Alicia Ngoc Phan

BSTA 477/677 – Winter 2021

Tutorial 4 - March 6th, 2021

- Linear Regression
    - Data partition
    - Selection methods
    - Linear Regression Training results
    - Linear Regression Validation results
    - Output evaluation
    - Prediction interval
    - Generalized difference
- Durbin Watson test

Data used: Bike sharing data

Alicia Ngoc Phan

Linear Regression

Data partition

Reference to Tutorial 1 for data partition.

Selection methods

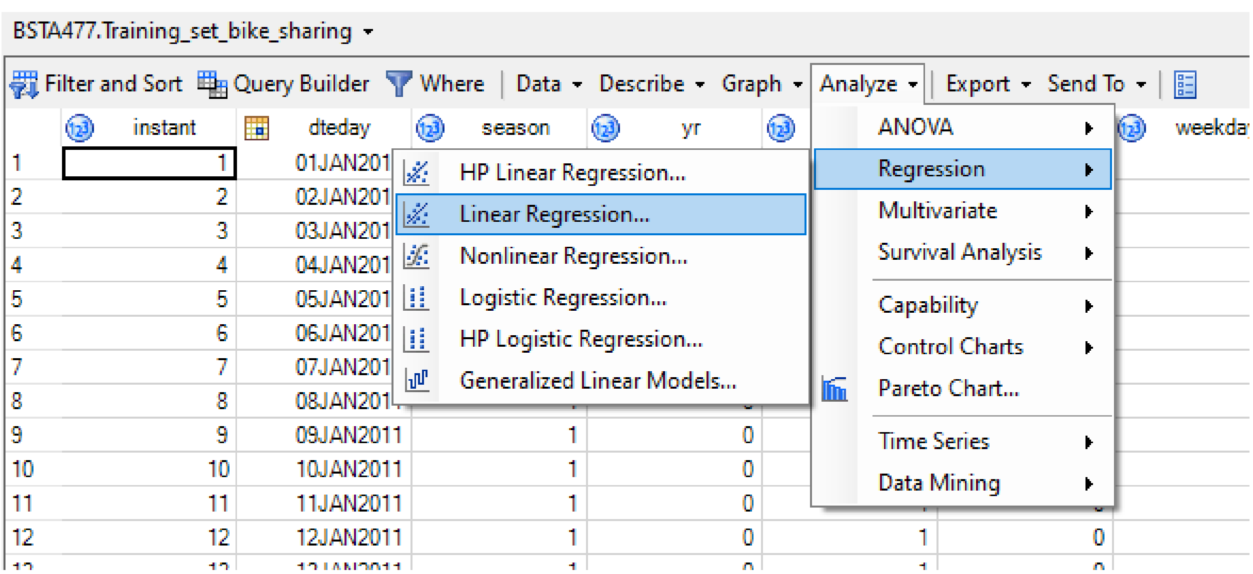
Resources:
- https://support.sas.com/resources/papers/proceedings/proceedings/sugi29/117-29.pdf
- https://blogs.sas.com/content/iml/2020/01/23/collinearity-regression-collin-option.html

To select predictors appropriately avoid two common mistakes:
- Plot the relationship between predictors and forecast variables to choose predictors.
- Fit all the predictors into the model and drop predictors that have p-value greater than 0.05.

Instead use the following techniques to select predictors and form the best model:
- Fit all potential regression models. Easy approach: Stepwise regression
- Evaluate the models based on Adjusted R-squared, cross-validation, AIC, BIC
- Evaluate Collinearity diagnostics
- Take into considerations of the project priorities.

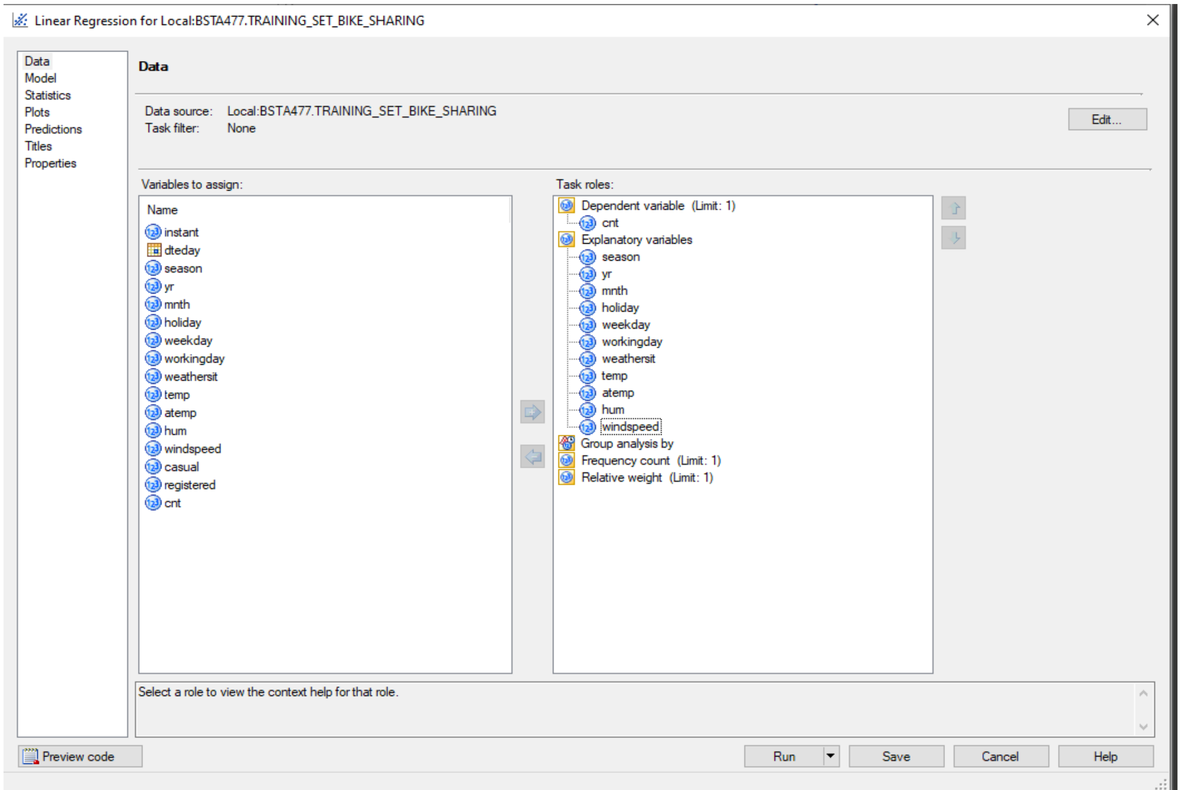
1. **Use stepwise regression on all predictors**

Choose the Linear regression task under the Analyze tab.

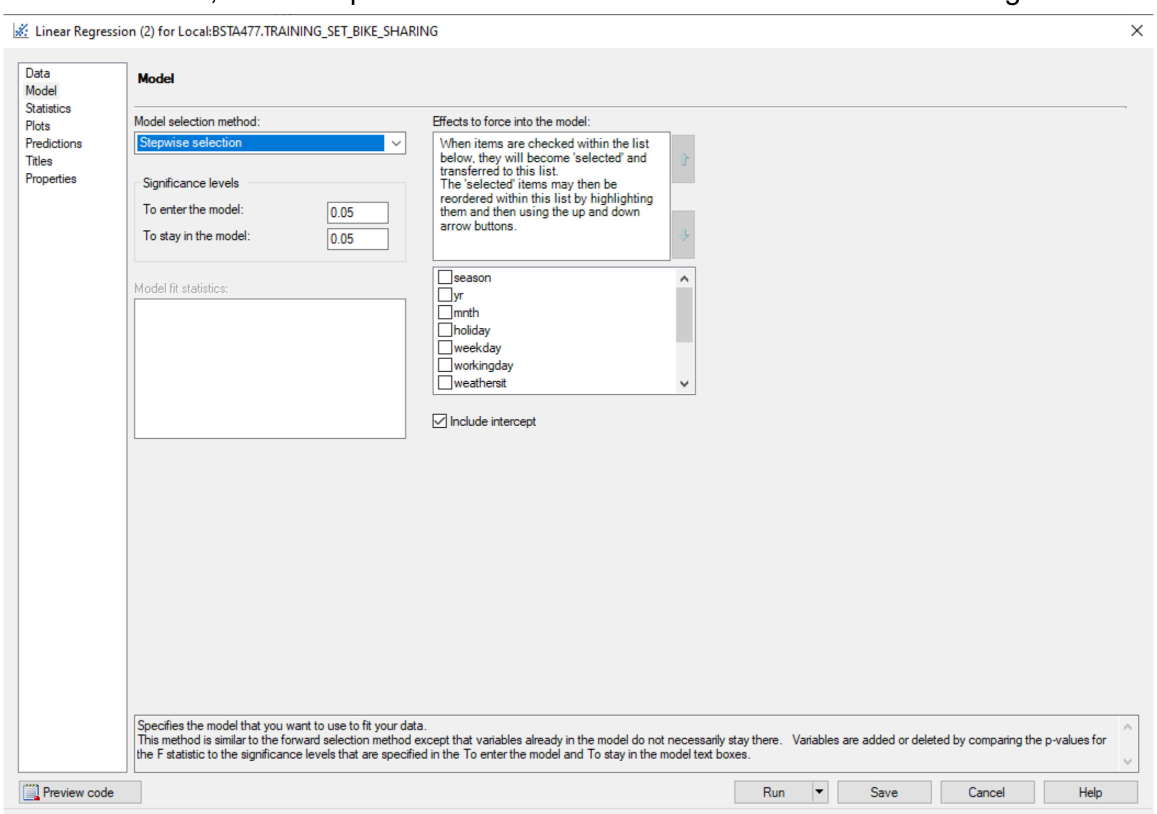


Alicia Ngoc Phan

In Data tab, select dependent and independent variables of the model

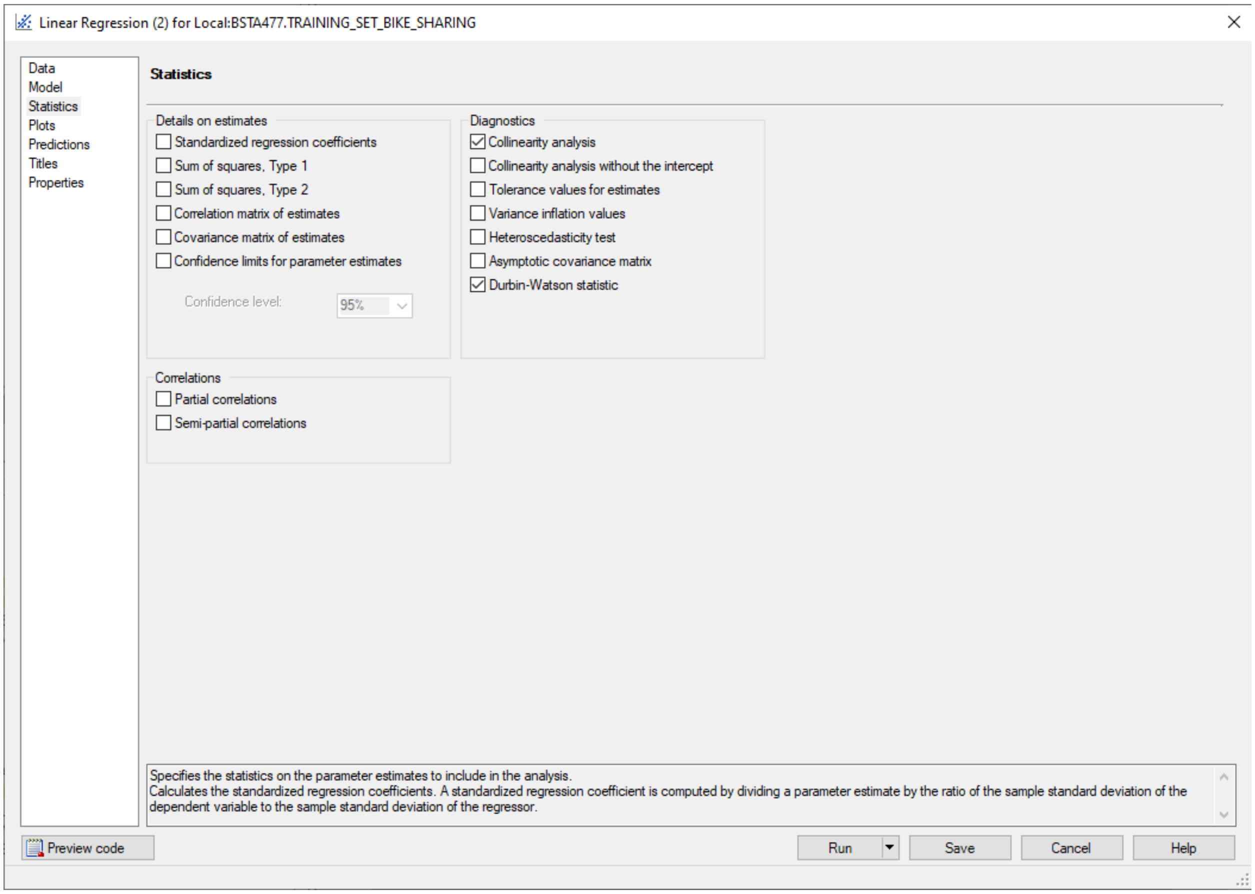


In the model tab, select stepwise selection for Model selection method. Add significant level



Alicia Ngoc Phan

To check for collinearity in predictors and autocorrelation of residuals, choose Collinearity analysis and Durbin Watson statistics in Diagnostics:



2. **Evaluate model selection results: Check the following statistics**
    - R squared: Goodness of fit statistics. This statistics measure the percentage of variance in the forecasting variable that the predictors were able to explain.

Result: ***R-Square = 0.7997***

=> 79%-80% of the variances in the bike rentals variable are explained by the predictors.

- Conduct F test: F test tests if any predictors in the linear regression model are significant. (Can use the p-value).

Ho: B1= B2 = B3….=0 (All parameter estimates are equal to zero, predictors are insignificant)
Ha: Bj != 0 (At least one of the predictors are significant)

Decision matrix:

- P-value < significant value => Reject Ho, the model is significant
- P-value > significant value => Accept Ho, the model is insignificant because predictors are insignificant.
-

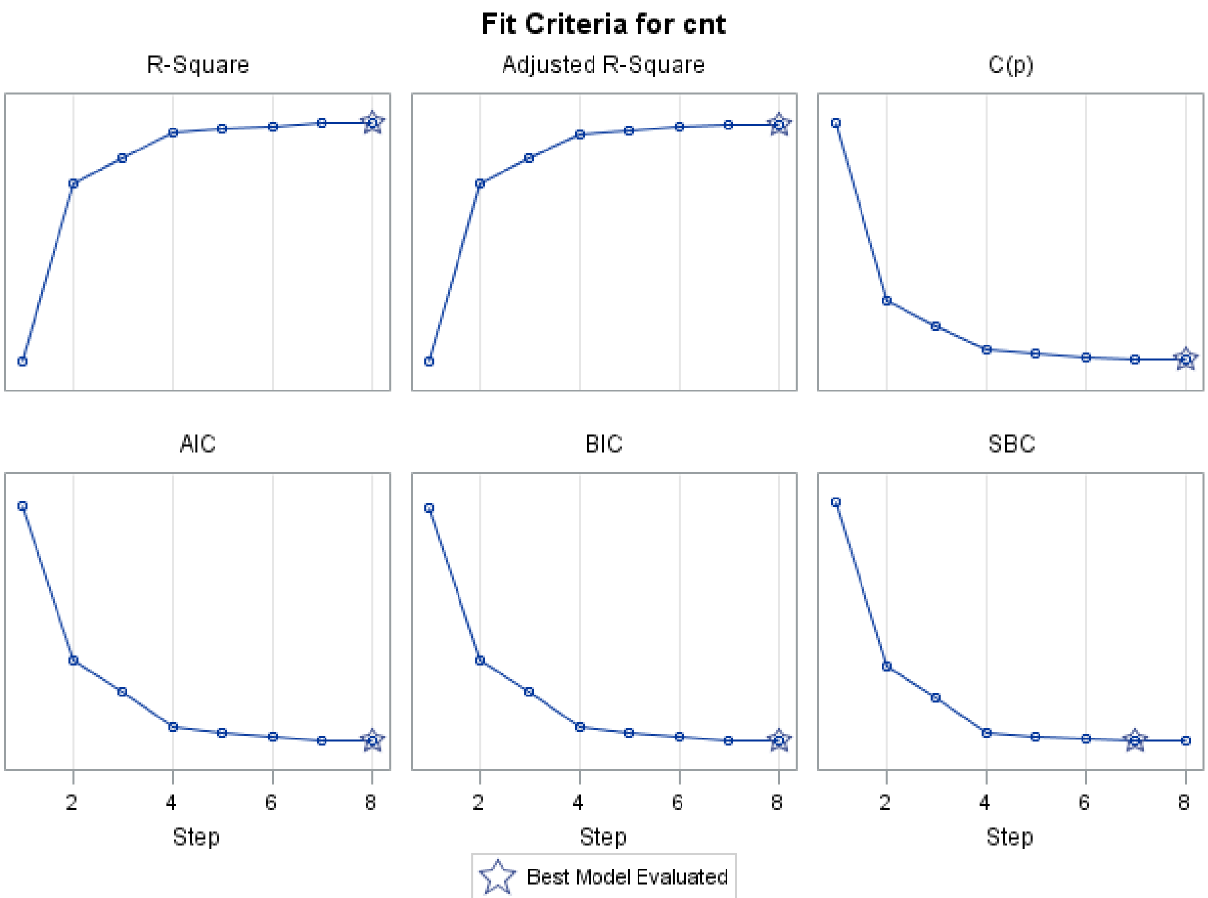
Result:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1499703847	187462981	287.42	<.0001
Error	576	375679414	652221		
Corrected Total	584	1875383261			

Set a significant level at 0.05, we can see that p-value <0.0001, which is less than 0.05. We conclude that there is enough evidence to reject Ho. And the model is significant.

- Adjusted R-squared, AIC, BIC, SBC

Result:



The final model chosen by SAS EG is considered the best overall model through assessing the results of R-squared, Adjusted R-squared, C(p), AIC, BIC, SBC. However, based on the project priorities, the team can choose the model that is best for the project.

For example, the lowest AIC usually indicates the best model; however, AIC sometimes let too many predictors in the model. If the cost of collecting predictors is high, then this would not be a good assessment using AIC. Instead, the team can choose the model that has the lowest BIC as BIC penalizes the model for adding more predictors.

With the situation above, we can see that generally, the final model is best, with lowest values in both AIC and BIC and highest R squared and adjusted R squared.

- Collinearity diagnostics

Collinearity Diagnostics (intercept adjusted)

			Proportion of Variation							
Number	Eigenvalue	Condition Index	season	yr	holiday	weekday	weathersit	atemp	hum	windspeed
1	1.91429	1.00000	0.08501	0.03768	0.00258	1.275259E-7	0.03644	0.04007	0.08522	0.05584
2	1.45899	1.14545	0.06291	0.00322	0.00148	0.00269	0.15425	0.13655	0.04792	0.02734
3	1.10051	1.31888	0.00090494	0.02906	0.43252	0.41121	0.00010243	0.00708	0.00119	0.00160
4	1.03621	1.35919	0.03685	0.51803	0.00005600	0.08342	0.01439	0.04418	0.02608	0.05627
5	0.89544	1.46213	0.00146	0.03705	0.54765	0.45761	0.01055	0.01713	0.00049330	0.00389
6	0.85861	1.49316	0.03938	0.01295	0.01429	0.02370	0.02888	0.12901	0.00003388	0.70321
7	0.42742	2.11629	0.71587	0.27010	0.00135	0.00428	0.06269	0.42028	0.10478	0.00658
8	0.30853	2.49088	0.05761	0.09190	0.00007193	0.01710	0.69270	0.20570	0.73429	0.14526

To assess the collinearity between predictors, we evaluate the collinearity diagnostics. We should look at the Condition index to identify high values in this column (usually would be higher than 30). However, the values in the Condition index in the result are quite small, thus, choose the highest values possible: 2.11 and 2.49.

We scan the rows of the highest values of the Condition index and assess if any predictor has a high proportion that contributes to collinearity (> 0.5 is significant). Usually, we would look for pairs of high values in the row.

We can see that, on line 7, season has high value of 0.7 and the second highest is atemp at 0.42. From this, we can see that the season and atemp (temperature variable) have a linear relationship, however, small one as atemp value is not significant.

On line 8, the variable weathersit and hum (humidity) has high values of proportion of variation, both are higher than 0.5. Therefore, we can conclude that these two values have strong linear relationships.

Alicia Ngoc Phan

Linear Regression Training results

From the results of predictor selection - stepwise and your chosen significance level:

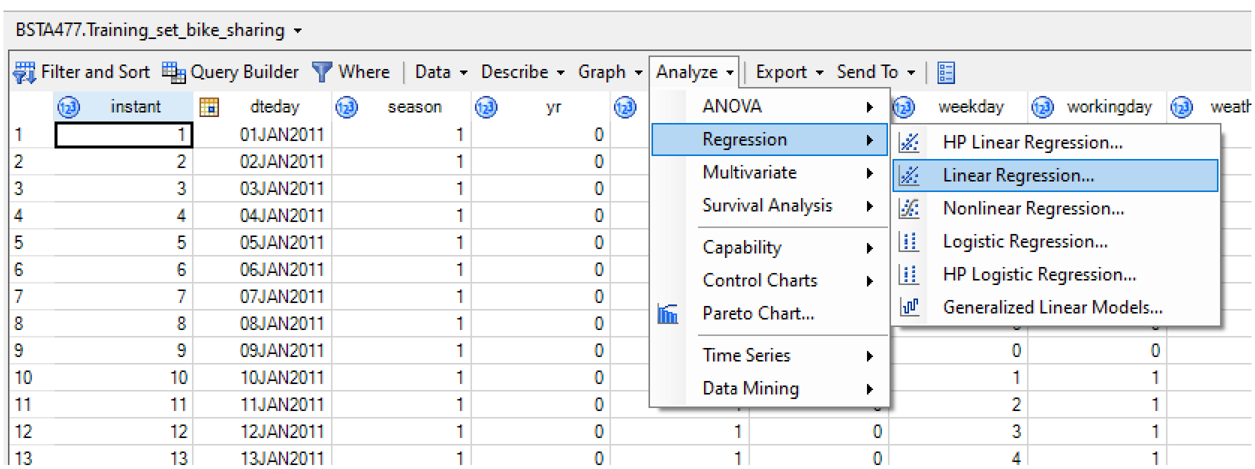
Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	atemp		1	0.4771	0.4771	922.021	531.92	<.0001
2	yr		2	0.2410	0.7181	231.221	497.65	<.0001
3	weathersit		3	0.0342	0.7523	135.042	80.11	<.0001
4	season		4	0.0328	0.7851	42.7221	88.56	<.0001
5	windspeed		5	0.0061	0.7912	27.2366	16.87	<.0001
6	hum		6	0.0040	0.7951	17.8398	11.19	0.0009
7	weekday		7	0.0031	0.7982	10.9190	8.88	0.0030
8	holiday		8	0.0014	0.7997	8.7937	4.13	0.0427

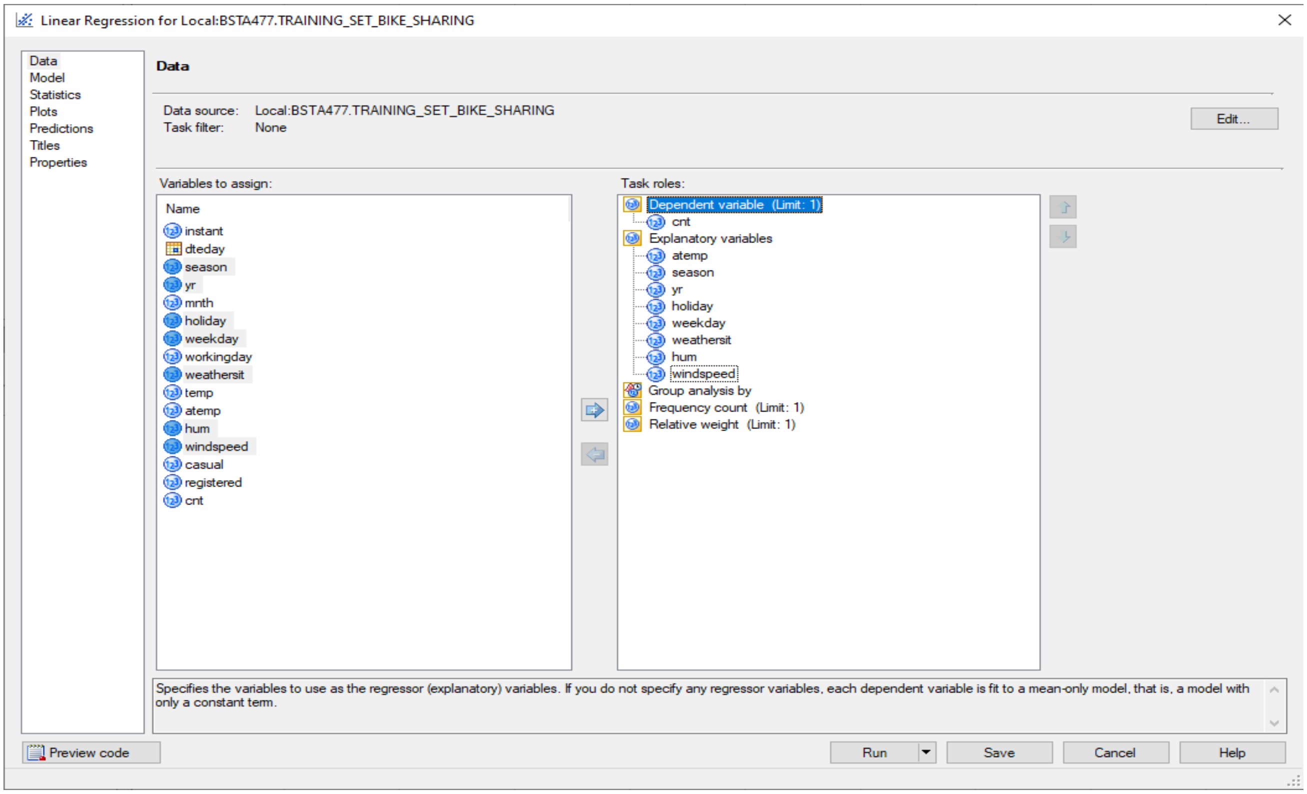
For example, with chosen significance level of 0.05, all the predictors from the selected final model seemed to be significant. Therefore, from 11 variables (season, yr, mnth, holiday, weekday, workingday, weathersit, temp,atemp, hum, windspeed), we narrowed down to 8 predictors for the model.

Note: Do keep in mind that the collinearity still exists (the season and atemp predictors). With this, we will evaluate the residuals of the fitted model for the performance and address if there is an issue.

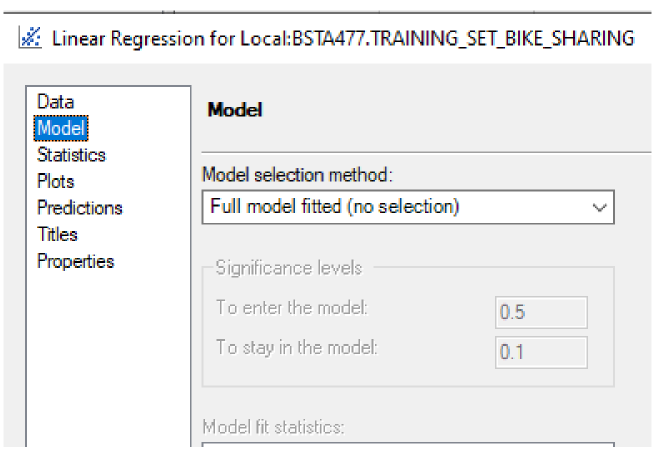
1. **Fit the model**:
    a. After choosing appropriate predictors, fit the model with the chosen predictors (8 predictors). Follow the following steps and click run.



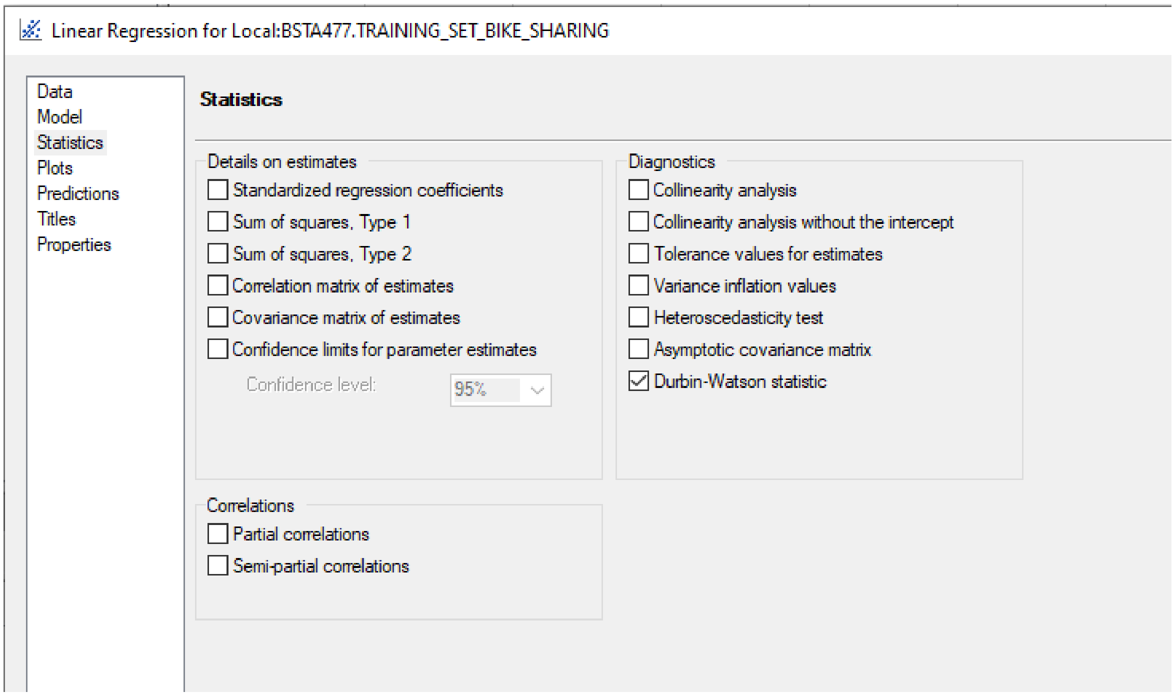
Alicia Ngoc Phan



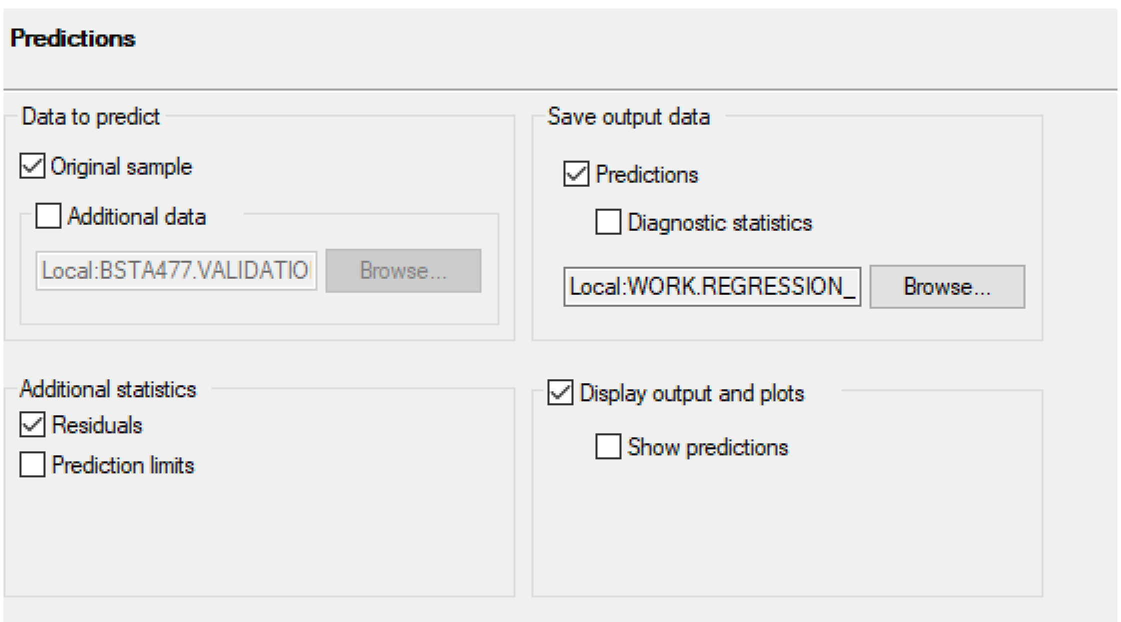
b. Select full model fitted:



c. Choose Diagnostics: Durbin Watson statistics (and others if necessary)



d. Save prediction data: Select “Original Sample” in Data to predict. Select Predictions and Save data as “Work.regression_training” in Save output data.



e. Click run.

2. **Residual diagnostics**: After running the model. Conduct residuals diagnostics to ensure:

- <u>Residuals don’t have autocorrelation</u>: Check using Durbin Watson test or Ljung Box test (please check the Durbin Watson test instruction down below. For the Ljung Box test, please check Tutorial 2 for instructions).

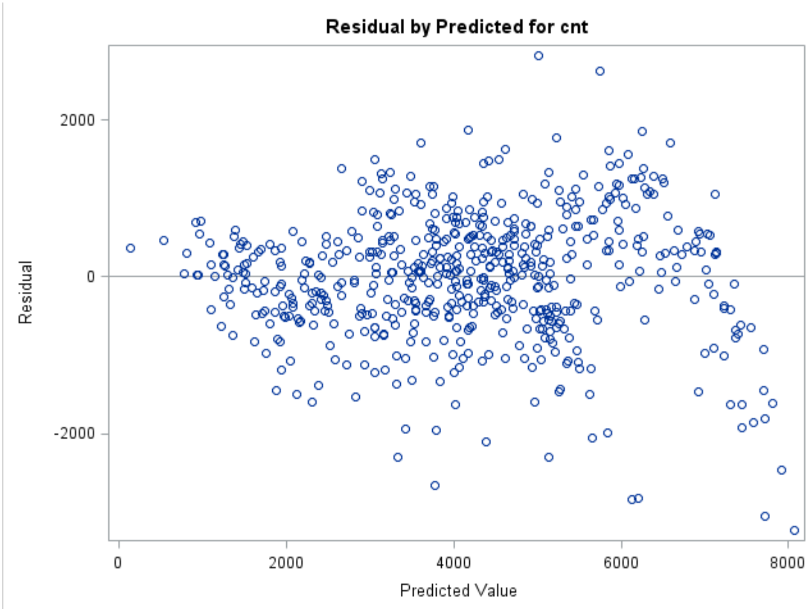
- Residuals mean equals to zero: Check using proc means with the residuals

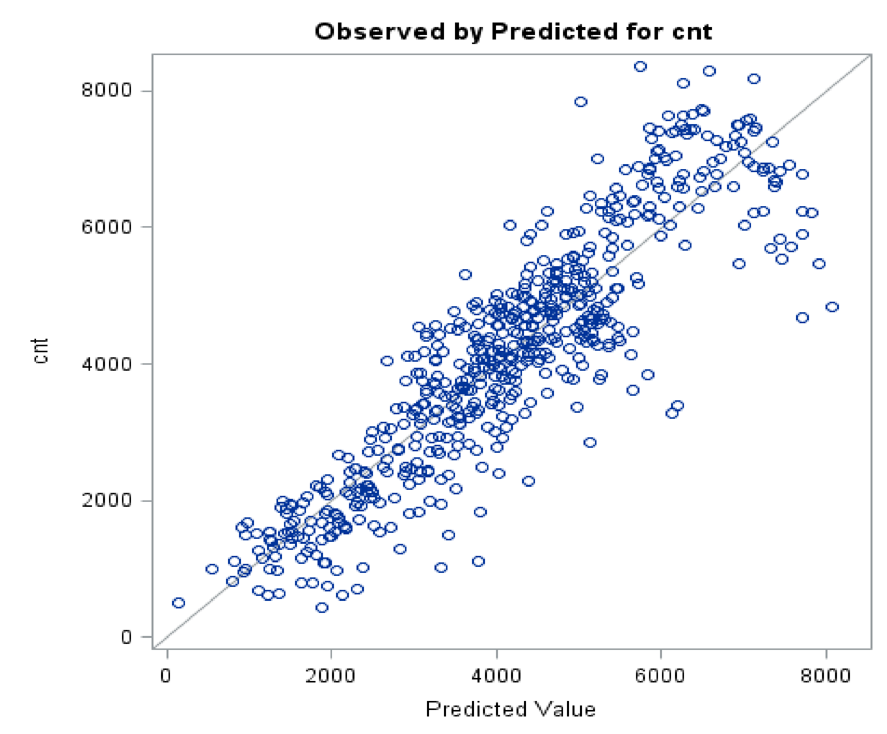
```
proc means data=work.regression_training maxdec=2;
 var residual_cnt;
 run;
```

Result: Mean of residuals is close to zero. (as the mean is rounded up 2 decimals)

Analysis Variable : residual_cnt Residual				
N	Mean	Std Dev	Minimum	Maximum
585	-0.00	802.05	-3229.97	2819.77

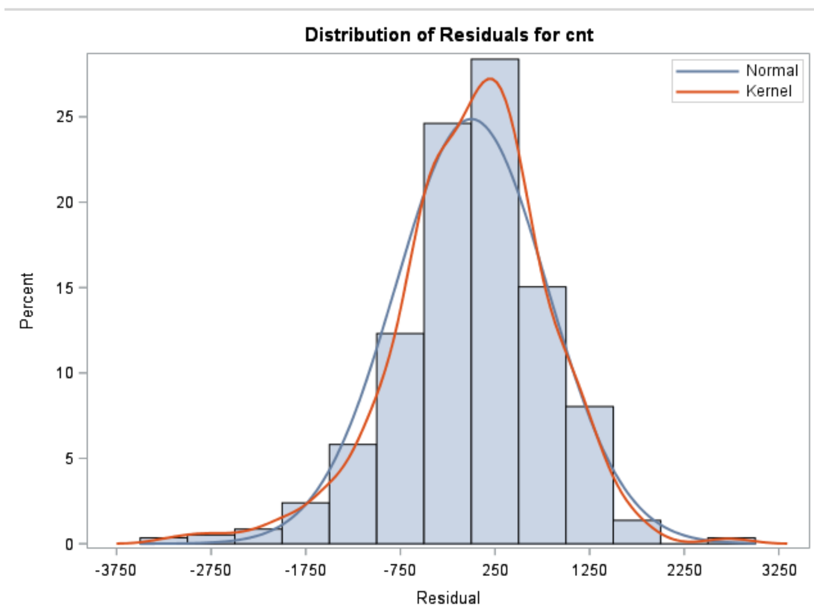
- Residuals have constant variance: Check by looking at the following graphs





We can see that the residuals have slightly constant variance at the beginning; however, larger variance increasingly. This indicates that residuals variance is not constant. We might address this by adding squared predictors or using Box Cox transformation.

- Residuals are normally distributed: Check by looking at residuals histogram



3. **Calculate other error terms**: As SAS do not compute other error terms beside RMSE and MSE, we compute MAPE, MPE separately. (MSE and RMSE calculation are included below for completion of error terms - Read important note below)

```
data training_regression_result;
  set work.regression_training;
  abs = abs(residual_cnt);
  square = residual_cnt**2;
  proportion = residual_cnt/cnt;
  abs_proportion = abs/cnt;
  run;

proc summary data=work.training_regression_result;
  var abs square proportion abs_proportion;
  output out=total_train_errors sum= / autoname;
  run;

data training_error_result;
  set total_train_errors;
  MAE = abs_Sum/585;
  MSE = square_Sum/(585-8-1);
  RMSE = sqrt(MSE);
  MPE = proportion_Sum/585;
  MAPE = abs_proportion_Sum/585;
  run;

proc print data=training_error_result; run;
```

Notes: The calculation for MSE and RMSE for regression incorporates the number of predictors in the regression model. Therefore, the MSE = SSE / (n-k-1), where SSE is the sum of squared errors, k: number of predictors. The remaining error terms are calculated as normal.

=> 585 in the code above is the number of observations in the training set.
=> MSE calculation = SSE/ (585 -8 -1), where 8 is the number of predictors in the current model.

Result:

TRAINING_ERROR_RESULT

Filter and Sort | Query Builder | Where | Data | Describe | Graph | Analyze | Export | Send To

	abs_Sum	square_Sum	proportion_Sum	abs_proportion_Sum	MAE	MSE	RMSE	MPE	MAPE
1	357564.59657	375679414.11	-37.60189898	113.86100014	611.22153259	652221.20506	807.6021329	-0.06427675	0.1946341883

Alicia Ngoc Phan

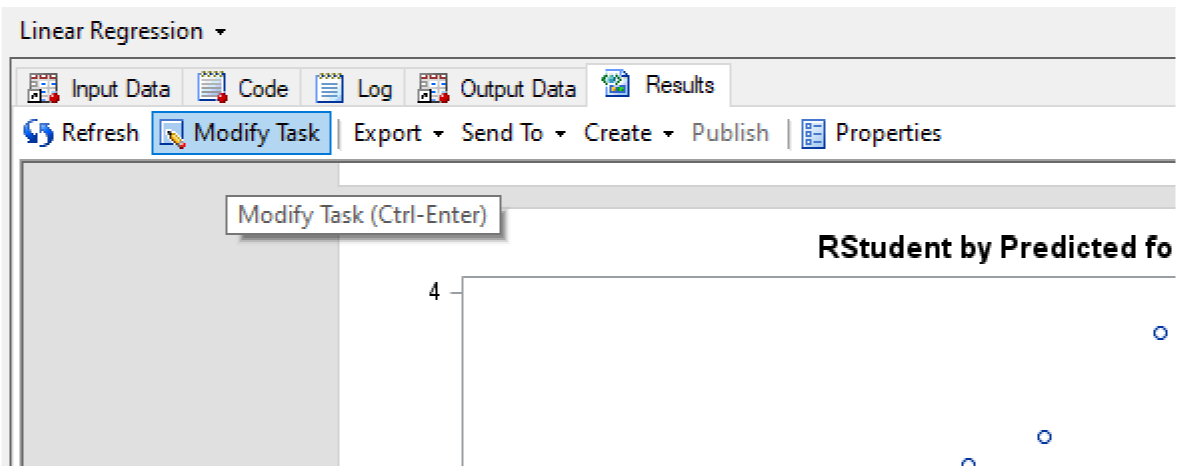
Linear Regression Validation results

After the regression model is fitted, choose between ex-ante and ex-post forecasts (based on your project).

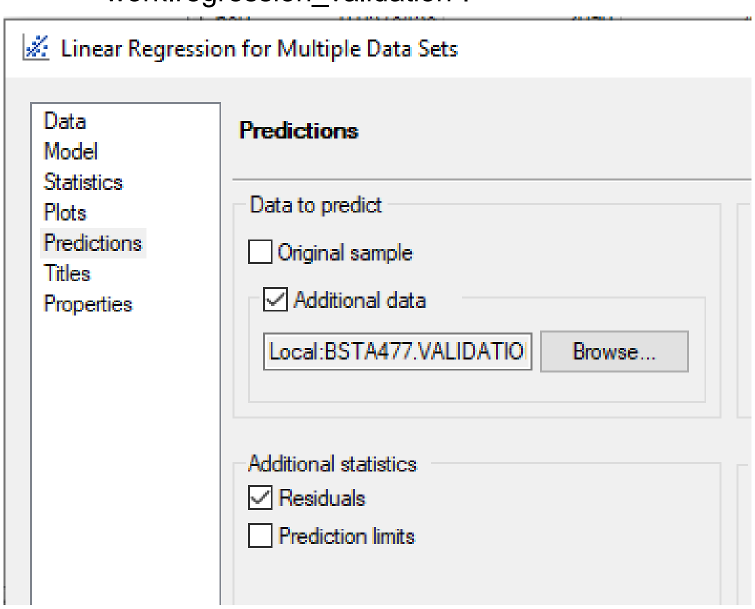
1. Ex-ante forecasts: Obtain forecasts of the predictors to forecast the forecasting variable
2. Ex-post forecasts: Obtain the actual observations of the predictors to forecast the forecasting variable.

For the current bike sharing data, we are conducting an ex-post forecast, as we obtained the actual observations of the predictors. We forecast using the data from the validation set:

1. **Forecast bike rentals for validation set:**
    - Use the same task with the training linear regression. Just click on modify task:



    - Click on additional data in the Prediction tab, unchosen the original sample and browse the validation set data. Change the save output data to "work.regression_validation".



2. **Calculate error terms for validation set**: As SAS do not calculate error terms for the validation set, we can calculate them as follow:

```
data validation_regression_result;
set work.regression_validation;
residual = cnt - predicted_cnt;
abs = abs(residual);
square = residual**2;
proportion = residual/cnt;
abs_proportion = abs/cnt;
run;

proc summary data=validation_regression_result;
var abs square proportion abs_proportion;
output out=total_validation_errors sum= / autoname;
run;

data validation_error_result;
set total_validation_errors;
MAE = abs_Sum/146;
MSE = square_Sum/(146-8-1);
RMSE = sqrt(MSE);
MPE = proportion_Sum/146;
MAPE = abs_proportion_Sum/146;
run;

proc print data=validation_error_result; run;
```

__Note:__ The calculation of MSE is similar to the training set; however, the number of observations changed. The total number of observations of the validation set is 146 observations. => MSE = SSE/ (146 - 8 -1), where 8 is the number of predictors in the current model.

Result:

VALIDATION_ERROR_RESULT

Filter and Sort | Query Builder | Where | Data | Describe | Graph | Analyze | Export | Send To

	abs_Sum	square_Sum	proportion_Sum	abs_proportion_Sum	MAE	MSE	RMSE	MPE	MAPE
1	124625.36054	190322590.07	-211.6022982	230.90652937	853.59835987	1389215.9859	1178.6500693	-1.449330809	1.5815515711

Output evaluation

We compare the error terms of both training and validation sets.
Training set:

Alicia Ngoc Phan

TRAINING_ERROR_RESULT

Filter and Sort | Query Builder | Where | Data | Describe | Graph | Analyze | Export | Send To

	abs_Sum	square_Sum	proportion_Sum	abs_proportion_Sum	MAE	MSE	RMSE	MPE	MAPE
1	357564.59657	375679414.11	-37.60189898	113.86100014	611.22153259	652221.20506	807.6021329	-0.06427675	0.1946341883

Validation set:

VALIDATION_ERROR_RESULT

Filter and Sort | Query Builder | Where | Data | Describe | Graph | Analyze | Export | Send To

	abs_Sum	square_Sum	proportion_Sum	abs_proportion_Sum	MAE	MSE	RMSE	MPE	MAPE
1	124625.36054	190322590.07	-211.6022982	230.90652937	853.59835987	1389215.9859	1178.6500693	-1.449330809	1.5815515711

Prediction interval

Formulate prediction interval:

$$\hat{y}_{T+h|T} \pm c\hat{\sigma}_h$$

Ex: Formulate 95% prediction interval

$$\hat{y}_{T+h|T} \pm 1.96\hat{\sigma}_h,$$

Multipliers to be used for prediction intervals

Percentage	Multiplier
50	0.67
55	0.76
60	0.84
65	0.93
70	1.04
75	1.15
80	1.28
85	1.44
90	1.64
95	1.96
96	2.05
97	2.17
98	2.33
99	2.58

Example: Formulate the 95% prediction interval for the bike rentals on Jan5,2011.

Alicia Ngoc Phan

From this, we need 2 information: predicted bike rentals on Jan 5, 2011 and RMSE of training set

- Predicted value of bike rentals on Jan5, 2011 = 1962.34

Linear Regression

Input Data (2) | Code | Log | Output Data | Results

Modify Task | Filter and Sort | Query Builder | Where | Data | Describe | Graph | Analyze | Ex

	dteday	predicted_cnt	season	yr	mnth	holiday	we
1	01JAN2011	2051.2473965	1	0	1	0	
2	02JAN2011	1619.5702758	1	0	1	0	
3	03JAN2011	1503.8208239	1	0	1	0	
4	04JAN2011	1731.7166968	1	0	1	0	
5	05JAN2011	1962.343035	1	0	1	0	
6	06JAN2011	2176.1724549	1	0	1	0	
7	07JAN2011	1415.8178029	1	0	1	0	
8	08JAN2011	939.50403305	1	0	1	0	
9	09JAN2011	788.0055437	1	0	1	0	
10	10JAN2011	1301.1607485	1	0	1	0	
11	11JAN2011	1103.3715734	1	0	1	0	
12	12JAN2011	1147.7734193	1	0	1	0	
13	13JAN2011	1274.4052205	1	0	1	0	
14	14JAN2011	1867.558412	1	0	1	0	

- RMSE = 807.6

Root MSE	807.60213	**R-Square**	0.7997
Dependent Mean	4158.93162	**Adj R-Sq**	0.7969
Coeff Var	19.41850		

- 95% Prediction interval: 1962.34 +/- 807.6*1.96 = [379.44; 3,545.24]

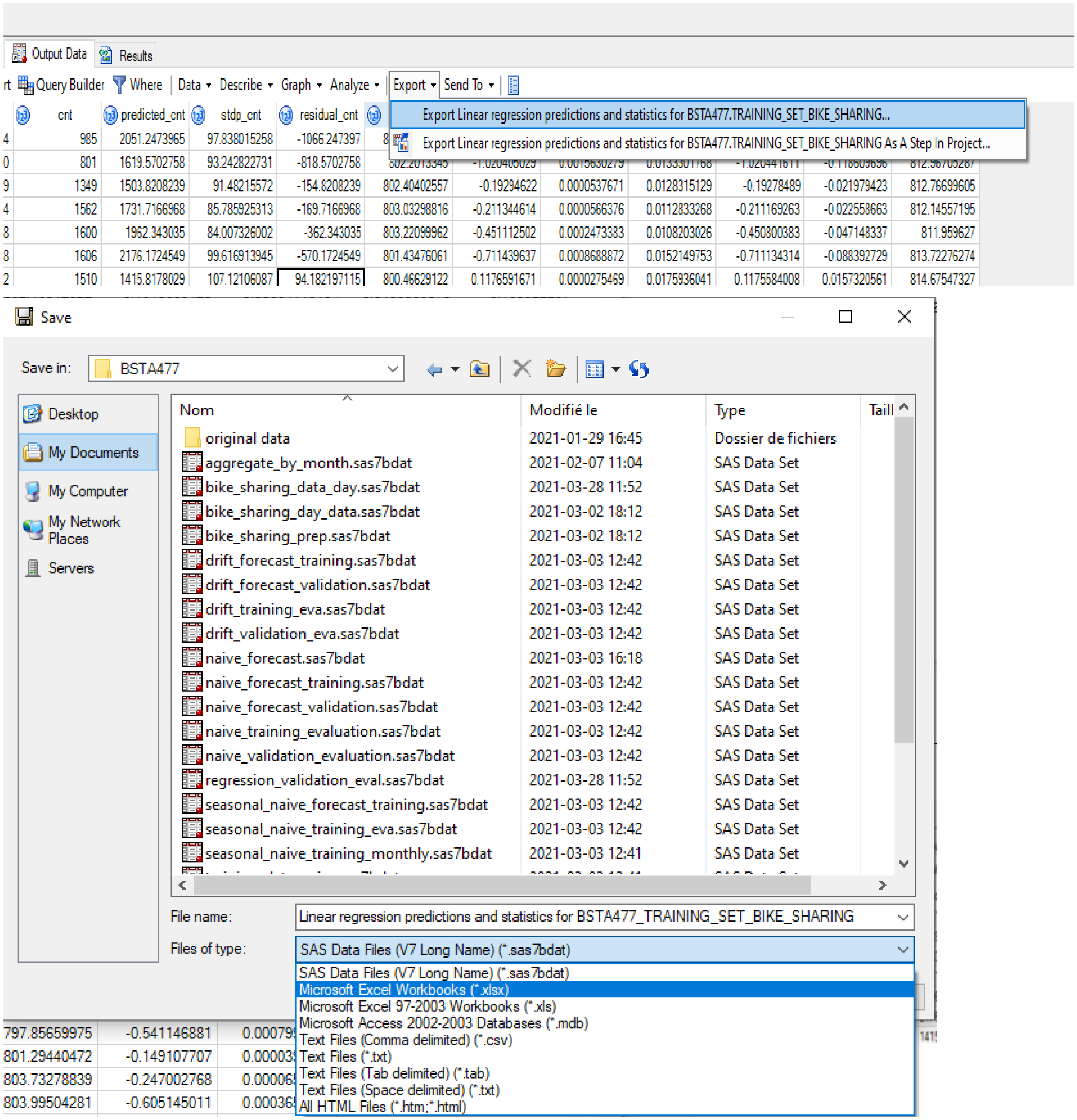
Generalized difference

Estimate the regression coefficient B1 using generalized difference:

1. Export residuals to Excel
2. Calculate autocorrelation at lag 1. Obtain the Autocorrelation coefficient at lag 1
3. Calculate the new Yt and Xt with the autocorrelation coefficient.
4. Obtain new simple linear regression as normal (fitted the regression as shown previously)

Step 1:

Alicia Ngoc Phan



Step 2: Calculate autocorrelation coefficient at lag 1

Reference: [Calculate Autocorrelation in Excel]

Alicia Ngoc Phan

Use VAR function and COVAR function to calculate autocorrelation coefficient.

= VAR (all observations of the residuals)

residual_cnt	stdr_cnt	student_c	cookd_cnt	h_cnt	rstudent_	dffits_cnt	stdi_cnt			
-1066,247397	801,6539	-1,33006	0,002928	0,014676	-1,33095	-0,16244	813,5069			
-818,5702758	802,2013	-1,02041	0,001563	0,01333	-1,02044	-0,11861	812,9671			
-154,8208239	802,404	-0,19295	5,38E-05	0,012832	-0,19278	-0,02198	812,767			
-169,7166968	803,033	-0,21134	5,66E-05	0,011283	-0,21117	-0,02256	812,1456			
-362,343035	803,221	-0,45111	0,000247	0,01082	-0,4508	-0,04715	811,9596		VAR	=VAR(S2:S586)
-570,1724549	801,4348	-0,71144	0,000869	0,015215	-0,71113	-0,08839	813,7228		COVAR	VAR(number1; [number2]; ...)
94,18219712	800,4663	0,117659	2,75E-05	0,017594	0,117558	0,015732	814,6755			
19,49596695	801,0318	0,024339	1,08E-06	0,016205	0,024317	0,003121	814,1194		ACF	0,493116499
33,9944563	798,4706	0,042574	4,63E-06	0,022486	0,042538	0,006452	816,6315			
19,83925151	802,2705	0,024729	9,06E-07	0,01316	0,024707	0,002853	812,8988			
159,6284266	801,9105	0,19906	6,27E-05	0,014045	0,198894	0,023739	813,2539			
14,22658073	801,3807	0,017753	5,46E-07	0,015348	0,017737	0,002214	813,776			
131,5947795	802,1091	0,164061	4,11E-05	0,013557	0,163922	0,019217	813,058			
-446,558412	801,6525	-0,55705	0,000514	0,01468	-0,55671	-0,06795	813,5082			
-459,4254936	799,6226	-0,57455	0,000736	0,019663	-0,57422	-0,08132	815,5035			
-597,0864818	802,0213	-0,74448	0,00086	0,013773	-0,74419	-0,08794	813,1447			
465,3200805	776,6222	0,599159	0,003246	0,075249	0,598825	0,17082	837,4368			
-422,6037551	801,7008	-0,52713	0,000456	0,014561	-0,5268	-0,06404	813,4606			
152,3306546	803,9521	0,189477	3,63E-05	0,009019	0,189319	0,018061	811,2357			
399,7793701	802,3612	0,498254	0,000362	0,012937	0,497928	0,057005	812,8093			
290,2314939	800,2816	0,362662	0,000269	0,018047	0,362388	0,049128	814,8569			
-349,4136584	799,2283	-0,43719	0,000447	0,02063	-0,43688	-0,06341	815,89			
32,01718733	800,2926	0,040007	3,26E-06	0,01802	0,039972	0,005415	814,8461			

= COVAR (series 1 ; series 2)

=COVAR(S3:S586;S2:S585)

Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
edicted	stdp_cnt	residual_cnt	stdr_cnt	student_c	cookd_cnt	h_cnt	rstudent_	dffits_cnt	stdi_cnt				
51,247	97,83802	-1066,247397	801,6539	-1,33006	0,002928	0,014676	-1,33095	-0,16244	813,5069				
619,57	93,24282	-818,5702758	802,2013	-1,02041	0,001563	0,01333	-1,02044	-0,11861	812,9671				
03,821	91,48216	-154,8208239	802,404	-0,19295	5,38E-05	0,012832	-0,19278	-0,02198	812,767				
31,717	85,78593	-169,7166968	803,033	-0,21134	5,66E-05	0,011283	-0,21117	-0,02256	812,1456				
62,343	84,00733	-362,343035	803,221	-0,45111	0,000247	0,01082	-0,4508	-0,04715	811,9596		VAR	643286,668	
76,172	99,61691	-570,1724549	801,4348	-0,71144	0,000869	0,015215	-0,71113	-0,08839	813,7228		COVAR	=COVAR(S3:S586;S2:S585)	
15,818	107,1211	94,18219712	800,4663	0,117659	2,75E-05	0,017594	0,117558	0,015732	814,6755			COVAR(array1; array2)	
39,504	102,8069	19,49596695	801,0318	0,024339	1,08E-06	0,016205	0,024317	0,003121	814,1194		ACF	0,493116499	
8,0055	121,1028	33,9944563	798,4706	0,042574	4,63E-06	0,022486	0,042538	0,006452	816,6315				
01,161	92,64542	19,83925151	802,2705	0,024729	9,06E-07	0,01316	0,024707	0,002853	812,8988				
03,372	95,71153	159,6284266	801,9105	0,19906	6,27E-05	0,014045	0,198894	0,023739	813,2539				
47,773	100,0511	14,22658073	801,3807	0,017753	5,46E-07	0,015348	0,017737	0,002214	813,776				
74,405	94,03273	131,5947795	802,1091	0,164061	4,11E-05	0,013557	0,163922	0,019217	813,058				
67,558	97,84913	-446,558412	801,6525	-0,55705	0,000514	0,01468	-0,55671	-0,06795	813,5082				
07,425	113,2468	-459,4254936	799,6226	-0,57455	0,000736	0,019663	-0,57422	-0,08132	815,5035				
01,086	94,77915	-597,0864818	802,0213	-0,74448	0,00086	0,013773	-0,74419	-0,08794	813,1447				
4,6799	221,5382	465,3200805	776,6222	0,599159	0,003246	0,075249	0,598825	0,17082	837,4368				
05,604	97,45236	-422,6037551	801,7008	-0,52713	0,000456	0,014561	-0,5268	-0,06404	813,4606				
97,669	76,69562	152,3306546	803,9521	0,189477	3,63E-05	0,009019	0,189319	0,018061	811,2357				
27,221	91,85732	399,7793701	802,3612	0,498254	0,000362	0,012937	0,497928	0,057005	812,8093				

Series 1: 2nd observation - last observation of residuals
Series 2: 1st observation - second to last observation of residuals

Autocorrelation coefficient = COVAR / SQRT(VAR*VAR)

Alicia Ngoc Phan

VAR	643286,668
COVAR	317215,2697
ACF	=AC7/SQRT(AC6*AC6)
	SQRT(number)

VAR	643286,668
COVAR	317215,2697
ACF	0,493116499

=> Autocorrelation coefficient at lag 1 = 0.493

Step 3: Calculate the new Yt and Xt with the autocorrelation coefficient

Create a new training set based on new variables adjusted to the autocorrelation coefficient.

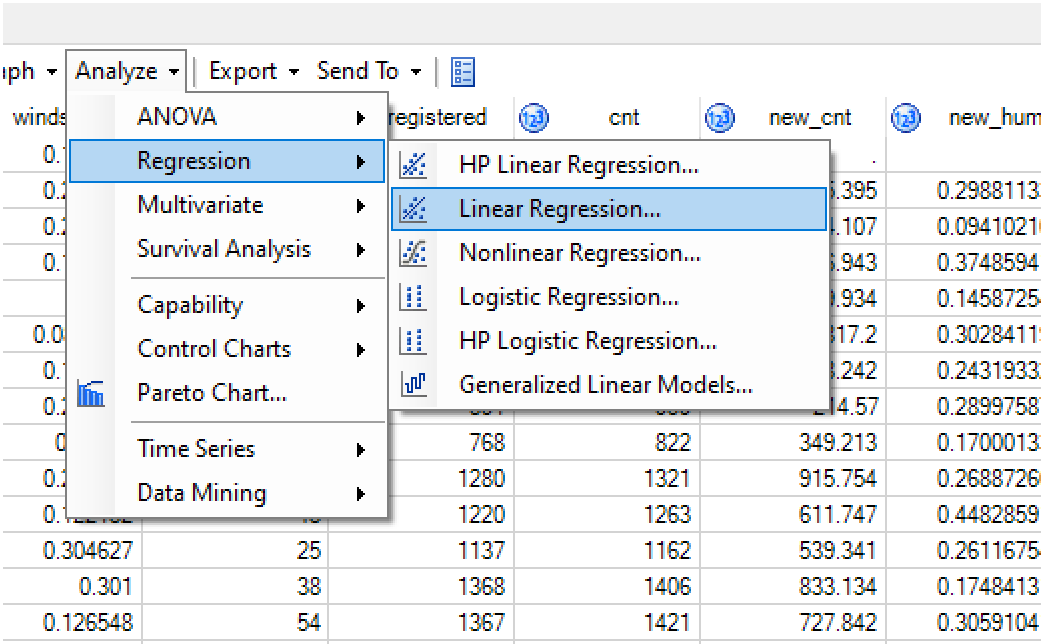
Example: Create new dependent variable (y or cnt here) and numerical independent variables (x or hum - humidity here)

```
data new_regression_training;
set bsta477.training_set_bike_sharing;
new_cnt = cnt - 0.493*lag(cnt);
new_hum = hum- 0.493*lag(hum);
run;
```

Note:

- Do this to all numerical predictors that were selected from the previous analysis

Step 4: Run the linear regression with the new selected independent and dependent variables.



All the steps are the same as doing simple linear regression.

Obtain a new regression equation and compare it to the original regression.

Durbin Watson test

Reference: Durbin Watson Table
Note: We cannot use the Durbin Watson test if we have lagged dependent variables.
We have the following Durbin Watson statistics from the Regression training result:

Linear Regression Training Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: cnt

Durbin-Watson D	1.010
Number of Observations	585
1st Order Autocorrelation	0.493

We have the Durbin Watson statistics DW = 1.010

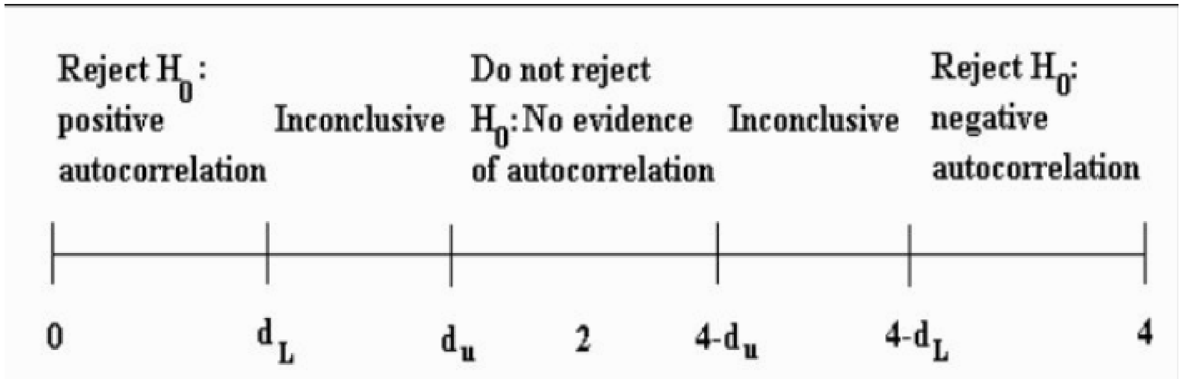
We have our hypothesis:

Alicia Ngoc Phan

Ho: Residuals are not correlated (The residuals are white noise, there is no information left in model)
Ha: Residuals are correlated.

Find the dU (Durbin Watson upper bound) and dL (Durbin watson lower bound) by using two information: number of predictors k and number of observations.
=> Current example: k =8, n=585
=> dL = 1.831, dU = 1.89

Decision rule:



- Reject Ho if DW < dL or DW > 4-dL
- Do not reject Ho if dU < DW < 4- Du

Decision:
As DW = 1.010 < dL =1.831, we have enough evidence to reject Ho. This means that the errors are positively correlated.

With positively correlated residuals, there is information left in the dependent variable that is not explained by the predictors.