77

Assignment 01

Question 1 (20 marks)

（a）(10 marks) Use SAS to study the distribution of the number of registered users per day (registered) by season. Obtain measures of location, dispersion, skewness and kurtosis. Obtain a boxplot, histogram and a quantilequantile plot. Also carry out Normal Goodness-of-fit tests. What are the key features of these distributions?

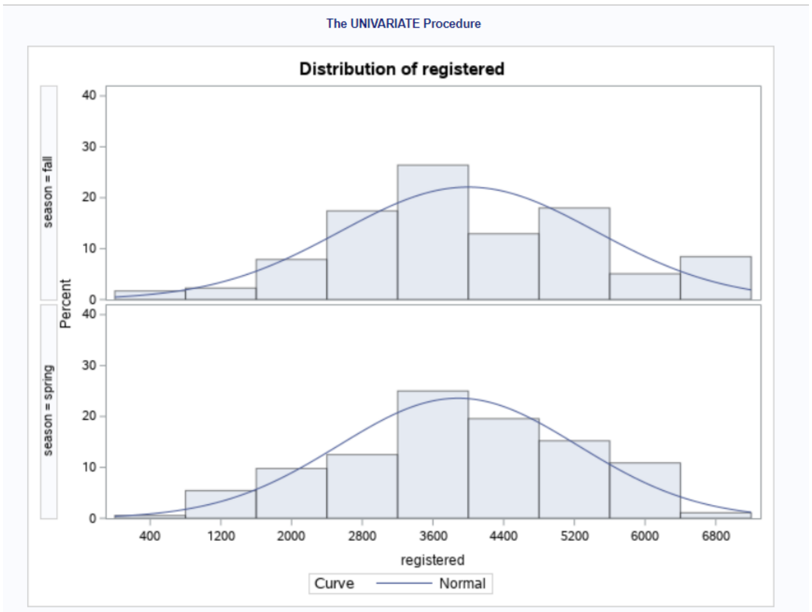
The MEANS Procedure

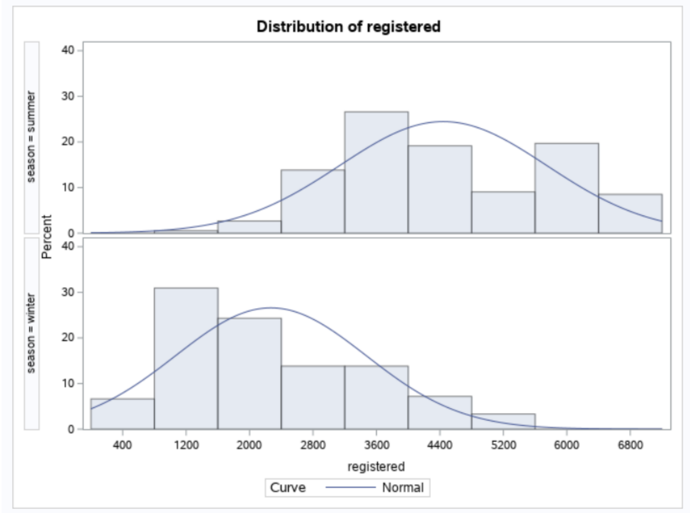
Analysis Variable : registered													
season	N Obs	N	Minimum	Maximum	Mean	Median	Lower Quartile	Upper Quartile	Quartile Range	Variance	Std Dev	Skewness	Kurtosis
fall	178	178	20.0000000	6946.00	3999.05	3815.00	2928.00	5080.00	2152.00	2087396.64	1444.78	0.0428496	-0.2907145
spring	184	184	674.0000000	6456.00	3886.23	3844.00	3006.00	4948.50	1942.50	1831625.59	1353.38	-0.1391497	-0.7125236
summer	188	188	889.0000000	6917.00	4441.69	4110.50	3474.50	5670.50	2196.00	1702051.48	1304.63	0.1495646	-0.8482097
winter	181	181	416.0000000	5315.00	2269.20	1867.00	1379.00	3162.00	1783.00	1440647.47	1200.27	0.6466041	-0.5071581

The above table gives the daily registered statistics for the four seasons. From the data point of view, the difference between the minimums of different seasons is very large. The minimum value of the fall season is the smallest, only 20; while the minimum value of summer is the largest, which is 889. For maximum, the value of fall and summer are not much different, both are close to 7000; while the maximum value of spring is around 6456, and the maximum value of winter is the smallest, which is 5315.

Although the fall quarter has the minimum user registered each day, the average and median values are not the lowest among the four. Winter has the lowest mean value (2269.20) and the lowest median value (1867.00).

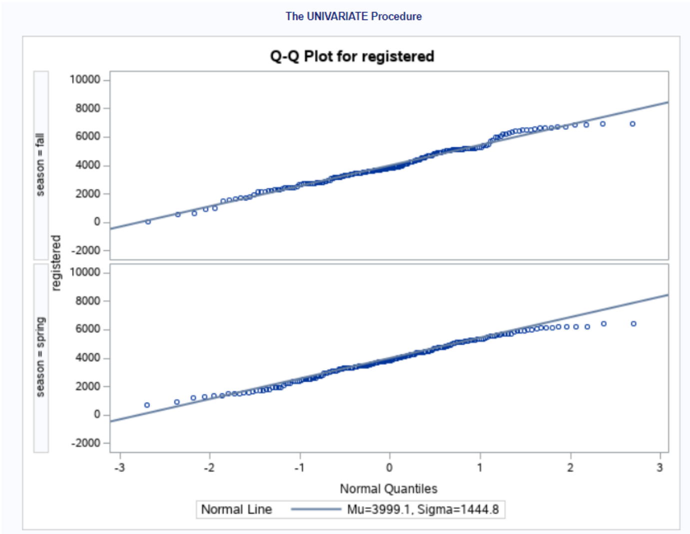
Judging from the value of Quartile Range, the data in the middle 50 of summar has the largest degree of dispersion, but the gap with fall is not large, while the degree of dispersion in winter is the smallest. But it is also necessary to consider the standard deviation. The degree of dispersion of fall is the largest (Std Dev=1444.78), followed by spring and summer (1353.38 and 1304.63), and the lowest degree of dispersion is 1200.27.

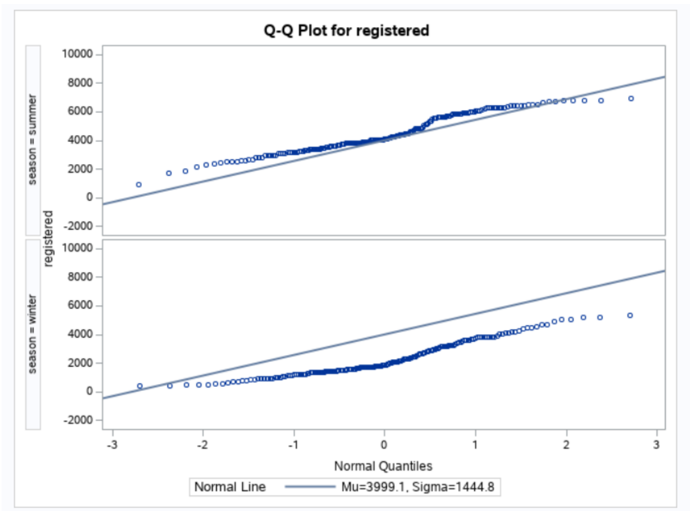




From the histogram, the distribution of the number of daily registered users in fall and spring is close to the normal distribution (but not strictly normal distribution).

The frequency of the number of daily registered users in summer is somewhat negatively skewed, while the frequency of the number of daily registered users in winter shows a significant positive skewed.





It is not difficult to draw the same conclusion from the above Q-Q diagram.

The following figure shows the results of the normality detection of different seasons. It can be found that when season=fall, the values of 4 p-values are all greater than 0.05; when season=spring, there are three p-values. The value is greater than 0.05; when season=summer and season=winter, the four p-values are all less than 0.05.

What conclusion? Which seasons can be assume normal?

This is a very favorable proof of the above conclusion.

The UNIVARIATE Procedure
Variable: registered
season = fall

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.985214	Pr < W	0.0573
Kolmogorov-Smirnov	D	0.065002	Pr > D	0.0660
Cramer-von Mises	W-Sq	0.108452	Pr > W-Sq	0.0892
Anderson-Darling	A-Sq	0.718851	Pr > A-Sq	0.0624

The UNIVARIATE Procedure
Variable: registered
season = summer

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.958596	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.109303	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.570268	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.153981	Pr > A-Sq	<0.0050

The UNIVARIATE Procedure
Variable: registered
season = spring

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.982232	Pr < W	0.0193
Kolmogorov-Smirnov	D	0.051655	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.065815	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.600269	Pr > A-Sq	0.1202

The UNIVARIATE Procedure
Variable: registered
season = winter

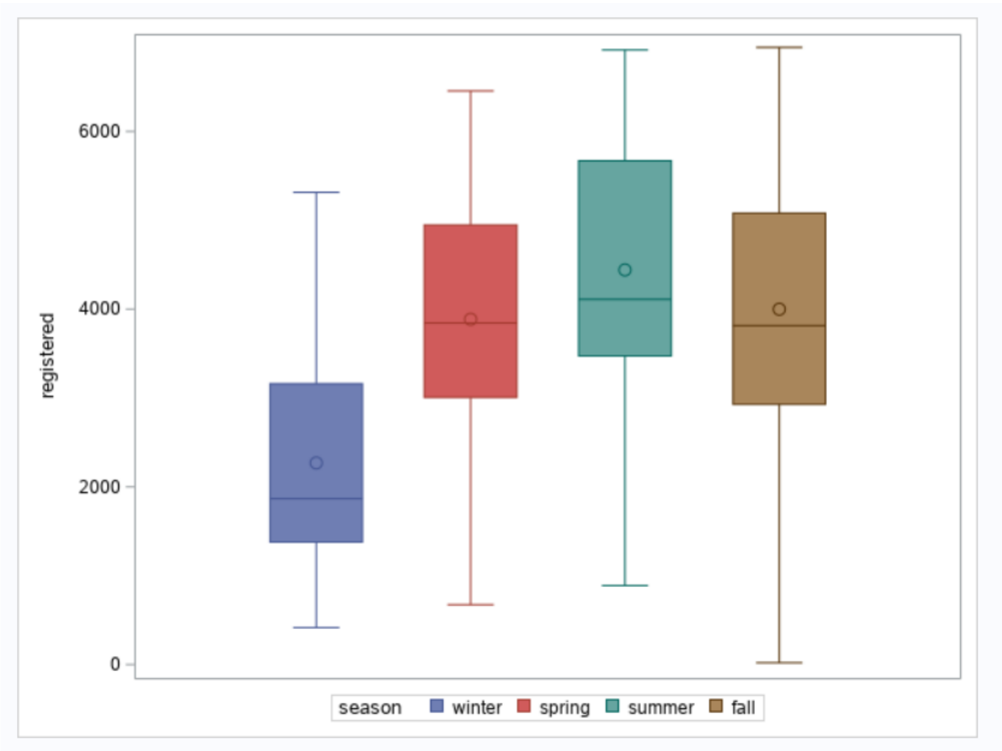
Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.9391	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.139193	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.72481	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.911387	Pr > A-Sq	<0.0050

The figure below is a box plot. It is not difficult to find that when season=winter, the value of IQR is significantly lower than the other three. When season=fall.

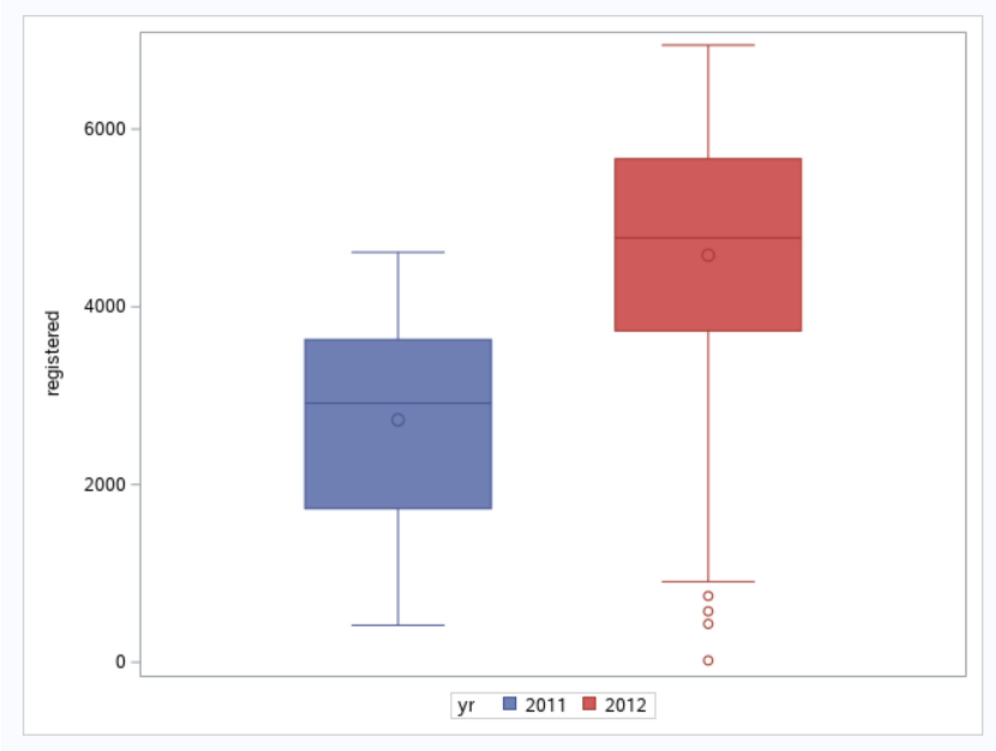
From the picture, there are no outliers.

9



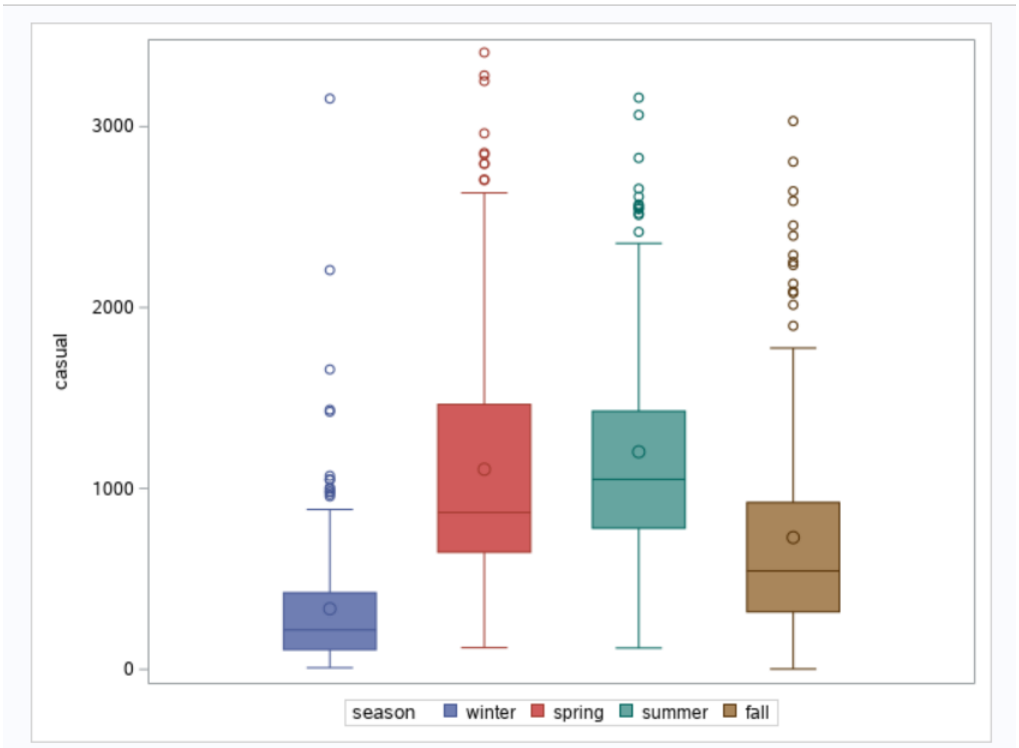
(b) (10 marks) Now use SAS to obtain boxplots of registered by season, and by yr, respectively. Similarly, obtain boxplots of casual by season and yr. What do the boxplots suggests about the pattern and trend, if any, of bike rentals?

It is not difficult to see from the above figure that the change in the number of registrations in a single day shows seasonality, showing the lowest overall in winter, then rising in spring, reaching the highest in summer, and starting in fall fall back. The middle 50% of fall and the middle 50% of spring are similar, but the dispersion of the first 25% and the last 25% of fall is higher than that of fall's.

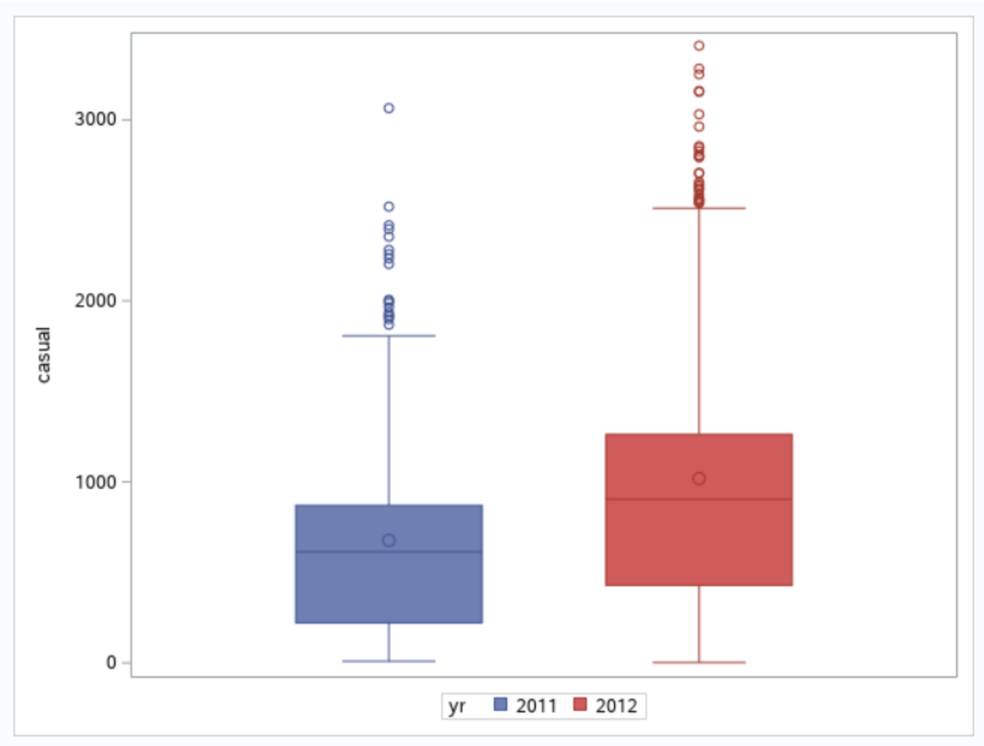


It is not difficult to see from the above figure that the registration in 2012 generally showed an upward trend relative to 2011. Also, the 2012 boxplots pointed out that there were outliers in that year. For some reason, one day in 2012 had an unusually high number of signups and was lower than the 2011 minimum.

low?



The above picture shows the casual users of each day in different seasons. It can be found that the trend and the trend of the number of registrations in different seasons are basically the same. It is important to note that, unlike the number of signups each day, there are a lot of outliers when using the variable casual users.



From the chart above, the trend of casual users is rising, higher in 2012 than in 2011. And 2012 has more outliers than 2011, and these outliers are all higher than the inner limit, which shows that a large number of temporary users appeared at some time in 2012.

Question 2 (60 marks)

(a) (8 marks) Obtain a Pearson correlation matrix relating variables registered, atemp, temp, hum and windspeed. Also obtain a scatterplot matrix of the same variables. Discuss the relationships.

The CORR Procedure

5 Variables:	registered atemp temp hum windspeed

Pearson Correlation Coefficients, N = 731 Prob > |r| under H0: Rho=0					
	registered	atemp	temp	hum	windspeed
registered	1.00000	0.54419 <.0001	0.54001 <.0001	-0.09109 0.0138	-0.21745 <.0001
atemp	0.54419 <.0001	1.00000	0.99170 <.0001	0.13999 0.0001	-0.18364 <.0001
temp	0.54001 <.0001	0.99170 <.0001	1.00000	0.12696 0.0006	-0.15794 <.0001
hum	-0.09109 0.0138	0.13999 0.0001	0.12696 0.0006	1.00000	-0.24849 <.0001
windspeed	-0.21745 <.0001	-0.18364 <.0001	-0.15794 <.0001	-0.24849 <.0001	1.00000

Spearman Correlation Coefficients, N = 731 Prob > |r| under H0: Rho=0					
	registered	atemp	temp	hum	windspeed
registered	1.00000	0.53188 <.0001	0.53117 <.0001	-0.09322 0.0117	-0.20298 <.0001
atemp	0.53188 <.0001	1.00000	0.99255 <.0001	0.13965 0.0002	-0.16899 <.0001
temp	0.53117 <.0001	0.99255 <.0001	1.00000	0.12990 0.0004	-0.14715 <.0001
hum	-0.09322 0.0117	0.13965 0.0002	0.12990 0.0004	1.00000	-0.23901 <.0001
windspeed	-0.20298 <.0001	-0.16899 <.0001	-0.14715 <.0001	-0.23901 <.0001	1.00000

Pearson Correlation Statistics (Fisher's z Transformation)									
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		p Value for H0:Rho=0
registered	atemp	731	0.54419	0.61009	0.0003727	0.54393	0.490773	0.593052	<.0001
registered	temp	731	0.54001	0.60417	0.0003699	0.53975	0.486268	0.589203	<.0001
registered	hum	731	-0.09109	-0.09134	-0.0000624	-0.09103	-0.162468	-0.018636	0.0137
registered	windspeed	731	-0.21745	-0.22098	-0.0001489	-0.21731	-0.285325	-0.147112	<.0001
atemp	temp	731	0.99170	2.74034	0.0006792	0.99169	0.990397	0.992810	<.0001
atemp	hum	731	0.13999	0.14091	0.0000959	0.13989	0.068071	0.210275	0.0001
atemp	windspeed	731	-0.18364	-0.18575	-0.0001258	-0.18352	-0.252673	-0.112505	<.0001
temp	hum	731	0.12696	0.12765	0.0000870	0.12688	0.054869	0.197573	0.0006
temp	windspeed	731	-0.15794	-0.15928	-0.0001082	-0.15784	-0.227746	-0.086313	<.0001
hum	windspeed	731	-0.24849	-0.25380	-0.0001702	-0.24833	-0.315168	-0.179040	<.0001

Spearman Correlation Statistics (Fisher's z Transformation)									
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		p Value for H0:Rho=0
registered	atemp	731	0.53188	0.59277	0.0003643	0.53162	0.477516	0.581710	<.0001
registered	temp	731	0.53117	0.59177	0.0003638	0.53091	0.476746	0.581050	<.0001
registered	hum	731	-0.09322	-0.09349	-0.0000639	-0.09316	-0.164561	-0.020786	0.0116
registered	windspeed	731	-0.20298	-0.20584	-0.0001390	-0.20285	-0.271368	-0.132278	<.0001
atemp	temp	731	0.99255	2.79476	0.0006798	0.99254	0.991383	0.993549	<.0001
atemp	hum	731	0.13965	0.14057	0.0000956	0.13955	0.067725	0.209943	0.0001
atemp	windspeed	731	-0.16899	-0.17062	-0.0001157	-0.16887	-0.238468	-0.097555	<.0001
temp	hum	731	0.12990	0.13064	0.0000890	0.12982	0.057847	0.200442	0.0004
temp	windspeed	731	-0.14715	-0.14823	-0.0001008	-0.14705	-0.217251	-0.075344	<.0001
hum	windspeed	731	-0.23901	-0.24372	-0.0001637	-0.23885	-0.306065	-0.169270	<.0001

Hypothesis tests are based on

1. H0: r=0

2. H1: r≠0 and α=0.05

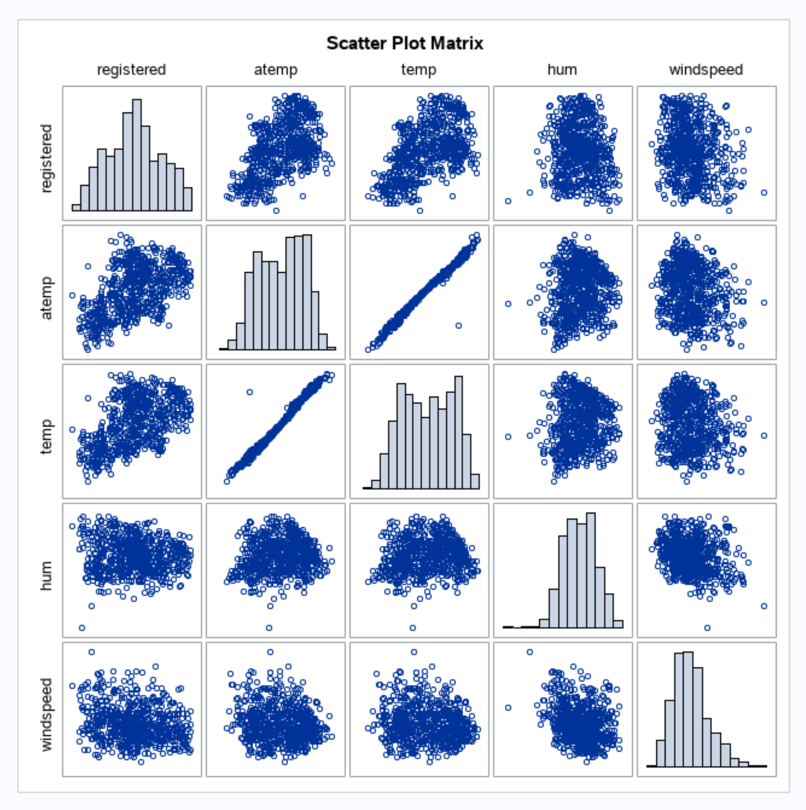
All variable pairs are significant on the 5% level for correlation coefficients.

We can know from the output:

1. temp has a significant positive correlattation with atemp while r = 0.99170 and p-value < 0.001
2. Windspeed has a weak negative correlation with hum while r = -0.24849 and p-value < 0.001
3. registered has a relative positive correlation with temp and atemp, both r ≈ 0.53 and p-value < 0.001
4. p-value of hum and windspeed is less than 0.001, otherwise is more than 0.001 ?

For registered , hum and windspeed variables pairs, there is non-linear patterns between registered and those variables, so the spearman correlation coefficient will be better.

From Fisher's Z Transformation output, we have 95% confident limits shows that smallest margin between temp and atemp. We have 95% confidence that the population correlation coefficient between temo and atemp is between 0.99 and 0.993, which is a very large effect.



From the above figure, there is a very obvious linear relationship between temp and atemp, and there is a more significant linear relationship between temp, atemp and registered. The relationship before the variables outside this is non-linear.

It is not difficult to see from the histogram that the distribution of temp and atemp is bimodal, while the distribution of hum is left skewed and windspeed is right skewed.

(b) (12 marks) In this question, we investigate observations where workingday=1. Fit a simple regression model relating registered on working days to atemp, with registered as the dependent variable. Discuss the fitted relationship and the goodness of fit. Examine residual plots and influence diagnostics and comment on the residual patterns.

The REG Procedure
Model: MODEL1
Dependent Variable: registered

Number of Observations Read	731
Number of Observations Used	731

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	526282237	526282237	306.72	<.0001
Error	729	1250829735	1715816		
Corrected Total	730	1777111972			

Root MSE	1309.89153	R-Square	0.2961
Dependent Mean	3656.17237	Adj R-Sq	0.2952
Coeff Var	35.82685		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1184.63986	149.20595	7.94	<.0001
atemp	1	5210.31247	297.50190	17.51	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: registered

Number of Observations Read	500
Number of Observations Used	500

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	374373355	374373355	218.42	<.0001
Error	498	853563854	1713984		
Corrected Total	499	1227937210			

how strong is the model?

Root MSE	1309.19198	R-Square	0.3049
Dependent Mean	3978.25000	Adj R-Sq	0.3035
Coeff Var	32.90874		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	1387.81754	184.79649	7.51	<.0001	1024.74068	1750.89440
atemp	1	5395.27265	365.06002	14.78	<.0001	4678.02501	6112.52030

According to the table, we can get simple linear regression equation:

$$Registered = 1387.82 + 5395.27 * attemp$$

or:

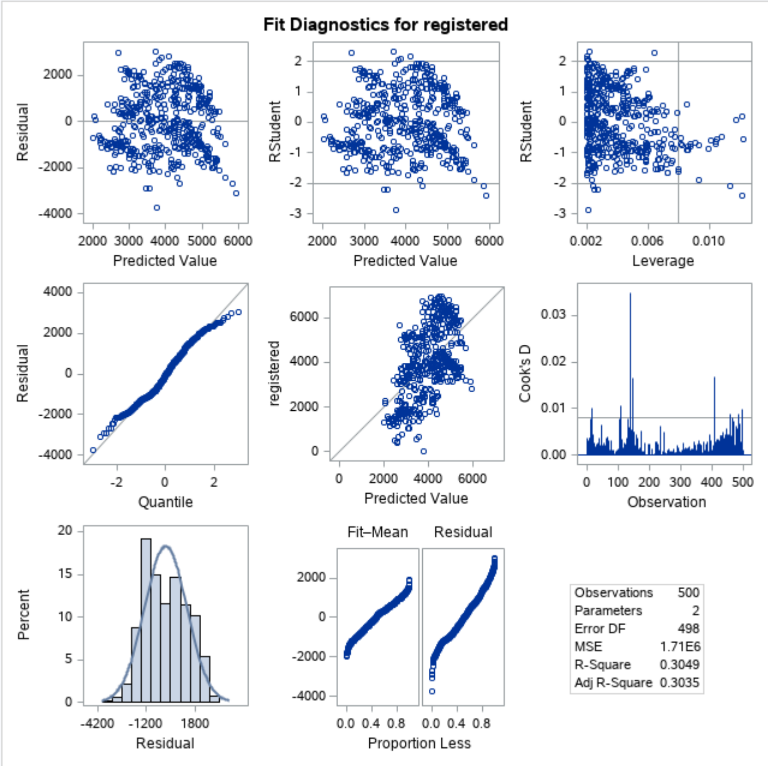
$$Registered = 1387.82 + 5395.27 * attemp$$

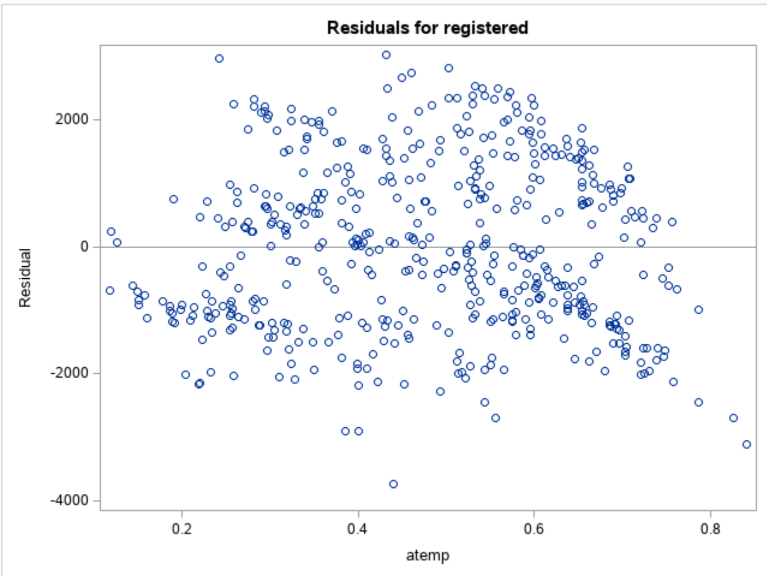
Statistically speaking, for every additional unit of atemp, there is an average increase of 5395 daily registrations. Additionally, we are 95% confident that the daily growth rate of signups is between 4678 and 6113.

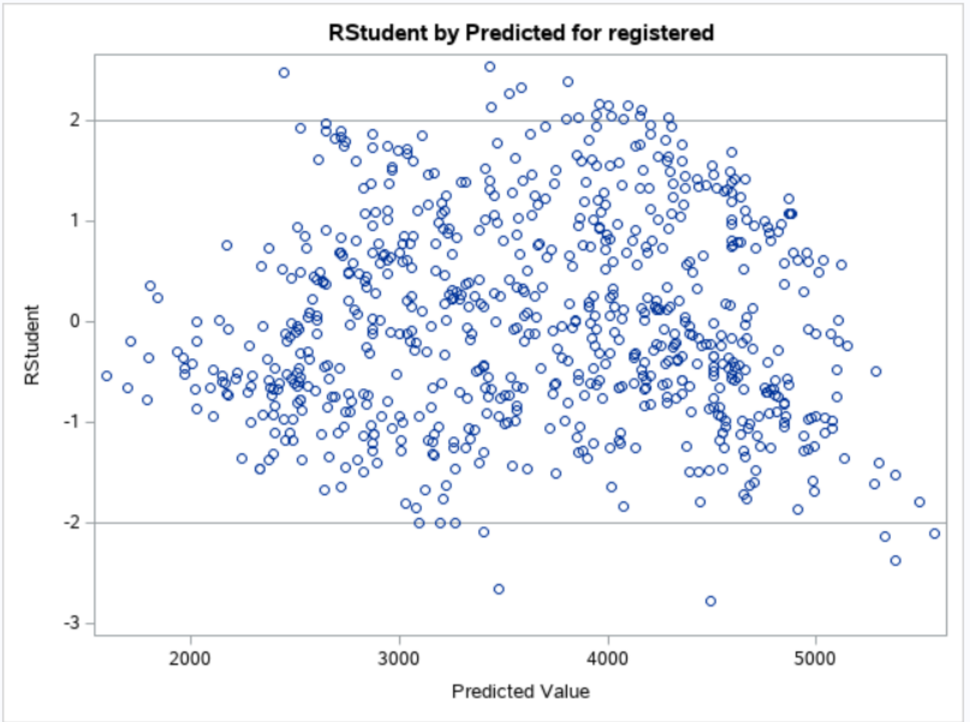
From the parameter estimation table, significant at the 5% level (p-value < 0.001), slope t=14.78, degrees of freedom=498. And F=218.42, and at the same time p-value<0.0001 also confirms this.

what is confirmed? Model is statistically signfiicant; significant relationship between atemp and registered...

The REG Procedure
Model: MODEL1
Dependent Variable: registered







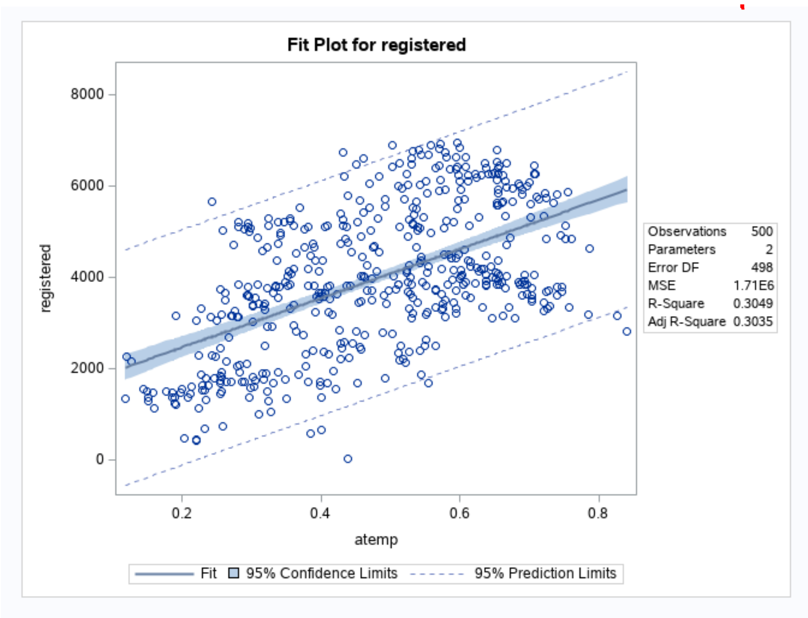
The residual value and the predicted value satisfy the linearity and error independence. According to the residual of the predicted value, the residual and predicted value do not show an obvious pattern or direction. However, homoscedasticity behaves differently, requiring further testing.

The QQ plot shows that the residual predicted values have a curvilinear pattern that does not satisfy normality. The histogram also shows that the residual distribution is skewed.

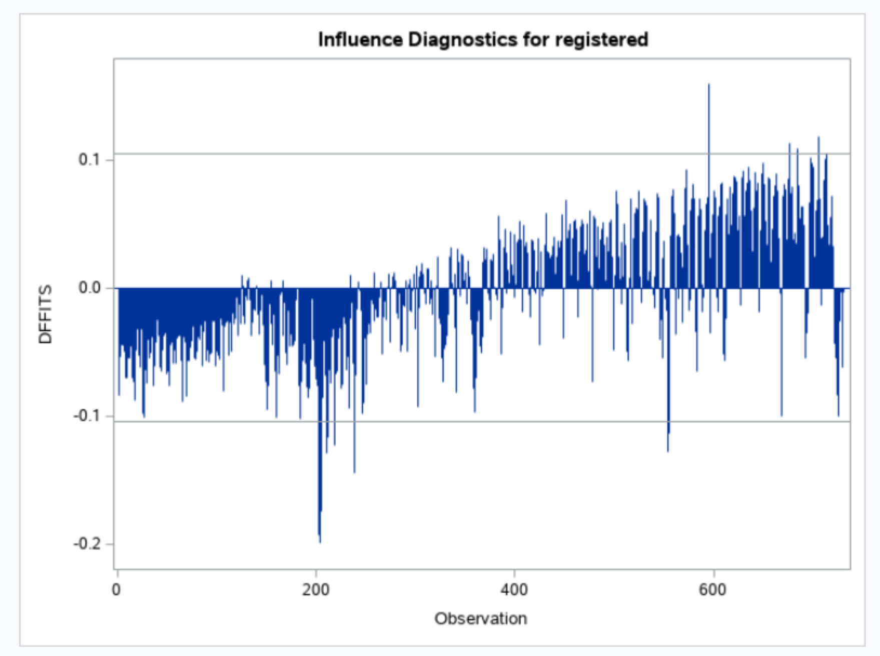
bimodal?

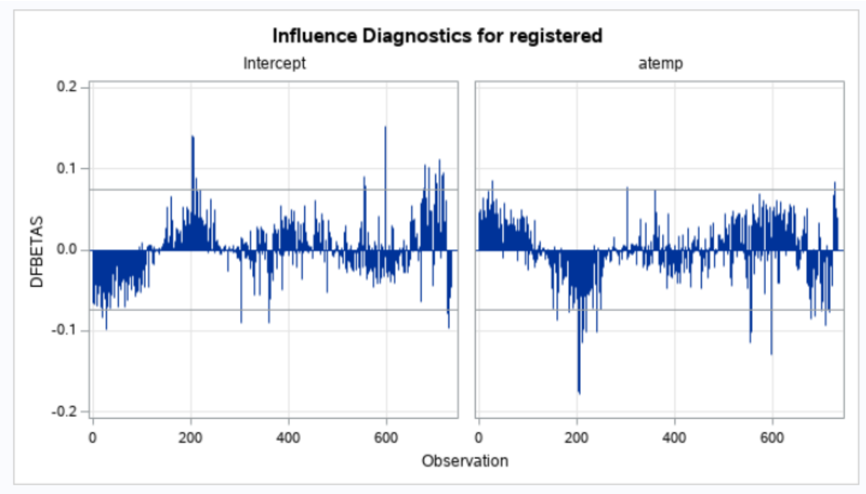
The residuals of the predicted value plots also confirm evidence of uneven vertical distribution. As the predicted value increases, the error variance increases from small to large. Violates the constant residual variance. There is still some information in the current model that has not yet been explained. The residual plot for predictor atemp shows a nonlinear relationship. The independence assumption is not satisfied.

?



The coefficient of determination = 0.3 indicates that there is a problem of under-fitting to the data, and the variable atemp occupies the weight of the change in the number of daily registrations. Adj = 0.3035 shows that the model has good generalization ability. There is a lot of data in the picture that is outside the 95% prediction interval.





There are some outliers outside 2 times the standard deviation, the 0.5% studentized residuals with absolute values > 2.5 do not appear, so there is no need to worry about these outliers. 9

From the results of hypothesis checking, further improvements to the model are required.

(20 marks) In this question, we investigate observations where workingday=1. Extend your multiple regression model for registered on working day by including the numerical and categorical predictors. In building your model consider as many potential explanatory variables as possible (you may need to define additional dummy variables). You can use stepwise selection to help you find the most parsimonious (simplest) model with the highest R-square. Be sure to check for collinearity and keep in mind that neither casual nor count should be used as explanatory variables for the total number of users. Summarise how your final model was obtained, including rationale for any modelling decisions you have made, and indicate why that final was considered the 'best'. Report and interpret your final model in detail, including a discussion of model diagnostics. Are there any observations that may require further inspection due to their influence on the model?

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/c0d6eee3-4f88-47c9-b9eb-4b5e7ae6ab38/analysisreport_citysharedbike_zhengzheng_eric_wang_resize(%E6%9C%89%E9%81%93%E6%96%87%E6%A1%A3%E7%BF%BB%E8%AF%91%E7%BB%93%E6%9E%9C).zip

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/c8d143b0-2e37-476a-82bc-40d4a9f5724c/analysisreport_citysharedbike_zhengzheng_eric_wang_resize.pdf

Correlation analysis between model residual and temp, hum, windspeed

The CORR Procedure

3 With Variables:	temp hum windspeed
1 Variables:	registered_residual

Pearson Correlation Coefficients, N = 500 Prob > |r| under H0: Rho=0	registered_residual
temp	0.02661 0.5528
hum	-0.23737 <.0001
windspeed	-0.14779 0.0009

Spearman Correlation Coefficients, N = 500 Prob > |r| under H0: Rho=0	registered_residual
temp	0.02313 0.6059
hum	-0.24681 <.0001
windspeed	-0.15040 0.0007

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		H0:Rho=Rho0 Rho0	H0:Rho=Rho0 p Value
registered_residual	temp	500	0.02661	0.02661	0.0000267	0.02658	-0.061253	0.114006	0	0.5530
registered_residual	hum	500	-0.23737	-0.24198	-0.0002378	-0.23714	-0.318217	-0.152627	0	<.0001
registered_residual	windspeed	500	-0.14779	-0.14888	-0.0001481	-0.14764	-0.232324	-0.060737	0	0.0009

Spearman Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		H0:Rho=Rho0 Rho0	H0:Rho=Rho0 p Value
registered_residual	temp	500	0.02313	0.02314	0.0000232	0.02311	-0.064714	0.110574	0	0.6060
registered_residual	hum	500	-0.24681	-0.25201	-0.0002473	-0.24658	-0.327193	-0.162397	0	<.0001
registered_residual	windspeed	500	-0.15040	-0.15155	-0.0001507	-0.15025	-0.234850	-0.063399	0	0.0007

From the pearson correlation coefficient test, it can be saw that model residual has negative correlations with hum while $r = 0.23737$ and $p < 0.001$, also has negative correlations with windspeed while $r = 0.14779$ and $p \leq 0.001$.

As for temp, its r value is only 0.02661 and its p value is 0.5530, which is not sifnificant at 5% level. H0 for correlation coefficient = 0 can not be rejected. There is no enough evidence to prove that residual has a neither positive relationship or negative relationship with temp. The same conclusion can be gained from the spearman correlation test result.

So the linear regression model will be expaned by using hum and windspeed.

The REG Procedure
Model: MODEL1
Dependent Variable: registered

Number of Observations Read	500
Number of Observations Used	500

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	460775747	153591916	99.30	<.0001
Error	496	767161463	1546696		
Corrected Total	499	1227937210			

Root MSE	1243.66253	R-Square	0.3752
Dependent Mean	3978.25000	Adj R-Sq	0.3715
Coeff Var	31.26155		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	1	3788.82986	366.65517	10.33	<.0001	0	0	3068.44109	4509.21864
atemp	1	5386.36617	352.54194	15.28	<.0001	0.55125	1.03346	4693.70647	6079.02587
hum	1	-2705.99644	407.54203	-6.64	<.0001	-0.24372	1.06965	-3506.71802	-1905.27486
windspeed	1	-3648.02224	750.44534	-4.86	<.0001	-0.17961	1.08377	-5122.46593	-2173.57855

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	atemp	hum	windspeed
1	3.76466	1.00000	0.00160	0.00637	0.00301	0.00827
2	0.15180	4.97996	0.00036367	0.15378	0.02097	0.59293
3	0.06743	7.47192	0.01673	0.70653	0.28060	0.07143
4	0.01611	15.28874	0.98131	0.13332	0.69542	0.32737

what does this number indicate?

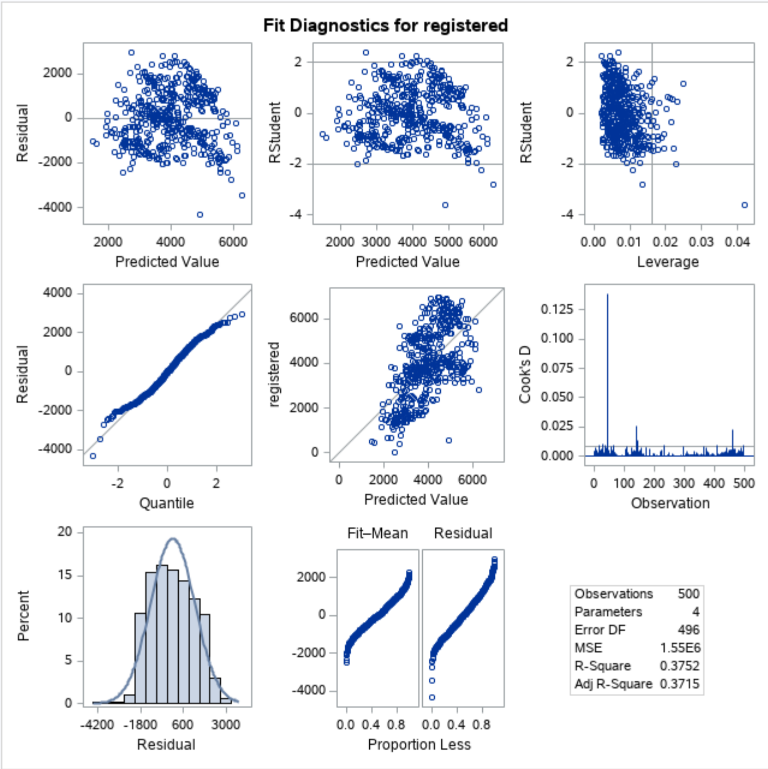
The above table shows the result of analysis of variance and R-square for multiple linear regression model with atemp, hum and windspeed.

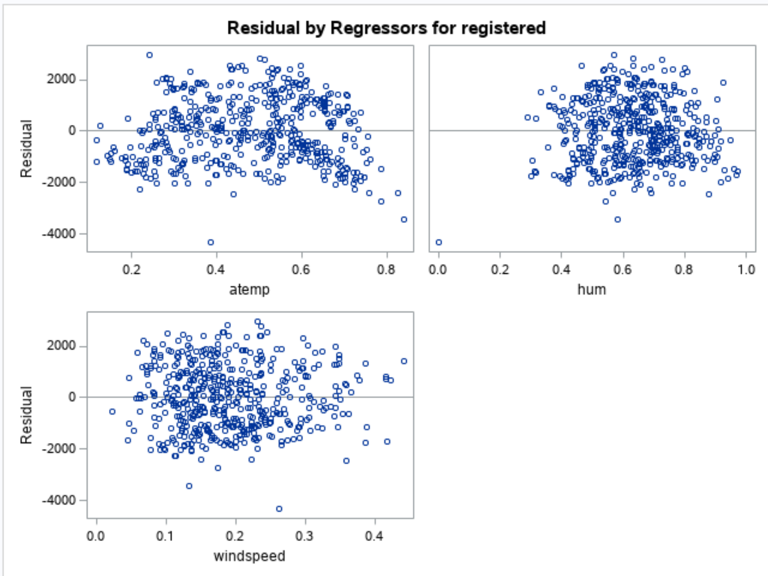
F value is 99.30 with p value < 0.001. Model's DF is 3 with 496 degrees of freedom. The R-square value of this model is 0.3752 which is improved from the original linear regression. Adj R-Sq value is 0.3715 which is also increased from the original linear regression. The generalization ability of the model has decreased, but overall it is still very good.

According to the estimated parameter table:

$$registered = 3788 + 5386 * atemp - 2705 * hum - 3648 * windspeed$$

Statistically speaking, on average, registered each day will increase 5386 fro each unit increase for atemp, decrease 2705 for each unit increase for hum and decrease 3648 for each unit increase for windspeed.

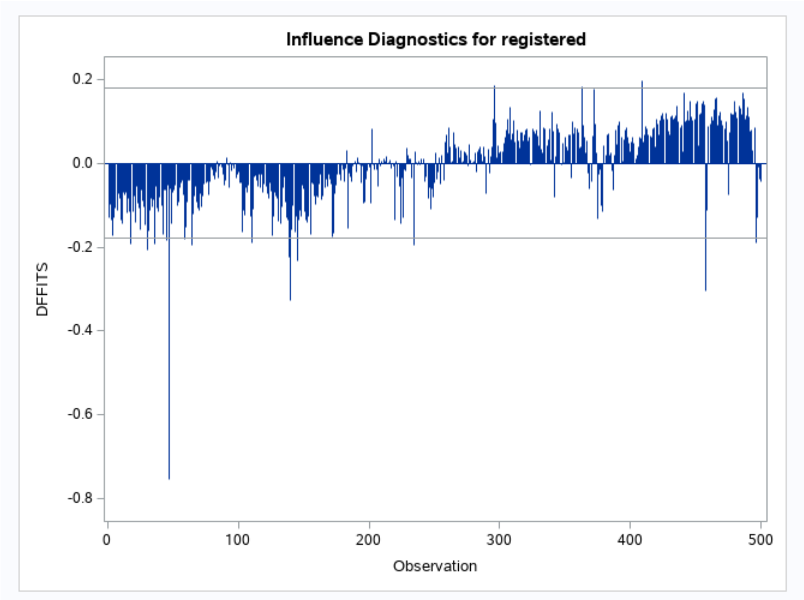


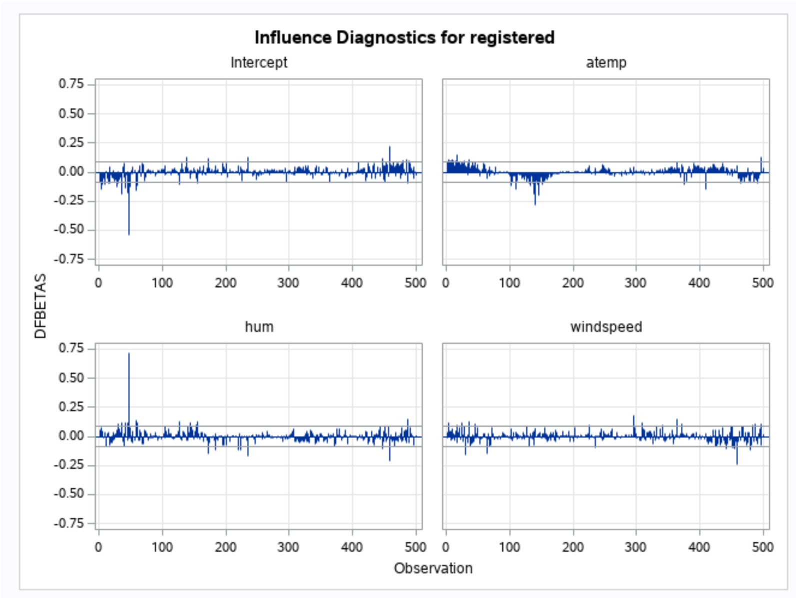


There is an unequal vertical distribution from left to right, which does not satisfy the constant error variance.

Decomposed into residuals according to the regressor plot, the residuals of atemp show non-linear relationships and unequal vertical error diffusion. Further research is needed on the variable temperature.

No patterns related to humidity and wind speed were found, and error variance spread seems to be good for humidity and wind speed.





Based on the updated equation, use atemp, hum and windspeed as the dependent variables for the registered prediction, and convert the season, month, etc. into dummy variables, and also need to create a dummy variable for P_Holoday to reduce multicollinearity.

?

The REG Procedure
Model: MODEL1
Dependent Variable: registered

R-Square Selection Method

Number of Observations Read	500
Number of Observations Used	500

Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	0.5252	0.5242	573.0477	dteday
1	0.4291	0.4280	789.1868	yr
1	0.3049	0.3035	1068.952	atemp
2	0.7200	0.7189	136.2782	atemp dteday
2	0.7183	0.7171	140.2793	dteday temp
2	0.6996	0.6984	182.2707	atemp yr
3	0.7614	0.7599	45.2049	atemp hum dteday
3	0.7574	0.7559	54.1369	hum dteday temp
3	0.7384	0.7368	96.9517	atemp dteday yr
4	0.7716	0.7697	24.2237	atemp hum windspeed dteday
4	0.7700	0.7681	27.9006	hum windspeed dteday temp
4	0.7686	0.7667	30.9587	atemp hum dteday yr
5	0.7805	0.7783	6.1549	atemp hum windspeed dteday yr
5	0.7789	0.7767	9.7602	hum windspeed dteday yr temp
5	0.7720	0.7697	25.2321	atemp hum windspeed dteday temp
6	0.7810	0.7784	7.0000	atemp hum windspeed dteday yr temp

From the above table, the last R-square = 0.7805 and R-square = 0.7810 are very close, and the cp values of the two are also very close to the number of predictors, so here are two candidate models. One of them has 5 variables, while the other has 6 variables. The difference is whether the temp variable is included or not.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	959042632	159840439	293.06	<.0001
Error	493	268894578	545425		
Corrected Total	499	1227937210			

Root MSE	738.52902	R-Square	0.7810
Dependent Mean	3978.25000	Adj R-Sq	0.7784
Coeff Var	18.56417		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	1	-61939	6308.76903	-9.82	<.0001	0	0	-74335	-49544
atemp	1	3106.55960	1423.85197	2.18	0.0296	0.31793	47.80493	308.99301	5904.12619
hum	1	-2263.60173	249.44512	-9.07	<.0001	-0.20387	1.13636	-2753.70839	-1773.49508
windspeed	1	-2403.91088	455.33338	-5.28	<.0001	-0.11835	1.13143	-3298.54423	-1509.27754
dteday	1	3.43889	0.33727	10.20	<.0001	0.46198	4.62159	2.77623	4.10154
yr	1	633.20446	140.77424	4.50	<.0001	0.20203	4.54173	356.61298	909.79593
temp	1	1353.01668	1259.00129	1.07	0.2830	0.15601	47.44386	-1120.65336	3826.68672

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation						
			Intercept	atemp	hum	windspeed	dteday	yr	temp
1	6.21850	1.00000	6.942543E-7	0.00005117	0.00102	0.00280	6.745897E-7	0.00165	0.00005833
2	0.46457	3.65862	6.330131E-7	0.00007729	0.00212	0.00525	4.620811E-7	0.21417	0.00009112
3	0.19545	5.64063	0.00000199	0.00224	0.00000522	0.34024	0.00000178	4.413483E-7	0.00291
4	0.09689	8.01122	0.00001967	0.00190	0.16539	0.34218	0.00001890	0.00019057	0.00321
5	0.02345	16.28410	0.00029777	0.00006007	0.79532	0.26286	0.00027605	0.01268	0.00056890
6	0.00113	74.24691	0.00002287	0.99567	0.00096459	0.02552	0.00002507	0.00007702	0.99221
7	0.00001350	678.75684	0.99966	0.00000382	0.03518	0.02115	0.99968	0.77124	0.00094502

From the results of the data, the value of R-Square has been raised to 0.7810, while the value of Adj R-Square has become 0.7784.

But by observing the data in the Parameter estimate table, we can find that there is multicollinearity between the VIF values of temp and atemp between 47.80493 and 47.44386. According to the initial list, there is a very strong positive linear correlation between atemp and temp. The p-value of temp is significantly higher than the p-value of atemp, so we can consider eliminating the temp variable from the model.

The REG Procedure
Model: MODEL1
Dependent Variable: registered

Number of Observations Read	500
Number of Observations Used	500

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	958412706	191682541	351.33	<.0001
Error	494	269524503	545596		
Corrected Total	499	1227937210			

Root MSE	738.64481	R-Square	0.7805
Dependent Mean	3978.25000	Adj R-Sq	0.7783
Coeff Var	18.56708		

Parameter Estimates									
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	1	-62175	6305.94200	-9.86	<.0001	0	0	-74565	-49785
atemp	1	4619.37679	213.90478	21.60	<.0001	0.47275	1.07857	4199.10144	5039.65214
hum	1	-2273.11422	249.32711	-9.12	<.0001	-0.20473	1.13493	-2762.98658	-1783.24187
windspeed	1	-2331.18800	450.34752	-5.18	<.0001	-0.11477	1.10644	-3216.02078	-1446.35523
dteday	1	3.44847	0.33720	10.23	<.0001	0.46327	4.61836	2.78595	4.11100
yr	1	630.54488	140.77456	4.48	<.0001	0.20118	4.54033	353.95416	907.13559

Collinearity Diagnostics								
			Proportion of Variation					
Number	Eigenvalue	Condition Index	Intercept	atemp	hum	windspeed	dteday	yr
1	5.29794	1.00000	9.605805E-7	0.00304	0.00141	0.00400	9.330229E-7	0.00234
2	0.45683	3.40548	0.00000108	0.00332	0.00305	0.00982	8.399E-7	0.21293
3	0.15172	5.90933	3.675082E-7	0.13848	0.01996	0.59009	4.102529E-7	0.00098415
4	0.07036	8.67715	0.00001602	0.75083	0.17781	0.09800	0.00001518	0.00021489
5	0.02314	15.13058	0.00030980	0.06843	0.76292	0.27779	0.00028688	0.01227
6	0.00001351	626.21294	0.99967	0.03589	0.03485	0.02031	0.99970	0.77126

We can find that the values of R-square and adj r-square have not changed much.

Based on the data from parameter estimates, we can now modify the formula to:

$$registered = -62175 + 4619 * atemp - 2273 * hum - 2331 * windspeed + 3 * dteday + 630 * yr$$

The REG Procedure
Model: stepwise
Dependent Variable: registered

Number of Observations Read	500
Number of Observations Used	500

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	999515006	124939376	268.56	<.0001
Error	491	228422204	465218		
Corrected Total	499	1227937210			

Root MSE	682.06916	R-Square	0.8140
Dependent Mean	3978.25000	Adj R-Sq	0.8109
Coeff Var	17.14495		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-4572.28892	10499	-0.44	0.6634	0
summer	1	-29.16419	109.92547	-0.27	0.7909	2.51112
fall	1	485.22553	139.25637	3.48	0.0005	3.82321
winter	1	-798.03400	114.31462	-6.98	<.0001	2.56178
atemp	1	4080.69634	324.45205	12.58	<.0001	2.91020
hum	1	-2324.71915	233.67563	-9.95	<.0001	1.16916
windspeed	1	-2483.50358	421.45702	-5.89	<.0001	1.13646
dteday	1	0.40731	0.56076	0.73	0.4680	14.97904
yr	1	1745.91345	214.91795	8.12	<.0001	12.41077

It can be found that the p-values of the three variables Intercept, summer and dteday are greater than 0.05, and all other variables are statistically significant at the 5% level. In addition, by checking the VIF, it can be found that the VIF coefficients of dteday and yr are both higher than 10, there is multilinearity between them, and the two variables dteday and yr will be considered to be removed.

why removing both of them. Should remove one only.

The REG Procedure
Model: stepwise
Dependent Variable: registered

Number of Observations Read	500
Number of Observations Used	500

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	563607643	93934607	69.71	<.0001
Error	493	664329567	1347524		
Corrected Total	499	1227937210			

Root MSE	1160.82922	R-Square	0.4590
Dependent Mean	3978.25000	Adj R-Sq	0.4524
Coeff Var	29.17939		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4259.12965	420.91974	10.12	<.0001	0
summer	1	-164.06908	164.12316	-1.00	0.3180	1.93255
fall	1	699.16983	161.35289	4.33	<.0001	1.77203
winter	1	-628.73707	188.56318	-3.33	0.0009	2.40642
atemp	1	5116.55822	549.16825	9.32	<.0001	2.87841
hum	1	-3341.06375	389.80514	-8.57	<.0001	1.12321
windspeed	1	-3203.81083	716.16836	-4.47	<.0001	1.13292

model will be:

$$registered = 4259 - 164 * summer + 699 * fall - 628 * winter + 5116 * atemp - 3341 * hum - 3203 * windspeed$$

Then, we will use month dummy variable:

The REG Procedure
Model: MODEL1
Dependent Variable: registered

Number of Observations Read	500
Number of Observations Used	500

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	593747160	34926304	26.54	<.0001
Error	482	634190049	1315747		
Corrected Total	499	1227937210			

Root MSE	1147.06015	R-Square	0.4835
Dependent Mean	3978.25000	Adj R-Sq	0.4653
Coeff Var	28.83328		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4346.78037	472.25103	9.20	<.0001	0
summer	1	-111.89860	324.59371	-0.34	0.7304	7.74168
fall	1	498.52651	380.62940	1.31	0.1909	10.09920
winter	1	-746.26590	312.59575	-2.39	0.0174	6.77312
atemp	1	5864.31505	738.68357	7.94	<.0001	5.33363
hum	1	-3902.14851	409.84545	-9.52	<.0001	1.27166
windspeed	1	-3139.85229	715.29851	-4.39	<.0001	1.15746
January	1	182.72598	296.45059	0.62	0.5379	2.45799
February	1	117.04449	283.02578	0.41	0.6794	2.18915
April	1	-271.48812	322.19196	-0.84	0.3999	2.90339
May	1	109.87670	337.33882	0.33	0.7448	3.39919
June	1	-425.37574	357.75650	-1.19	0.2350	3.82311
July	1	-800.66525	450.72486	-1.78	0.0763	5.81135
August	1	-249.07843	431.09747	-0.58	0.5637	5.89960
September	1	445.87549	392.10850	1.14	0.2561	4.30019
October	1	190.77983	405.36196	0.47	0.6381	4.80461
November	1	40.71795	407.05875	0.10	0.9204	4.63436
December	1	203.00320	341.52492	0.59	0.5525	3.33656

Model can be modified as：

$$registered = 4346 - 111 * summer + 498 * fall - 746 * winter + 5864 * atemp - 3902 * hum - 3139 * windspeed + 182 * January + 117 * February - 371 * April + 109 * May - 425 * June - 800 * July - 249 * August + 445 * September + 190 * October + 40 * November + 203 * December$$

Some wrong steps in model selection.
Did not interpret in details the final model: model equation, how good is the fit, LINE assumptions, influential observations?

13

The REG Procedure
Model: MODEL1
Dependent Variable: registered

Number of Observations Read	500
Number of Observations Used	500

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	596444540	28402121	21.50	<.0001
Error	478	631492670	1321114		
Corrected Total	499	1227937210			

Root MSE	1149.39740	R-Square	0.4857
Dependent Mean	3978.25000	Adj R-Sq	0.4631
Coeff Var	28.89204		

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

spring =	Intercept - summer - fall - winter
Friday =	Intercept - Monday - Tuesday - Wednesday - Thursday

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	B	4328.74883	479.37068	9.03	<.0001
summer	B	-120.97895	325.44754	-0.37	0.7103
fall	B	507.17563	381.67470	1.33	0.1845
winter	B	-742.67836	313.34959	-2.37	0.0182
spring	0	0	.	.	.
atemp	1	5799.68065	742.30095	7.81	<.0001
hum	1	-3886.42534	413.54876	-9.40	<.0001
windspeed	1	-3152.90880	717.69384	-4.39	<.0001
January	1	172.32912	297.23278	0.58	0.5623
February	1	106.75140	283.73481	0.38	0.7069
April	1	-259.41667	323.03608	-0.80	0.4223
May	1	118.58017	338.23073	0.35	0.7261
June	1	-401.73889	359.20465	-1.12	0.2640
July	1	-763.12171	452.85386	-1.69	0.0926
August	1	-218.93359	432.92448	-0.51	0.6133
September	1	458.40961	393.58535	1.16	0.2447
October	1	190.47546	406.50749	0.47	0.6396
November	1	32.48830	408.13598	0.08	0.9366
December	1	197.21167	342.50688	0.58	0.5650
Monday	B	-102.36742	166.83098	-0.61	0.5398
Tuesday	B	51.70724	161.53457	0.32	0.7490
Wednesday	B	99.70691	161.36933	0.62	0.5369
Thursday	B	100.39539	161.52279	0.62	0.5345
Friday	0	0	.	.	.

(d) (20 marks) In this question, we investigate observations where workingday=0. Build a multiple regression model for registered on non-working day, similar to question (c).

The REG Procedure
Model: MODEL1
Dependent Variable: registered

Number of Observations Read	231
Number of Observations Used	231

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	127044640	127044640	112.77	<.0001
Error	229	257996886	1126624		
Corrected Total	230	385041526			

Root MSE	1061.42544	R-Square	0.3300
Dependent Mean	2959.03463	Adj R-Sq	0.3270
Coeff Var	35.87067		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	912.80997	204.95738	4.45	<.0001
atemp	1	4430.47971	417.21707	10.62	<.0001

Where workingday = 0, its r-square = 0.3300 and adj R-square = 0.3270.

Its model will be:

$registered = 912 + 4430 * atemp$

The REG Procedure
Model: MODEL1
Dependent Variable: registered

R-Square Selection Method

Number of Observations Read	231
Number of Observations Used	231

Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	0.3300	0.3270	18.2217	atemp
1	0.3181	0.3151	22.5580	temp
1	0.0669	0.0629	114.4821	windspeed
2	0.3474	0.3417	13.8249	atemp windspeed
2	0.3469	0.3411	14.0291	atemp hum
2	0.3417	0.3359	15.9210	temp atemp
3	0.3744	0.3661	5.9558	atemp hum windspeed
3	0.3659	0.3575	9.0568	temp atemp hum
3	0.3653	0.3569	9.2813	temp hum windspeed
4	0.3825	0.3715	5.0000	temp atemp hum windspeed

From the results, the model with the best behavior indicated by the red arrow in the picture. Therefore, we will use four variables temp, atemp, num and windspeed.

The REG Procedure
Model: MODEL1
Dependent Variable: registered

Number of Observations Read	231
Number of Observations Used	231

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	147267682	36816920	34.99	<.0001
Error	226	237773844	1052097		
Corrected Total	230	385041526			

Root MSE	1025.71763	R-Square	0.3825
Dependent Mean	2959.03463	Adj R-Sq	0.3715
Coeff Var	34.66393		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2110.64571	473.94061	4.45	<.0001	0
temp	1	-8436.96717	4907.36741	-1.72	0.0869	184.66108
atemp	1	13836	5520.63203	2.51	0.0129	187.48944
hum	1	-1647.34497	491.91041	-3.35	0.0010	1.11296
windspeed	1	-2366.61719	961.62271	-2.46	0.0146	1.23602

The P-value is less than 0.001 and the corresponding F-statistic is F = 34 with 4 and 226 degrees of freedom.

The coefficient of determination R^2 is 0.3825, indicating that the chosen variables together explain 38.25% of overall variability in registered on non-workingday.

A updated model could be gained as below:

$$registered = 2110 - 8436 * temp + 13836 * atemp - 1647 * hum - 2366 * windspeed$$

All variables are highly statistically significant except for temp. The p-value for that coefficient estimate is 0.0869, which means that the hypothesis may beta for temp is zero can not be rejected.

The variance inflation for temp is 184.66108 and it is 187.48944 for atemp, which indicates that these two variables are highly correlated and there is multicollinearity in that model. One of those two variables is enough for the model. Because the p-value of temp is 0.0869, which is higher than the p-value of atemp, so temp variable is removed from the model.

The REG Procedure
Model: MODEL1
Dependent Variable: registered

Number of Observations Read	231
Number of Observations Used	231

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	144157891	48052630	45.28	<.0001
Error	227	240883635	1061161		
Corrected Total	230	385041526			

Root MSE	1030.12688	R-Square	0.3744
Dependent Mean	2959.03463	Adj R-Sq	0.3661
Coeff Var	34.81294		

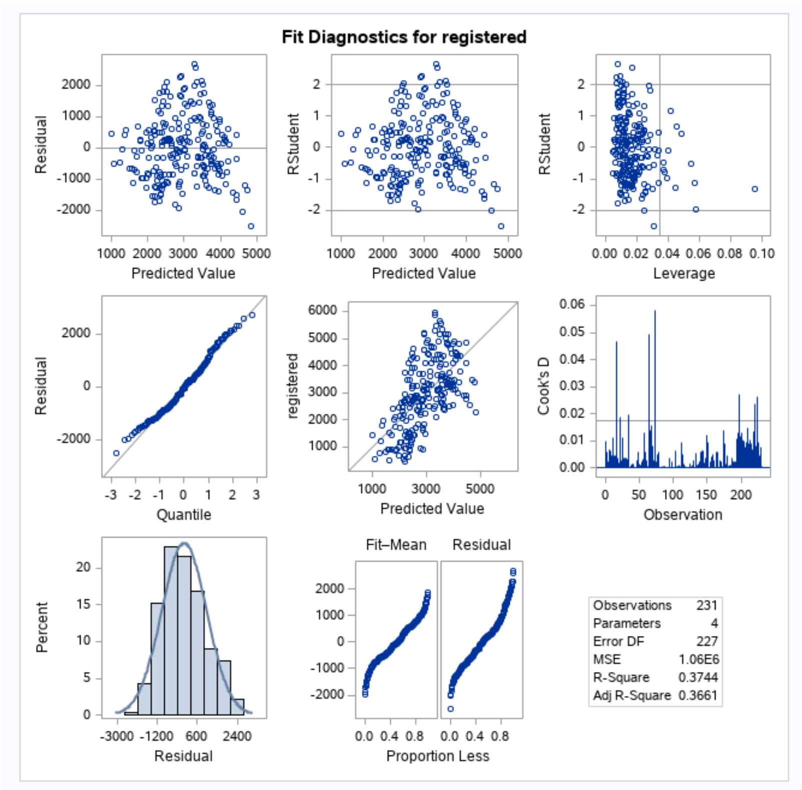
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2450.34944	432.63987	5.66	<.0001	0
atemp	1	4371.95630	419.57427	10.42	<.0001	1.07372
hum	1	-1530.52883	489.28951	-3.13	0.0020	1.09172
windspeed	1	-2893.12811	915.47320	-3.16	0.0018	1.11066

When we remove the temp variable, we can see that the values of R-square and Adj R-sq have not changed much. The remaining variables can explain 37.44% of overall variability in registered on non-workingday.

The model can be updated as below:

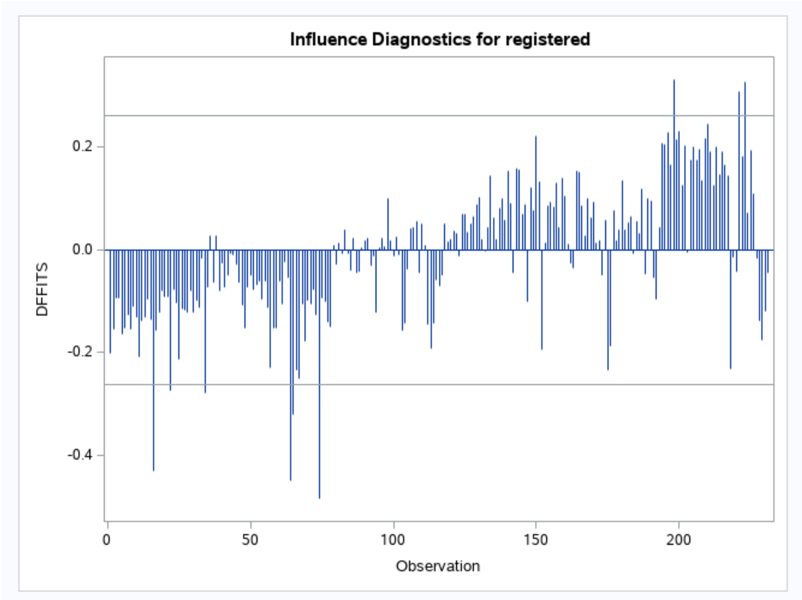
$$registered = 2450 + 4371 * atemp - 1530 * hum - 2893 * windspeed$$

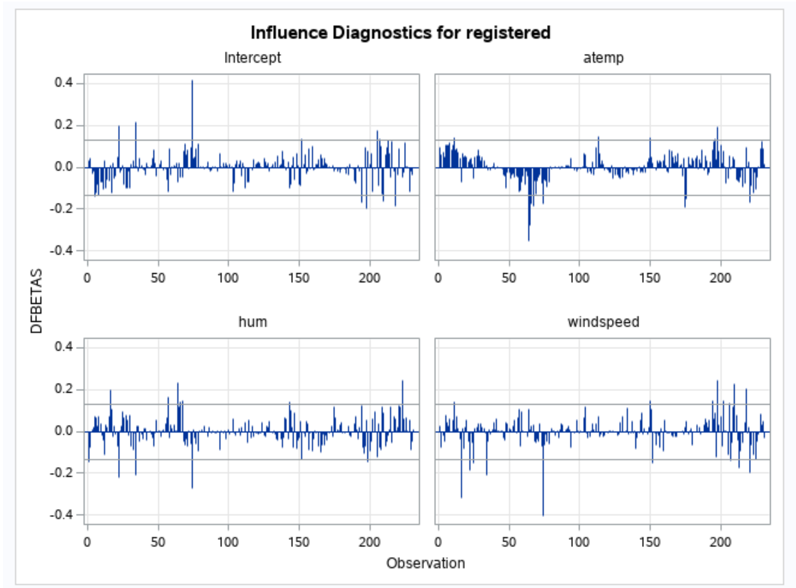
Those coefficient are relatively highly statistically significant according to their p-values. Statistically speaking, on average, a 1 unit increase in atemp will increase registered by 4371, a 1 unit increase in hum will decrease registered by 1530, and windspeed will decrease by 2893 for every 1 unit increase in windspeed.



The graph above shows the regression diagnostics, and the residuals plot shows that there are quite a few outliers and influential observations.

From the histogram and the Q-Q plot, residuals are slightly skewed to the right hence non-Normal.





By looking at the DFFITS plot, it is not difficult to see that there are many observations with a DFITS value 2 to 3 times higher than 0.23. These observations may have larger implications.

The REG Procedure
Model: MODEL1
Dependent Variable: registered

Number of Observations Read	231
Number of Observations Used	231

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	162280236	11591445	11.24	<.0001
Error	216	222761289	1031302		
Corrected Total	230	385041526			

Root MSE	1015.53053	R-Square	0.4215
Dependent Mean	2959.03463	Adj R-Sq	0.3840
Coeff Var	34.31966		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2634.31257	439.34605	6.00	<.0001	0
temp	1	3838.51378	407.80920	9.41	<.0001	1.30095
April	1	111.50056	247.98972	0.45	0.6534	1.08939
March	1	12.13114	269.42637	0.05	0.9641	1.10852
May	1	79.07054	254.55010	0.31	0.7564	1.09556
hum	1	-1652.50526	493.56052	-3.35	0.0010	1.14303
windspeed	1	-3505.50042	931.37391	-3.76	0.0002	1.18287
October	1	789.60291	247.29979	3.19	0.0016	1.08333
December	1	151.33250	262.47132	0.58	0.5648	1.27528
Monday	1	-19.81540	281.83922	-0.07	0.9440	1.08031
Tuesday	1	-1525.11765	1044.12926	-1.46	0.1456	1.05254
Wednesday	1	540.81352	1030.05688	0.53	0.6001	1.02436
Thursday	1	-1481.64887	735.04933	-2.02	0.0451	1.03872
Friday	1	408.17150	735.39841	0.56	0.5794	1.03971
Saturday	1	214.91366	140.42143	1.53	0.1274	1.09504

12

Similar issues with previous question
Lack comparison b/w working and non-working

Conslusion

Write a summary of your findings from Questions 1 and 2. Keep the technical details of the analyses that led you to these conclusions to the absolute minimum. Rather, focus on practical significance and present your findings in non-specialist terms. One to two paragraphs (up to a page) will be sufficient.

The number of daily registered users in a year will move within a very large range. In the most extreme cases, there are only 20 registered users a day, and in the most cases, there will be nearly 7,000 registered users. The three elements of atemp, hum and windspeed have obvious influence on whether to register.

In terms of years, registrations in 2012 were generally higher than those in 2011. Seasonally, single-day signups in summer are generally higher than in the other three seasons, while fall and spring sign ups closely follow the climate and are generally about the same.

Temporary users have a greater impact on registered users, but this variable is excluded here. Seasons have an effect on registered users, because the temperature is different in different seasons, the windspeed is different, and people have different atemp because of it. In addition, working days, weekends and holidays need to be considered. User registrations will be up on Fridays, likely due to people needing to use their bikes on weekends. And people's registrations for public holidays didn't rise as much as originally thought.

16