

Historical Data RSM Tutorial

(Part 1 – The Basics)

Introduction

In this tutorial you will see how the tool of regression in Design-Expert® software, intended for response surface methods (RSM), can be applied to historical data. We don't recommend you work with such happenstance variables if there's any possibility of performed a designed experiment. However, if you feel you must, why not take advantage of how easy Design-Expert makes it to develop predictive models and graph responses, as you will see by doing this tutorial. It will be assumed that at this stage you've mastered many of the program features by completing the preceding tutorials. At the very least you ought to first do the one-factor RSM tutorials, basic and advanced, prior to starting this one.

The historical data for this tutorial, shown below, comes from the U.S. Bureau of Labor Statistics via James Longley (An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User, *Journal of the American Statistical Association*, 62 (1967): 819-841). As discussed in *RSM Simplified* (Mark J. Anderson and Patrick J. Whitcomb, Productivity, Inc., New York: Chapter 2), it presents some interesting challenges for regression modeling.

Run #	A: Prices (1954 =100)	B: GNP	C: Unemp.	D: Military Armed Forces	E: Pop. People >14	F: Time Year	Employ. Total
1	83	234289	2356	1590	107608	1947	60323
2	88.5	259426	2325	1456	108632	1948	61122
3	88.2	258054	3682	1616	109773	1949	60171
4	89.5	284599	3351	1650	110929	1950	61187
5	96.2	328975	2099	3099	112075	1951	63221
6	98.1	346999	1932	3594	113270	1952	63639
7	99	365385	1870	3547	115094	1953	64989
8	100	363112	3578	3350	116219	1954	63761
9	101.2	397469	2904	3048	117388	1955	66019
10	104.6	419180	2822	2857	118734	1956	67857
11	108.4	442769	2936	2798	120445	1957	68169
12	110.8	444546	4681	2637	121950	1958	66513
13	112.6	482704	3813	2552	123366	1959	68655
14	114.2	502601	3931	2514	125368	1960	69564
15	115.7	518173	4806	2572	127852	1961	69331
16	116.9	554894	4007	2827	130081	1962	70551


Longley data on U.S. economy from 1947-1962

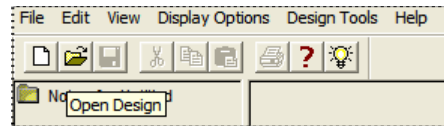
Assume that the objective for analysis of this data is to predict future employment as a function of leading economic indicators – the factors labeled A through F in the table above. Longley's goal was different: He wanted to test regression software circa 1967

for round-off error due to highly-correlated inputs. Will Design-Expert be up to the challenge? We will see!

Let's begin by setting up this "experiment" (quotes added to emphasize that this is not really an experiment, but rather an after-the-fact analysis of happenstance data).

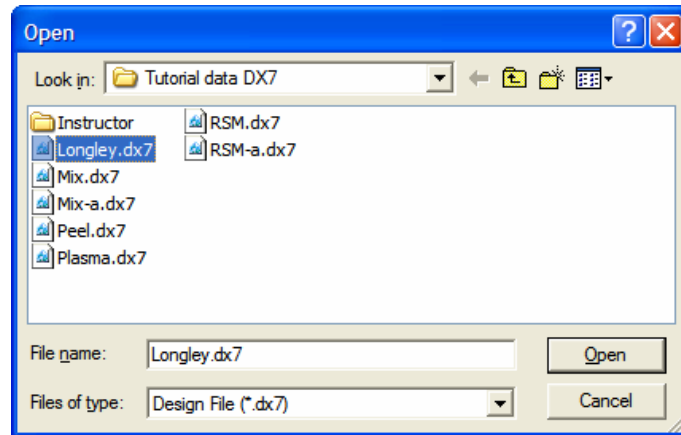
Design the "Experiment"

Start Design-Expert. You will then see the main menu and icon bar. To save you time typing stuff, we will re-build a previously saved design rather than enter it from scratch. Using your mouse, press the Open Design icon  (or select File, Open Design).




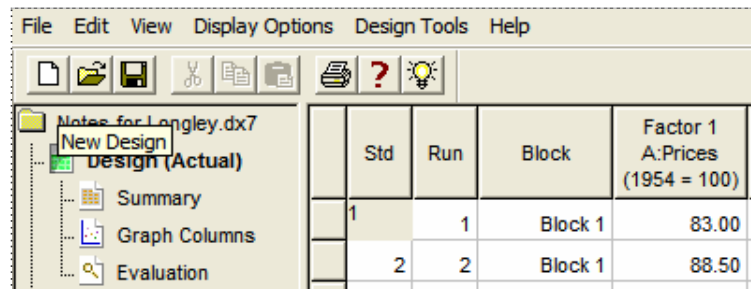
Main menu and Tool bar – Open Design icon highlighted

The file name is **Longley.dx7**. Click on it and press **Open**.



Opening the Longley data

The data should now appear on your screen. To re-build this design (and thus see how it was created), press the blank-sheet icon  on the left of the toolbar (or select File, New Design).



New Design icon

When prompted by Design-Expert to “Use previous design info,” click **Yes**.



Re-using previous design

Now you see how this design was created via the Response Surface tab and Historical Data option.

 A screenshot of the "Historical Data Design" dialog box. On the left is a tree view with categories: Factorial, Combined, Mixture, Response Surface, Central Composite, Box-Behnken, One Factor, Miscellaneous, D-Optimal, Distance-Based, User-Defined, and Historical Data (which is highlighted). The main area contains the following text: "Design for importing data that already exists. Specify the factor names and levels. The level names are case sensitive and imported data must be pasted into the design layout." Below this text are two dropdown menus: "Numeric Factors: 5 (1 to 10)" and "Categoric Factors: 0 (0 to 10)". At the bottom is a table with 6 rows of data.

	Name	Units	Min	Max
A:	Prices	(1954 = 100)	83	116.9
B:	GNP		234289	554894
C:	Unemployment		1870	4806
D:	Military	Armed Forces	1456	3594
E:	Population	People >14	107608	130081
F:	Time	Year	1947	1962

Setting up design on historical data

Note that for each of the 6 numeric factors we entered the name, units and range from minimum (“Min”) to maximum (“Max”). Before moving ahead you must also tell Design-Expert how many rows of data you want to type or paste into the design layout. In this case there are 16 rows.

 A screenshot of a dialog box showing the "Rows (2-32767):" label followed by a text input field containing the number "16".

Entry for rows

Press **Continue** to accept all the entries on this screen. You now see details on the response(s) – in this case only the one we will study.

Historical Data Design

Responses: 1

Name	Units
Employment	Total

Response entry

Press **Continue** to see the resulting design layout in run order (ignore the column labeled “Std” because there will be no standard order for happenstance data).

A Peculiarity on Pasting Data

You could now type in all the data for factor levels and resulting responses, row-by-row. (Don’t worry: We won’t make you do this!) However, in most cases the data will already be available via a Microsoft Window’s based spreadsheet. Then simply drag over this data, copy it to the Window’s clipboard, and Edit, Paste (or right-click and Paste as shown below) into the design layout within Design-Expert after first dragging the top row or over all the destination cells (as shown below).

Run	Block	Factor 1 A:Prices (1954 = 100)	Factor 2 B:GNP	Factor 3 C:Unemploye	Factor 4 D:Military Armed Forces	Factor 5 E:Population People >14	Factor 6 F:Time Year	Response 1 Employment Total
1	Block 1							
2	Block 1							
3	Block 1							
4	Block 1							
5	Block 1							
6	Block 1							
7	Block 1							
8	Block 1							
9	Block 1							
10	Block 1							
11	Block 1							
12	Block 1							
13	Block 1							
14	Block 1							
15	Block 1							
16	Block 1							

The correct way to Paste data into Design-Expert (destination cells pre-selected)

If you simply click the upper left cell in the empty run-sheet, the program will warn you that it cannot accept this and specify the number of rows and columns that must be made available for the data to be pasted into.

Run	Block	Factor 1 A:Prices (1954 = 100)	Factor 2 B:GNP	Factor 3 C:Unemploye	Factor 4 D:Military Armed Forces	Factor 5 E:Population People >14
1	Block 1					
2	Block 1					
3	Block 1					


Error

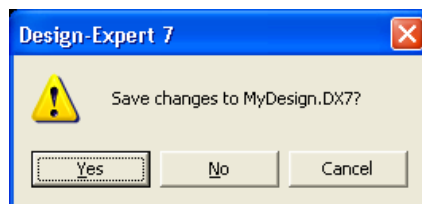
Could not paste 16R x 7C table into 1R x 1C area.

Wrong way to paste data into Design-Expert software

In this case it pops up an error message about trying to cram 16 rows and 7 columns into only one cell (1 row by 1 column).

Analyze the Results

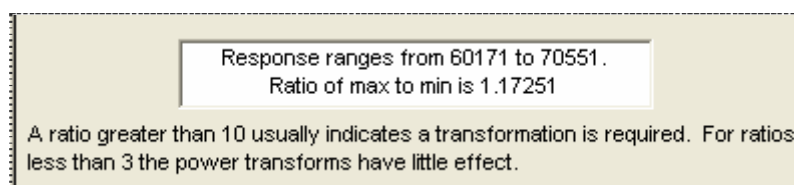
Normally you'd save your work at this stage, but since we already did this, simply re-open our file: Press the Open Design icon  and double-click **Longley.dx7**. Click **No** to pass on the opportunity to save what you did previously.



Last chance to save (say "No" in this case)

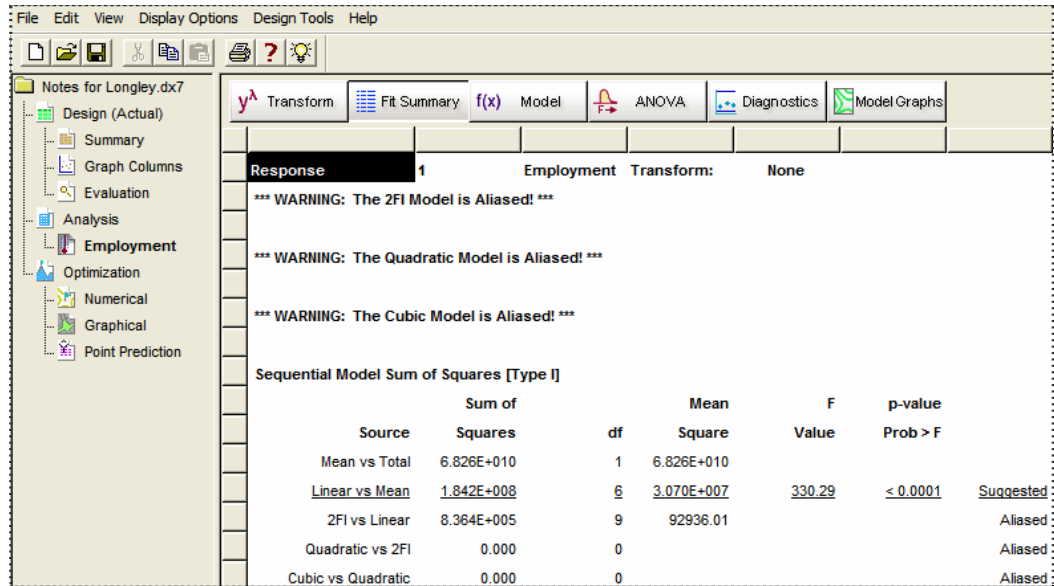
Before we get started, be forewarned that you will now get exposed to quite a number of statistics related to least squares regression and analysis of variance (ANOVA). If you are coming into this cold, pick up a copy of *RSM Simplified* and keep it handy. For a good guided tour of these statistics for RSM analysis, attend the Stat-Ease workshop RSM for Process Optimization. Details on this computer-intensive hands-on class, including what's needed a prerequisite, can be found at www.statease.com.

Under the **Analysis** node click the branch labeled **Employment**. Design-Expert then displays a screen for transforming the response. However, as noted by the program, the range of response in this case is so small that there would be little advantage to applying any transformation.



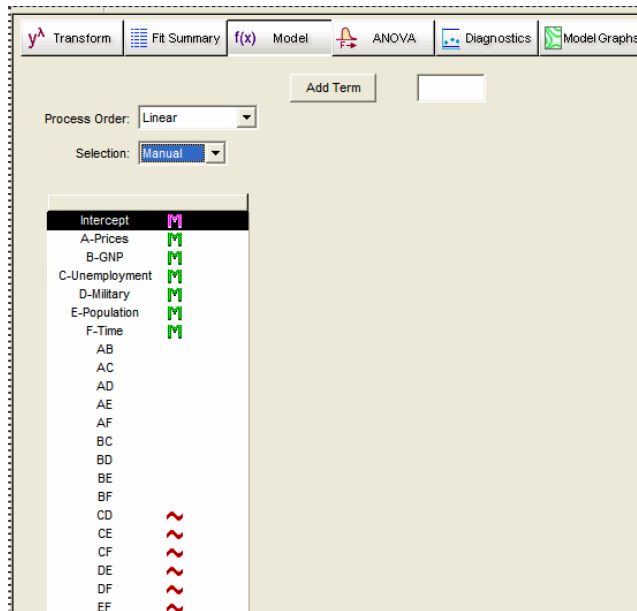
Information about the response shown on the Transformation screen

Go ahead and press **Fit Summary**. Design-Expert then evaluates each degree of the model from the mean on up. In this case, the best that can be done is linear. Anything above that becomes aliased.



Fit Summary – only the linear model possible in this case

You may as well press on to **Model**.



The linear model chosen

It's all set the way the software suggested. Notice that many of the two-factor interactions cannot be estimated due to aliasing symbolized by the red tildas (~). Hold on to your hats (because this data is really just a lot of hot air!) and press **ANOVA** for the analysis of variance.

Response	1 Employment				
ANOVA for Response Surface Linear Model					
Analysis of variance table [Partial sum of squares - Type III]					
Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	1.842E+008	6	3.070E+007	330.29	< 0.0001
A-Prices	2923.98	1	2923.98	0.031	0.8631
B-GNP	1.063E+005	1	1.063E+005	1.14	0.3127
C-Unemployment	1.590E+006	1	1.590E+006	17.11	0.0025
D-Military	2.161E+006	1	2.161E+006	23.25	0.0009
E-Population	4748.95	1	4748.95	0.051	0.8262
F-Time	1.499E+006	1	1.499E+006	16.13	0.0030
Residual	8.364E+005	9	92936.01		
Cor Total	1.850E+008	15			

Analysis of variance (ANOVA)

Notice that although the overall model is significant, some terms are not.*

*(Here are some statistical details on how Design-Expert does analysis of variance. You may have noticed that this ANOVA is labeled as “[Partial sum of squares - Type III]. This approach to ANOVA, done by default, causes the total sums-of-squares (SS) for the terms to come up short of the overall model when analyzing data from a non-orthogonal array, such as historical data. If you want SS terms to add up to the model SS, go to Edit, Preferences and change the default to Sequential (Type I). However, we do not recommend this approach because it favors the first term put into the model. For example, in this case the ANOVA by partial SS (Type III -- the default of DX) for the response (employment total) calculates prob>F p-value for A as 0.8631 (F=0.031) as seen above, which is not significant. Recalculating ANOVA by sequential sum of squares (Type I) changes the p to <0.0001 (F=1876), which looks highly significant, but only because this term (main effect of factor A) is fit first. That simply is not correct.)

Assuming Factor A (population) is least significant of all as indicated by the default ANOVA (partial SS), let’s see what happens with it removed. However, before we do, scroll down and look at some statistics (shown below) that will help us compare what happens before and after reducing the model.

Std. Dev.	304.85	R-Squared	0.9955
Mean	65317.00	Adj R-Squared	0.9925
C.V. %	0.47	Pred R-Squared	0.9844
PRESS	2.887E+006	Adeq Precision	53.075
The "Pred R-Squared" of 0.9844 is in reasonable agreement with the "Adj R-Squared" of 0.9925.			
"Adeq Precision" measures the signal to noise ratio. A ratio greater than 4 is desirable. Your ratio of 53.075 indicates an adequate signal. This model can be used to navigate the design space.			

Model statistics

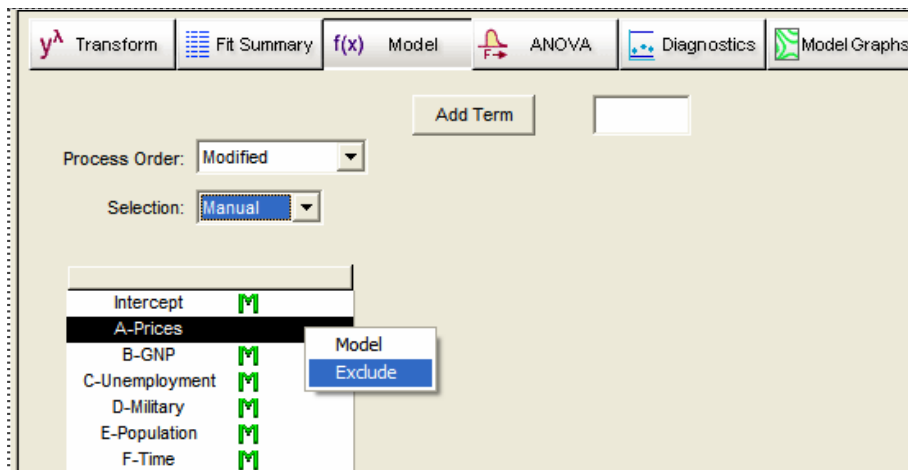
While you’re at it, also look at the coefficient estimates.

Factor	Coefficient	DF	Standard	95% CI		VIF
	Estimate		Error	Low	High	
Intercept	64763.53	1	224.55	64255.56	65271.49	
A-Prices	255.30	1	1439.31	-3000.64	3511.24	135.53
B-GNP	-5741.90	1	5368.69	-17886.73	6402.92	1788.51
C-Unemployment	-2965.70	1	716.97	-4587.60	-1343.80	33.62
D-Military	-1104.52	1	229.06	-1622.69	-586.35	3.59
E-Population	-574.23	1	2540.27	-6320.72	5172.26	399.15
F-Time	13718.64	1	3416.09	5990.91	21446.37	758.98

Coefficient estimates for linear model

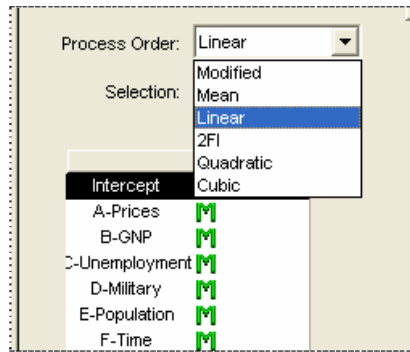
Notice the huge VIF's (variance inflation factor). A value of 1 is ideal (orthogonal), but a VIF less than 10 is generally accepted. VIF's above 1000, such as that observed for factor B (GNP), indicate severe multicollinearity in the model coefficients (that's bad!). In the follow-up tutorial (Part 2) based on this same Longley data, we will delve more into this and other statistics generated by Design-Expert for purposes of design evaluation. For now, try right-clicking any of the VIF results to access context-sensitive Help, or go to Help on the main menu and search on this statistic. You will likely find some details there.

Press back to **Model** and via a right-click on **A-Prices** and **Exclude** it, or simply double-click on this term to take off the model ("M") designation.



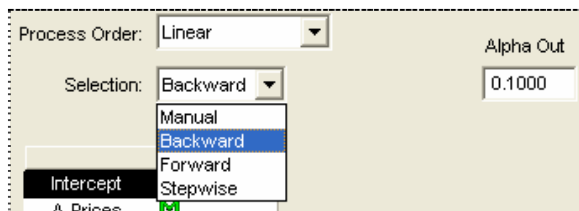
Excluding an insignificant term

You can now go back to ANOVA, look for the next least significant term, exclude it, etc. However, this backward elimination process can be done automatically in Design-Expert. Here's how. First, reset the **Process Order** to **Linear**.



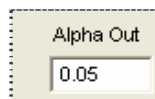
Resetting model to linear

Now change the **Selection** to **Backward**.



Specifying backward stepwise regression

Notice that a new field called “Alpha Out” appears. By default the program will remove the least significant term step-by-step so long as it exceeds the risk level (symbolized with the Greek letter alpha by statisticians) of 0.1 (estimated by the p-value). Let’s be a bit more conservative by changing **Alpha Out** to **0.05**.



Changing the risk level alpha for taking out model terms via backward selection

Now press **ANOVA** to see what happens.

y^x Transform Fit Summary f(x) Model ANOVA Diagnostics Model Graphs					
Response	1	Employment			
Backward Elimination Regression with Alpha to Exit = 0.050					
Forced Terms	Intercept				
		Coefficient	t for H ₀		
Removed		Estimate	Coeff=0	Prob > t	R-Squared MSE
A-Prices		255.30	0.18	0.8631	0.9955 83934.80
E-Population		-871.25	-0.48	0.6416	0.9954 78061.86

Results from backward regression

Not surprisingly the program first removed A and then E – that’s it. Take a look at the ANOVA table that follows to see that all the other terms come out significant.

ANOVA for Response Surface Reduced Linear Model					
Analysis of variance table [Partial sum of squares - Type III]					
Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	1.842E+008	4	4.604E+007	589.76	< 0.0001
B-GNP	4.647E+005	1	4.647E+005	5.95	0.0328
C-Unemployment	4.049E+006	1	4.049E+006	51.87	< 0.0001
D-Military	2.381E+006	1	2.381E+006	30.50	0.0002
F-Time	1.898E+006	1	1.898E+006	24.31	0.0004
Residual	8.587E+005	11	78061.86		
Cor Total	1.850E+008	15			

ANOVA for backward-reduced model

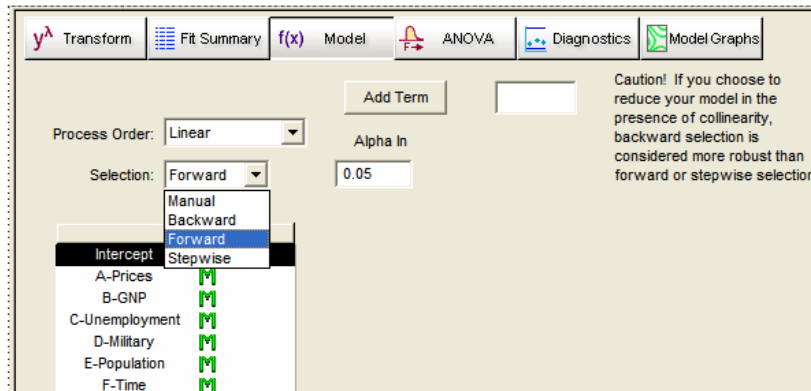
You may’ve noticed that in the full model, factor B had a much higher p-value than what’s shown above. This instability is typical of models based on historical data. Scroll down and view the model statistics and coefficients.

Std. Dev.	279.40	R-Squared	0.9954
Mean	65317.00	Adj R-Squared	0.9937
C.V. %	0.43	Pred R-Squared	0.9892
PRESS	1.998E+006	Adeq Precision	68.654

Factor	Coefficient		Standard Error	95% CI		VIF
	Estimate	df		Low	High	
Intercept	64820.68	1	159.08	64470.56	65170.81	
B-GNP	-6442.63	1	2640.62	-12254.59	-630.67	515.12
C-Unemployment	-3065.76	1	425.68	-4002.67	-2128.85	14.11
D-Military	-1084.65	1	196.41	-1516.95	-652.35	3.14
F-Time	14155.57	1	2870.75	7837.10	20474.05	638.13

Backward reduced model statistics and coefficients

Now let’s try a different regression approach – building the model from the ground (mean) up, rather than from tearing things down from the top (all terms in chosen polynomial). Press **Model**, re-set the **Process Order** to **Linear** and this time ask for a **Selection** based on **Forward** stepwise regression. To provide a fair comparison of this forward approach with that done earlier going backward, change **Alpha In** to **0.05**.



Forward selection (remember to re-set the model to the original process order first!)

Heed the caution put up by the program – this approach may not work as well for this highly-collinear set of factors. See what happens now in **ANOVA**.

y^ Transform Fit Summary f(x) Model ANOVA Diagnostics Model Graphs					
Response	1	Employment			
Forward Regression with Alpha to Enter = 0.050					
Forced Terms	Intercept				
	Coefficient	t for H ₀			
Added	Estimate	Coeff=0	Prob > t	R-Squared	MSE
B-GNP	5570.88	20.37	<0.0001	0.9674	4.312E+005
C-Unemployment	-797.97	-2.99	0.0105	0.9807	2.753E+005

Results from forward regression

Surprisingly, factor B now comes in first as the single most significant factor. Then comes factor C. That's it! The next most significant factor evidently does not achieve the alpha-in significance threshold of $p < 0.05$.

Take a look at the ANOVA table that follows to see that all the other terms come out significant.

ANOVA for Response Surface Reduced Linear Model					
Analysis of variance table [Partial sum of squares - Type III]					
Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	1.814E+008	2	9.071E+007	329.50	< 0.0001
B-GNP	1.347E+008	1	1.347E+008	489.31	< 0.0001
C-Unemployment	2.457E+006	1	2.457E+006	8.92	0.0105
Residual	3.579E+006	13	2.753E+005		
Cor. Total	1.850E+008	15			

ANOVA for forward-reduced model

Scroll down and view the model statistics and coefficients.

Std. Dev.	524.70	R-Squared	0.9807
Mean	65317.00	Adj R-Squared	0.9777
C.V. %	0.80	Pred R-Squared	0.9726
PRESS	5.077E+006	Adeq Precision	49.446

The "Pred R-Squared" of 0.9726 is in reasonable agreement with the "Adj R-Squared" of 0.9777.

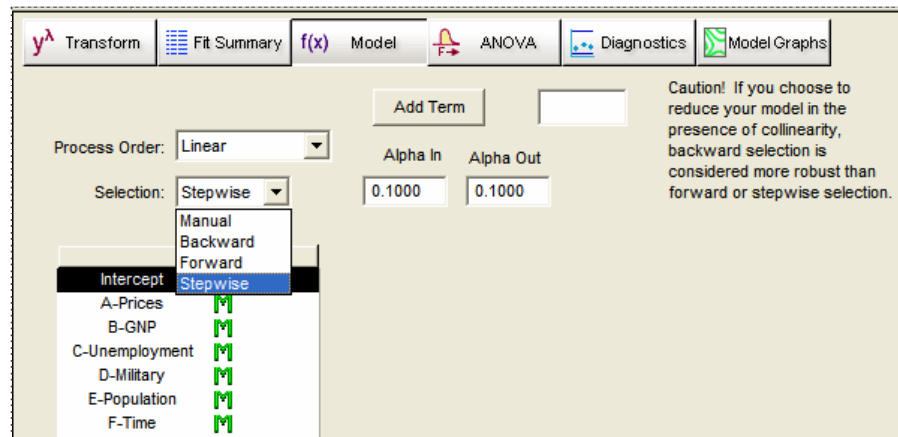
"Adeq Precision" measures the signal to noise ratio. A ratio greater than 4 is desirable. Your ratio of 49.446 indicates an adequate signal. This model can be used to navigate the design space.

Factor	Coefficient		Standard Error	95% CI		VIF
	Estimate	df		Low	High	
Intercept	65499.19	1	132.91	65212.06	65786.31	
B-GNP	6065.90	1	274.22	5473.48	6658.32	1.58
C-Unemployment	-797.97	1	267.11	-1375.02	-220.91	1.58

Forward reduced model statistics and coefficients

This simpler model scores very high on all measures of R-squared, but it falls a bit short of what was achieved in the model derived from the backward regression.

Finally, go back to the **Model**, re-set the **Process Order** to **Linear** and check out the last model **Selection** option offered by Design-Expert software: **Stepwise**.



As you might infer from seeing both Alpha In and Alpha Out now displayed, the stepwise algorithm involves elements of forward selection with bits of backward added in for good measure. For details search program Help, but consider this – terms that pass the alpha test in (via forward regression) may later (after further terms are added) become disposable according to the alpha test out (via backward selection). If this seems odd, just look back at how the p-value for factor B changed depending what other factors were chosen along with it for modeling.

If you want to see what happens with this selection method, press ANOVA. The results depend on what you do with Alpha In and Alpha Out, which default back to 0.1.

As you see on the cautionary message displayed for both forward and stepwise (in essence an enhancement on forward) approaches, we favor the backward approach if you

decide to make use of an automated selection method. Ideally an analyst will also be an expert on the subject matter, or have such a person readily accessible. Then they could do model reduction via the manual method filtered not only by the statistics, but also common sense of someone with profound knowledge of the system.

This concludes part 1 of our exploration of the Longley data set. In Part 2 we will dig further under the covers of Design-Expert to see some interesting aspects in the residual analysis under Diagnostics, and also see what can be gleaned from its sophisticated tools under Design, Evaluation.