Correlation and Simple 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

Correlation

For this example, we will be investigating the question "Is there any relationship between runs (Number of runs) and at_bats (Number of at bats)?". The two variables are both numerical.

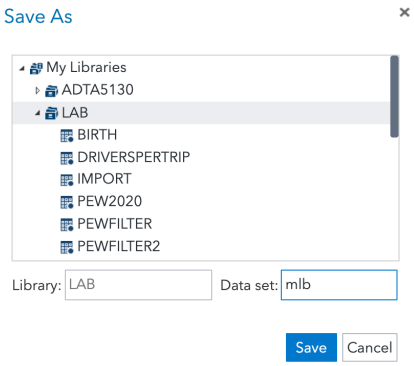
If we perform a formal test, the hypotheses are:

H_0: the correlation coefficient between runs and at_bats is 0.

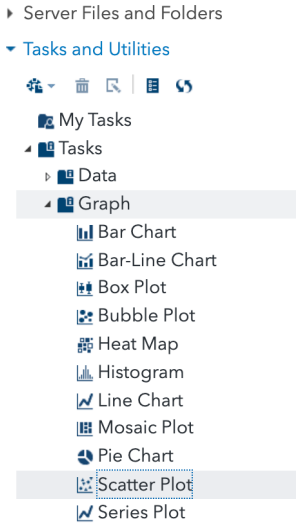
H_A: the correlation coefficient between runs and at_bats is not 0.

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

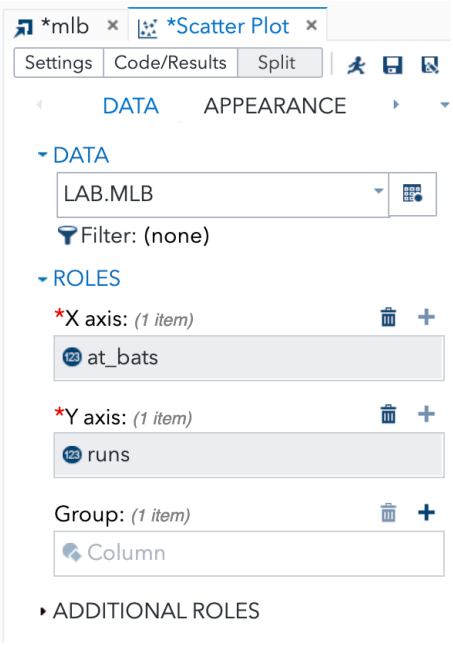
2. Import the mlb.xlsx from SAS folder to a selected SAS library.



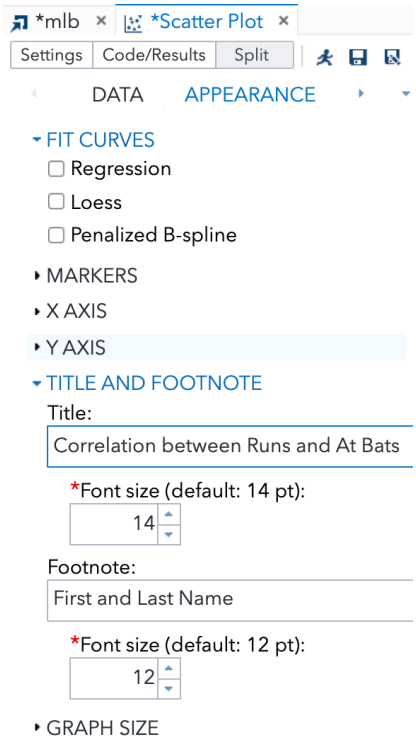
3. We will draw a scatter plot prior to any hypothesis testing.
   a. Under Tasks and Utilities, expand Tasks, expand Graph, then select Scatter Plot.



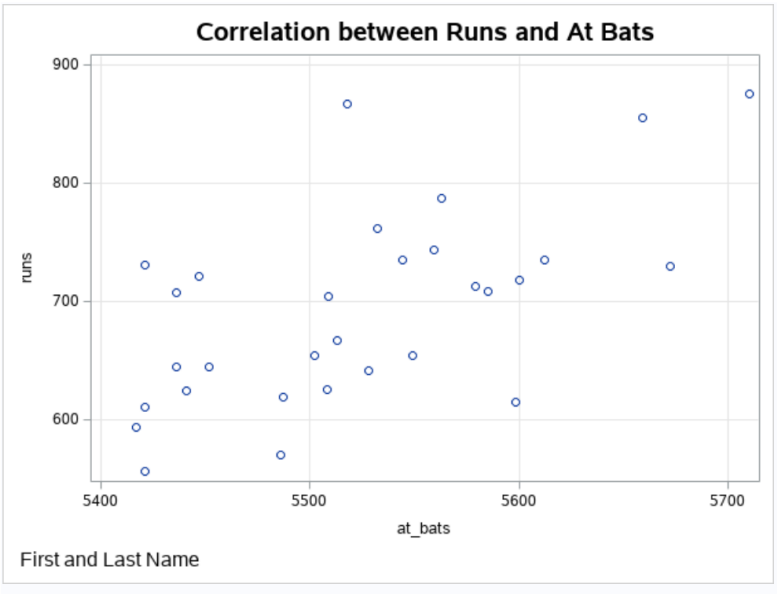
b. In Data tab of the Scatter Plot window, select data to analyze and variables on x and y axes. By convention, it is generally expected that you set the x- axis to the independent variable and the y- to be the dependent variable. But switching the placement of these variables doesn't change the correlation analysis results.



c. Now, move to the Appearance tab. Expand Title and Footnote to add you first and Last name to the graph. Title the graph "Correlation between Runs and At Bats".

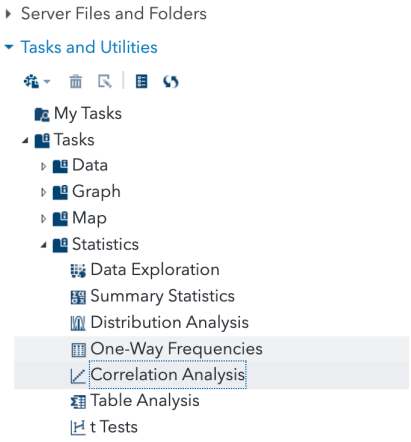


d. Click Run. The following graph should be produced.

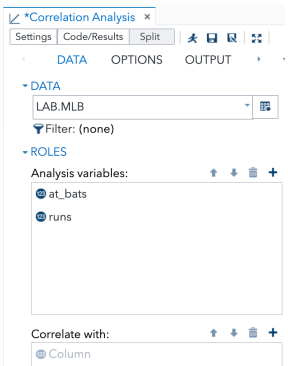


There appears to be a moderate positive relationship between runs and at_bats.

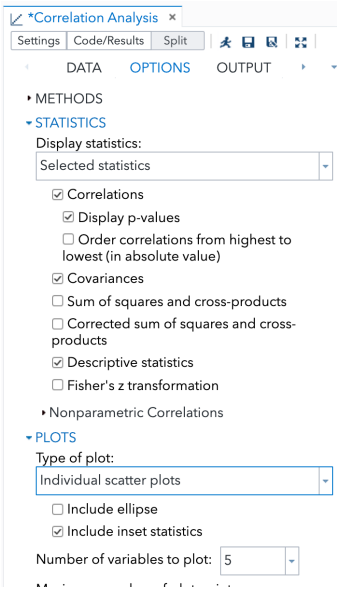
4. We can then perform a hypothesis testing on the correlation coefficient.
   a. Under Tasks and Utilities, expand Tasks, expand Statistics, then select Correlation Analysis.



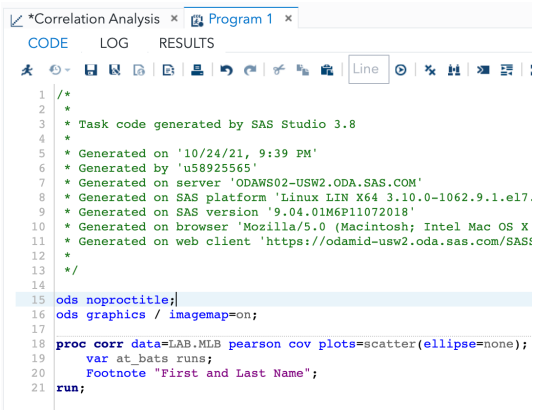
   b. In the Data tab of Correlation Analysis window, select data and Analysis variables. Click + button to set at_bats and runs as variables for correlation analysis.



c. Proceed to the Options tab, under statistics, for Display statistics, choose “Selected statistics”. Then select Correlations, display p- value, covariances, and descriptive statistics. Then, under Plots, choose “individual scatter plots”, and select include inset statistics.



d. Move to the Code/Result view and choose Code tab. Click Edit to modify the SAS code. When the SAS program view is shown, add ‘footnote "First and Last Name";’ between var and run; statements.



e. Click run to obtain the following output.

2 Variables:	at_bats runs

Covariance Matrix, DF = 29

		at_bats	runs
at_bats	at_bats	6379.706897	4022.724138
runs	runs	4022.724138	6802.800000

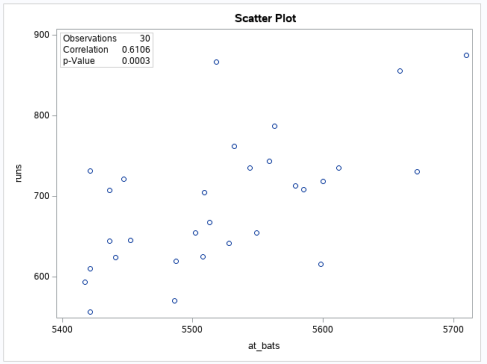
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
at_bats	30	5524	79.87307	165705	5417	5710	at_bats
runs	30	693.60000	82.47909	20808	556.00000	875.00000	runs

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	at_bats	runs
at_bats at_bats	1.00000	0.61063 0.0003
runs runs	0.61063 0.0003	1.00000

First and Last Name



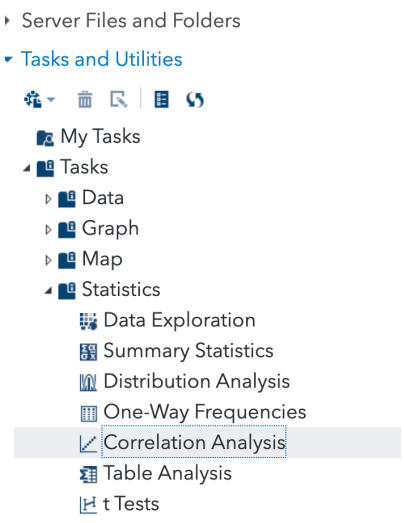
First and Last Name

Scatter plot produced by Correlation Analysis gives additional inset statistics unavailable under the graph task. When you have the Pearson correlation results, first check the scatter plot to see if there are any outliers in the data as these statistics are very sensitive to outliers. In our case, no apparent outliers stand out.

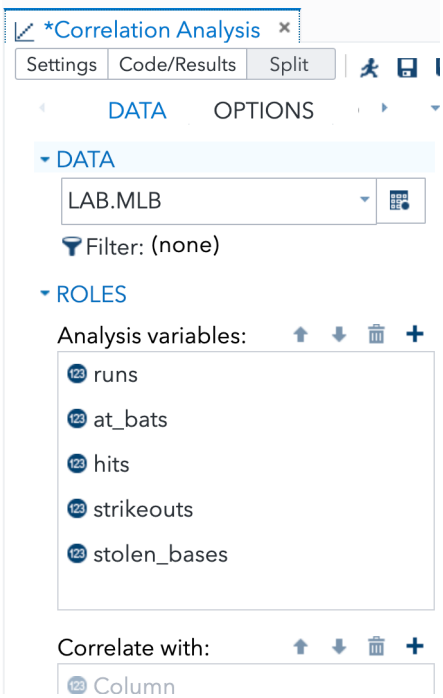
The covariance table shows the covariance between the two variables. The correlation table shows the correlation on the top and p-value on the bottom of the cell. When the two variables are NOT independent, neither of variance nor correlation is zero. The results show that the covariance is 4023 and the correlation is 0.61 with a p value of 0.0003. Therefore, the linear correlation is positive and significant between at_bats and runs. Keep in mind that Pearson correlation coefficients lie between -1 and +1, inclusive. The coefficient value shows both of direction and strength of the relationship.

Now, we can conclude that there is a significantly positive correlation of 0.61 between at_bats and runs with a p-value of 0.0003.

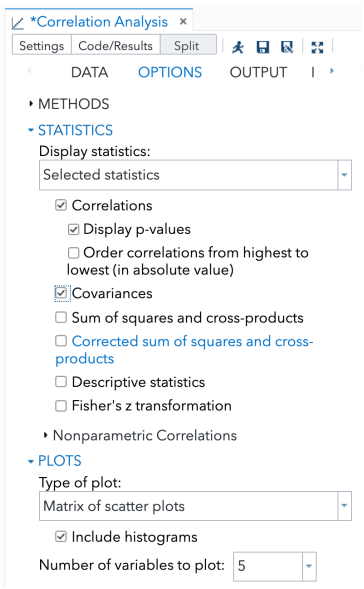
5. We can also perform a correlation analysis between multiple variables by using a pairwise correlation matrix.
    a. Under Tasks and Utilities, expand Tasks, expand Statistics, then select Correlation Analysis.



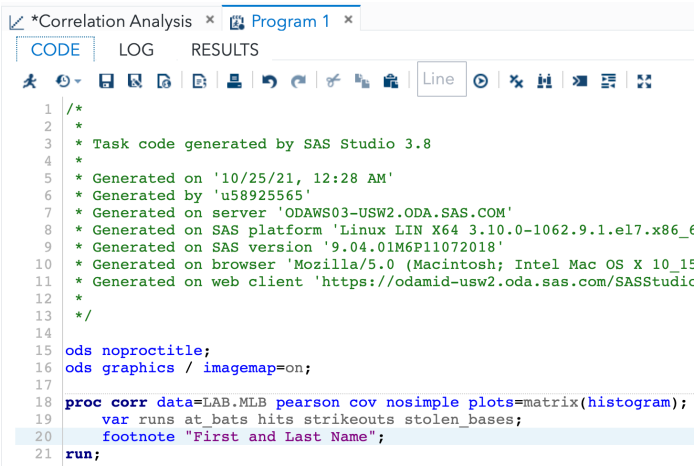
b. In the Data tab of Correlation Analysis window, select data and Analysis variables. Click + button to set runs, at_bats, hits, strikeouts, and stolen_bases as variables for correlation analysis.



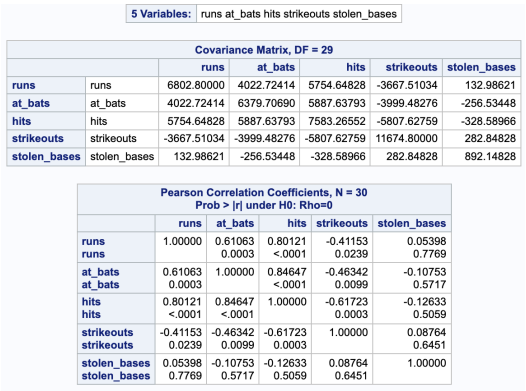
c. Proceed to the Options tab, under statistics, for Display statistics, choose “Selected statistics”. Then select Correlations, display p- value, covariances, and descriptive statistics. Then, under Plots, choose “Matrix of scatter plots”, and select ‘Include histograms’.



d. Move to the Code/Result view and choose Code tab. Click Edit to modify the SAS code. When the SAS program view is shown, add ‘footnote "First and Last Name";’ between var and run; statements.



e. Click Run.



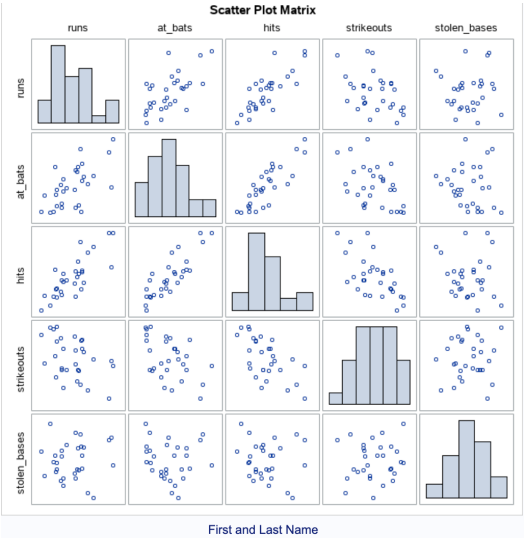
5 Variables: runs at_bats hits strikeouts stolen_bases

Covariance Matrix, DF = 29

		runs	at_bats	hits	strikeouts	stolen_bases
runs	runs	6802.80000	4022.72414	5754.64828	-3667.51034	132.98621
at_bats	at_bats	4022.72414	6379.70690	5887.63793	-3999.48276	-256.53448
hits	hits	5754.64828	5887.63793	7583.26552	-5807.62759	-328.58966
strikeouts	strikeouts	-3667.51034	-3999.48276	-5807.62759	11674.80000	282.84828
stolen_bases	stolen_bases	132.98621	-256.53448	-328.58966	282.84828	892.14828

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	runs	at_bats	hits	strikeouts	stolen_bases
runs	1.00000	0.61063	0.80121	-0.41153	0.05398
runs		0.0003	<.0001	0.0239	0.7769
at_bats	0.61063	1.00000	0.84647	-0.46342	-0.10753
at_bats	0.0003		<.0001	0.0099	0.5717
hits	0.80121	0.84647	1.00000	-0.61723	-0.12633
hits	<.0001	<.0001		0.0003	0.5059
strikeouts	-0.41153	-0.46342	-0.61723	1.00000	0.08764
strikeouts	0.0239	0.0099	0.0003		0.6451
stolen_bases	0.05398	-0.10753	-0.12633	0.08764	1.00000
stolen_bases	0.7769	0.5717	0.5059	0.6451	



In the Pearson correlation coefficients table, the intersection of each row and column shows the correlation coefficient on the top and the p-value on the bottom. You only need to look at the upper or lower triangle of the matrix due to symmetry. For example, the correlation coefficient between runs and hits is 0.801 with a p-value of less than 0.0001. What does the p-value mean? Here, we have a single sample of 30 observations out of different possible samples. If we draw another sample of 30 observations from the same population, we will end up with a different correlation coefficient. The p-value shows the probability of having a correlation coefficient as high or higher than 0.801 if the true correlation coefficient is zero. This probability is very low meaning it is very UNLIKELY to have this observed correlation if the null hypothesis is true. Therefore, we can reject the null. We can also conclude that other significant linear relationships are between runs and at_bats, runs and strikeouts, at_bats and hits, at_bats and strikeouts, hits and strikeouts.

Now, we explain the squared correlation coefficient. As an example, I will take the correlation coefficient between runs and at_bats. As shown in the previous tables, the correlation coefficient is R=0.6106. Therefore, $R^2=0.6106^2$ = 0.372868 which means that the variable at_bats can explain 37% of the variation in runs.

Simple Linear Regression

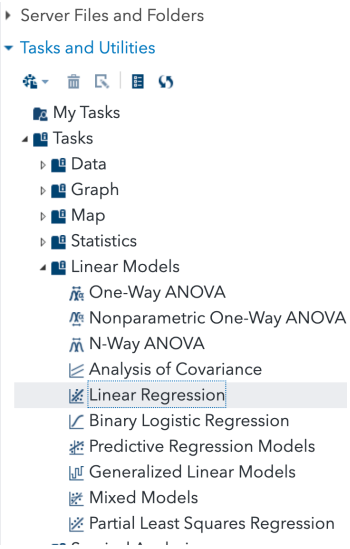
For this example, we will be investigating the question "Is at_bats (Number of at bats) a significant variable to predict runs (Number of runs)?". Essentially, this is the same as asking "Is there any relationship between runs (Number of runs) and at_bats (Number of at bats)?". But we will build a simple linear regression $Runs = \alpha + \beta at\_bats + \varepsilon$ to answer this question.

The hypotheses are:

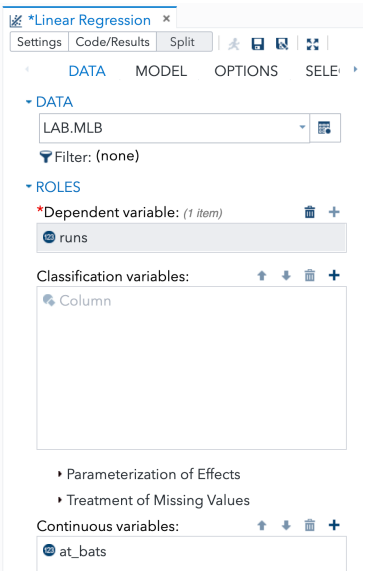
H_0: the coefficient parameter β is 0.

H_A: the coefficient parameter β is not 0.

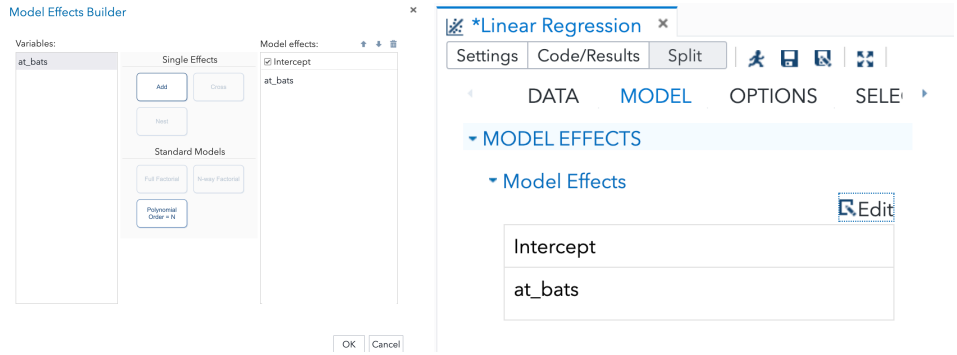
1. We will use mlb.xlsx.
2. Under Tasks and Utilities, expand Tasks, expand Linear Models, then select Linear Regression.



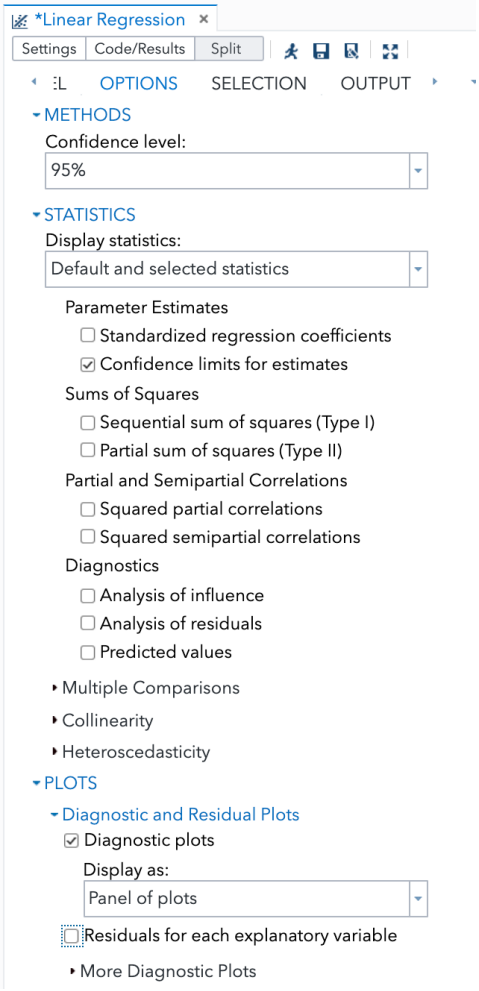
3. In the Linear Regression window, in DATA tab, select the correct library and data source. Add runs as the Dependent variable and at_bats into Continuous variables.



In MODEL tab, Click ‘Edit’ under Model Effects. When the Model Effects Builder opens, select at_bats to model effects and click OK. Now at_bats should show up under Intercept in MODEL tab.



4. In OPTIONS tab, select Default and selected statistics for Display statistics. Check the box in front of Confidence limits for estimates. Make sure Diagnostic plots displays as Panel of plots. You can uncheck the box of Residuals for each explanatory variable.



5. Click run to obtain the results.

Model: MODEL1
Dependent Variable: runs runs

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

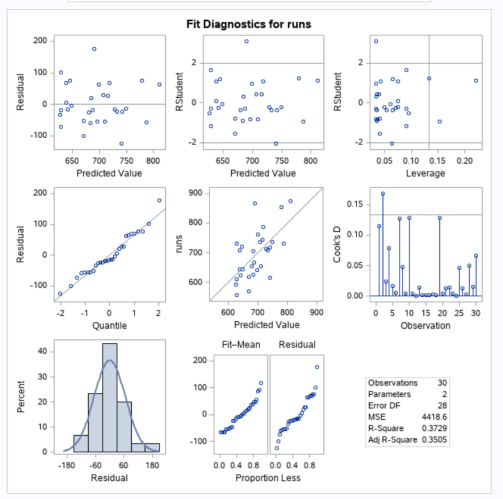
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	73559	73559	16.65	0.0003
Error	28	123722	4418.63816		
Corrected Total	29	197281			

Root MSE	66.47284	R-Square	0.3729
Dependent Mean	693.60000	Adj R-Sq	0.3505
Coeff Var	9.58374		

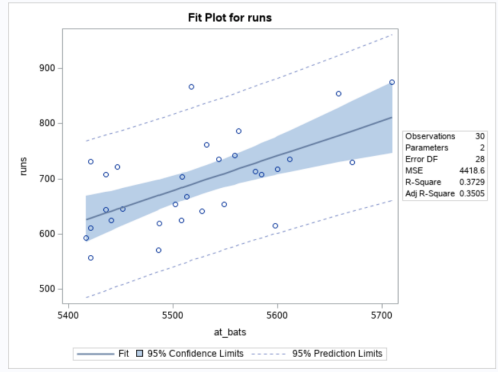
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	-2789.24289	853.69572	-3.27	0.0029	-4537.95930	-1040.52647
at_bats	at_bats	1	0.63055	0.15454	4.08	0.0003	0.31399	0.94711

From these tables, we can find the R-square value, and parameter estimates with confidence intervals and p-values. Due to a small p-value, we can conclude that the number of at bats is a significant predictor for the number of runs. This is consistent with the fact that the confidence interval of the coefficient parameter excludes zero.



The diagnostic panel help to validate the assumptions of the regression model. We can see that the residuals are roughly normally distributed with mean zero, the variance of residuals is constant. There might be some potential outlier according to the residual plot and Cook's D plot. But the data point above Cook's D of 4/n is not too much higher than the cutoff. Also, its Cook's D is not over 1.



The fit plot shows the confidence interval and prediction for the number of runs given any specific value of the predictor number of at bats.