STAT 512 APPLIED REGRESSION ANALYSIS

CARBON MONOXIDE FROM A FREEWAY

FINAL TERM PROJECT – SPRING 2015 : STAT 512

-RATIK DUGAR

Table of Contents

Introduction

Hourly carbon monoxide (CO) averages were recorded on summer weekdays at a measurement station in Los Angeles. The station was established by the Environmental Protection Agency as part of a larger study to assess the effectiveness of the catalytic converter. It was located about 25 feet from the San Diego Freeway, which in this particular area is located at 145 degrees north. It was located such that winds from 145 to 325 degrees (which in the summer are the prevalent wind directions during the daylight hours) transport the CO emissions from the highway toward the measurement station. Aggregate measurements were recorded for each hour of the day 1 to 24.

Hour	-	hour of the day, from midnight to midnight
CO	-	average summer weekday CO concentration (parts per million)
TD	-	average weekday traffic density (traffic count/traffic speed)
WS	-	average perpendicular wind-speed component,

Discussion on Problems

QUESTION 1.

***Use any technique we learned in class to determine if the response variable and/or any of the explanatory variables need to be transformed. If you decide that a variable needs to be transformed then transform it and keep it in the full model for the rest of the questions*.**

SOLUTION:

In order to check for the needed transformation of response variable, box-cox transformation code was run in SAS to see whether there is any need to transform response variable, y or not. Since this analysis suggested an optimal value of λ equal to 0.75, the response variable is not transformed as λ is closer to 1.

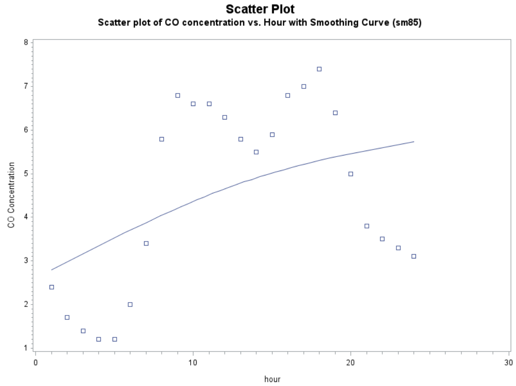
The scatter plots for all the explanatory variables i.e. hour, traffic, wind were The two explanatory variables i.e. hour and wind were transformed based on observing the diagnostic plots for the regression analysis. The potential violations of the assumptions that were found are:

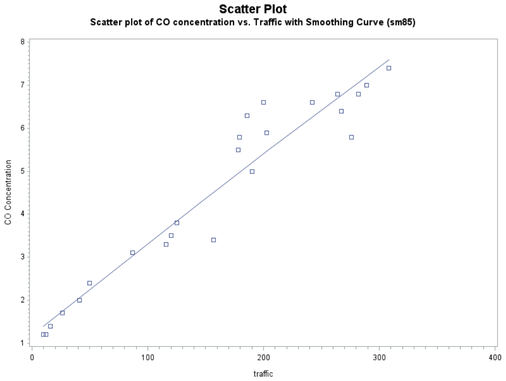
1. Errors terms are dependent. (due to the cyclic nature of the plot of residual vs observations)
2. The residual plot against wind indicates that this term should probably be quadratic in nature.

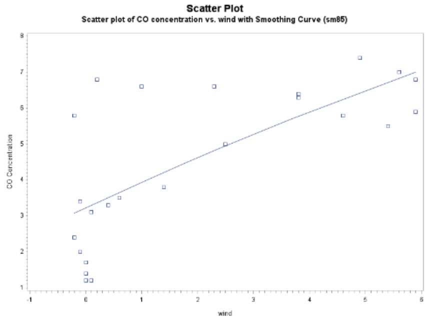
In order to fix these violations, a quadratic term of wind i.e. wind^2 and cosine term of hour i.e. cos ((4 * pi)/24 * Hour) were created to do transformations for wind and hour. These new terms were introduced into the model. The guidance to do so was taken from the data description on the source website (from where dataset was taken).

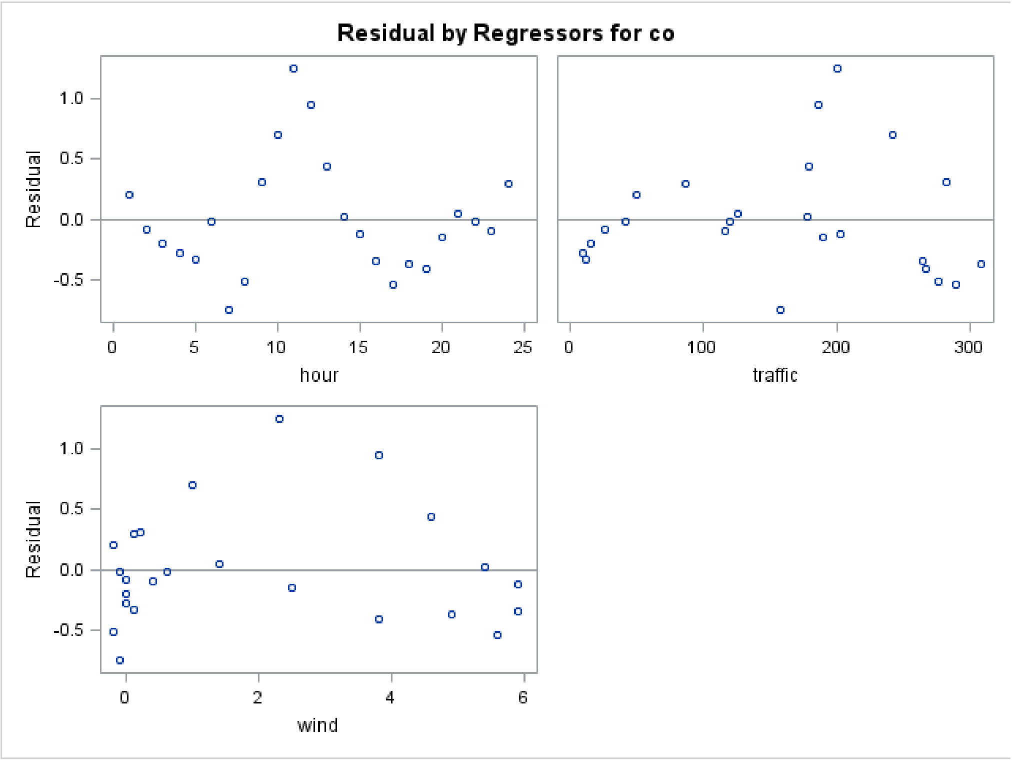
Through proc univariate procedure, the qq-plot was studied for normality assumption and the data was found to be fairly normal.

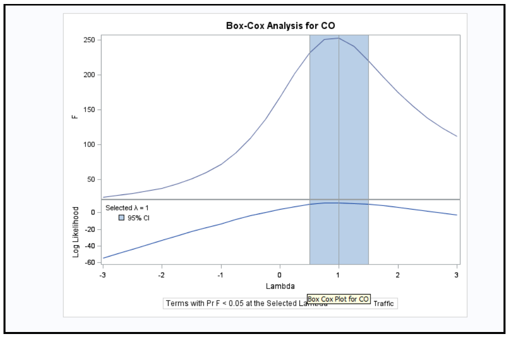
OUTPUT FROM SAS:



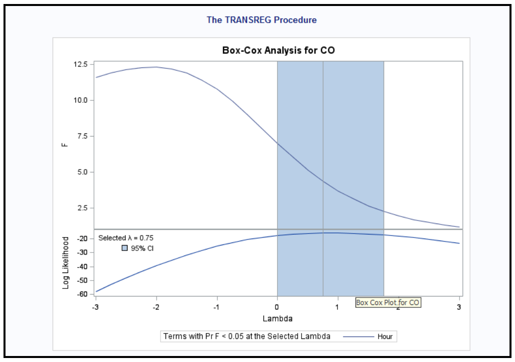




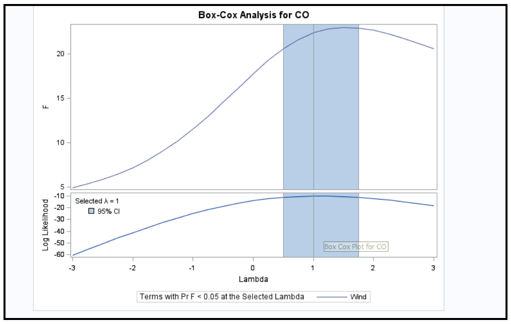




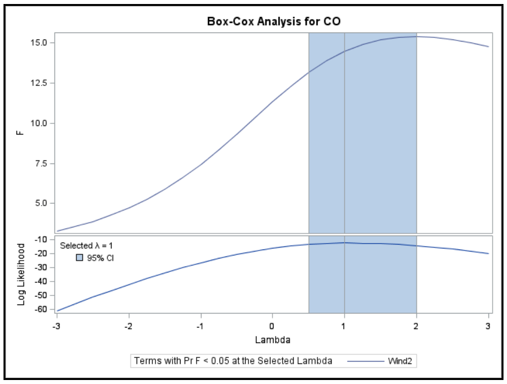
No need to transform Y.



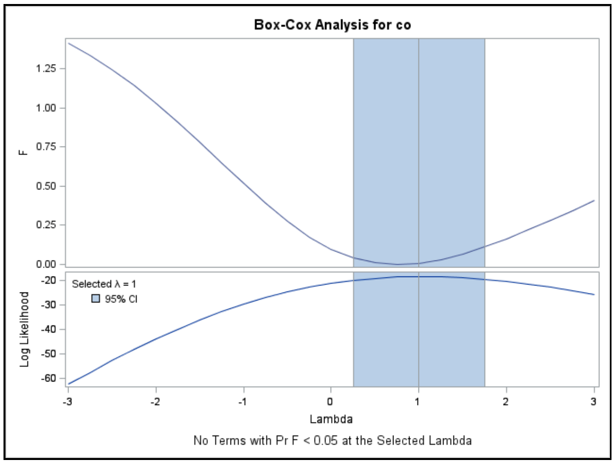
Box-cox procedure suggestsλ=0.75 but p value is greater than 0.05 so it is not needed to do transformation for response variable with Hour.



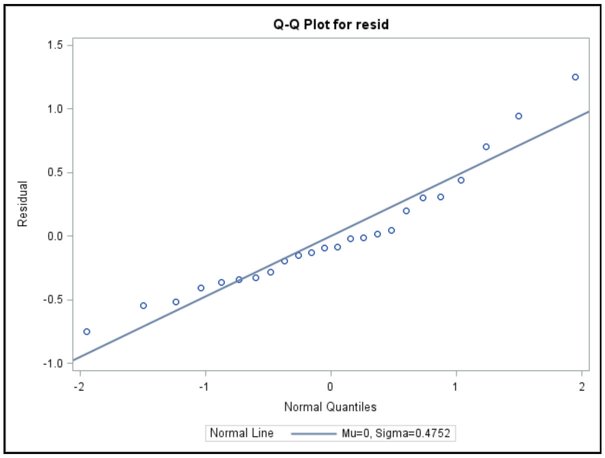
Not needed to transform Y for Wind



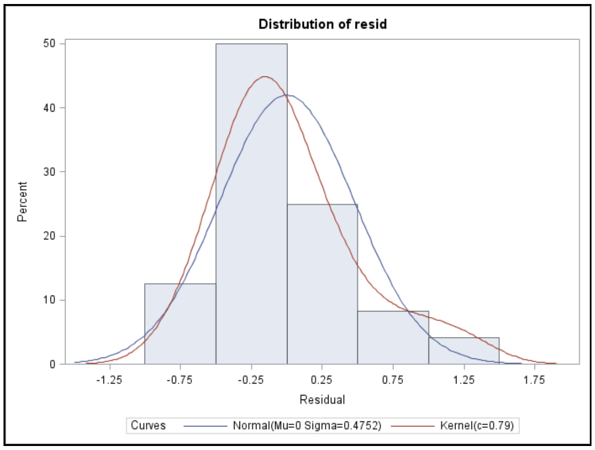
Even for Wind^2, transformation is not required.



No transformation of response variable required for new transformed variable “thour”.



The qq-plot for normality assumption shows that there is some problem with normality which has been fixed after the transformation of explanatory variables. The new plots for new best model have been presented in Problem 8.



Histogram shows that data is slightly skewed on right.

QUESTION 2.

Select a predictor and use a piecewise simple linear regression to model the relationship with the response variable. Be smart about choosing the point at which the two piece change slope, if none of the predictors has nonlinear relationship with the predicted variable then the center should be your point where the two pieces meet. Determine whether the two lines are parallel, the same or have the same intercept.

SOLUTION:

Background

Segmented regression, also known as piecewise regression or 'broken-stick regression', is a method in regression analysis in which the independent variable is partitioned into intervals and a separate line segment is fit to each interval. Segmented regression analysis can also be performed on multivariate data by partitioning the various independent variables. Segmented regression is useful when the independent variables, clustered into different groups, exhibit different relationships between the variables in these regions. The boundaries between the segments are breakpoints. Segmented linear regression is segmented regression whereby the relations in the intervals are obtained by linear regression.

From the previous problem, it was found that there exists that wind has a quadratic relationship with the response variable. So the explanatory variable “wind” was chosen to model the relationship with the response variable.

Same line: H_0 : β_1=β_3=0   Ha: at least one is not zero

p value= 0.2662 : fail to reject so β_1=β_3=0   This is same line.

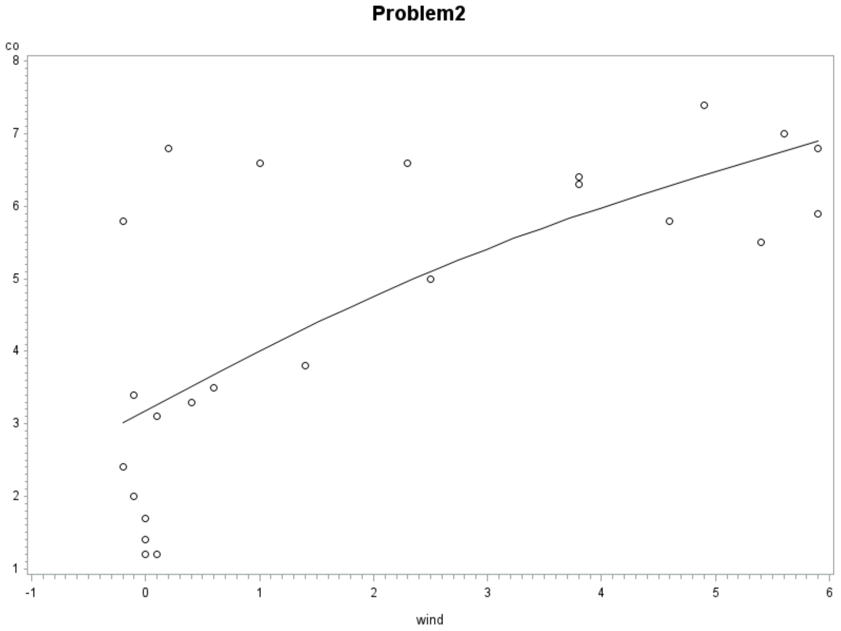
Parallel: H_0 : β_3=0   Ha: β_3 is not zero

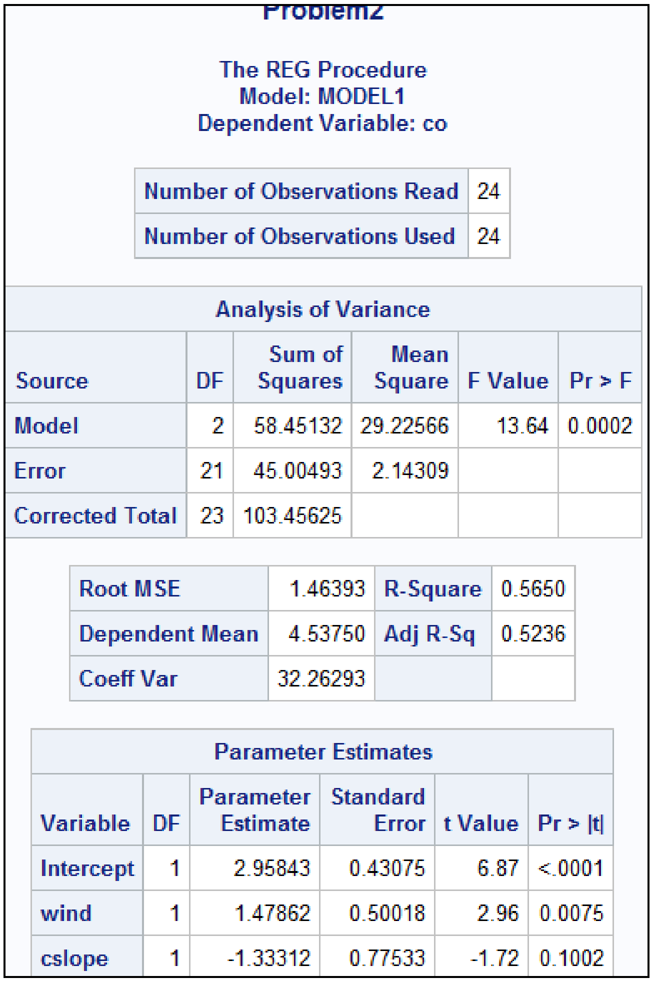
P value= 0.9115 : fail to reject so β_3=0   This is parallel.

Intercept: H_0 : β_1=0   Ha: β_1 is not zero

P value=0.6622 : fail to reject null, So β_1=0   The intercept is the same.

OUTPUT FROM SAS:





Problem2

The REG Procedure
Model: MODEL1
Dependent Variable: co

Number of Observations Read	24
Number of Observations Used	24

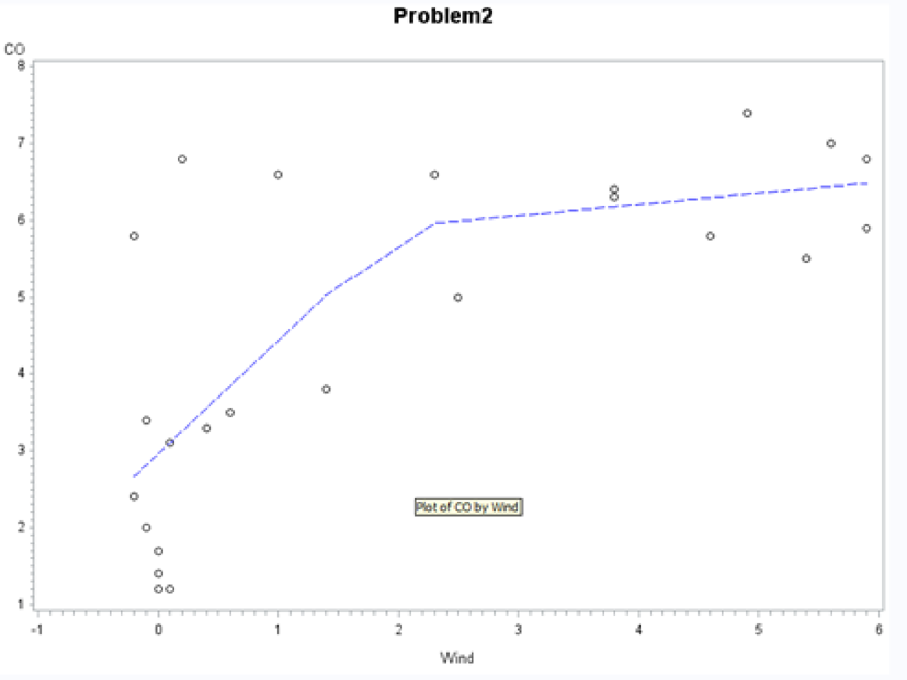
Analysis of Variance

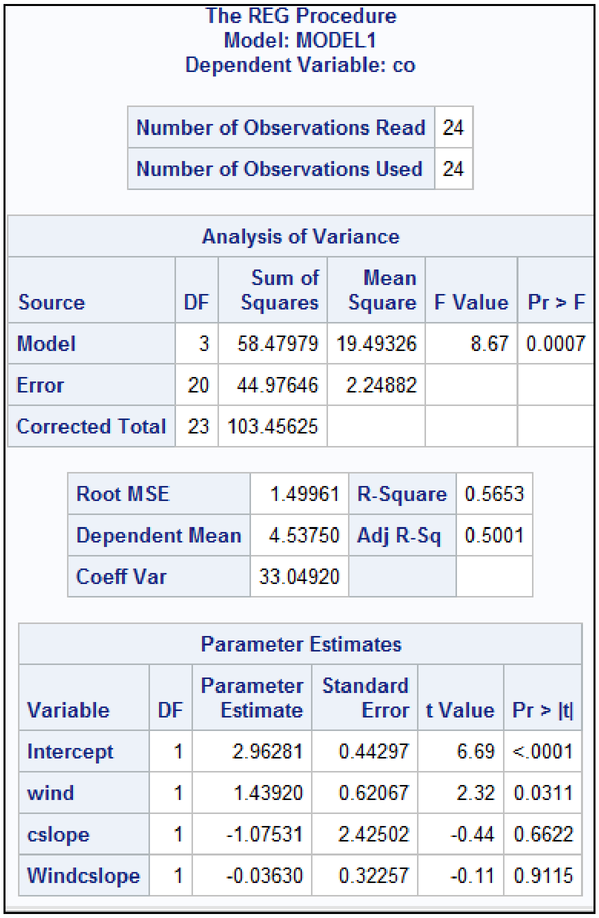
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	58.45132	29.22566	13.64	0.0002
Error	21	45.00493	2.14309		
Corrected Total	23	103.45625			

Root MSE	1.46393	R-Square	0.5650
Dependent Mean	4.53750	Adj R-Sq	0.5236
Coeff Var	32.26293		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.95843	0.43075	6.87	<.0001
wind	1	1.47862	0.50018	2.96	0.0075
cslope	1	-1.33312	0.77533	-1.72	0.1002




The REG Procedure
Model: MODEL1
Dependent Variable: co

Number of Observations Read	24
Number of Observations Used	24

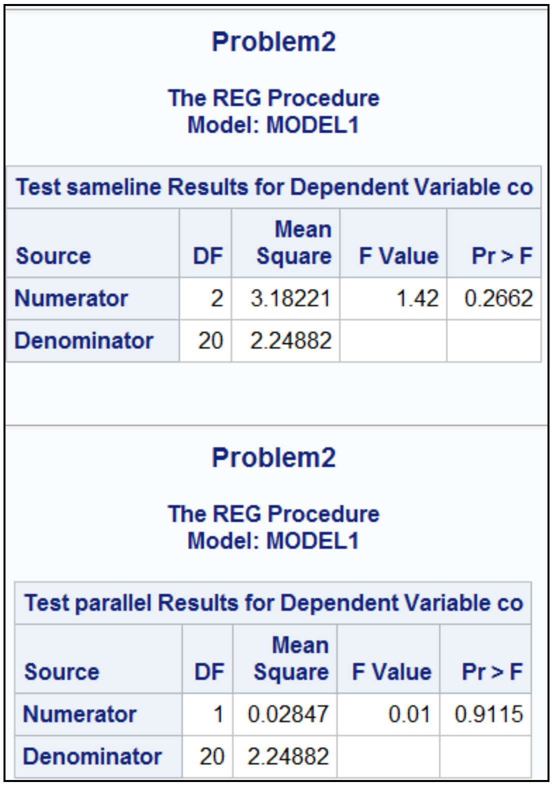
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	58.47979	19.49326	8.67	0.0007
Error	20	44.97646	2.24882		
Corrected Total	23	103.45625			

Root MSE	1.49961	R-Square	0.5653
Dependent Mean	4.53750	Adj R-Sq	0.5001
Coeff Var	33.04920		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.96281	0.44297	6.69	<.0001
wind	1	1.43920	0.62067	2.32	0.0311
cslope	1	-1.07531	2.42502	-0.44	0.6622
Windcslope	1	-0.03630	0.32257	-0.11	0.9115



Problem2

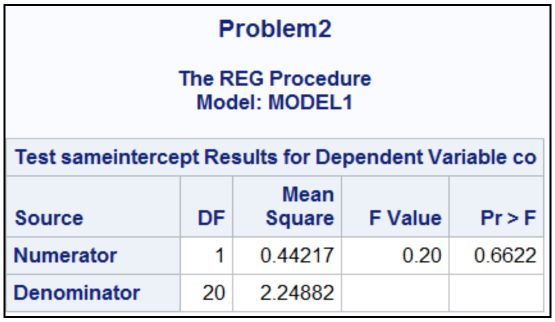
The REG Procedure
Model: MODEL1

Test sameline Results for Dependent Variable co				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	3.18221	1.42	0.2662
Denominator	20	2.24882		

Problem2

The REG Procedure
Model: MODEL1

Test parallel Results for Dependent Variable co				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.02847	0.01	0.9115
Denominator	20	2.24882		



Problem2

The REG Procedure
Model: MODEL1

Test sameintercept Results for Dependent Variable co				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.44217	0.20	0.6622
Denominator	20	2.24882		

QUESTION 3.

In this question you will illustrate some of the ideas related to the extra sums of squares.

a) Create a variable called SUM, which equals to the summation of any two predictors and run the following two regression models without the variables you used to create SUM:

i. Predict the response using all the explanatory variables;

ii. Predict the response using all the explanatory variables including SUM.

Calculate the extra sum of squares for the comparison of these two analyses. Use it to construct the F-statistic (in other words, the general linear test statistic) for testing the null hypothesis that the coefficient of the SUM variable is zero in the model with all predictors. What are the degrees of freedom for this test statistic?

b) Use the test statement in proc reg to obtain the same test statistic. Give the statistic, degrees of freedom, p-value and conclusion.

c) Compare the test statistic and p-value from the test statement with the individual t-test for the coefficient of the SUM variable in the full model. Explain the relationship.

<u>SOLUTION:</u>

(a) A new variable SUM was created by adding two predictors i.e. traffic and wind. The two regression models were run; one with all the explanatory variables and the second with all explanatory variables excluding SUM.
SUM=traffic+wind

Response with all explanatory variables: (values given in output)

$\hat{\mathbf{y}} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5$

Response with all explanatory variables including SUM: (values in output)

$\hat{\mathbf{y}} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_0 + \beta_6X_6$

$$F = \frac{SSR(X_2, X_3, X_4, X_5, X_6)}{1} \div \frac{SSE(X_1, X_2, X_3, X_4, X_5, X_6)}{18} = 665.57$$

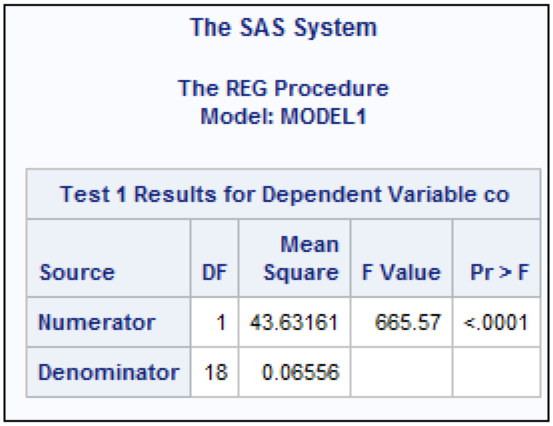
$$F = (SSR/1)/MSE(F) \sim F_{1,n-p} = 665.57$$

Degrees of freedom: 1, 18
Ho: $\beta_{sum} = 0$, the coefficient of the SUM variable is zero in the model with all predictors.
Ha: $\beta_{sum} \neq 0$, the coefficient of the SUM variable is not zero
Fs (1,18): 665.57
P <0.0001 reject Ho. Based on enough statistical evidence, we conclude that the slope for the SUM is not 0 or the parameter estimate for term SUM is significant with all other predictors.

<u>(b)</u> **<u>Using the “test” statement in proc reg:</u>**
F-statistic: 665.57 > 4.41
Degree of freedom: (1, 18)
p-value: <.0001
Conclusion: the parameter for the term SUM is significant with all other predictors in model.

(c) **Individual t-test statistic** = 25.80 which is the square root of our previous F Statistic = 665. P-value<0.0001 = p-value from previous F test. So it corresponds to the fact that t^2=F

OUTPUT FROM SAS:



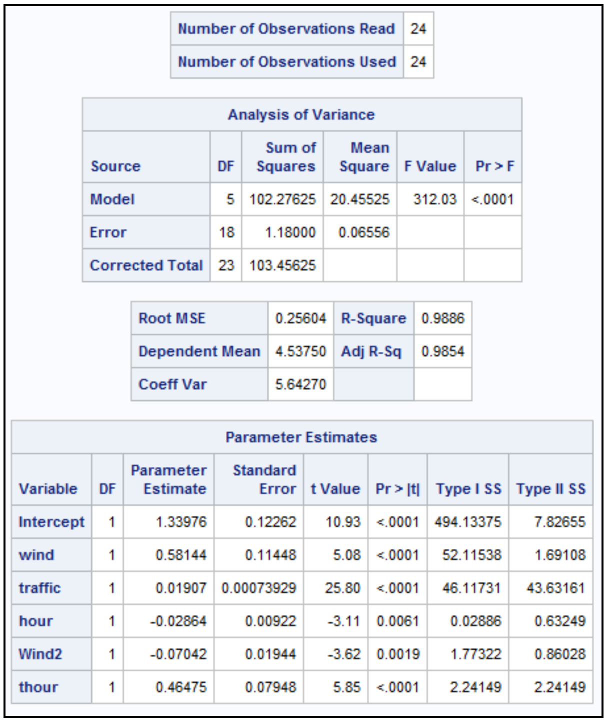
The SAS System

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable co

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	43.63161	665.57	<.0001
Denominator	18	0.06556		

(a) Prediction of response using all the explanatory variables



Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance

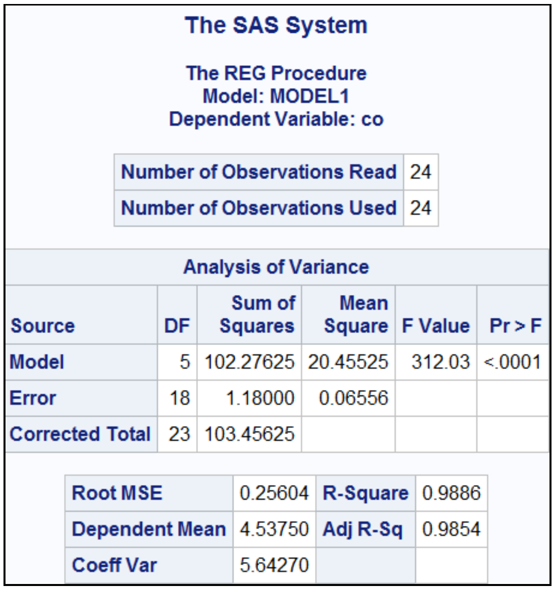
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	102.27625	20.45525	312.03	<.0001
Error	18	1.18000	0.06556		
Corrected Total	23	103.45625			

Root MSE	0.25604	R-Square	0.9886
Dependent Mean	4.53750	Adj R-Sq	0.9854
Coeff Var	5.64270		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	1.33976	0.12262	10.93	<.0001	494.13375	7.82655
wind	1	0.58144	0.11448	5.08	<.0001	52.11538	1.69108
traffic	1	0.01907	0.00073929	25.80	<.0001	46.11731	43.63161
hour	1	-0.02864	0.00922	-3.11	0.0061	0.02886	0.63249
Wind2	1	-0.07042	0.01944	-3.62	0.0019	1.77322	0.86028
thour	1	0.46475	0.07948	5.85	<.0001	2.24149	2.24149

(b) Prediction of response using all the explanatory variables including SUM



The SAS System

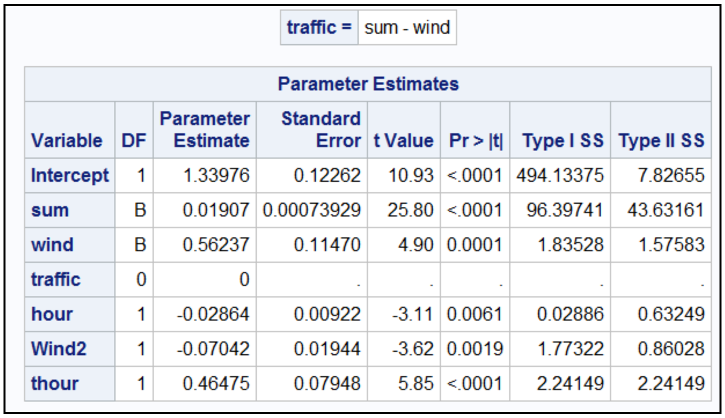
The REG Procedure
Model: MODEL1
Dependent Variable: co

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	102.27625	20.45525	312.03	<.0001
Error	18	1.18000	0.06556		
Corrected Total	23	103.45625			

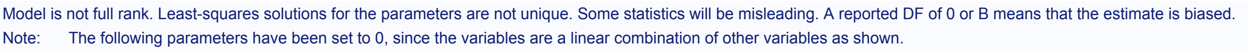
Root MSE	0.25604	R-Square	0.9886
Dependent Mean	4.53750	Adj R-Sq	0.9854
Coeff Var	5.64270		



traffic =	sum - wind

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	1.33976	0.12262	10.93	<.0001	494.13375	7.82655
sum	B	0.01907	0.00073929	25.80	<.0001	96.39741	43.63161
wind	B	0.56237	0.11470	4.90	0.0001	1.83528	1.57583
traffic	0	0	.	.	.	.	.
hour	1	-0.02864	0.00922	-3.11	0.0061	0.02886	0.63249
Wind2	1	-0.07042	0.01944	-3.62	0.0019	1.77322	0.86028
thour	1	0.46475	0.07948	5.85	<.0001	2.24149	2.24149



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.

QUESTION 4.

Run the regression to predict the response using the predictors including all transformed variables, but not the variable SUM. Put the variables in any order you wish in the model statement. Use the SS1 and SS2 options in the model statement. Add the Type I sums of squares for the predictors. Do the same for the Type II sums of squares. Do either of these sum to the model sum of squares? Are there any predictors for which the two sums of squares (Type I and Type II) are the same? Explain why.

SOLUTION:

Summation of Type I SS = 102.27 and Summation of Type II SS = 49.0565 (please refer to SAS output below).

Σ Type I SS = SSM (Model Sum of Squares)

52.11+ 46.11+ 0.028+1.77+2.24 = 102.27

We changed the order of variables in the model and rerun the regression and we were able to verify the this again that sum of Type I SS is equal to SSM:

18.98+76.92+2.354+1.77+2.24= 102.27

For the last predictor in the model, *the two sums of squares (Type I and Type II) are the same*

The type I SS = type II SS for the last variable which in this case is the transformed variable hour (transhour)

For variables: X1,X2,X3,X4,X5

TYPE I SS:

SS_1 (X1), SS_1 (X2|X1), SS_1(X3|X2 X1), SS_1(X4|X3 X2 X1), ***SS_1(X5|X4 X3 X2 X1)***

TYPE II SS:

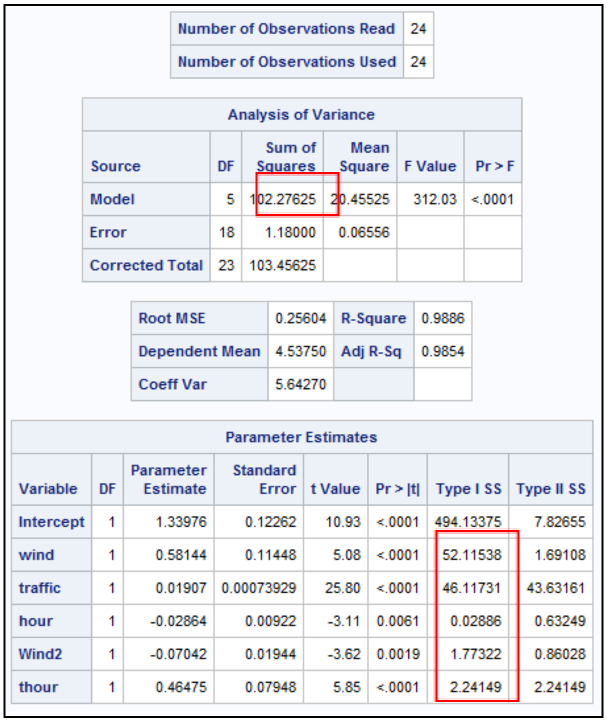
SS_2(X1|X2 X3 X4 X5), SS_2(X2|X1 X3 X4 X5), SS_2(X3|X1 X2 X4 X5), SS_2(X4|X1 X2 X3 X5), ***SS_2(X5|X1 X2 X3 X4)***

From the above description of Type I SS and Type II SS, it can be seen that the last value for both is the same.

Hence, Sum of Squares for last variable X5 is equal for both type I and type II.

SS_1(X5|X4 X3 X2 X1)= SS_2(X5|X1 X2 X3 X4)

OUTPUT FROM SAS:



Number of Observations Read	24
Number of Observations Used	24

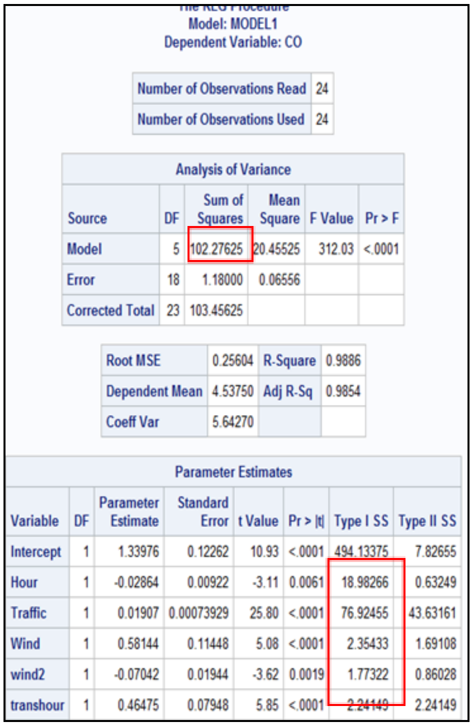
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	102.27625	20.45525	312.03	<.0001
Error	18	1.18000	0.06556		
Corrected Total	23	103.45625			

Root MSE	0.25604	R-Square	0.9886
Dependent Mean	4.53750	Adj R-Sq	0.9854
Coeff Var	5.64270		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	1.33976	0.12262	10.93	<.0001	494.13375	7.82655
wind	1	0.58144	0.11448	5.08	<.0001	52.11538	1.69108
traffic	1	0.01907	0.00073929	25.80	<.0001	46.11731	43.63161
hour	1	-0.02864	0.00922	-3.11	0.0061	0.02886	0.63249
Wind2	1	-0.07042	0.01944	-3.62	0.0019	1.77322	0.86028
thour	1	0.46475	0.07948	5.85	<.0001	2.24149	2.24149



The REG Procedure
Model: MODEL1
Dependent Variable: CO

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	102.27625	20.45525	312.03	<.0001
Error	18	1.18000	0.06556		
Corrected Total	23	103.45625			

Root MSE	0.25604	R-Square	0.9886
Dependent Mean	4.53750	Adj R-Sq	0.9854
Coeff Var	5.64270		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	1.33976	0.12262	10.93	<.0001	494.13375	7.82655
Hour	1	-0.02864	0.00922	-3.11	0.0061	18.98266	0.63249
Traffic	1	0.01907	0.00073929	25.80	<.0001	76.92455	43.63161
Wind	1	0.58144	0.11448	5.08	<.0001	2.35433	1.69108
wind2	1	-0.07042	0.01944	-3.62	0.0019	1.77322	0.86028
transhour	1	0.46475	0.07948	5.85	<.0001	2.24149	2.24149

QUESTION 5.

Run the regression to predict the response using a variety of variables, including SUM as an explanatory variable, you should have at least 10 different models. Summarize the results by making a table giving the percentage of variation explained (R^2) by each model.

SOLUTION:

No.	Variables	R^2
1	Hour sum	0.9319
2	Hour traffic sum	0.9498
3	Hour traffic	0.9270
4	Sum	0.9318
5	Traffic wind	0.9495
6	Hour traffic wind2	0.9392
7	Hour	0.1835
8	Wind2	0.3966
9	Traffic	0.9267
10	Wind	0.5037
11	Hour wind2	0.4520
12	Traffic wind2	0.9391
13	Hour traffic wind wind2	0.9669
14	Hour thour traffic wind wind2	0.9886
15	Thour traffic wind wind2	0.9825
16	Thour traffic wind2	0.971
17	Thour traffic wind	0.9775
18	Hour thour traffic wind	0.9803

OUTPUT FROM SAS: *Not required*

QUESTION 6.

Use the Cp criterion to select the best subset of variables for your data (i.e. use the options " / selection = cp b;"). Use the original and transformed variables, not SUM. Summarize the results and explain your choice of the best model.

SOLUTION:

Background

In statistics, Mallows's Cp named for Colin Lingwood Mallows, is used to assess the fit of a regression model that has been estimated using ordinary least squares. It is applied in the context of model selection, where a number of predictor variables (p-1) are available for predicting some outcome, and the goal is to find the best model involving a subset of these predictors.
The basic idea is to compare subset models with the full model. The full model is good at prediction, but if there is multi-collinearity, our interpretations of the parameter estimates may not make sense. A subset model is good if there is not substantial "bias" in the predicted values (relative to the full model).
The Cp criterion looks at the ratio of error SS for the model with p variables to the MSE of the full model, then adds a penalty for the number of variables.

$$\text{Cp} = \frac{SSEp}{MSE\ (full)} - (n - 2p)$$

SSE is based on a specific choice of p − 1 variables (p is the number of regression coefficients including the intercept); while MSE is based on the full set of variables. A model is good according to this criterion if **Cp ≤ p**. We may choose the smallest model for which **Cp ≤ p**, so a benefit of this criterion is that it can achieve for us a "good" model containing as few variables as possible. One might also choose to pick the model that minimizes Cp.

SOLUTION

For Cp:
Now after running the SAS Analysis for Cp selection criteria, the following models have been reported:

Number in Model	**C(p)**	**R-Square**	**Parameter Estimates**						
			Intercept	**hour**	**traffic**	**wind**	**wind2**	**thour**	**p**
5	**6**	0.9886	1.33976	-0.0286	0.01907	0.58144	-0.0704	0.46475	6
4	**13.6482**	0.9825	1.13566	.	0.01861	0.45532	-0.0517	0.43631	
4	**17.1229**	0.9803	1.2919	-0.0183	0.01953	0.18113	.	0.53477	
3	**19.4664**	0.9775	1.15683	.	0.01912	0.16736	.	0.50154	
3	**29.7437**	0.971	1.15516	.	0.01985	.	0.02396	0.53512	
4	**29.7963**	0.9722	1.24317	-0.012	0.02019	.	0.02489	0.55856	
4	**38.1923**	0.9669	1.38068	-0.0224	0.01797	0.73701	-0.0981	.	
3	**42.1896**	0.9631	1.21669	.	0.01765	0.62935	-0.0818	.	

2	**50.3998**	0.9567	1.09997	.	0.02157	.	.	0.51803
3	**51.8155**	0.957	1.14661	-0.0065	0.02179	.	.	0.53041
2	**61.6819**	0.9495	1.27446	.	0.01829	0.17475	.	.
3	**63.2415**	0.9498	1.31897	-0.0057	0.0184	0.17919	.	.
2	**78.0377**	0.9391	1.27398	.	0.01921	.	0.02226	.
3	**80.0057**	0.9392	1.26232	0.0015	0.01917	.	0.02216	.
1	**95.6395**	0.9267	1.21905	.	0.02083	.	.	.
2	**97.1552**	0.927	1.175	0.00581	0.02065	.	.	.
3	**668.183**	0.5665	2.81953	0.01871	.	1.44724	-0.1536	.
4	**669.57**	0.5669	2.81336	0.01918	.	1.46172	-0.1568	-0.0602
2	**670.493**	0.5637	2.98241	.	.	1.55065	-0.1685	.
3	**672.082**	0.564	2.98074	.	.	1.56464	-0.1713	-0.0492
2	**733.5**	0.5238	2.77693	0.04688	.	0.5885	.	.
3	**734.598**	0.5244	2.78529	0.04563	.	0.59212	.	0.07138
2	**761.878**	0.5058	3.23483	.	.	0.6527	.	0.13464
1	**763.169**	0.5037	3.24241	.	.	0.6489	.	.
3	**846.498**	0.4535	2.78542	0.07275	.	.	0.0922	0.1137
2	**846.753**	0.452	2.77287	0.0747	.	.	0.0909	.
2	**925.261**	0.4023	3.56625	.	.	.	0.10626	0.22331
1	**932.189**	0.3966	3.5833	.	.	.	0.10439	.
2	**1264.34**	0.1874	2.9073	0.13042	.	.	.	-0.1857
1	**1268.59**	0.1835	2.93152	0.12848	.	.	.	.
1	**1557.59**	0.0004	4.5375	.	.	.	.	-0.0553

Summary of the results and Choice of the best model

In the light of criteria set for Cp method of the selection of a good model, the model where Cp=6=p (highlighted in yellow in Table above) satisfies the criteria requirements. In all of the other cases, Cp value is more than p and violates the criteria of a good model. So all the models can be rejected except the full model. From Cp selection criteria, it can be reported that our full model is the best model.

QUESTION 7.

Use the stepwise option to report the best subset of variables for your data (i.e. use the options / selection = stepwise;). Use the original and transformed variables, not SUM. Summarize the results and explain your choice of the best model.

SOLUTION:

BACKGROUND

In statistics, stepwise regression includes regression models in which the choice of predictive variables is carried out by an automatic procedure. [1][2][3][4] Usually, this takes the form of a sequence of F-tests or t-tests, but other techniques are possible, such as adjusted R-square, Akaike information criterion, Bayesian information criterion, Mallows's Cp, PRESS, or false discovery rate.

The main approach used here in this SAS Analysis is:

- Forward selection, which involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable that improves the model the most, and repeating this process until none improves the model.

Output from SAS

- ➢ F-tests for the variables under consideration are conducted and the variable with the largest F-statistic value is considered as the first candidate for addition. If the p-value is less than α, the variable is added. In the first case, explanatory variable Traffic was considered because it has the highest F value of 278.24 and its p-value is significant (<.0001).

- ➢ In second step, another variable i.e. thour was added to the model and it was found to be significant with p value = 0.001 (<0.05) and F value 14.50.

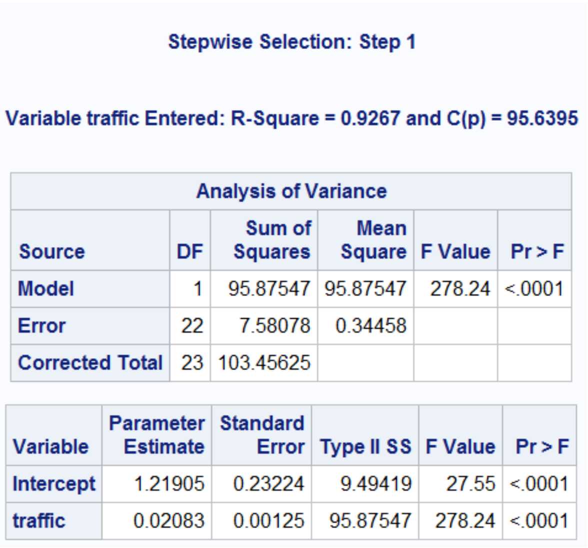
- ➢ In third step, another variable i.e. wind was found to be significant having F value 18.57 and p value 0.0003 (<0.05).

- ➢ In fourth step, another variable i.e. wind^2 was found to be significant having F value 5.37 and p value 0.03 (<0.05)

- ➢ In fifth and last step, another variable i.e. hour was found to be significant having F value 9.65 and p value 0.0061 (< 0.05)

- ➢ All variables left in the model are significant at the 0.1500 level and also 95% confidence level. All variables have been entered into the model.

- ➢ So, our full model has been reported as the best model.

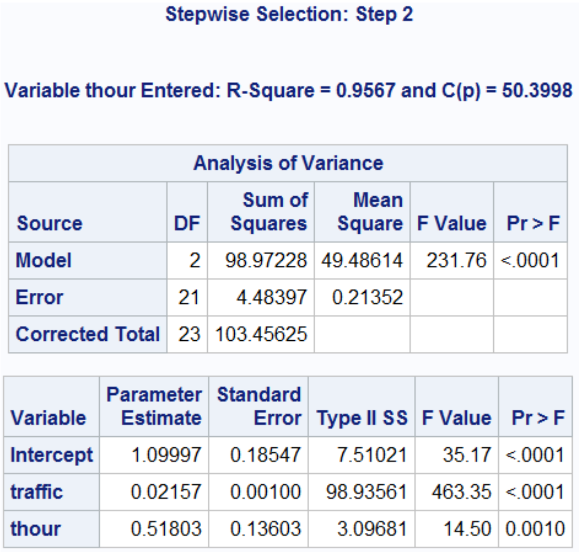


Stepwise Selection: Step 1

Variable traffic Entered: R-Square = 0.9267 and C(p) = 95.6395

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	95.87547	95.87547	278.24	<.0001
Error	22	7.58078	0.34458		
Corrected Total	23	103.45625			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.21905	0.23224	9.49419	27.55	<.0001
traffic	0.02083	0.00125	95.87547	278.24	<.0001

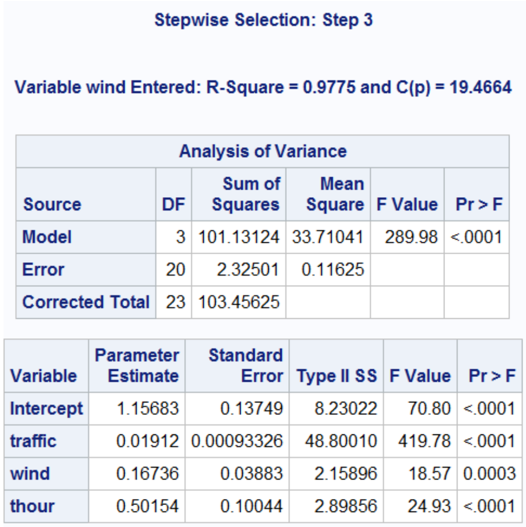


Stepwise Selection: Step 2

Variable thour Entered: R-Square = 0.9567 and C(p) = 50.3998

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	98.97228	49.48614	231.76	<.0001
Error	21	4.48397	0.21352		
Corrected Total	23	103.45625			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.09997	0.18547	7.51021	35.17	<.0001
traffic	0.02157	0.00100	98.93561	463.35	<.0001
thour	0.51803	0.13603	3.09681	14.50	0.0010

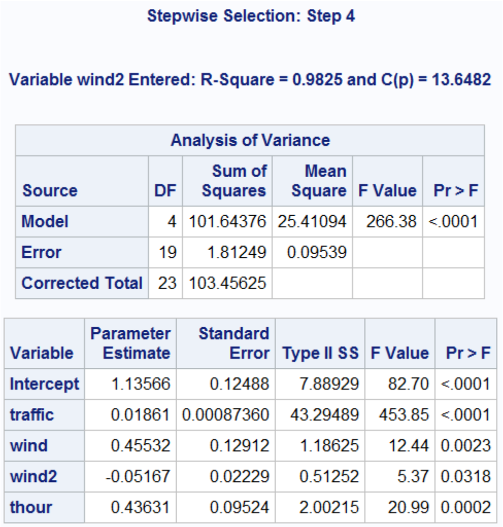


Stepwise Selection: Step 3

Variable wind Entered: R-Square = 0.9775 and C(p) = 19.4664

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	101.13124	33.71041	289.98	<.0001
Error	20	2.32501	0.11625		
Corrected Total	23	103.45625			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.15683	0.13749	8.23022	70.80	<.0001
traffic	0.01912	0.00093326	48.80010	419.78	<.0001
wind	0.16736	0.03883	2.15896	18.57	0.0003
thour	0.50154	0.10044	2.89856	24.93	<.0001

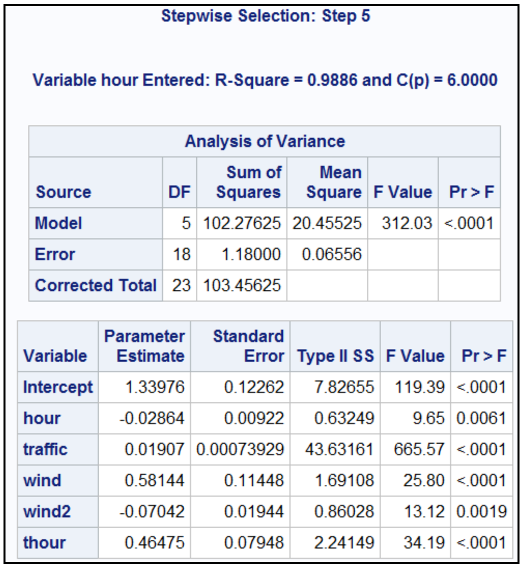


Stepwise Selection: Step 4

Variable wind2 Entered: R-Square = 0.9825 and C(p) = 13.6482

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	101.64376	25.41094	266.38	<.0001
Error	19	1.81249	0.09539		
Corrected Total	23	103.45625			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.13566	0.12488	7.88929	82.70	<.0001
traffic	0.01861	0.00087360	43.29489	453.85	<.0001
wind	0.45532	0.12912	1.18625	12.44	0.0023
wind2	-0.05167	0.02229	0.51252	5.37	0.0318
thour	0.43631	0.09524	2.00215	20.99	0.0002

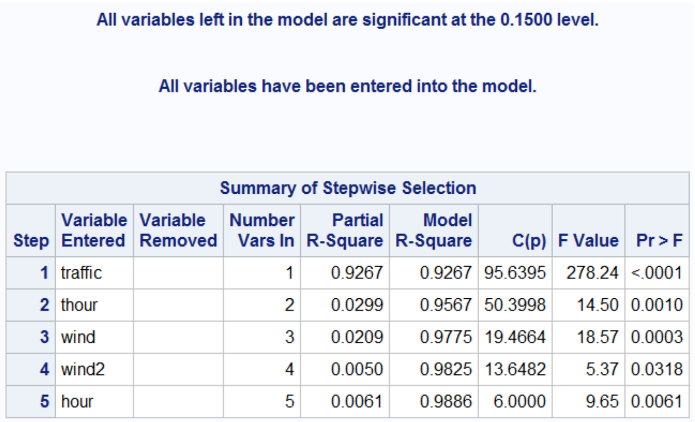


Stepwise Selection: Step 5

Variable hour Entered: R-Square = 0.9886 and C(p) = 6.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	102.27625	20.45525	312.03	<.0001
Error	18	1.18000	0.06556		
Corrected Total	23	103.45625			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.33976	0.12262	7.82655	119.39	<.0001
hour	-0.02864	0.00922	0.63249	9.65	0.0061
traffic	0.01907	0.00073929	43.63161	665.57	<.0001
wind	0.58144	0.11448	1.69108	25.80	<.0001
wind2	-0.07042	0.01944	0.86028	13.12	0.0019
thour	0.46475	0.07948	2.24149	34.19	<.0001



All variables left in the model are significant at the 0.1500 level.

All variables have been entered into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	traffic		1	0.9267	0.9267	95.6395	278.24	<.0001
2	thour		2	0.0299	0.9567	50.3998	14.50	0.0010
3	wind		3	0.0209	0.9775	19.4664	18.57	0.0003
4	wind2		4	0.0050	0.9825	13.6482	5.37	0.0318
5	hour		5	0.0061	0.9886	6.0000	9.65	0.0061

QUESTION 8.

Check the assumptions of this "best" model using all the usual plots (you know what they are by now). Explain in detail whether or not each assumption appears to be substantially violated*.*

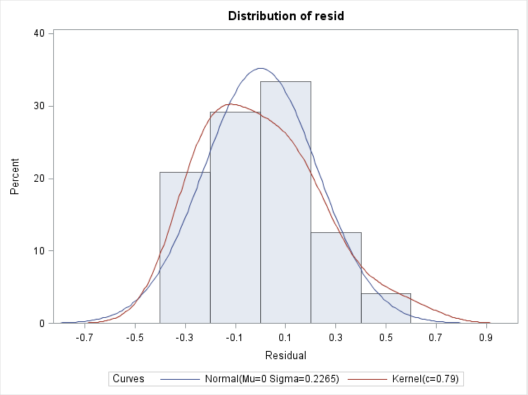
SOLUTION:

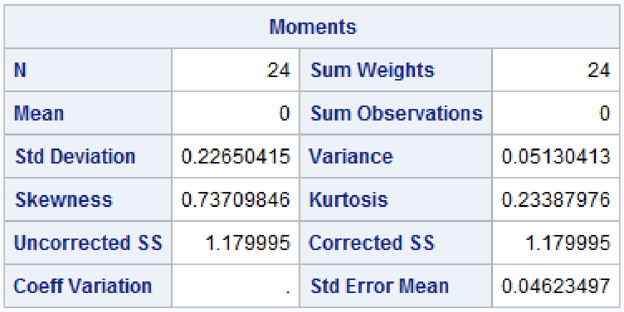
There are 4 major assumptions for the best model: Linearity (verified by residual or scatter plot), Independence (verified by sequence plot), Constant Variance (verified by scatter or residual plots) and Normality (verified by qq-plot).

- **NORMALITY (QQ-PLOT) AND HISTOGRAM:**

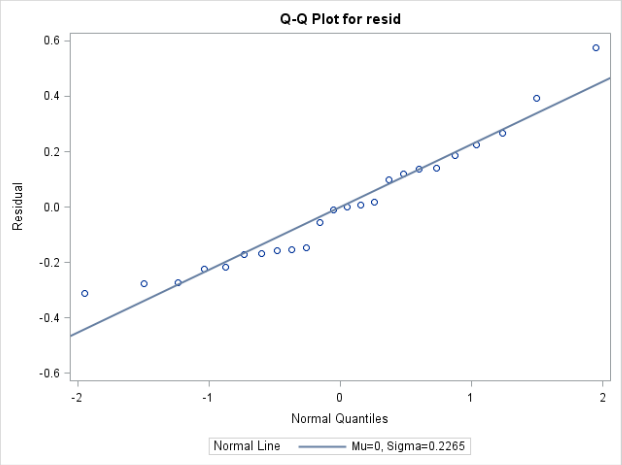
The histogram shows that there is a negligibly small skewness on right (calculations show only 0.737 of skewness on right.)

The qq-plot shows that the residuals are approximately normal which an outlier towards the top and tail.





Moments			
N	24	Sum Weights	24
Mean	0	Sum Observations	0
Std Deviation	0.22650415	Variance	0.05130413
Skewness	0.73709846	Kurtosis	0.23387976
Uncorrected SS	1.179995	Corrected SS	1.179995
Coeff Variation	.	Std Error Mean	0.04623497



Normally it is strongly discouraged to delete outlier as sometimes outliers convey a very important or unusual information about data; or sometimes it may be due to bad observations or rough data collection process. It is recommended to use "robust methods (temporarily eliminating extreme observations at both ends of sample)" to handle outliers.

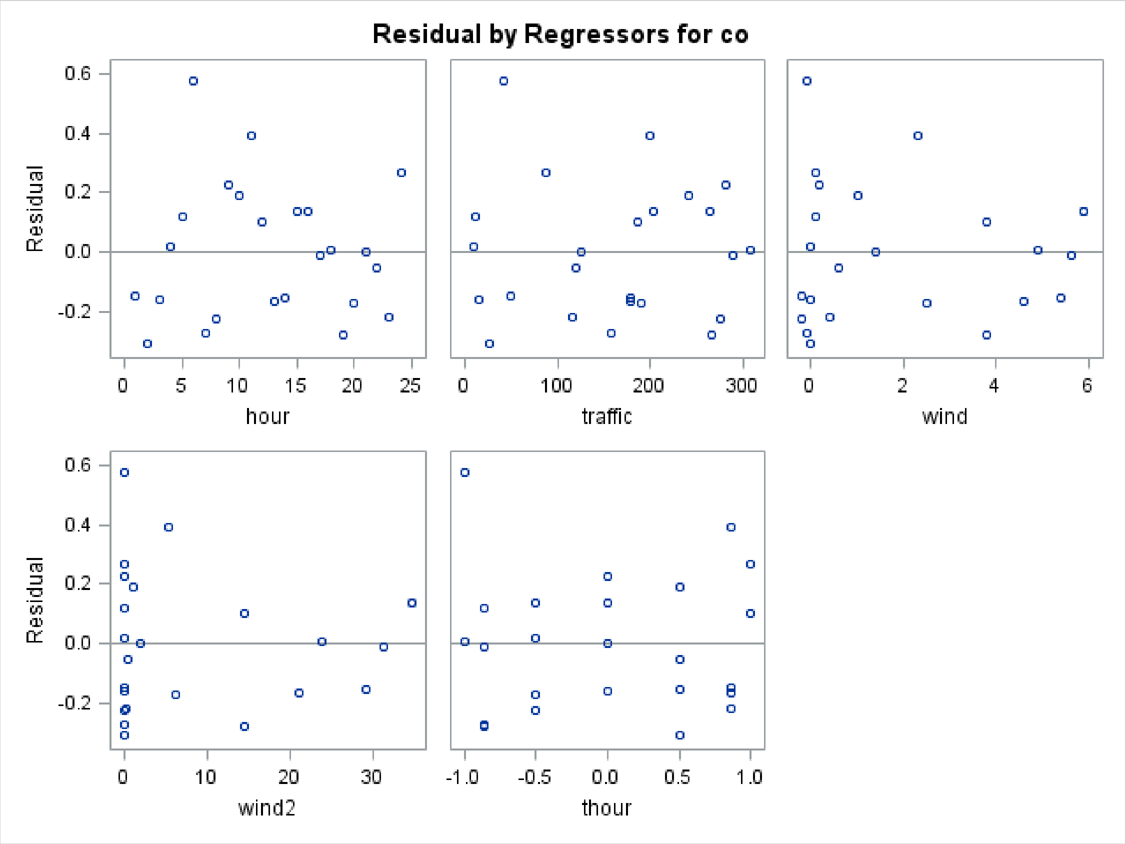
- **LINEARITY and CONSTANT VARIANCE (RESIDUAL/SCATTER PLOTS):**

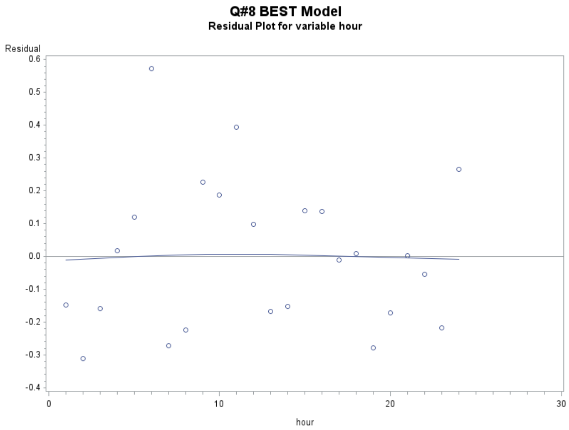
The residuals plots don't show any patterns; for the reason the assumption of linearity is not violated.

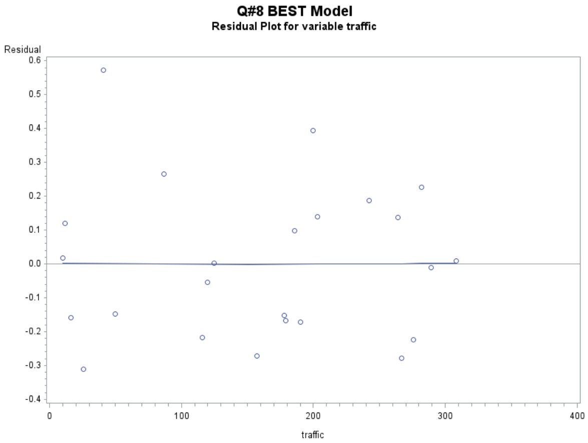
The residuals plot show that variance has got some issue which can be fixed through weighted regression or variance stabilizing transformation. Box-cox transformations didn't help in stabilizing variance. We tried different transformations for response variable like log y or y^0.75 but these transformations didn't put any significant impact on results. Secondly, from the literature review of this dataset, we found that it would be needed to create three more variables (with sine, cosine relationships) for hour but since we wanted to keep the model simple, we didn't create and include more variables.

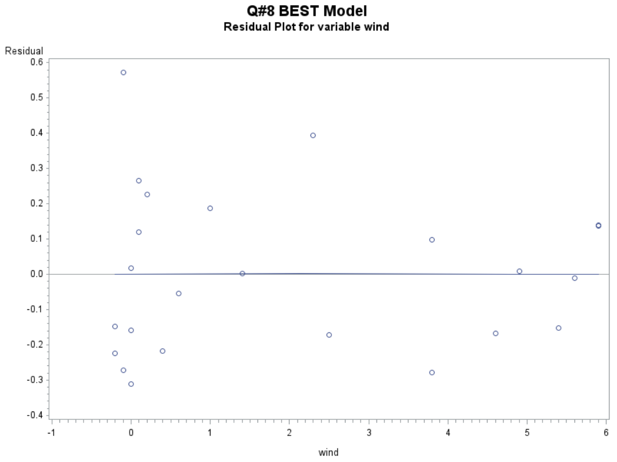
Thirdly, the number of observations is very less. Also, it is highly recommended to implement weighted linear regression or variance stabilizing transformation to cater for constant variance violation.

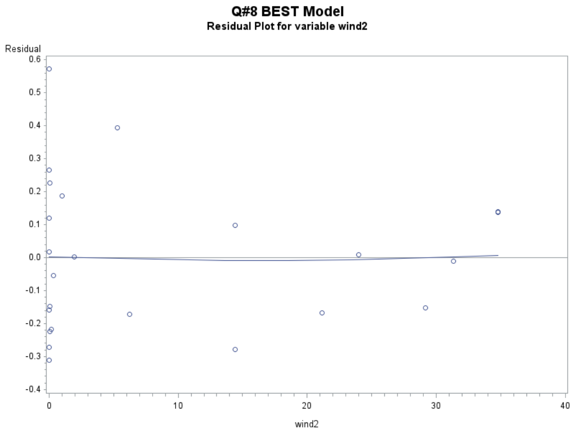
There can be seen one outlier for residual = 0.58. Some researchers recommend to delete outliers or influential observations while others strongly discourage the deletion of outliers, rather, they recommend to use robust methods (temporary elimination of extreme observations at both ends).

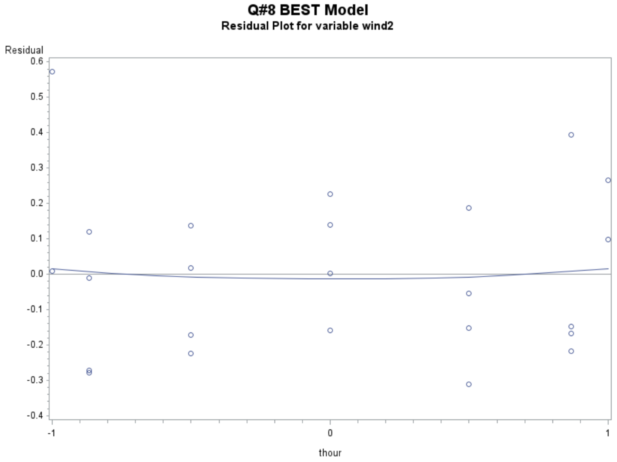








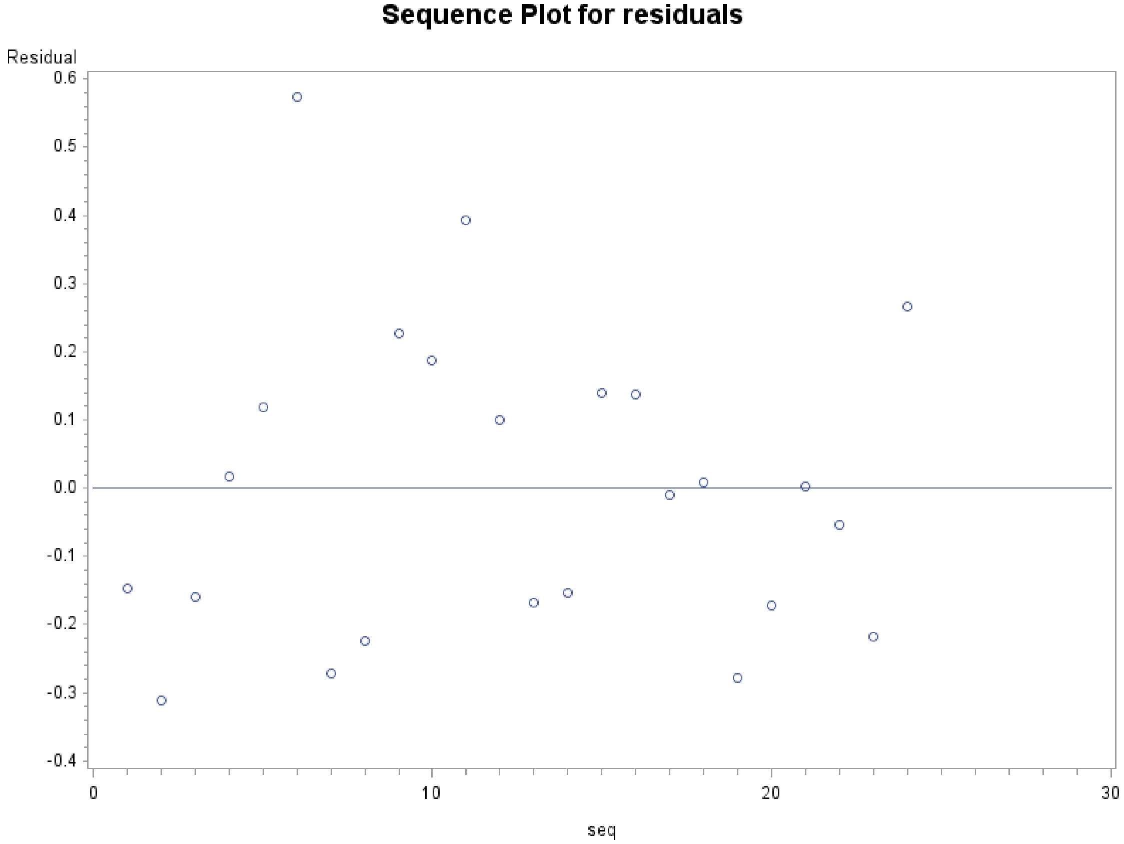




- **INDEPENDENCE (SEQUENCE PLOTS):**

Random samples consist of independent and identically distributed (i.i.d.) observations. This means that the observations all come from the same parent population and are independent of one other. With a sequence plot, you can check the identically distributed aspect of a random sample by looking for evidence of stability in the plot. Stability is supported when both the mean of the observations and the amount of variation among observations appear to be constant over time.

From the sequence plots below, the assumption of independence of error terms doesn't appear to be violated as the amount of variation appear constant.



QUESTION 9.

Use the “best" model to predict the response variable. Examine other diagnostics such as (but not necessarily exclusively) studentized and studentized-deleted residuals, Cook's D, tolerance or vif, and partial residual plots. Explain any problems such as outliers, highly influential observations or multicollinearity that these diagnostics point out. (Do not include in your output any tables of values for all observations. Use plots and verbal summaries instead. You may include values for a few selected individuals if you wish.)

SOLUTION:

Best model predicted as:

$$CO = 1.33976 - 0.02864\ hour + 0.46475\ thour + 0.01907\ traffic + 0.58144\ wind - 0.07042\ wind^2$$

By using Bonferroni correction, t_{n-p-1} (1-alpha/2n)= t_{17}(.99896)= 3.627 > |3.0912| . Magnitudes of Studentized-deleted residuals for all observations lower than 3.627, thus no outliers based on this test.

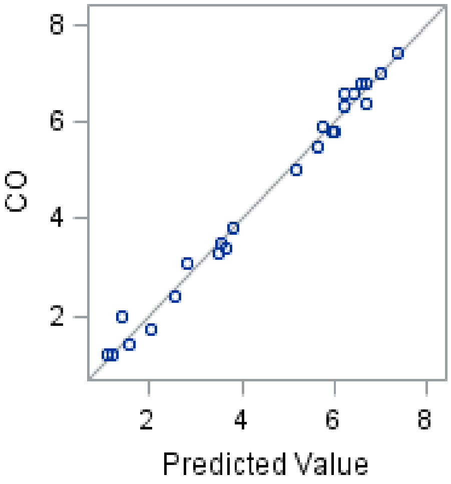
$F_{p,n-p} = F_{6,18}(0.5) = 0.93>0.313$

Given, the Cook’s D is smaller than the 50th percentile of F and Studentized-Deleted Residual is smaller than the test statistic from the Bonferroni correction, observation 6 is not considered to be an influential point or outlier. Thus, there is no influential point or outlier found from the dataset.

The partial plots show that a linear pattern exists for all 5 predictors. So each predictor makes a positive contribution in predicting y(CO).

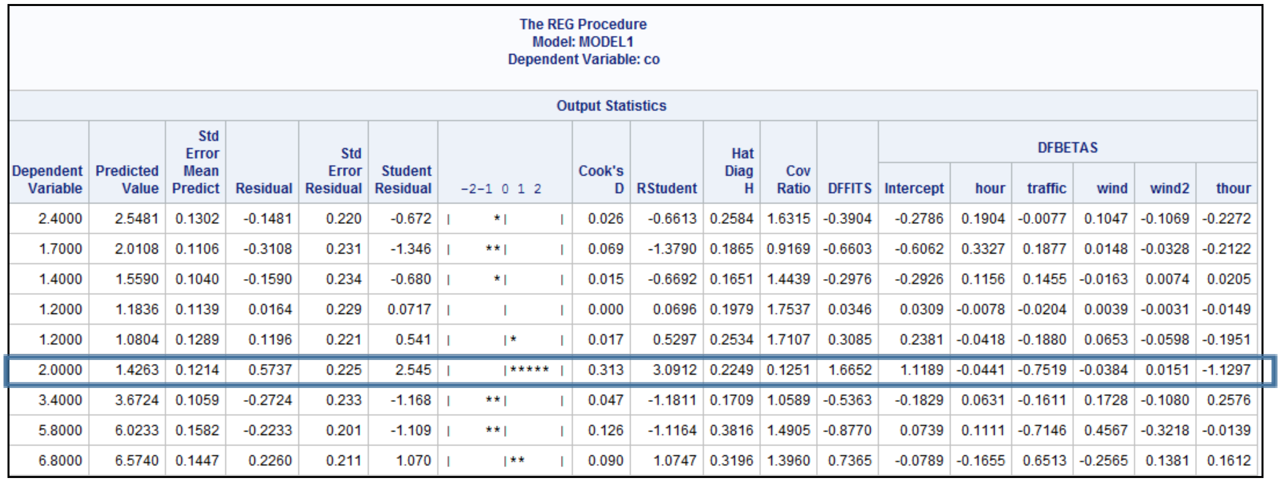
The tolerance for variables Wind and Wind2 are smaller than 0.1 and consequently the Variance inflation is >10 which puts the 2 variables at a risk of multicollinearity. The table above shows that test statistics for all 5 predictor variables is significant and the entire model F statistic is significant as well so we can say that multicollinearity isn’t a problem.

There is only 1 model with 5 variables and it is the best model evident from the smallest C(p) and largest R^2.



OUTPUT FROM SAS:

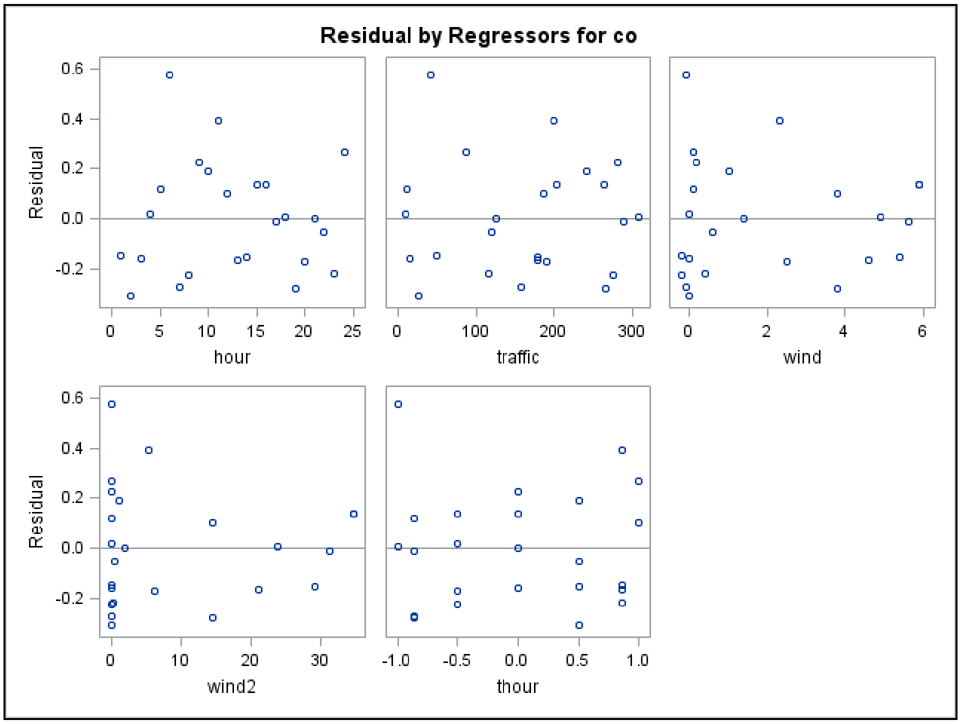
(1): From “r influence”: Cook’s D, Studentized and studentized deleted residuals



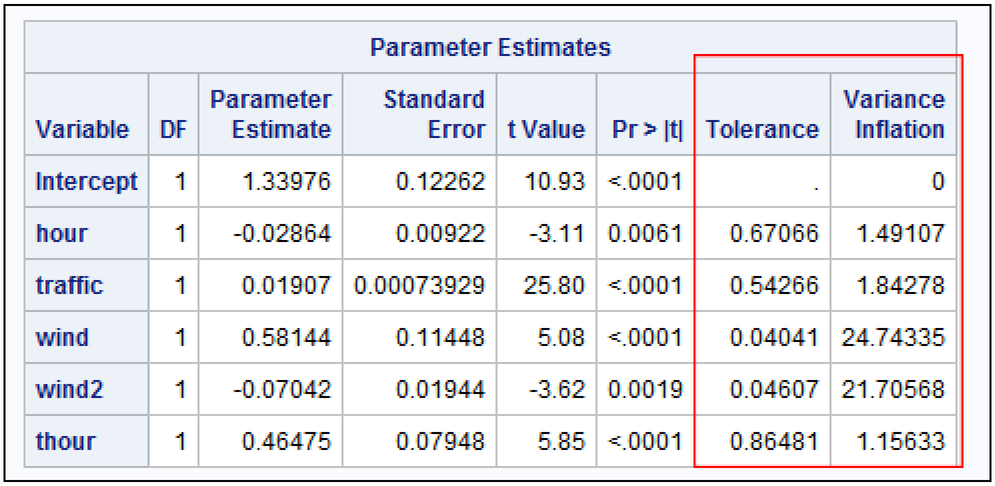
The REG Procedure
Model: MODEL1
Dependent Variable: co

Output Statistics

												DFBETAS					
Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	hour	traffic	wind	wind2	thour
2.4000	2.5481	0.1302	-0.1481	0.220	-0.672	| *| |	0.026	-0.6613	0.2584	1.6315	-0.3904	-0.2786	0.1904	-0.0077	0.1047	-0.1069	-0.2272
1.7000	2.0108	0.1106	-0.3108	0.231	-1.346	| **| |	0.069	-1.3790	0.1865	0.9169	-0.6603	-0.6062	0.3327	0.1877	0.0148	-0.0328	-0.2122
1.4000	1.5590	0.1040	-0.1590	0.234	-0.680	| *| |	0.015	-0.6692	0.1651	1.4439	-0.2976	-0.2926	0.1156	0.1455	-0.0163	0.0074	0.0205
1.2000	1.1836	0.1139	0.0164	0.229	0.0717	| | |	0.000	0.0696	0.1979	1.7537	0.0346	0.0309	-0.0078	-0.0204	0.0039	-0.0031	-0.0149
1.2000	1.0804	0.1289	0.1196	0.221	0.541	| |* |	0.017	0.5297	0.2534	1.7107	0.3085	0.2381	-0.0418	-0.1880	0.0653	-0.0598	-0.1951
2.0000	1.4263	0.1214	0.5737	0.225	2.545	| |***** |	0.313	3.0912	0.2249	0.1251	1.6652	1.1189	-0.0441	-0.7519	-0.0384	0.0151	-1.1297
3.4000	3.6724	0.1059	-0.2724	0.233	-1.168	| **| |	0.047	-1.1811	0.1709	1.0589	-0.5363	-0.1829	0.0631	-0.1611	0.1728	-0.1080	0.2576
5.8000	6.0233	0.1582	-0.2233	0.201	-1.109	| **| |	0.126	-1.1164	0.3816	1.4905	-0.8770	0.0739	0.1111	-0.7146	0.4567	-0.3218	-0.0139
6.8000	6.5740	0.1447	0.2260	0.211	1.070	| |** |	0.090	1.0747	0.3196	1.3960	0.7365	-0.0789	-0.1655	0.6513	-0.2565	0.1381	0.1612

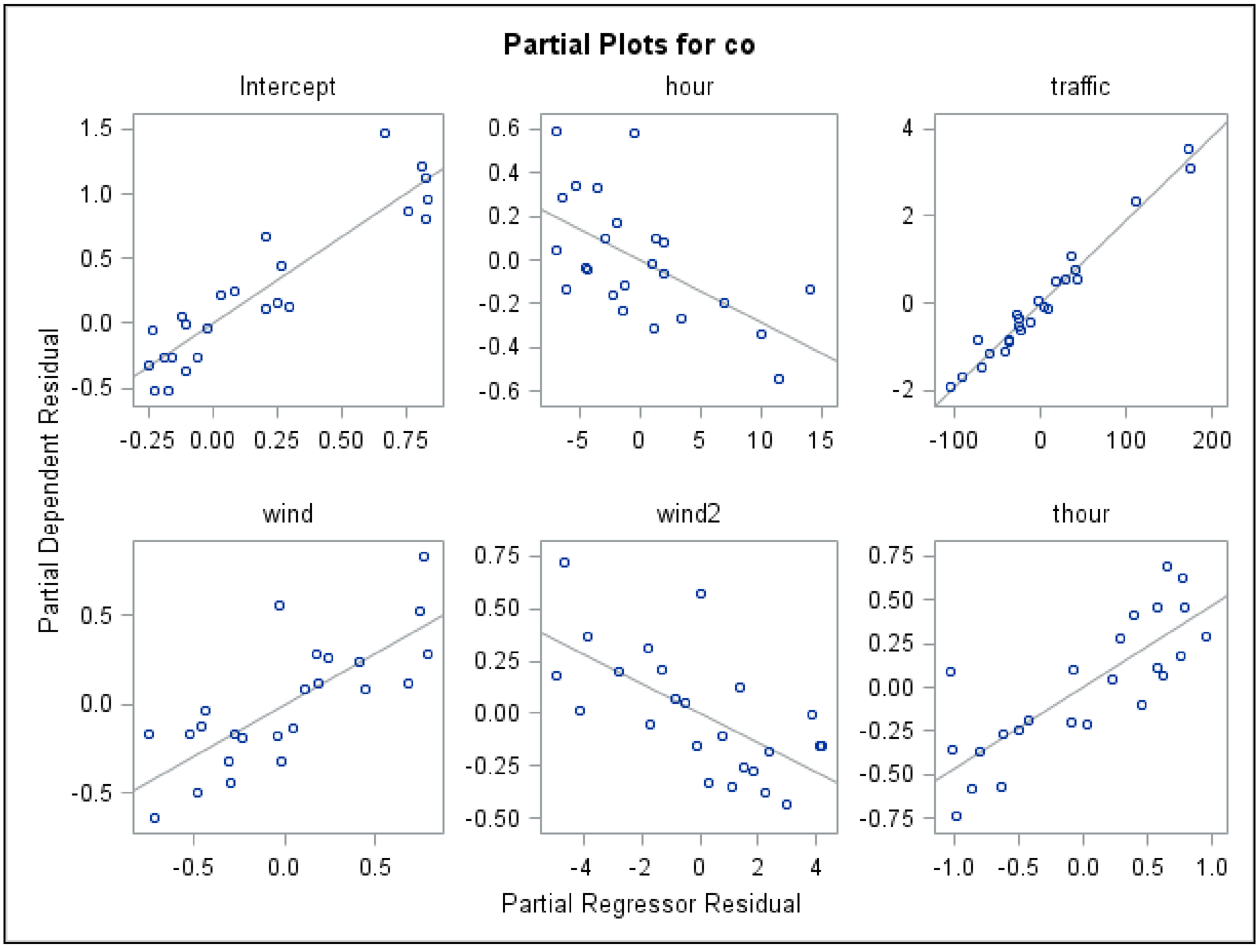


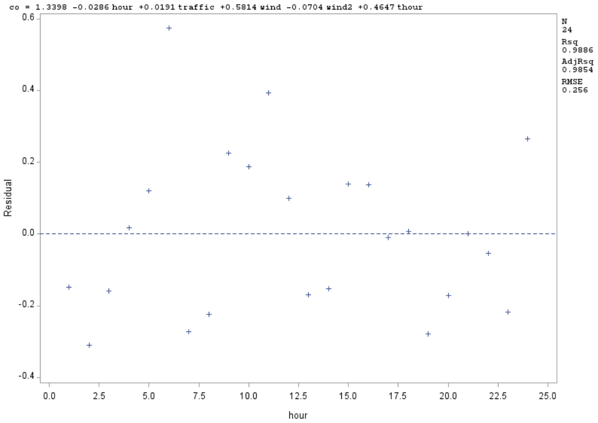
(2) Tolerance and Variance Inflation:

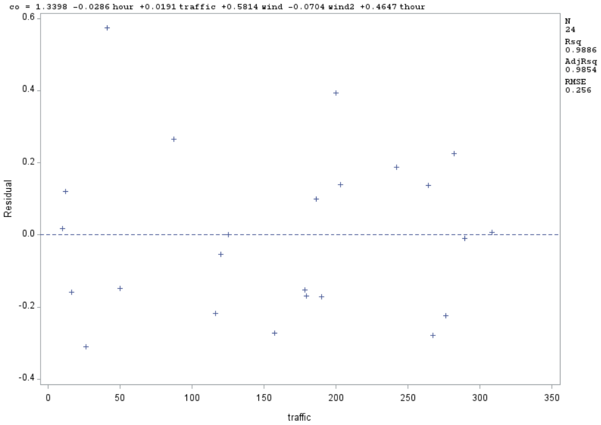


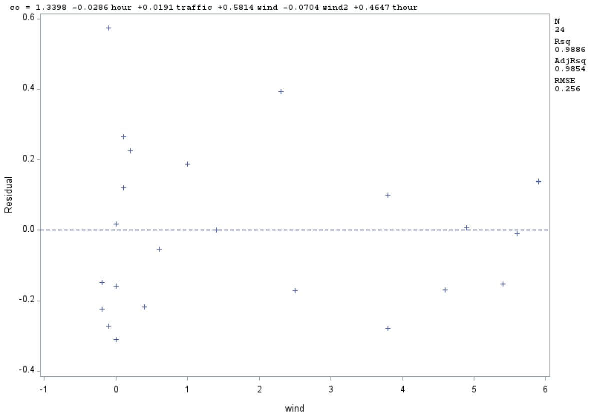
Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	1.33976	0.12262	10.93	<.0001	.	0
hour	1	-0.02864	0.00922	-3.11	0.0061	0.67066	1.49107
traffic	1	0.01907	0.00073929	25.80	<.0001	0.54266	1.84278
wind	1	0.58144	0.11448	5.08	<.0001	0.04041	24.74335
wind2	1	-0.07042	0.01944	-3.62	0.0019	0.04607	21.70568
thour	1	0.46475	0.07948	5.85	<.0001	0.86481	1.15633

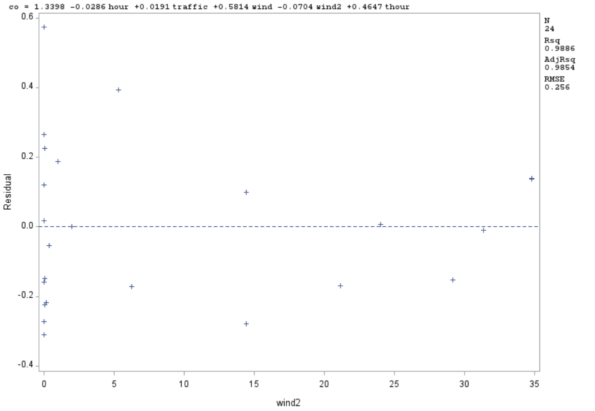
(3) Partial Plots

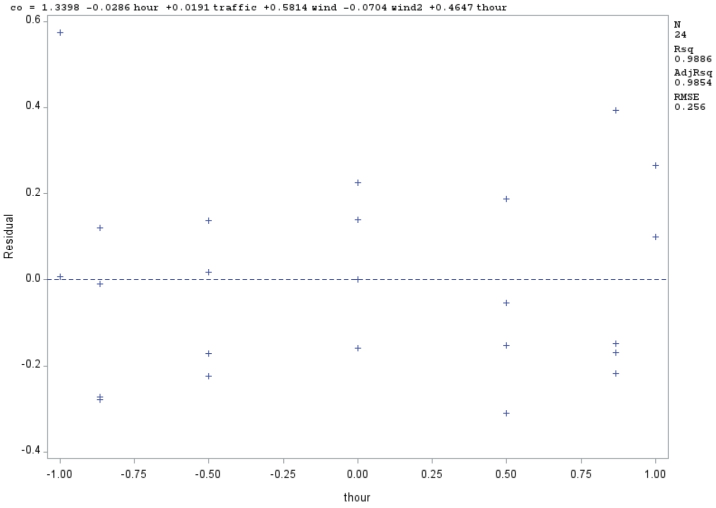



co = 1.3398 -0.0286 hour +0.0191 traffic +0.5814 wind -0.0704 wind2 +0.4647 thour
N
24
Rsq
0.9886
AdjRsq
0.9854
RMSE
0.256
Residual
0.6
0.4
0.2
0.0
-0.2
-0.4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
hour


co = 1.3398 -0.0286 hour +0.0191 traffic +0.5814 wind -0.0704 wind2 +0.4647 thour
N
24
Rsq
0.9886
AdjRsq
0.9854
RMSE
0.256
Residual
0.6
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
250
300
350
traffic


co = 1.3398 -0.0286 hour +0.0191 traffic +0.5814 wind -0.0704 wind2 +0.4647 thour
N
24
Rsq
0.9886
AdjRsq
0.9854
RMSE
0.256
Residual
0.6
0.4
0.2
0.0
-0.2
-0.4
-1
0
1
2
3
4
5
6
wind


co = 1.3398 -0.0286 hour +0.0191 traffic +0.5814 wind -0.0704 wind2 +0.4647 thour
N
24
Rsq
0.9886
AdjRsq
0.9854
RMSE
0.256
Residual
0.6
0.4
0.2
0.0
-0.2
-0.4
0
5
10
15
20
25
30
35
wind2


co = 1.3398 -0.0286 hour +0.0191 traffic +0.5814 wind -0.0704 wind2 +0.4647 thour
N
24
Rsq
0.9886
AdjRsq
0.9854
RMSE
0.256
Residual
0.6
0.4
0.2
0.0
-0.2
-0.4
-1.00
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00
thour

QUESTION 10.

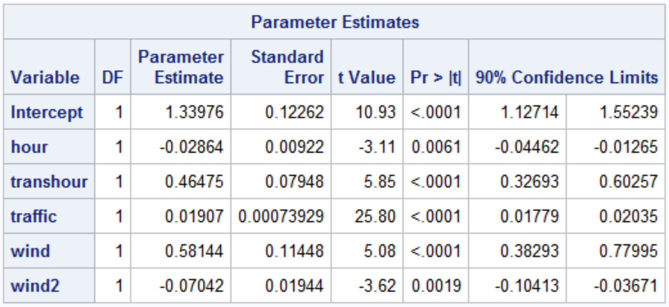
For the “best" model report the following:

(a) Equation of the regression model.
(b) 90% confidence interval for the mean of the response variable
(c) 90% prediction interval for individual observations.
(d) 90% confidence intervals for the regression coefficients.

SOLUTION:

(a) Equation of the regression model is:

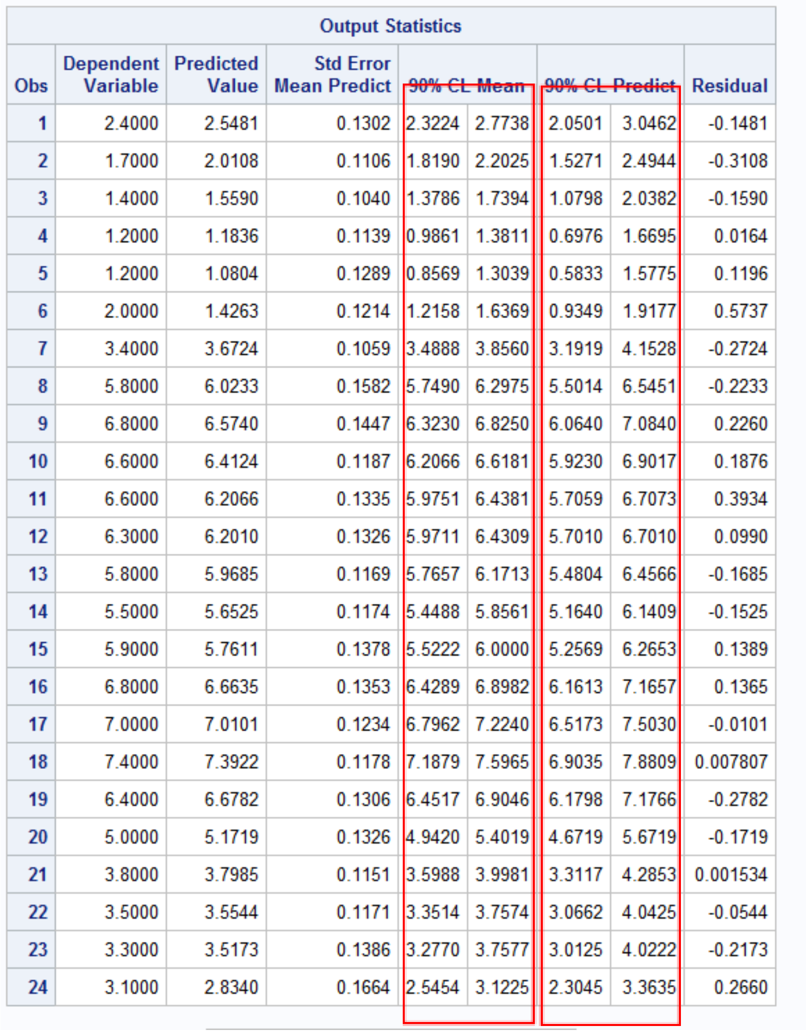
CO = 1.33976 - 0.02864 hour + 0.46475 thour + 0.01907 traffic + 0.58144 wind - 0.07042 wind^2



Parameter Estimates

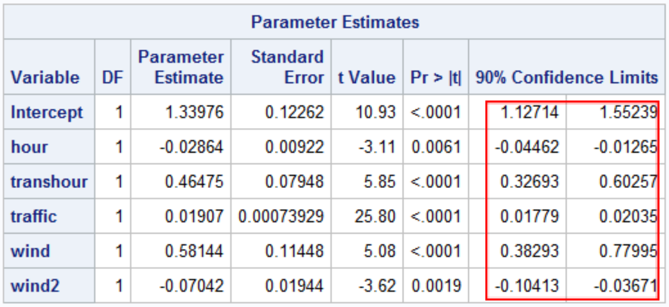
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
Intercept	1	1.33976	0.12262	10.93	<.0001	1.12714	1.55239
hour	1	-0.02864	0.00922	-3.11	0.0061	-0.04462	-0.01265
transhour	1	0.46475	0.07948	5.85	<.0001	0.32693	0.60257
traffic	1	0.01907	0.00073929	25.80	<.0001	0.01779	0.02035
wind	1	0.58144	0.11448	5.08	<.0001	0.38293	0.77995
wind2	1	-0.07042	0.01944	-3.62	0.0019	-0.10413	-0.03671

(b) 90% confidence interval for the ***mean of the response variable and individual observations*** are reported below encircled in red boxes:



Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	90% CL Mean		90% CL Predict		Residual
1	2.4000	2.5481	0.1302	2.3224	2.7738	2.0501	3.0462	-0.1481
2	1.7000	2.0108	0.1106	1.8190	2.2025	1.5271	2.4944	-0.3108
3	1.4000	1.5590	0.1040	1.3786	1.7394	1.0798	2.0382	-0.1590
4	1.2000	1.1836	0.1139	0.9861	1.3811	0.6976	1.6695	0.0164
5	1.2000	1.0804	0.1289	0.8569	1.3039	0.5833	1.5775	0.1196
6	2.0000	1.4263	0.1214	1.2158	1.6369	0.9349	1.9177	0.5737
7	3.4000	3.6724	0.1059	3.4888	3.8560	3.1919	4.1528	-0.2724
8	5.8000	6.0233	0.1582	5.7490	6.2975	5.5014	6.5451	-0.2233
9	6.8000	6.5740	0.1447	6.3230	6.8250	6.0640	7.0840	0.2260
10	6.6000	6.4124	0.1187	6.2066	6.6181	5.9230	6.9017	0.1876
11	6.6000	6.2066	0.1335	5.9751	6.4381	5.7059	6.7073	0.3934
12	6.3000	6.2010	0.1326	5.9711	6.4309	5.7010	6.7010	0.0990
13	5.8000	5.9685	0.1169	5.7657	6.1713	5.4804	6.4566	-0.1685
14	5.5000	5.6525	0.1174	5.4488	5.8561	5.1640	6.1409	-0.1525
15	5.9000	5.7611	0.1378	5.5222	6.0000	5.2569	6.2653	0.1389
16	6.8000	6.6635	0.1353	6.4289	6.8982	6.1613	7.1657	0.1365
17	7.0000	7.0101	0.1234	6.7962	7.2240	6.5173	7.5030	-0.0101
18	7.4000	7.3922	0.1178	7.1879	7.5965	6.9035	7.8809	0.007807
19	6.4000	6.6782	0.1306	6.4517	6.9046	6.1798	7.1766	-0.2782
20	5.0000	5.1719	0.1326	4.9420	5.4019	4.6719	5.6719	-0.1719
21	3.8000	3.7985	0.1151	3.5988	3.9981	3.3117	4.2853	0.001534
22	3.5000	3.5544	0.1171	3.3514	3.7574	3.0662	4.0425	-0.0544
23	3.3000	3.5173	0.1386	3.2770	3.7577	3.0125	4.0222	-0.2173
24	3.1000	2.8340	0.1664	2.5454	3.1225	2.3045	3.3635	0.2660

(c) 90% confidence intervals for the regression coefficients.



Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
Intercept	1	1.33976	0.12262	10.93	<.0001	1.12714	1.55239
hour	1	-0.02864	0.00922	-3.11	0.0061	-0.04462	-0.01265
transhour	1	0.46475	0.07948	5.85	<.0001	0.32693	0.60257
traffic	1	0.01907	0.00073929	25.80	<.0001	0.01779	0.02035
wind	1	0.58144	0.11448	5.08	<.0001	0.38293	0.77995
wind2	1	-0.07042	0.01944	-3.62	0.0019	-0.10413	-0.03671

References

- ➢ Mallows, C. L. (1973). "Some Comments on CP". Technometrics 15 (4): 661–675. doi:10.2307/1267380. JSTOR 1267380.

- ➢ Gilmour, Steven G. (1996). "The interpretation of Mallows's Cp-statistic". Journal of the Royal Statistical Society, Series D 45 (1): 49–56. JSTOR 234841

- ➢ Kutner et al., (2013) “Applied Linear Statistical Models” by (5th edition), ISBN: 0-07-238688-6.

- ➢ Introduction to SAS. UCLA: Statistical Consulting Group. Retrieved from: http://www.ats.ucla.edu/stat/sas/notes2/ (accessed November 24, 2014).

- ➢ Efroymson,M. A. (1960) "Multiple regression analysis," Mathematical Methods for Digital Computers, Ralston A. and Wilf,H. S., (eds.), Wiley, New York.

- ➢ Hocking, R. R. (1976) "The Analysis and Selection of Variables in Linear Regression," Biometrics, 32.

- ➢ Draper, N. and Smith, H. (1981) Applied Regression Analysis, 2d Edition, New York: John Wiley & Sons, Inc.

- ➢ SAS Institute Inc. (1989) SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2, Cary, NC: SAS Institute Inc.

APPENDIX

QUESTION 1:

```
data COC;
infile 'C:\Users\DELL\Desktop\COC.DAT';
input hour co traffic wind;
seq=_n_;
proc print data=COC;
run;
*Generate a scatterplot for hour with smooth curve fitted to the data;
proc sort data = coc; by hour; run;
symbol1 v=square i=sm85;
title1 'Scatter Plot';
title2 'Scatter plot of CO concentration vs. Hour with Smoothing Curve (sm85)';
axis1 label=('hour');
axis2 label=(angle=90 'CO Concentration');
proc gplot data=COC;
      plot co*hour / haxis=axis1 vaxis=axis2; run;

*Generate a scatterplot for traffic with smooth curve fitted to the data;
proc sort data = coc; by traffic; run;
symbol1 v=square i=sm85;
title1 'Scatter Plot';
title2 'Scatter plot of CO concentration vs. Traffic with Smoothing Curve (sm85)';
axis1 label=('traffic');
axis2 label=(angle=90 'CO Concentration');
proc gplot data=COC;
      plot co*traffic / haxis=axis1 vaxis=axis2; run;

*Generate a scatterplot for wind with smooth curve fitted to the data;
proc sort data = coc; by wind; run;
symbol1 v=square i=sm85;
title1 'Scatter Plot';
title2 'Scatter plot of CO concentration vs. wind with Smoothing Curve (sm85)';
axis1 label=('wind');
axis2 label=(angle=90 'CO Concentration');
proc gplot data=COC;
      plot co*wind / haxis=axis1 vaxis=axis2; run;

      *regression;

      proc reg data=COC;
model CO=hour traffic wind/clb p r;
output out=diag p=pred r=resid;
run;

*Plotting residuals versus explanatory variables;
proc gplot data=diag;
symbol1 v=circle;
title1 'Residual Plot for Linear Assupmptions Verification';
```

```
title2 'Residual plot';
plot resid*pred / haxis=axis1 vaxis=axis2  vref=0; run;
plot resid*hour / haxis=axis1 vaxis=axis2 vref=0;
plot resid*traffic / haxis=axis1 vaxis=axis2 vref=0;
plot resid*wind / haxis=axis1 vaxis=axis2 vref=0; run;

proc univariate data=diag plot normal;
var resid;
title1 'Study of CO Concentration vs hour, traffic, wind';
title2 'Residual plot';
histogram resid / normal kernel(L=2);
qqplot resid/normal (L=1 mu=est sigma=est); run;

***check for transformations;

proc transreg data = COC;
model boxcox(co)=identity(wind); run;
proc transreg data = COC;
model boxcox(co)=identity(hour); run;
proc transreg data = COC;
model boxcox(co)=identity(traffic); run;


***including transformed variable in the model;

data coc;
infile 'C:\Users\DELL\Desktop\coc.dat';
input hour  co traffic wind;
run;
data coc1;
set coc;
wind2=wind*wind;
thour= cos(0.523598776*hour)
run;
proc print data=coc1;
run;
```

QUESTION 2:

```
title Problem2;
symbol1 v=circle i=sm80 c=black;
proc sort data=COC; by Wind;
proc gplot data=COC;
plot CO*Wind; run;
data piecewise; set COC;
        if Wind le 2
                then cslope=0;
        if Wind gt 2
                then cslope=Wind-2; run;
proc print data=piecewise; run;
proc reg data=piecewise;
        model CO = Wind cslope;
        output out = pieceout p=COhat; run;
```

```
symbol1 v=circle i=none c=black;
symbol2 v=none i=join c=blue;
proc sort data=pieceout; by Wind; run;
proc gplot data=pieceout;
plot (CO COhat)*Wind/overlay;
run;
proc print data=piecewise;run;

symbol1 v=M i=sm70 c=black l=1;
symbol2 v=S i=sm70 c=black l=3;

proc sort data=piecewise;
        by Wind cslope; run;

proc gplot data=piecewise;
   plot CO*Wind=cslope;
run;
data piecewise; set piecewise;
   Windcslope=Wind*cslope;

proc print data=piecewise; run;

proc reg data=piecewise;
   model CO = Wind cslope Windcslope;
   sameline: test cslope, Windcslope;
   parallel: test Windcslope;
   sameintercept: test cslope; run;
```

QUESTION 3:

```
data trans; set CO1;
  sum=Wind+Traffic;
proc print data=trans;
run;
  proc reg data=trans;
      model CO = sum Wind Traffic Hour Wind2 thour;
      model CO = Wind Traffic Hour Wind2 thour;
            test SUM; run;
```

QUESTION 4:

```
proc reg data = CO1;
model CO1 = Hour Traffic Wind Wind2 thour/ ss1 ss2;
run;
```

QUESTION 5:

```
data co;
infile 'W:/cofreewy.dat';
input hour  co traffic wind;
run;
data co1;
set co;
sum=traffic+wind;
```

```
wind2=wind*wind;
thour=cos(0.523598776*hour);
run;
proc print data=co1;run;
proc reg data=co1;
model co=hour thour traffic wind;run;
model co=thour traffic wind wind2;
model co=thour traffic wind2;
model co=thour traffic wind;
model co=hour sum;
model co=hour traffic wind;
model co=hour traffic;
model co=sum;
model co=traffic wind;
model co=hour traffic wind2;
model co=hour;
model co=wind2;
model co=traffic;
model co=wind;
model co=hour wind2;
model co=traffic wind2;
model co=hour traffic wind wind2;
model co=thour hour traffic wind wind2;
run;
```

QUESTION 6, 7:

```
data coc;
infile 'C:\Users\DELL\Desktop\coc.dat';
input hour  co traffic wind;
run;
data coc1;
set coc;
wind2=wind*wind;
thour=cos(0.523598776*hour);
run;
proc print data=coc1;
run;
proc corr data = coc1;
  var hour traffic wind wind2 thour;
run;

*******Cp selection;
proc reg data=coc1;
model co=hour traffic wind wind2 thour/selection = cp b;
run;

*******Stepwise selection;
proc reg data=coc1;
model co=hour traffic wind wind2 thour/selection=stepwise;
run;
```

```
*******New coding after fixing correlation problem;

data copy; set coc1;
swind=wind;
drop wind2; run;
proc standard data=copy out=std mean=0;
var swind; run;
proc print data=std;
run;
data std; set std;
swind2=swind*swind;run;
proc reg data=std;
model co=hour traffic wind swind2/selection = cp b;
run;
proc corr data= std noprob;
var hour traffic wind swind2;
run;
```

QUESTION 8:

```
data coc;
infile 'C:\Users\DELL\Desktop\coc.dat';
input hour  co traffic wind;
run;
data coc1;
set coc;
wind2=wind*wind;
thour=cos(0.523598776*hour);
seq=_n_;
run;
proc print data=coc1;
run;
proc reg data=coc1;
model co=hour traffic wind wind2 thour/clb p r;
output out=diag p=pred r=resid;
run;

*********for hour;
proc sort data=diag; by hour; run;
symbol1 v=circle i=sm85;
title1'Q#8 BEST Model';
title2'Residual Plot for variable hour';
proc gplot data=diag;
plot resid*hour/vref=0;
run;

*********for traffic;
proc sort data=diag; by traffic;run;
title1'Q#8 BEST Model';
title2'Residual Plot for variable traffic';
proc gplot data=diag;
plot resid*traffic/vref=0;
```

```
run;
*******************for wind;
proc sort data = diag; by wind;run;
symbol1 v=circle i=sm85;
title1'Q#8 BEST Model';
title2'Residual Plot for variable wind';
proc gplot data=diag;
plot resid*wind/vref=0;
run;
***********************for wind2;
proc sort data = diag; by wind2;run;
symbol1 v=circle i=sm85;
title1'Q#8 BEST Model';
title2'Residual Plot for variable wind2';
proc gplot data=diag;
plot resid*wind2/vref=0;
run;

***********************for thour;
proc sort data = diag; by thour;run;
symbol1 v=circle i=smnn;
title1'Q#8 BEST Model';
title2'Residual Plot for variable wind2';
proc gplot data=diag;
plot resid*thour/vref=0;
run;

*****************for univariate command;
proc univariate data=diag;
title1'Q#8, BEST MODEL ASSUMPTIONS';
title2'qq-plot for residuals';
var resid;
histogram resid / normal kernel(L=2);
qqplot resid/normal (L=1 mu=est sigma=est);
run;


*residuals time series plots;
proc sort data = diag; by resid;run;
title1 'Sequence plot for residuals with smooth curve';
symbol1 v=circle;
proc gplot data=diag;
plot resid*seq;
run;

*sequence plot for y;

title1 'Sequence plot for co with smooth curve';
symbol1 v=circle i=sm85;
proc gplot data=COC1;
plot co*seq;
run;

*sequence plots for explanatory variables;
```

```
title1 'Sequence plot for hour with smooth curve';
symbol1 v=circle i=sm85;
proc gplot data=COC1;
plot hour*seq;
run;

title1 'Sequence plot for traffic with smooth curve';
symbol1 v=circle i=sm85;
proc gplot data=COC1;
plot traffic*seq;
run;

title1 'Sequence plot for wind with smooth curve';
symbol1 v=circle i=sm85;
proc gplot data=COC1;
plot wind*seq;
run;

title1 'Sequence plot for wind2 with smooth curve';
symbol1 v=circle i=sm85;
proc gplot data=COC1;
plot wind2*seq;
run;

title1 'Sequence plot for thour with smooth curve';
symbol1 v=circle i=sm85;
proc gplot data=COC1;
plot thour*seq;
run;

*Generate a scatterplot for hour with smooth curve fitted to the data;
proc sort data = coc1; by hour;run;
symbol1 v=square i=sm85;
title1 'Scatter Plot';
title2 'Scatter plot of CO concentration vs. Hour with Smoothing Curve
(sm85)';
axis1 label=('hour');
axis2 label=(angle=90 'CO Concentration');
proc gplot data=COC1;
      plot co*hour / haxis=axis1 vaxis=axis2; run;

*Generate a scatterplot for traffic with smooth curve fitted to the data;
proc sort data = coc1; by traffic;run;
symbol1 v=square i=sm85;
title1 'Scatter Plot';
title2 'Scatter plot of CO concentration vs. Traffic with Smoothing Curve
(sm85)';
axis1 label=('traffic');
axis2 label=(angle=90 'CO Concentration');
proc gplot data=COC1;
      plot co*traffic / haxis=axis1 vaxis=axis2; run;

*Generate a scatterplot for wind with smooth curve fitted to the data;
```

```
proc sort data = coc1; by wind;run;
symbol1 v=square i=sm85;
title1 'Scatter Plot';
title2 'Scatter plot of CO concentration vs. wind with Smoothing Curve
(sm85)';
axis1 label=('wind');
axis2 label=(angle=90 'CO Concentration');
proc gplot data=COC1;
      plot co*wind / haxis=axis1 vaxis=axis2; run;

*Plotting residuals versus explanatory variables;
proc sort data = coc1; by wind2;run;
symbol1 v=square i=sm85;
title1 'Scatter Plot';
title2 'Scatter plot of CO concentration vs. wind2 with Smoothing Curve
(sm85)';
axis1 label=('wind2');
axis2 label=(angle=90 'CO Concentration');
proc gplot data=COC1;
      plot co*wind2 / haxis=axis1 vaxis=axis2; run;

proc sort data = coc1; by thour;run;
symbol1 v=square i=sm85;
title1 'Scatter Plot';
title2 'Scatter plot of CO concentration vs. thour with Smoothing Curve
(sm85)';
axis1 label=('wind2');
axis2 label=(angle=90 'CO Concentration');
proc gplot data=COC1;
      plot co*thour / haxis=axis1 vaxis=axis2; run;
```

SEQUENCE PLOT FOR RESIDUALS:

```
data coc;
infile 'C:\Users\DELL\Desktop\coc.dat';
input hour co traffic wind;
run;
proc print data = coc; run;
data CO1; set COC;
        Wind2=Wind*Wind;
            thour=cos(0.523598776*hour);
            seq=_n_; run;
proc print data=CO1;run;

proc reg data=co1;
model co=hour traffic wind wind2 thour/clb p r;
output out=diag p=pred r=resid;run;

proc sort data = diag; by resid;run;
symbol1 v=circle i=rl;
title1'Sequence Plot for residuals';
proc gplot data=diag;
plot resid*seq;
```

```
run;
```

QUESTION 9:

```
proc reg data = CO1;
model CO = Hour Traffic Wind Wind2 transhour/r influence;
run;
proc reg data = CO1;
model CO = Hour Traffic Wind Wind2 transhour/tol vif;
run;
proc reg data = CO1;
model CO = Hour Traffic Wind Wind2 transhour/r partial;
plot r.*(Hour Traffic Wind Wind2 transhour);
run;
proc reg data = CO1;
model CO = Hour Traffic Wind Wind2 transhour/r vif;
plot r.*(Hour Traffic Wind Wind2 transhour); run;
```

QUESTION 10:

```
proc reg data=co1 alpha=0.1;
model co=hour transhour traffic wind wind2/clm clb cli;
run;
```