The mode is classified into "機車" and "汽車", no matter the respondent is passenger or driver. The trip purpose is condensed into six types, that is, "上班/學旅次", "出差旅次", "返家旅次", "購物旅次", "個人事務與休閒旅次", and "其他". Income is simply labeled as "30 萬以下" and "30 萬以上", which indicates whether low income group or not. Occupation is categorized into "第一級產業", "第二級產業", "第三級產業", and "無職業 (含學生/退休/家管)". Last, create a new variable that separate the districts into two parts, "舊高雄市" and "舊高雄縣". The model result is shown in Table 9. Also the residual plot is shown in Figure 10.

The interpretation of the model result is as the followings. First we can find that the household population has a significant effect on the trip distance. As for the transport mode, all other modes (自行車, 機車, 汽車) has a significant longer trip distance than walking. Age over 19 has also a significantly longer trip distance than the age under 18. The more scooter owns, the longer trip distance occurs. Female is inclined to have a lower trip distance than male, which is approximately $exp(-0.105) = 90\%$ of the trip distance of male. For the trip purpose, we can find that 返家旅次, 購物旅次, 個人事務與休閒旅次 have a significant shorter travel distance compared to 上班/學旅次. If the trip is not from home, the trip distance would increase by $exp(0.321) - 1 = 37.9\%$. If the income is over 300 thousand NTD per year, the trip distance is significantly longer. As for the industry category, the tertiary sector of the economy has a significant short distance compared to the one who has no job. But the primary and secondary industry has no obvious difference.

Table 9: Model Result of Trip Distance

	Dependent variable:
	log(trip_dis)
hh_pop_rev	−0.071***
	(0.018)
mode_class 自行車	0.984***
	(0.114)
mode_class 機車	1.394***
	(0.065)
mode_class 汽車	2.477***
	(0.078)
age_group19-64	0.656***
	(0.068)
age_group>=65	0.300***
	(0.072)
scooter_no	0.113***
	(0.022)
gender 女生	−0.105***
	(0.033)
trip_purp_rev 出差旅次	0.043
	(0.257)
trip_purp_rev 返家旅次	−0.253***
	(0.064)
trip_purp_rev 購物旅次	−0.723***
	(0.051)
trip_purp_rev 個人事務與休閒旅次	−0.493***
	(0.058)
HB 否	0.321***
	(0.056)
p_income_rev30 萬以上	0.223***
	(0.072)
occu_rev 第一級產業	−0.162
	(0.172)
occu_rev 第二級產業	0.037
	(0.089)
occu_rev 第三級產業	−0.247***
	(0.078)
city 原高雄市	−0.134***
	(0.032)
Constant	6.085***
	(0.103)
Observations	4,155
R^2	0.390
Adjusted R^2	0.387
Residual Std. Error	1.012 (df = 4136)
F Statistic	146.896*** (df = 18; 4136)
Note:	*p<0.1; **p<0.05; ***p<0.01

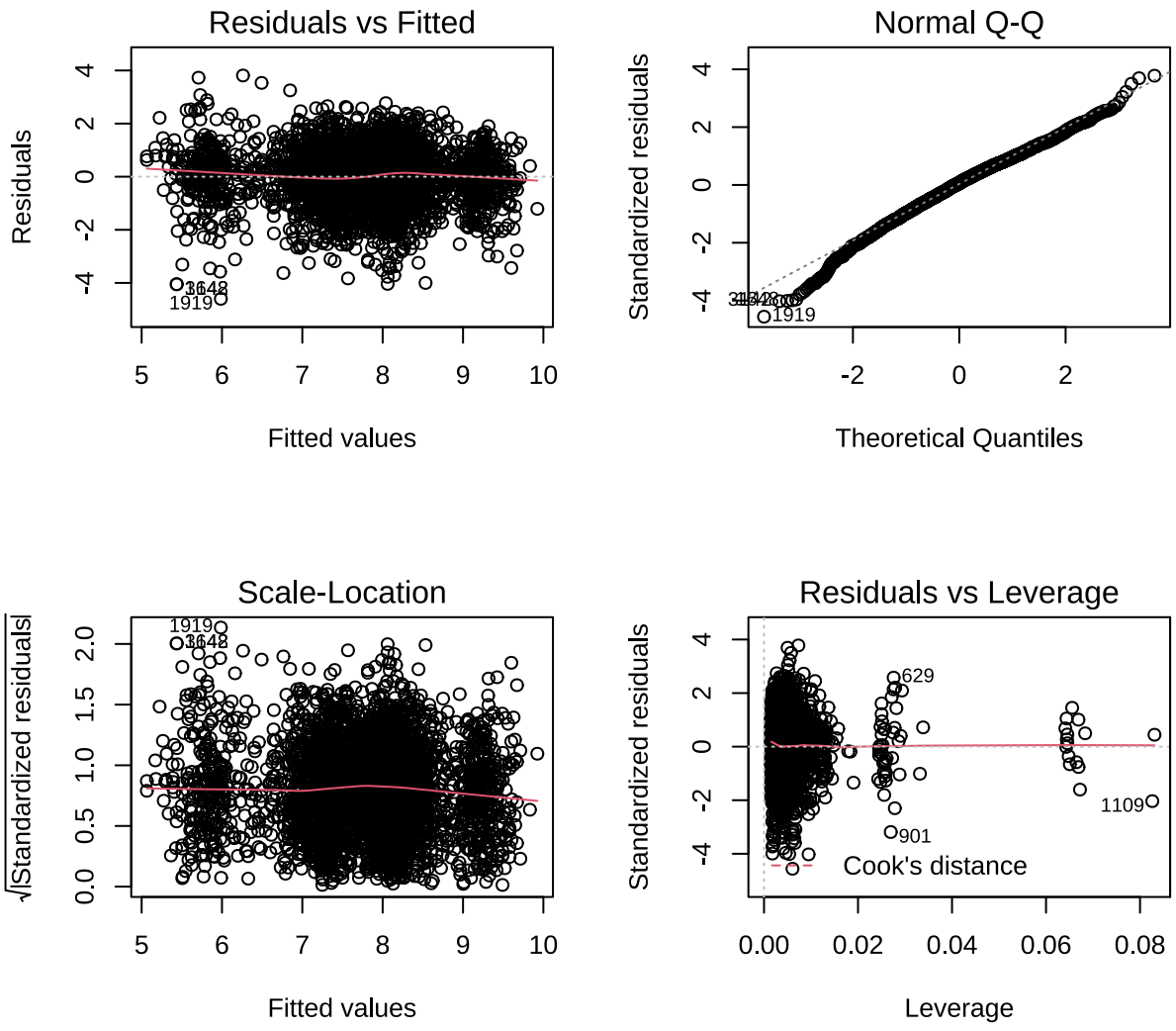

Figure 10: Residual Plot of Regression Model

Table 10: Prediction of Average Trip Distance for Each District

District	Trip Distance	Prediction of Trip Distance	District	Trip Distance	Prediction of Trip Distance
旗津區	3172.25	2056.53	茄萣區	11559.72	4343.65
小港區	5244.22	3331.59	永安區	5602.71	3644.98
鳳山區	4469.48	3476.83	彌陀區	4630.66	3955.28
鹽埕區	3504.30	2605.15	梓官區	5890.82	4169.51
鼓山區	4213.13	2973.87	大樹區	6232.64	3378.13
三民區	4207.41	3094.16	大社區	6139.51	3933.37
新興區	3385.97	2943.92	仁武區	4859.94	3693.87
前金區	3792.68	3135.65	燕巢區	7327.89	4428.88
苓雅區	3687.55	2903.99	阿蓮區	8555.80	3890.11
前鎮區	3674.39	2892.84	田寮區	15512.17	5101.91
左營區	4145.63	3174.85	旗山區	6401.18	3379.49
楠梓區	4643.06	3191.22	美濃區	10693.80	4044.38
林園區	7775.59	3904.71	六龜區	12072.49	3582.05
大寮區	6295.92	3499.49	甲仙區	347.17	1354.45
鳥松區	5339.36	3532.43	杉林區	3839.83	3885.19
岡山區	3419.80	3231.33	內門區	12408.67	5884.34
橋頭區	4969.22	4141.29	茂林區	282.00	2006.99
路竹區	3113.71	3144.37	桃源區	3216.33	1989.94
湖內區	9151.34	3783.67	那瑪夏區	2158.33	1033.09

Last, if the district is 原高雄市, the trip distance would drop $1 - exp(-0.134) = 12.5\%$ compared to the one is 原高雄縣. From the residual plot, we can find that the residual bounce randomly around the horizontal line, indicating that the model is not biased, and it is nearly normally distributed. The adjusted R square of the model is 0.387.

By using the linear regression model, we can calculate the average fitted values of each district, to predict the average trip distance of each district in Kaohsiung. The result is shown in Table 10. And the relationship between the real value and prediction value of average trip distance for each district is shown in Figure 11. Here we define the 20% is acceptable, marked green in the figure. And most of the prediction is underestimated, which is marked blue. Few of prediction is overestimated, marked red.

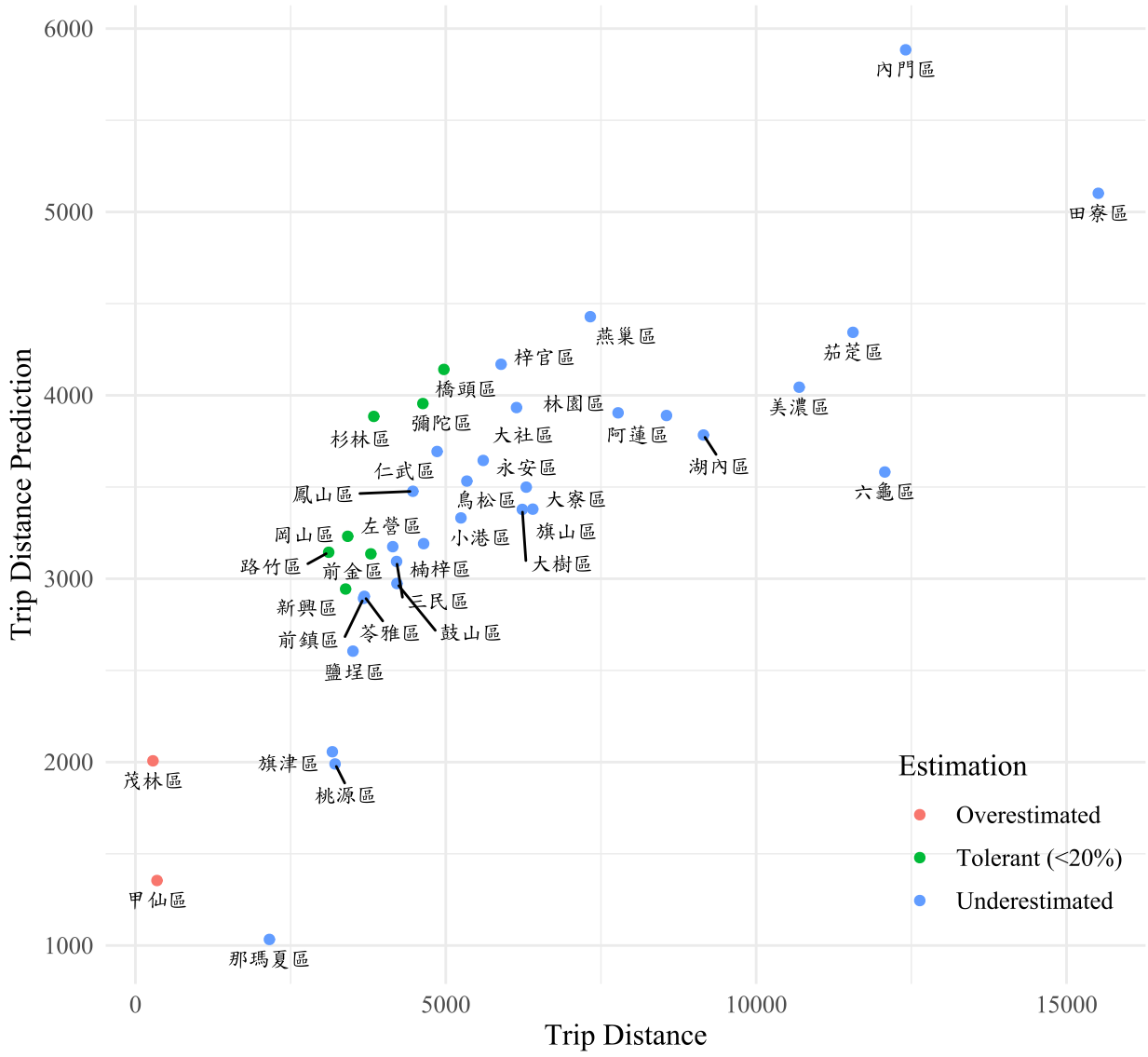

Figure 11: Plot of Prediction and Real Trip Ditance