Multiple Linear Regression in SAS® Studio

We are going to conduct a correlation analysis to begin exploring the relationship between numerical variables. Then we will build the simplest regression model a simple linear regression model.

Dataset Description

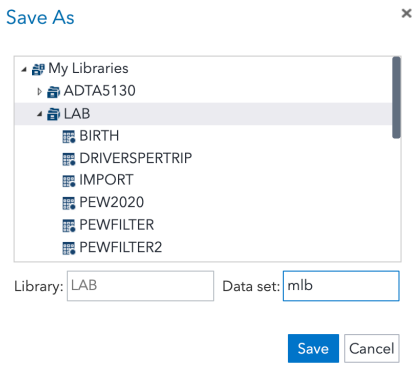
This dataset contains 30 observations and 12 variables. Data is collected from all 30 Major League Baseball teams from the 2011 season. We will use this data to analyze the relationships between wins, runs scored in a season, and several other player statistics.

- team: Team name.
- runs: Number of runs.
- at_bats: Number of at bats.
- hits: Number of hits.
- homerun: Number of home runs.
- bat_avg: Batting average.
- strikeouts: Number of strikeouts.
- stolen_bases: Number of stolen bases.
- wins: Number of wins.
- new_onbase: On base percentage, measure of how often a batter reaches base for any reason other than a fielding error, fielder's choice, dropped/uncaught third strike, fielder's obstruction, or catcher's interference.
- new_slug: Slugging percentage, popular measure of the power of a hitter calculated as the total bases divided by at bats.
- new_obs: On base plus slugging, calculated as the sum of these two variables

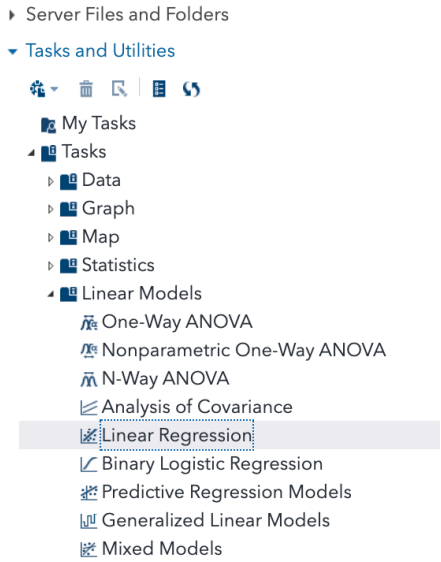
For this example, we will be developing a linear regression model to predict the number of runs using predictors in the data. There are multiple hypotheses tested by the regression model including the significance of the whole model as well as that of each predictor.

1. You can upload mlb.xlsx from your local computer to a selected folder in SAS

2. Import the mlb.xlsx from SAS folder to a selected SAS library.

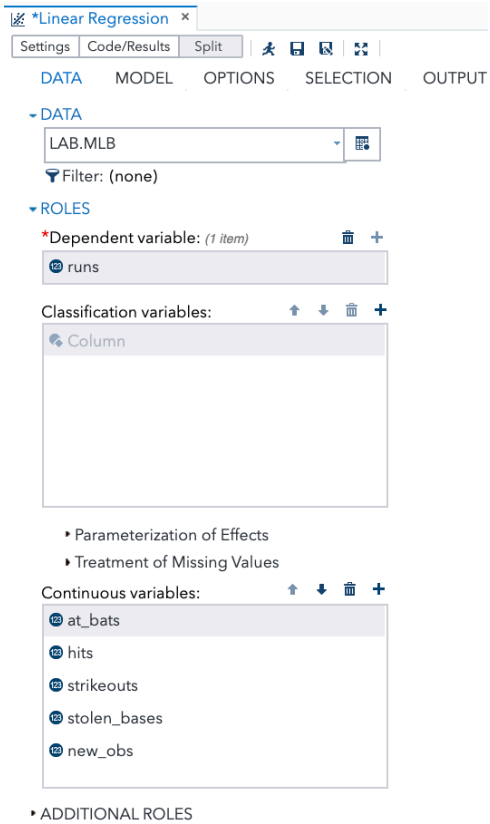


3. Under Tasks and Utilities, expand Tasks, expand Linear Models and select Linear Regression.



4. In the Data tab of the Linear Regression window, select data to analyze, Dependent variables to be runs, and place predictors at_bats, hits, strikeouts, stolen_bases and new_obs into the box of Continuous variables. If you have a categorical variable, you can place it into the box of Classification variables, and it will be recoded to dummy variables automatically by the program. **Many of actions in this instruction will require modifications in the existing regression procedure. Therefore, please make sure this**

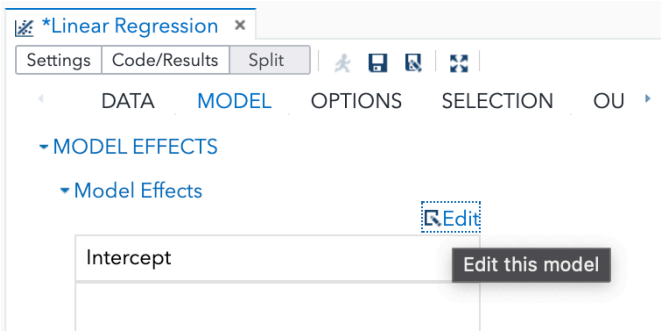
Linear Regression window remains during the whole session so we can easily adjust selections instead of restarting from the beginning.



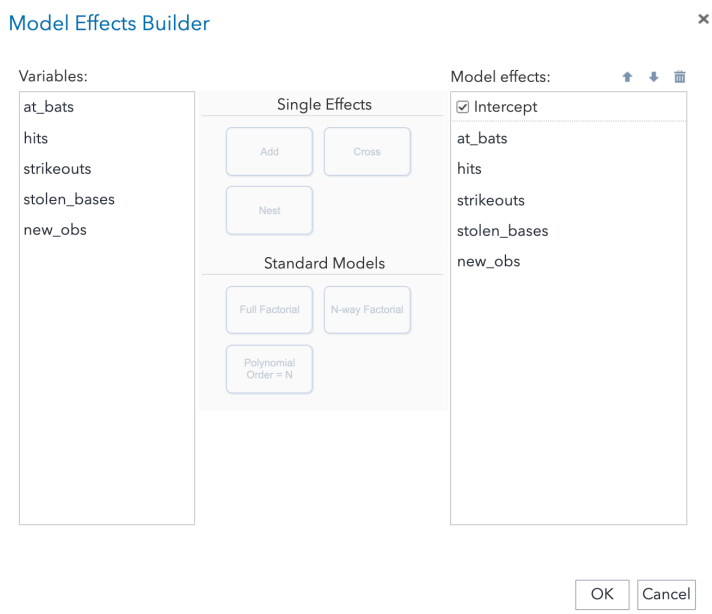
Main effects only

Let's only consider main effects with any interaction.

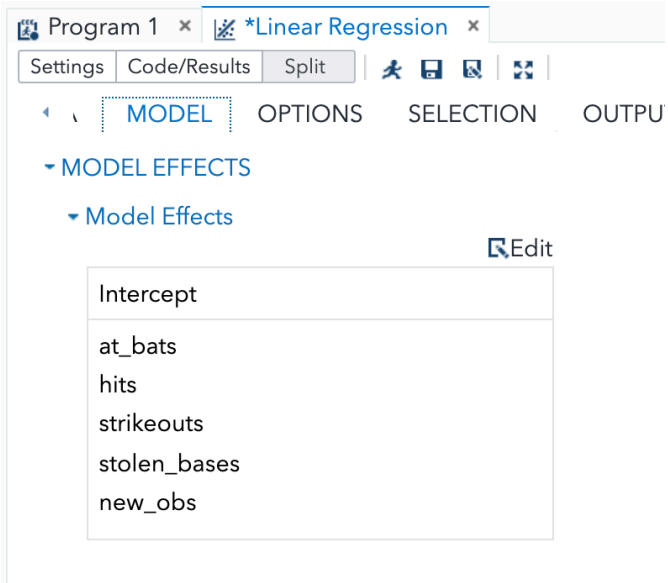
1. Move onto the Model tab of Linear regression window, click 'Edit' under Model Effects.



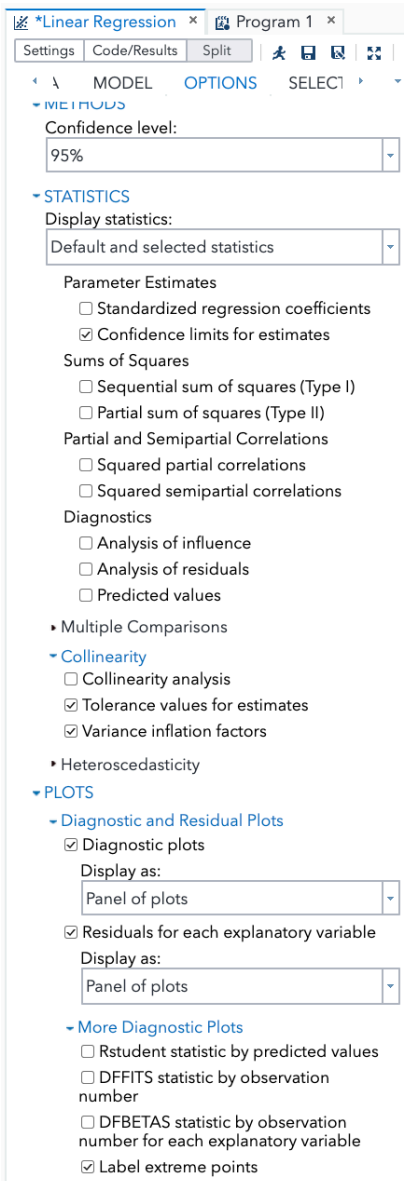
You can select main effects and interactions, i.e., cross, to be included in the Model Effects Builder window. For example, we can click Full Factorial to consider effects of all orders and click N-way Factorial to specify the highest order of effects needed. You can test the function by clicking those buttons. But for now, we will only consider all main effects so we will manually select all variables, add them to Model effects box, and click OK.



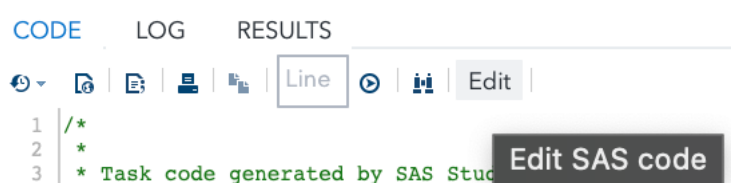
Now the predictors are added.



2. Move onto Options tab, under statistics, select Default and selected statistics to specify output needed. Under Parameter Estimates of STATISTICS, check Confidence limits of estimates. Under Collinearity of STATISTICS, check Tolerance values for estimates and Variance inflation factors. Under More Diagnostic Plots of PLOTS, check Label extreme points.



3. In the Code views, click Edit to open a program window to edit SAS code.



4. Add a footnote statement as below to include **your own first and last name** to the footnote.

*Linear Regression × Program 1 ×

CODE LOG RESULTS

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '10/31/21, 9:00 PM'
 * Generated by 'u58925565'
 * Generated on server 'ODAWS04-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) App
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?l
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc reg data=LAB.MLB alpha=0.05 plots(only label)=(diagnostics residuals
        observedbypredicted);
    model runs=at_bats hits strikeouts stolen_bases new_obs / clb tol vif;
    Footnote 'First and Last name';
    run;
quit;
```

5. Click Run and obtain the following output.

Model: MODEL1
Dependent Variable: runs runs

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	187697	37539	94.00	<.0001
Error	24	9584.18186	399.34091		
Corrected Total	29	197281			

Root MSE	19.98352	R-Square	0.9514
Dependent Mean	693.60000	Adj R-Sq	0.9413
Coeff Var	2.88113		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	-1207.03103	423.28863	-2.85	0.0088	.	0	-2080.65583	-333.40623
at_bats	at_bats	1	0.11285	0.09485	1.19	0.2458	0.23991	4.16818	-0.08291	0.30862
hits	hits	1	-0.17682	0.14506	-1.22	0.2347	0.08630	11.58779	-0.47621	0.12256
strikeouts	strikeouts	1	-0.01267	0.04566	-0.28	0.7837	0.56568	1.76779	-0.10692	0.08157
stolen_bases	stolen_bases	1	0.30084	0.12579	2.39	0.0250	0.97551	1.02511	0.04122	0.56045
new_obs	new_obs	1	2097.19394	187.25794	11.20	<.0001	0.22746	4.39628	1710.71254	2483.67533

First and Last name

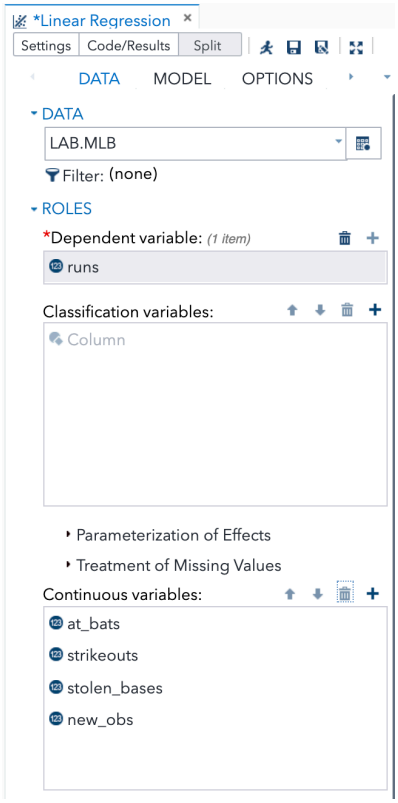
The Analysis of Variance shows the overall significance of the model. The Pr>F is the p-value and because it is <0.001 the model contains at least one significant predictor to predict the response. In other words, at least one predictor in the model has a coefficient parameter that is significantly different from 0. We can also read degree of freedom of the model, error, and total from the table which are 5 and 24 respectively. From the DF of model, we can tell there are 5 parameter estimates including 4 for predictors and 1 intercept. From the DF of corrected total, we can tell there are 30 observations because n-1=29.

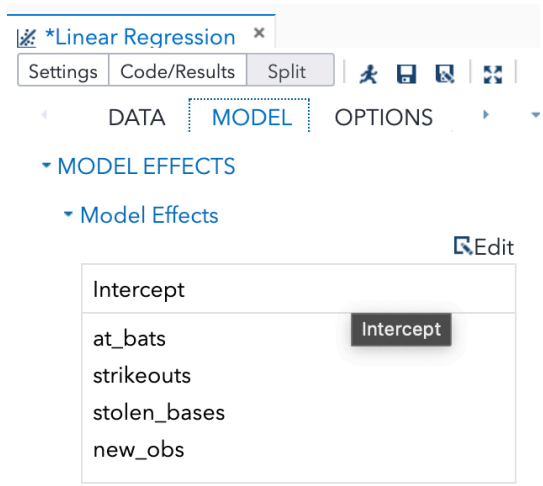
The Adj R-sq servs as a goodness of fit assessment for the model. The value 0.94 is very close 1 so the model is a good model.

The Parameter Estimates table shows point estimates, confidence interval, tolerance, and VIF for all coefficient parameters.

As we can see from VIF and tolerance, the variable hits showing high VIF (VIF>10) i.e., low tolerance. We will need to deal with the multicollinearity issue by remove this variable. In the case that multiple variables have VIF>10, we will need to remove them one at a time starting from the one with highest VIF.

6. Now, go back to Data tab of Linear regression window and remove hits from the box of Continuous variables. Move the Model tab to ensure hits is not in the model effects. Your options should be the same as before.





7. Now you need to add your own First and Last Name as footnote by editing the SAS code as earlier.
8. Run the program again. This time no variable has high VIF, and we can proceed to model diagnostic plots.

Model: MODEL1
Dependent Variable: runs runs

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	187104	46776	114.90	<.0001
Error	25	10178	407.10224		
Corrected Total	29	197281			

Root MSE	20.17677	R-Square	0.9484
Dependent Mean	693.60000	Adj R-Sq	0.9402
Coeff Var	2.90899		

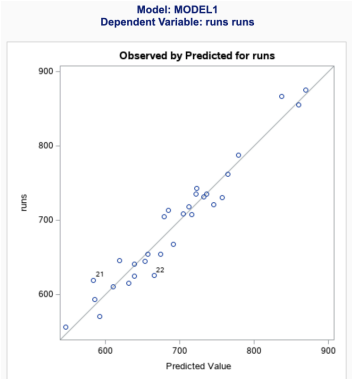
Parameter Estimates

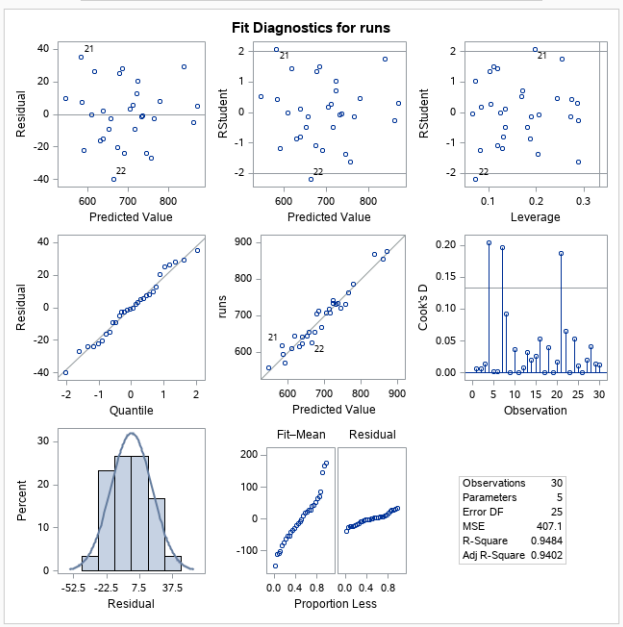
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	-878.73903	329.71629	-2.67	0.0133	.	0	-1557.80243	-199.67563
at_bats	at_bats	1	0.02543	0.06267	0.41	0.6884	0.56023	1.78498	-0.10365	0.15450
strikeouts	strikeouts	1	0.01464	0.04017	0.36	0.7187	0.74504	1.34221	-0.06810	0.09738
stolen_bases	stolen_bases	1	0.31689	0.12631	2.51	0.0190	0.98632	1.01387	0.05676	0.57702
new_obs	new_obs	1	1919.65909	118.83916	16.15	<.0001	0.57575	1.73686	1674.90526	2164.41292

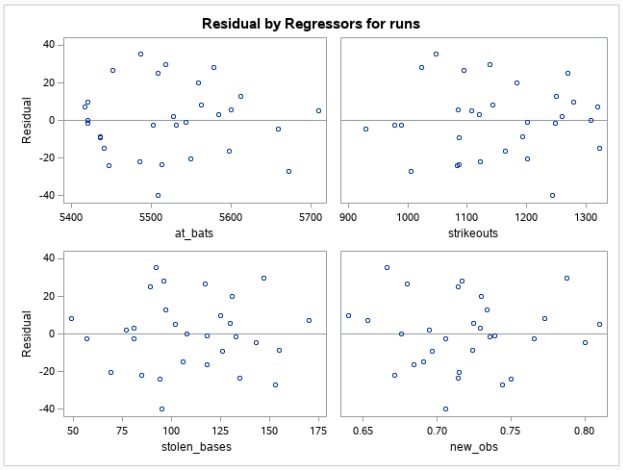
First and Last Name

9. There are multiple plots in the output. Fit Diagnostics for runs can be used to validate assumptions for the model. The residuals vs predicted value plots shows the residuals are distributed with constant variance and zero mean. Two extreme observations 21 and 22 are also labeled in the residual plots but

their residual values are not too different. The QQ plot of residuals shows a normal distribution. The Cook's D plot shows all Cook's D values are below 1 so no outlier exists. If using 4/n=4/30=0.13 as the cutoff, 3 data points are above the threshold. But since their Cook's D values are not much higher than the cutoff, we tend to include them.







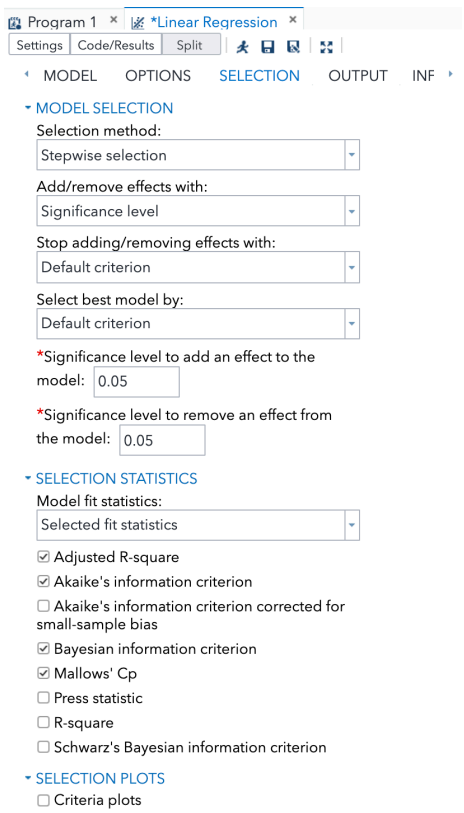
First and Last Name

10. The model's assumptions are satisfied. We can then move back to the Parameter estimates table to review individual significance. We can see that at_bats and strikeouts are not significant using the significance level of 0.05. In the final model we would want only significant predictors but we cannot remove both because removing one variable from the model can affect the whole model. But it is not easy for us to determine which one to remove first. We should try some automatic variable selection procedure from here.

Parameter Estimates										
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	-878.73903	329.71629	-2.67	0.0133	.	0	-1557.80243	-199.67563
at_bats	at_bats	1	0.02543	0.06267	0.41	0.6884	0.56023	1.78498	-0.10365	0.15450
strikeouts	strikeouts	1	0.01464	0.04017	0.36	0.7187	0.74504	1.34221	-0.06810	0.09738
stolen_bases	stolen_bases	1	0.31689	0.12631	2.51	0.0190	0.98632	1.01387	0.05676	0.57702
new_obs	new_obs	1	1919.65909	118.83916	16.15	<.0001	0.57575	1.73686	1674.90526	2164.41292

First and Last Name

11. In the Linear Regression window, keep all settings of Data, Model, and Options the same as above. Move to Selection tab, choose Stepwise selection as Selection method, Add/remove effects with Significance level. Under SELECTION STATISTICS, set Model fit statistics: Selected fit statistics, then select Adjusted R-Square, Akaike information criterion, Bayesian information criterion, and Mallows' Cp. Uncheck all plots under SELECTION PLOTS.



12. You can also add your own First and Last Name as footnote to the output. Note that you need to add two statements one in each regression-related procedure as below!

*Linear Regression × Program 1 ×

CODE LOG RESULTS

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '10/31/21, 10:45 PM'
 * Generated by 'u58925565'
 * Generated on server 'ODAWS04-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) Appl
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?lc
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc glmselect data=LAB.MLB outdesign(addinputvars)=Work.reg_design;
    model runs=at_bats strikeouts stolen_bases new_obs / showpvalues
        selection=stepwise

   (slentry=0.05 slstay=0.05 select=sl) stats=(adjrsq aic bic cp);
   footnote "First and Last Name";
run;

proc reg data=Work.reg_design alpha=0.05 plots(only label)=(diagnostics
        residuals observedbypredicted);
    ods select ParameterEstimates DiagnosticsPanel ResidualPlot
        ObservedByPredicted;
    model runs=&_GLSMOD / clb tol vif;
    footnote "First and Last Name";
    run;
quit;

proc delete data=Work.reg_design;
run;
```

13. Run the program and obtain results. The first part shows the settings of selection methods. The second part is Stepwise Selection Summary which shows an effect being entered and removed in each step along with each model's Adj R-square, AIC, BIC, Mallow's Cp, F statistic and P-value. The Stop Details shows the selection is done because next candidate for entry cannot be entered and the candidate for removal will stay.

Data Set	LAB.MLB
Dependent Variable	runs
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	30
Number of Observations Used	30

Dimensions	
Number of Effects	5
Number of Parameters	5

First and Last Name

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Adjusted R-Square	AIC	BIC	CP	F Value	Pr > F
0	Intercept		1	0.0000	297.7356	264.0994	456.5987	0.00	1.0000
1	new_obs		2	0.9326	217.7682	187.5689	5.5342	402.29	<.0001
2	stolen_bases		3	0.9441*	213.0846*	184.1459*	1.2364*	6.74	0.0151

* Optimal Value of Criterion

Selection stopped because the candidate for entry has SLE > 0.05 and the candidate for removal has SLS < 0.05.

Stop Details

Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	at_bats	0.7459	>	0.0500	(SLE)
Removal	stolen_bases	0.0151	<	0.0500	(SLS)

First and Last Name

The section of Selected Model provides the ANOVA table, model fit statistics and parameter estimates for the selected model. The model is selected using default criterion which is the last step of selection where all significant variables are included. You can choose other options such as Adj R-square which will provide you the model with highest Adj R-square according to the Stepwise Selection Summary.

Selected Model

The selected model is the model at the last step (Step 2).

Effects:	Intercept stolen_bases new_obs

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	187007	93504	245.73	<.0001
Error	27	10274	380.51096		
Corrected Total	29	197281			

Root MSE	19.50669
Dependent Mean	693.60000
R-Square	0.9479
Adj R-Sq	0.9441
AIC	213.08463
AICC	214.68463
BIC	184.14593
C(p)	1.23640
SBC	185.28823

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-731.170979	65.095055	-11.23	<.0001
stolen_bases	1	0.315400	0.121506	2.60	0.0151
new_obs	1	1933.385841	87.345649	22.13	<.0001

First and Last Name

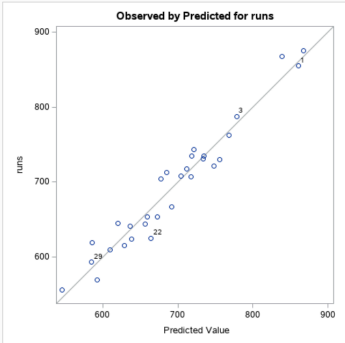
The Model: MODEL1 portion provides more details of the final model including the VIF, tolerance, confidence interval for all coefficients as we specified in Option tab. Now all predictors are significant. As shown earlier, plots especially the diagnostic plots will be used to check the assumptions of this final model. No severe violation of assumptions, including outliers, is detected.

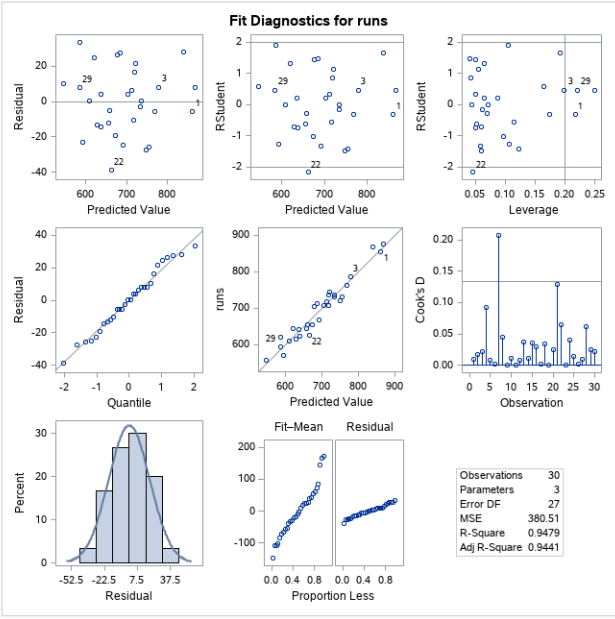
Model: MODEL1
Dependent Variable: runs runs

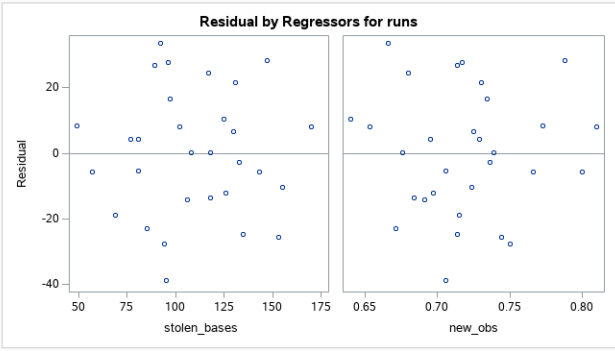
Parameter Estimates										
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	-731.17098	65.09505	-11.23	<.0001	.	0	-864.73500	-597.60696
stolen_bases	stolen_bases	1	0.31540	0.12151	2.60	0.0151	0.99618	1.00384	0.06609	0.56471
new_obs	new_obs	1	1933.38584	87.34565	22.13	<.0001	0.99618	1.00384	1754.16737	2112.60431

First and Last Name

Model: MODEL1
Dependent Variable: runs runs







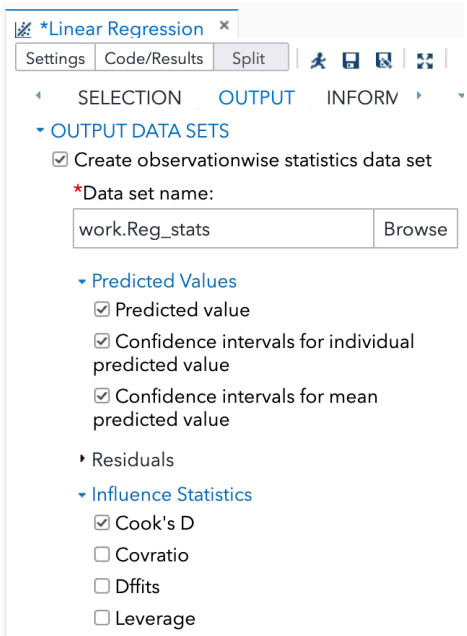
First and Last Name

14. We can write out the equation of the regression model. Using the above Parameter Estimates results.

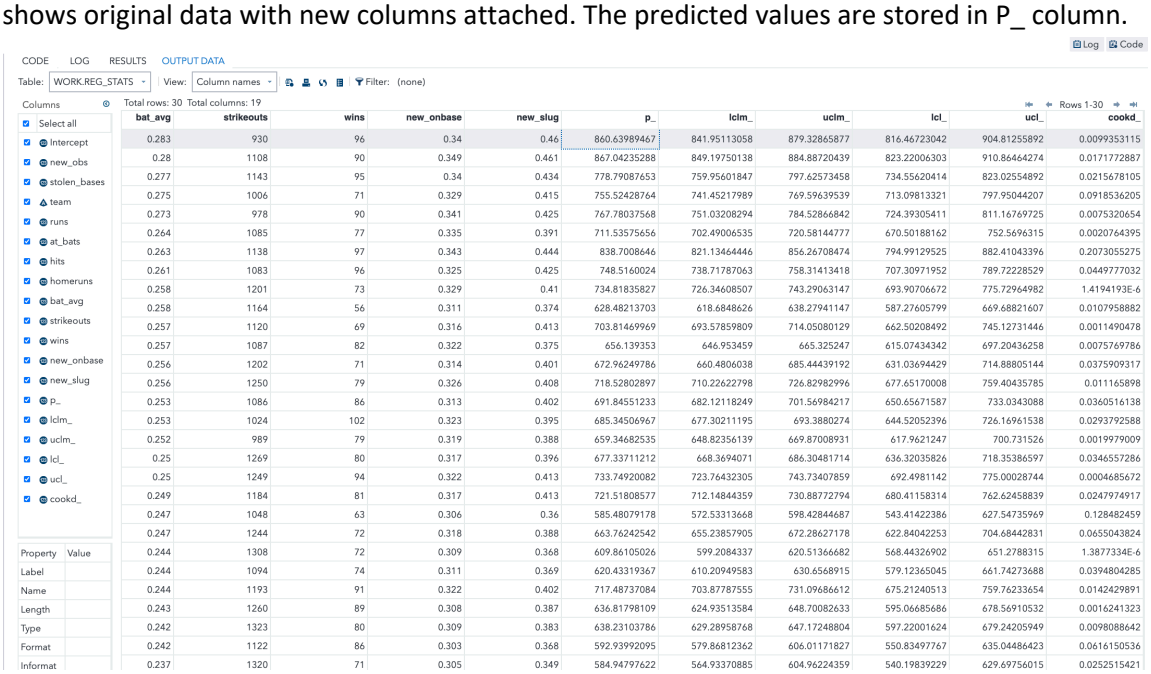
$$\hat{y} = -731.17 + 0.32stolen\_bases + 1933.39new\_obs$$

This final model equation can be deployed to predict the response number of runs now. For this data set, we can create columns to show each observation's predicted value, confidence interval, etc.

Move to output tab of Linear Regression window, check Create observation wise statistics data set.We can choose library and data set name to store the statistics but we will leave it as default here. Check Predicted value, Confidence intervals for individual predicted value, and Confidence interval for mean predicted value. Check Cook's D under Influence Statistics.



15. Run and you will obtain the same results as before. But a new tab 'Output Data' in the Results View shows original data with new columns attached. The predicted values are stored in P_ column.



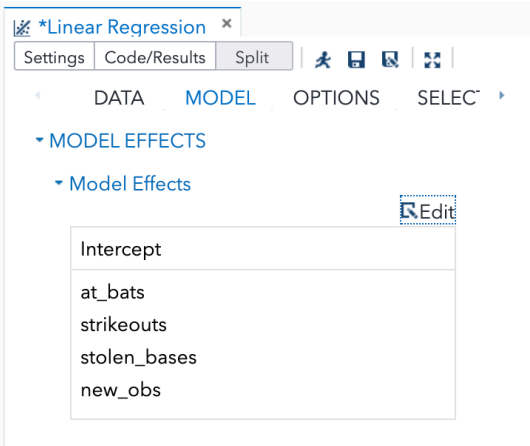
bat_avg	strikeouts	wins	new_onbase	new_slug	p_	lclm_	uclm_	lcl_	ucl_	cookd_
0.283	930	96	0.34	0.46	860.63989467	841.95113058	879.32865877	816.46723042	904.81255892	0.0099353115
0.28	1108	90	0.349	0.461	867.04235288	849.19750138	884.88720439	823.22006303	910.86464274	0.0171772887
0.277	1143	95	0.34	0.434	778.79087653	759.95601847	797.62573458	734.55620414	823.02554892	0.0215678105
0.275	1006	71	0.329	0.415	755.52428764	741.45217989	769.59639539	713.09813321	797.95044207	0.0918536205
0.273	978	90	0.341	0.425	767.78037568	751.03208294	784.52866842	724.39305411	811.16769725	0.0075320654
0.264	1085	77	0.335	0.391	711.53575656	702.49006535	720.58144777	670.50188162	752.5696315	0.0020764395
0.263	1138	97	0.343	0.444	838.7008646	821.13464446	856.26708474	794.99129525	882.41043396	0.2073055275
0.261	1083	96	0.325	0.425	748.5160024	738.71787063	758.31413418	707.30971952	789.72228529	0.0449777032
0.258	1201	73	0.329	0.41	734.81835827	726.34608507	743.29063147	693.90706672	775.72964982	1.4194193E-6
0.258	1164	56	0.311	0.374	628.48213703	618.6848626	638.27941147	587.27605799	669.68821607	0.0107958882
0.257	1120	69	0.316	0.413	703.81469969	693.57859809	714.05080129	662.50208492	745.12731446	0.0011490478
0.257	1087	82	0.322	0.375	656.139353	646.953459	665.325247	615.07434342	697.20436258	0.0075769786
0.256	1202	71	0.314	0.401	672.96249786	660.4806038	685.44439192	631.03694429	714.88805144	0.0375909317
0.256	1250	79	0.326	0.408	718.52802897	710.22622798	726.82982996	677.65170008	759.40435785	0.011165898
0.253	1086	86	0.313	0.402	691.84551233	682.12118249	701.56984217	650.65671587	733.0343088	0.0360516138
0.253	1024	102	0.323	0.395	685.34506967	677.30211195	693.3880274	644.52052396	726.16961538	0.0293792588
0.252	989	79	0.319	0.388	659.34682535	648.82356139	669.87008931	617.9621247	700.731526	0.0019979009
0.25	1269	80	0.317	0.396	677.33711212	668.3694071	686.30481714	636.32035826	718.35386597	0.0346557286
0.25	1249	94	0.322	0.413	733.74920082	723.76432305	743.73407859	692.4981142	775.00028744	0.0004685672
0.249	1184	81	0.317	0.413	721.51808577	712.14844359	730.88772794	680.41158314	762.62458839	0.0247974917
0.247	1048	63	0.306	0.36	585.48079178	572.53313668	598.42844687	543.41422386	627.54735969	0.128482459
0.247	1244	72	0.318	0.388	663.76242542	655.23857905	672.28627178	622.84042253	704.68442831	0.0655043824
0.244	1308	72	0.309	0.368	609.86105026	599.2084337	620.51366682	568.44326902	651.2788315	1.3877334E-6
0.244	1094	74	0.311	0.369	620.43319367	610.20949583	630.6568915	579.12365045	661.74273688	0.0394804285
0.244	1193	91	0.322	0.402	717.48737084	703.87787555	731.09686612	675.21240513	759.76233654	0.0142429891
0.243	1260	89	0.308	0.387	636.81798109	624.93513584	648.70082633	595.06685686	678.56910532	0.0016241323
0.242	1323	80	0.309	0.383	638.23103786	629.28958768	647.17248804	597.22001624	679.24205949	0.0098088642
0.242	1122	86	0.303	0.368	592.93992095	579.86812362	606.01171827	550.83497767	635.04486423	0.0616150536
0.237	1320	71	0.305	0.349	584.94797622	564.93370885	604.96224359	540.19839229	629.69756015	0.0252515421

This data set can also be opened from the Work library.

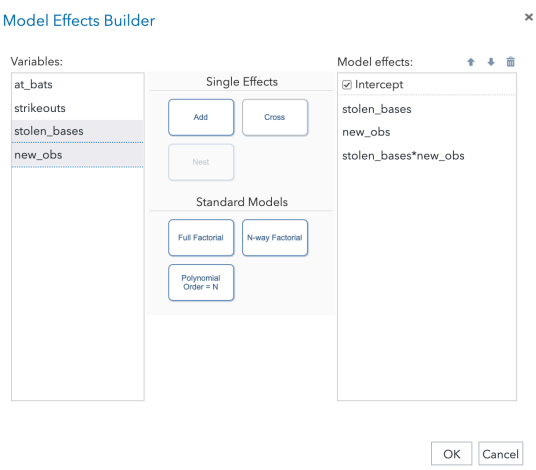
Main effects and interactions

Let's consider interactions.

1. Move back the Model tab of Linear regression window, click 'Edit' under Model Effects.

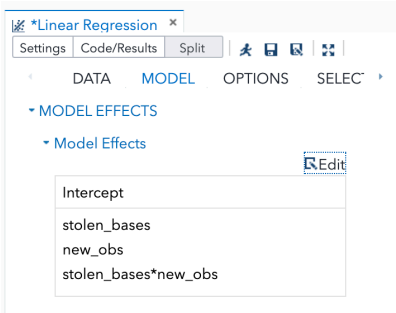


We will select main effects and interactions, i.e., cross, to be included in the Model Effects Builder window. We will only consider the interaction of the two selected main effects in the selection procedure. In the Model Effects Builder, select both Stolen_bases and new_obs simultaneously and click cross to add them to Model effects box. Make sure that the main effects in the box are only Stolen_bases and new_obs. Click OK.

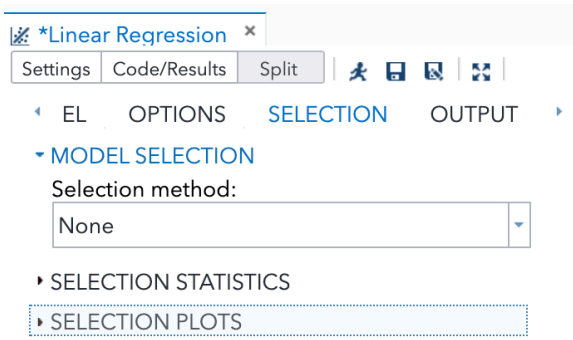


Note that we do not use N-way Factorial to re-select final model from all main effect and two-way interactions because the selection method will not ensure hierarchy of effects i.e., all main effects of significant interactions will be in the final model. For example, if var1 itself is not significant but var1*var2 is significant, the selection method will probably provide a model with interaction of var1*var2 but without var1. But we typically want to keep a lower ordered effects if a higher ordered effect involves it.

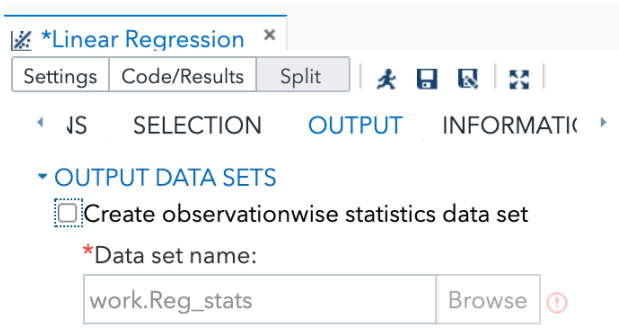
Now the Model effects are as below.



2. Leave the settings of Options tab as before. In the Selection tab, choose None for Selection method because we will not need to re-select model.



3. You can uncheck Create Observation wise statistics data set in the Output tab if you don't want to save the statistics until confirming that the interaction term is significant.



4. Run the program and obtain results. We can see that the interaction is not significant so we should not have to included it. Therefore, we will need to go back to the main effect model. Suppose you obtain significant interaction, then keep the interaction and its main effects in the model.

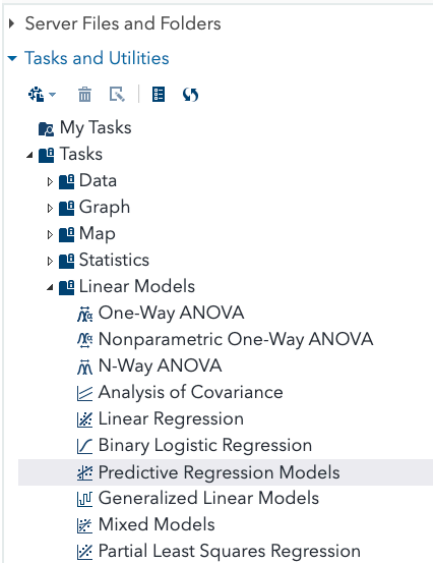
Model: MODEL1
Dependent Variable: runs runs

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Tolerance**	**Variance Inflation**
Intercept	Intercept	1	-765.35182	249.29625	-3.07	0.0050	.	0
stolen_bases	stolen_bases	1	0.60930	2.07003	0.29	0.7708	0.00356	280.78033
new_obs	new_obs	1	1980.26446	341.38880	5.80	<.0001	0.06767	14.77841
stolen_bases*new_obs	stolen_bases*new_obs	1	-0.40310	2.83408	-0.14	0.8880	0.00349	286.87235

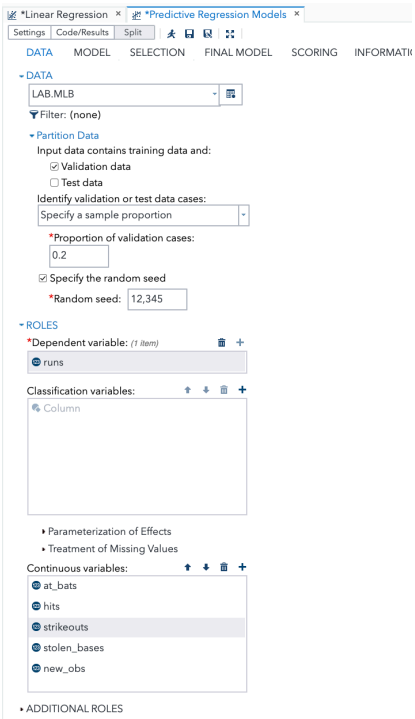
Main effects and interactions using another regression tool

We can use another tool to build the regression model.

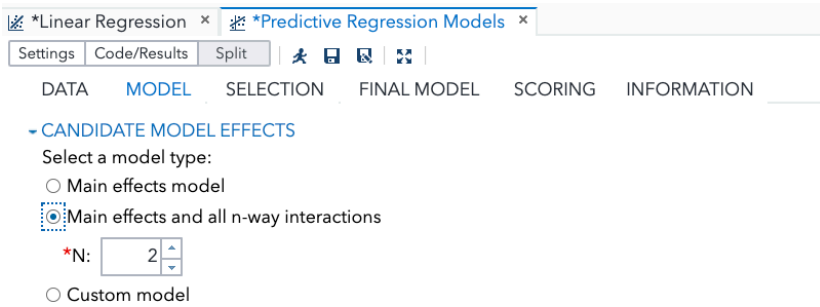
1. Under Tasks and Utilities, expand Tasks, expand Linear Models and select Predictive Regression Models



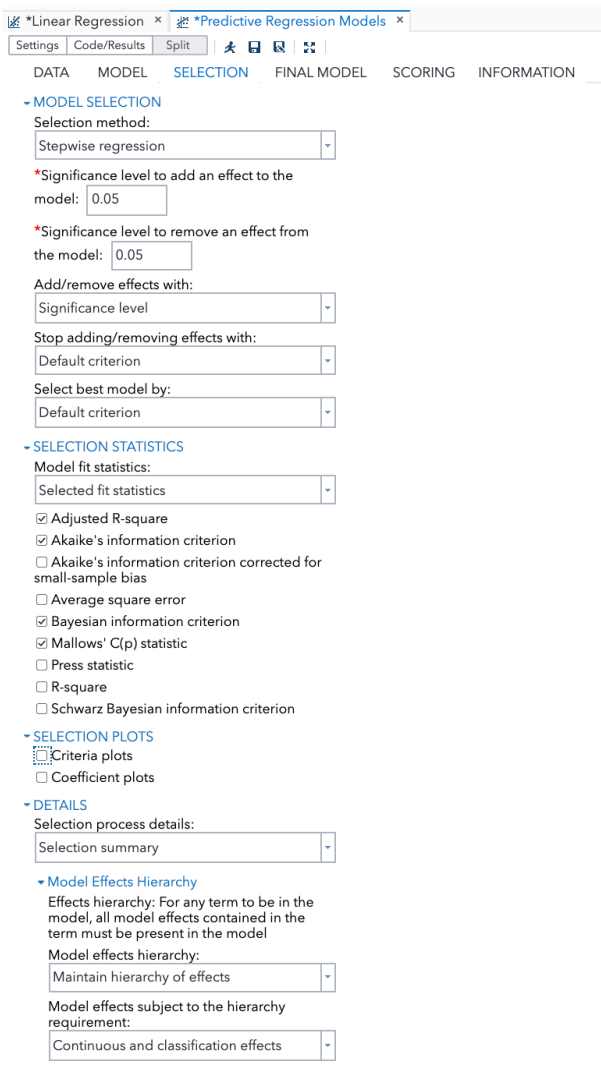
2. In the Data tab, you have additional place to specify data partitioning for validation. Check only Validation data, specify sample proportion of 0.2 to partition data into 80% training and 20% validation data. Set Random Seed to be 12345 to be able to reproduce the partitioning. Note that partitioning is not very meaningful for small dataset so this is only for illustration purpose here.
Set the variables as below. For this procedure we will have to check all assumptions including multicollinearity after final model is built after the selection is done



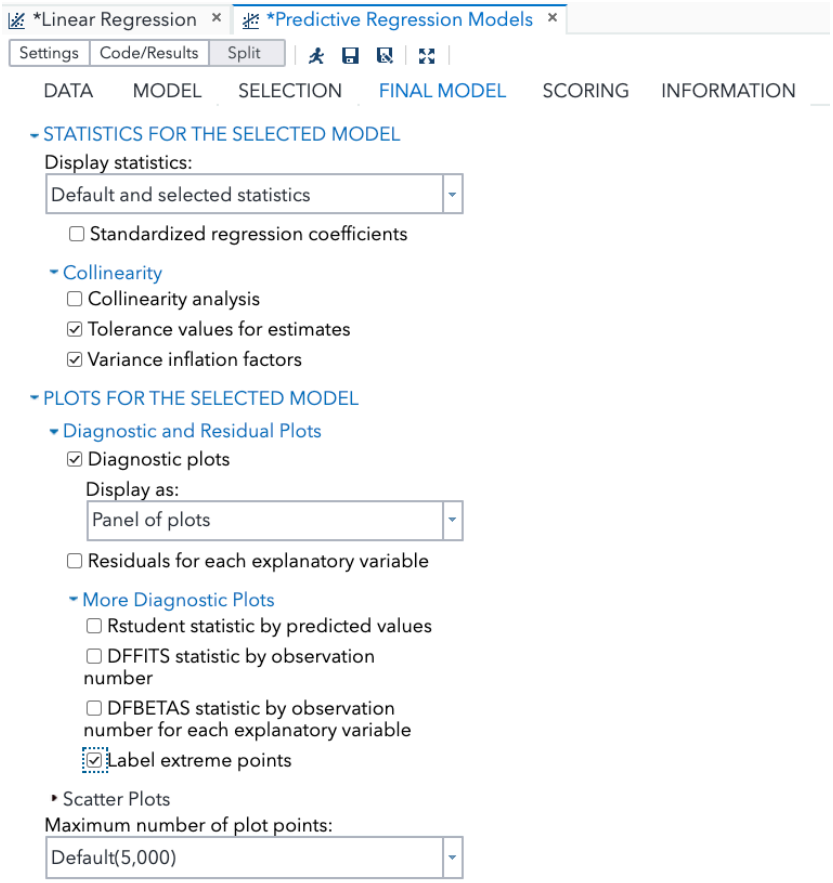
3. In Model tab, select Main effects and all n-way interactions, then specify N is 2 for all 2-way interactions to be considered.



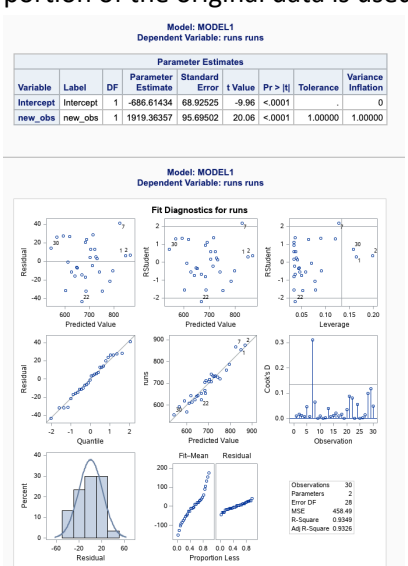
4. In Selection tab, follow the below settings. Make sure to Maintain hierarchy of effects for Continuous and Classification effects under Model Effects Hierarchy. This maintains the main effect as long as its interaction term is significant in the selection results.



5. In the Final model tab, follow below settings.



6. Run the program and obtain results. The final model contains only new_obs predictor and all assumptions are satisfied. The final model is different from using Linear Regression tool since only a portion of the original data is used. For large dataset, the model will be more likely to be the same.



Model: MODEL1
Dependent Variable: runs runs

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-686.61434	68.92525	-9.96	<.0001	.	0
new_obs	new_obs	1	1919.36357	95.69502	20.06	<.0001	1.00000	1.00000