Historical Data RSM Tutorial
Part 1 – The Basics

Introduction

In this tutorial you will see how the regression tool in Design-Expert® software, intended for response surface methods (RSM), is applied to historical data. We don't recommend you work with such happenstance variables if there's any possibility of performing a designed experiment. However, if you must, 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 is assumed that at this stage you've mastered many program features by completing preceding tutorials. At the very least you ought to first do the one-factor RSM tutorials, both 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, 2005: 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 the objective for analyzing this data is to predict future employment as a function of leading economic indicators – 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 this is not really an experiment but rather an after-the-fact analysis of happenstance data).

Design the "Experiment"

Click the Design-Expert icon that may appear on your desktop. You will see our handy new easy-start opening page. (To save you typing time, we will re-build a previously saved design rather than entering it from scratch.) Click the **Open Design** button as shown below.



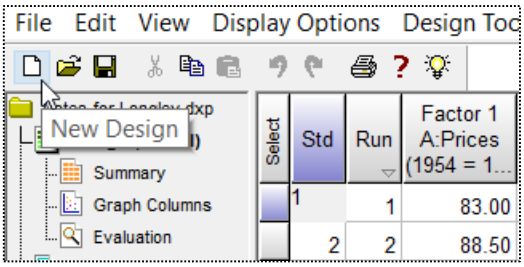
New easy-start page – Open Design

The file name is **Longley.dxp**. Double-click to open.



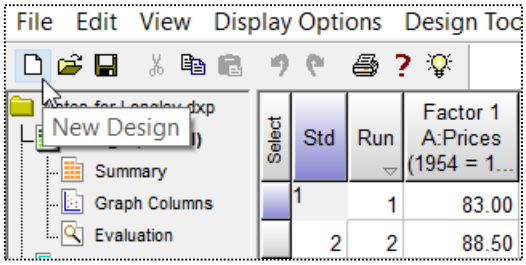
Opening the Longley data

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



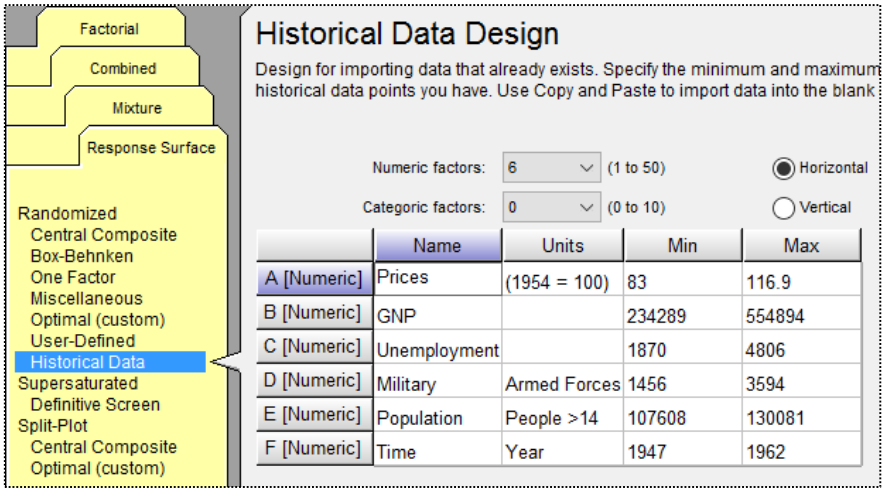
New Design icon

Click **Yes** when Design-Expert queries "Use previous design info?"



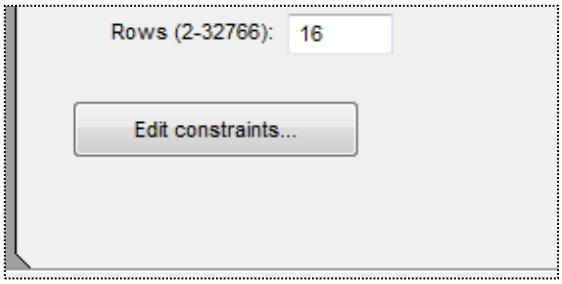
Re-using previous design

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



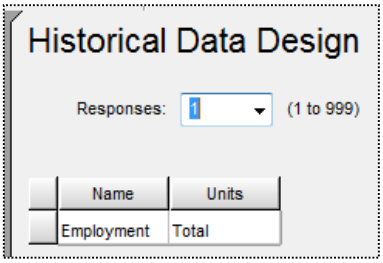
Setting up historical data design

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



Entry for rows

Press **Continue** to accept all entries on your screen. You now see response details – in this case only one response.

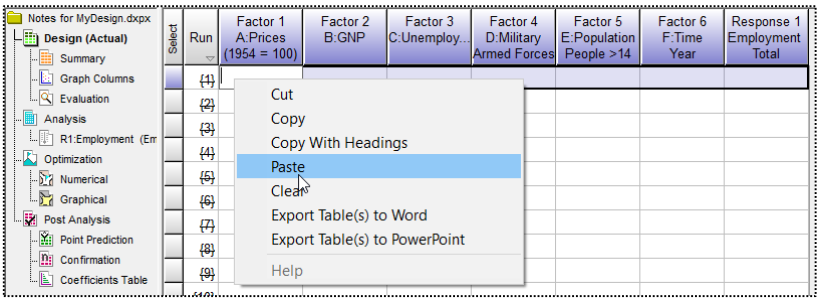


Response entry

Press **Continue** to see the resulting design layout in run order.

A Peculiarity on Pasting Data

You could now type in all data for factor levels and resulting responses, row-by-row. (Don't worry: We won't make you do this!) However, in most cases data is already available via a Microsoft Window-based spreadsheet. If so, simply click/drag these data, copy to Window's clipboard, and Edit, Paste (or right-click and Paste as shown below) into the design layout within Design-Expert. (Be sure, as shown below, to first click/drag the top row of all your destination cells.)

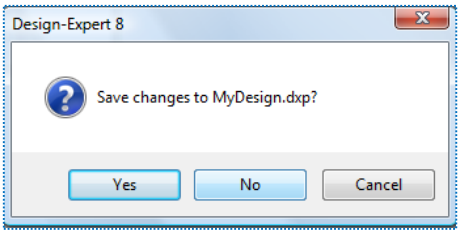


Correct way to paste data into Design-Expert (top-row of cells pre-selected)

If you simply click the upper left cell in the empty run sheet, the program only pastes one value.

Analyze the Results

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

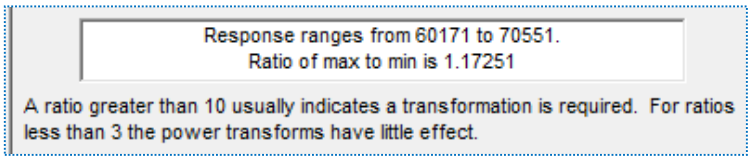


Last chance to save (say "No" in this case)

Before we get started, be forewarned you will encounter many statistics related to least squares regression and analysis of variance (ANOVA). If you are coming into this without previous knowledge, pick up a copy of *RSM Simplified* and keep it

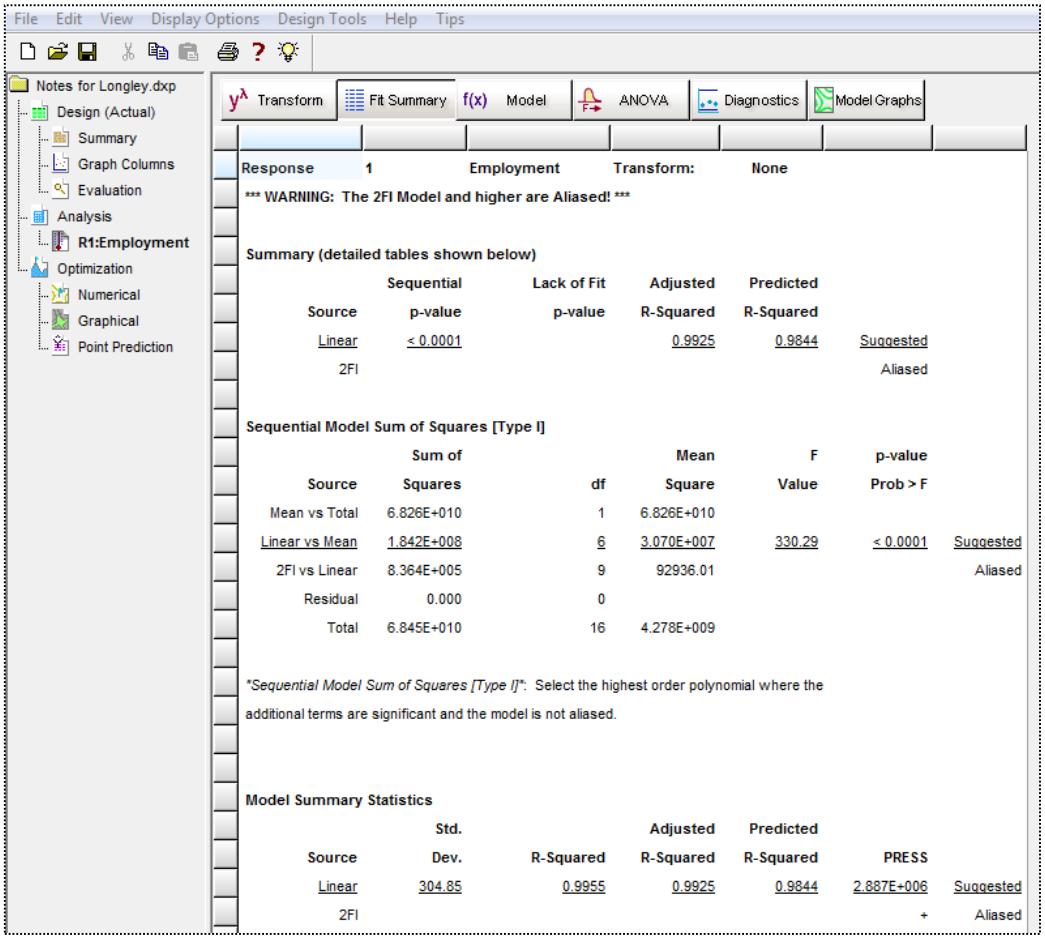
handy. For a good guided tour of statistics for RSM analysis, attend our Stat-Ease workshop titled RSM for Process Optimization. Details about this computer-intensive, hands-on class – including prerequisites – are at www.statease.com.

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



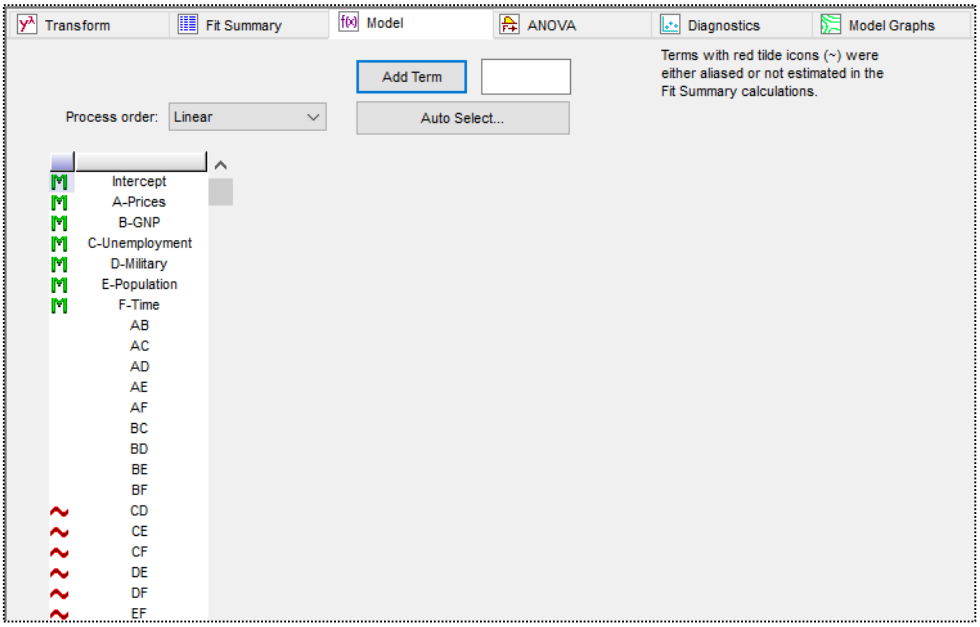
Information about the response shown on the Transformation screen

Press ***Fit Summary***. Design-Expert evaluates each degree of the model from the mean on up. In this case, the best that can be done is linear. Anything higher is aliased.



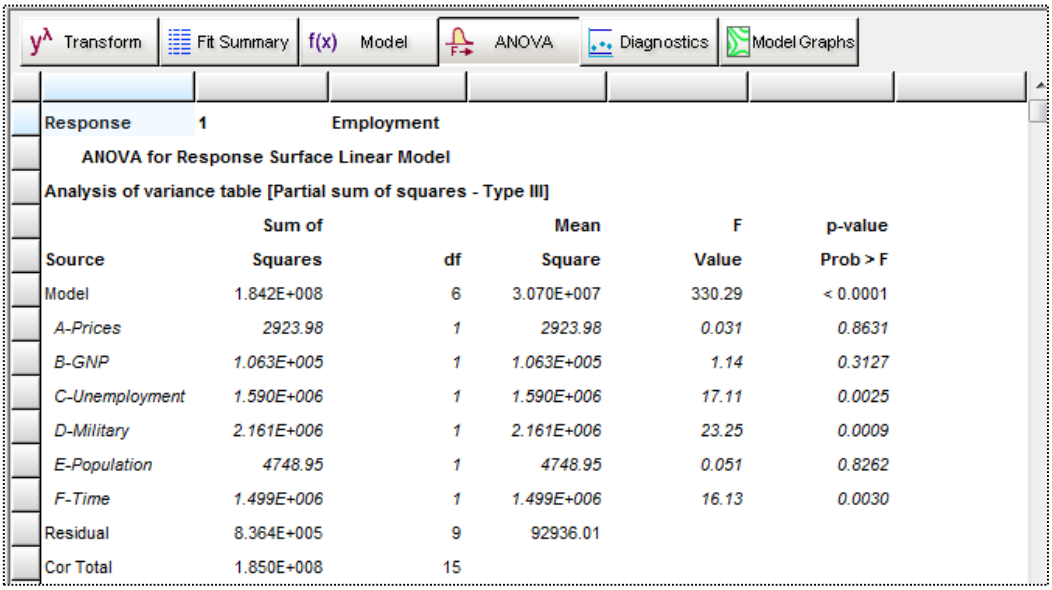
Fit Summary – only the linear model is possible here

Move on by pressing ***Model***.



Linear model is chosen

It's all set up how Design-Expert suggested. Notice many two-factor interactions can't be estimated due to aliasing – symbolized by a red tilde (~). Hold on to your hats (because this upcoming data is really a lot of hot air!) and press ***ANOVA*** (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			

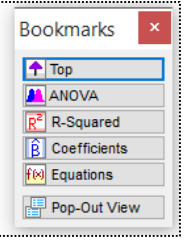
Analysis of variance (ANOVA)

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

☒ ***Some statistical details on how Design-Expert does analysis of variance***: You may have noticed this ANOVA is labeled "[Partial sum of squares - Type III]. This approach to ANOVA, done by default, causes 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 for Analysis and change the default to Sequential (Type I) for these numeric factors. However, we do not recommend this approach because it favors the first term put into the model. For example, in this case, 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. This simply is not correct.

Assuming Factor A (population) is least significant of all as indicated by default ANOVA (partial SS), let's see what happens with it removed. However, before we do, on the **Bookmarks,** click **R-Squared** and view statistics (shown below) to help us compare what happens before and after reducing the model.

Std. Dev.	304.85	R-Squared	0.9955
Mean	65317.00	Adj R-Squarec	0.9925
C.V. %	0.47	Pred R-Square	0.9844
PRESS	2.887E+006	Adeq Precisio	53.075
-2 Log Likelihc	219.23	BIC	238.64
		AICc	247.23



Model statistics

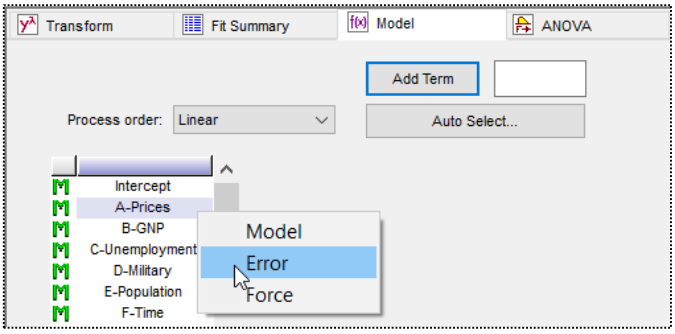
Also bookmark to the **Coefficients** estimates.

Factor	Coefficient Estimate	df	Standard Error	95% CI Low	95% CI High	VIF
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

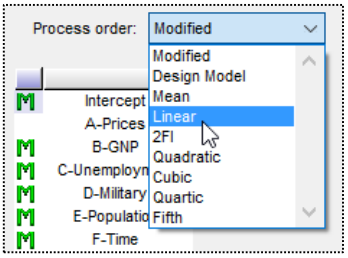
Notice the huge VIF (variance inflation factor) values. A value of 1 is ideal (orthogonal), but a VIF below 10 is generally accepted. A VIF above 1000, such as factor B (GNP), indicates severe multicollinearity in the model coefficients. (That's bad!). In the follow-up tutorial (Part 2) based on this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

Press ***Model*** again. Right-click **A-Prices** and **Exclude** it, or simply double-click this term to remove the "M" (model) designation.



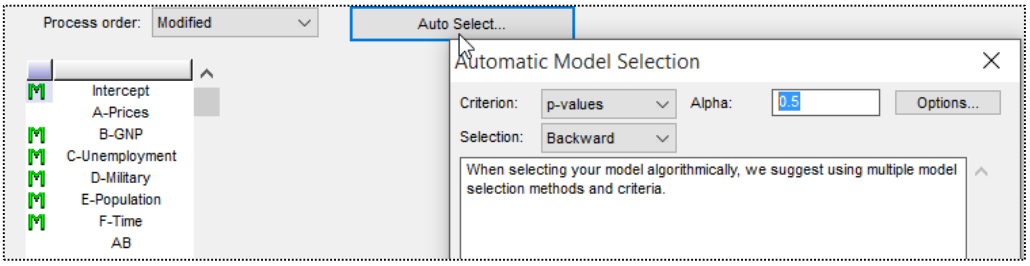
Excluding an insignificant term

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



Resetting model to linear

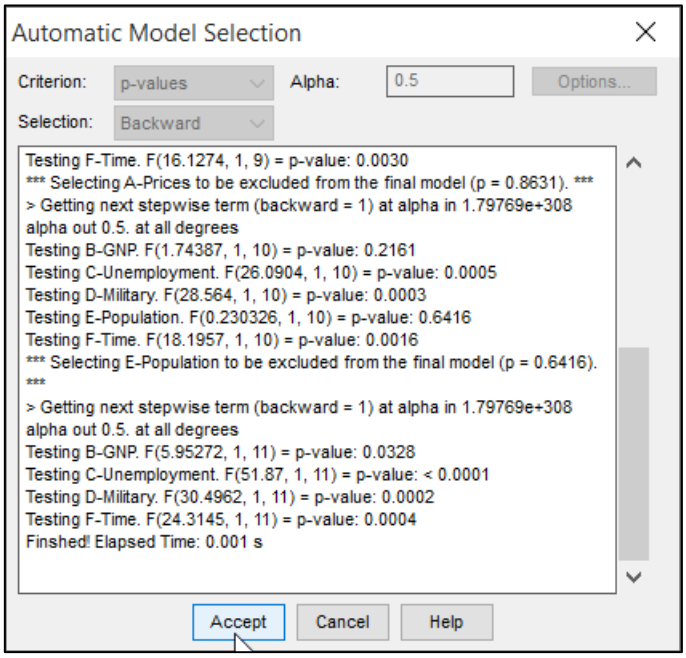
Now click on the **Autoselect…** button. Then change the selection to **Backward** and the Criterion to **p-value**.



Specifying backward regression

Notice a new field called "Alpha out" appears. By default the program removes the least significant term, step-by-step, as long as it exceeds the risk level (symbolized by statisticians with the Greek letter alpha) of 0.1 (estimated by p-value). Let's be a bit more conservative by changing **Alpha out** to **0.05**.

Now press the **Start** button to see what happens.



Backward regression results

The automatic selection is shown, step-by-step. Scroll up to see the whole thing if you like. For now, though, let's move on and see what model is left and check out the more user friendly "selection log" to see what was done. The Start button becomes an **Accept** button, so click on that and then you click on the ***ANOVA*** to see the resulting model.

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	significant
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

We are left with the same model we landed on by hands, but this was much easier. We also get a nice summary of how we got here. Click on the View menu and select "Show Selection Log".

Model Selection Log

Backward Elimination Regression

Criteria: Alpha to Exit = 0.050

Removed	F Value	p-value Prob > F	R-Squared	MSE
A-Prices	0.031	0.8631	0.9955	92936.01
E-Population	0.23	0.6416	0.9955	83934.80

Model Selection Log

Not surprisingly, the program first removed A and then E – that's it. All of the other terms on the ANOVA table come out significant. (Note: If you do not see the report of the model being "significant" change your View to Annotated ANOVA.)

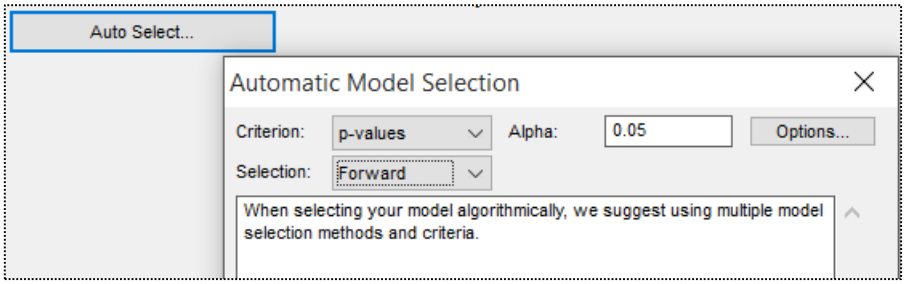
You may have noticed that in the full model, factor B had a much higher p-value than what's shown above. This instability is typical of models based on historical data. Scroll down the ANOVA table to view model statistics and coefficients (or click the R-squared Bookmark).

Std. Dev.	279.40	R-Squared	0.9954
Mean	65317.00	Adj R-Squarec	0.9937
C.V. %	0.43	Pred R-Square	0.9892
PRESS	1.998E+006	Adeq Precisio	68.654
-2 Log Likelihc	219.66	BIC	233.52
		AICc	235.66

Factor	Coefficient Estimate	df	Standard Error	95% CI Low	95% CI High	VIF
Intercept	64820.68	1	159.08	64470.56	65170.81	
B-GNP	-6442.63	1	2640.62	-12254.59	-630.67	515.12
C-Unemploym	-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 tearing terms down from the top (all terms in chosen polynomial). Press **Model**, then re-set **Process Order** to **Linear** and click the **Auto Select…** button. This time choose **p-values** as your criterion and leave **Forward** for the Selection method. To provide a fair comparison of this forward approach with that done earlier going backward, change **Alpha** to **0.05**.



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

Heed the text displayed by the program *(When reducing your model…)* because this approach may not work as well for this highly collinear set of factors. Press Start and then See what happens now in ***ANOVA***.

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	significant
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				

Results of 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.

On **Bookmarks,** click **R-Squared**.

Std. Dev.	524.70	R-Squared	0.9807
Mean	65317.00	Adj R-Squarec	0.9777
C.V. %	0.80	Pred R-Square	0.9726
PRESS	5.077E+006	Adeq Precisio	49.446
-2 Log Likelihc	242.49	BIC	250.81
		AICc	250.49

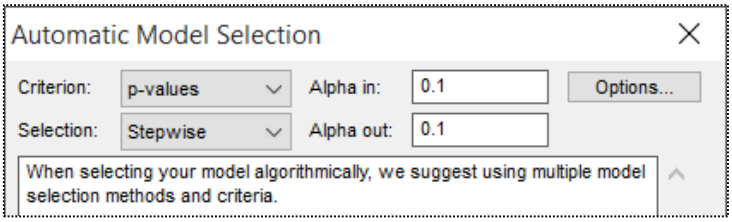
Bookmarks
Top
ANOVA
R-Squared
Coefficients
Equations
Pop-Out View

Factor	Coefficient Estimate	df	Standard Error	95% CI Low	95% CI High	VIF
Intercept	65499.19	1	132.91	65212.06	65786.31	
B-GNP	6065.90	1	274.22	5473.48	6658.32	1.58
C-Unemploym	-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 ***Model***, re-set **Process Order** to **Linear** and go to **Autoselect…** to try the last model **Selection** option offered by Design-Expert software: **Stepwise** (be sure to also choose **p-value** as your criterion). Note, AIC and BIC are newer model criterion that we will use in future tutorials.



Specifying stepwise regression

As you might infer from seeing both Alpha in and Alpha out now displayed, stepwise algorithms involve 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, look back at how factor B's p-value changed depending on which other factors were chosen with it for modeling. To see what happens with this forward-selection method, press **Start**, **Accept**, and then ***ANOVA*** again. Results depend on what you do with Alpha in and Alpha out – both which default back to 0.1000. With the defaults, the same model is selected by this method as the backwards selection chose.

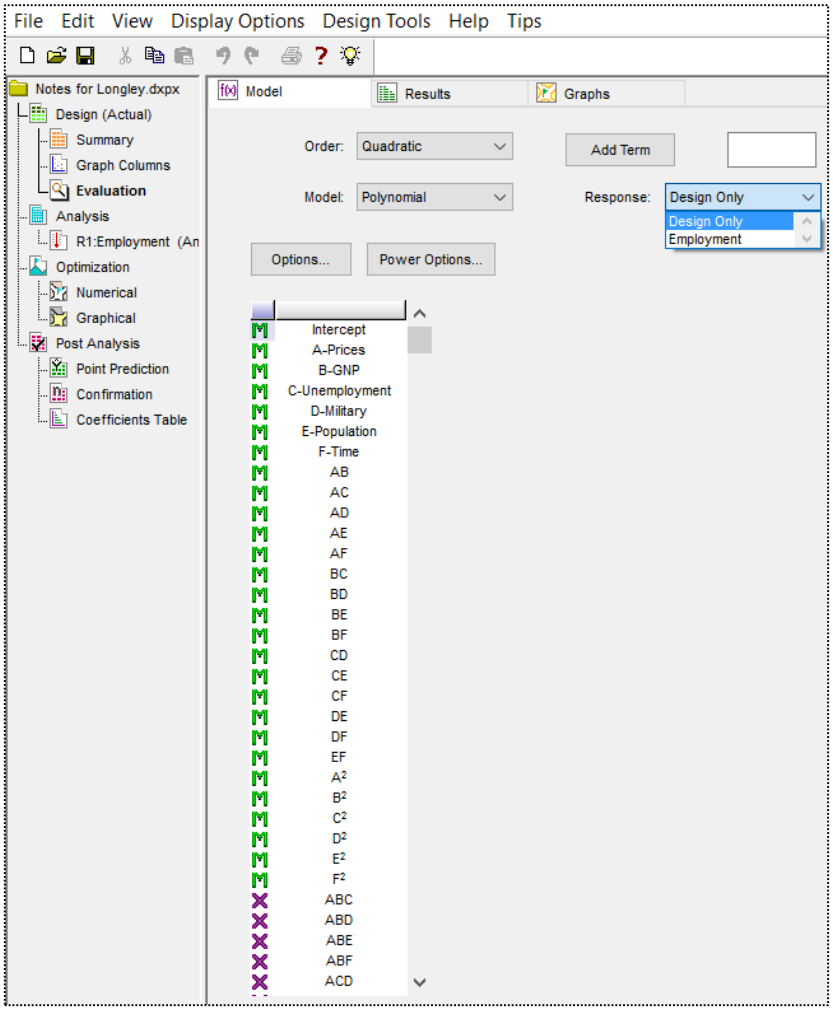
As you see in the message displayed for both forward and stepwise (in essence an enhancement of forward) approaches, we favor the backward approach if you decide to make use of an automated selection method. Ideally, an analyst is also a subject-matter expert, or such a person is readily accessible. Then they could do model reduction via the manual method filtered not only by the statistics, but also by simple common sense from someone with profound system knowledge.

This concludes part 1 of our Longley data-set exploration. In Part 2 we mine deeper into Design-Expert to see interesting residual analysis aspects within Diagnostics, and we also see what can be gleaned from its sophisticated tools within Design, Evaluation.

Part 2 – Advanced Topics

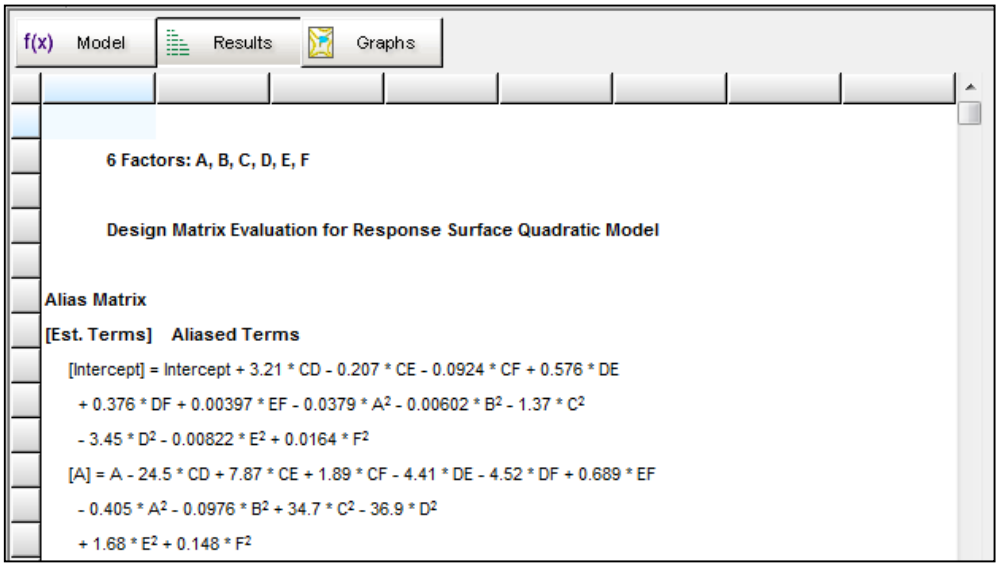
Design Evaluation

If you still have the Longley data active in Design-Expert software from Part 1 of this tutorial, continue on. If you exited the program, re-start it and use **Open Design** to open your data file (**Longley.dxp**). Under the **Design** branch of the program, click **Evaluation**. The software brings up a quadratic polynomial model by default, but, as you will see, the order must be downgraded to linear (we will get to the reason momentarily). The screen shot shows the Response field set at "Design Only" as opposed to the Employment response. In other words, it will evaluate the entire matrix of factors, regardless whether response data are present. The other option (response by response) comes in handy when experimenters end up with missing data, thus degrading the "designed-for" model.



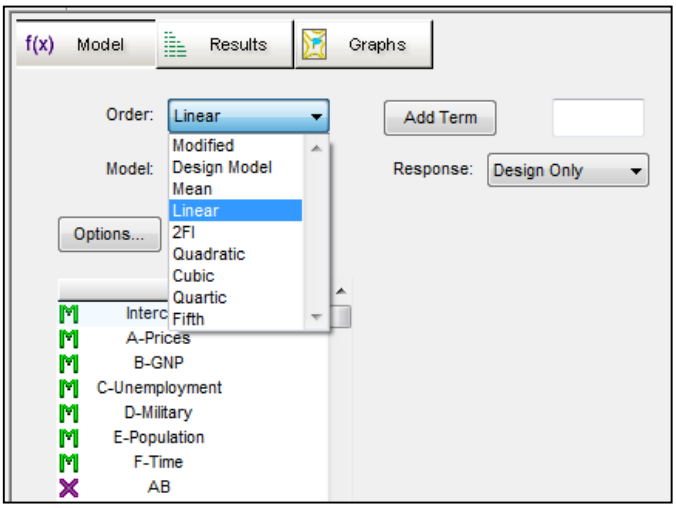
Design evaluation (design only)

Press the ***Results*** button.



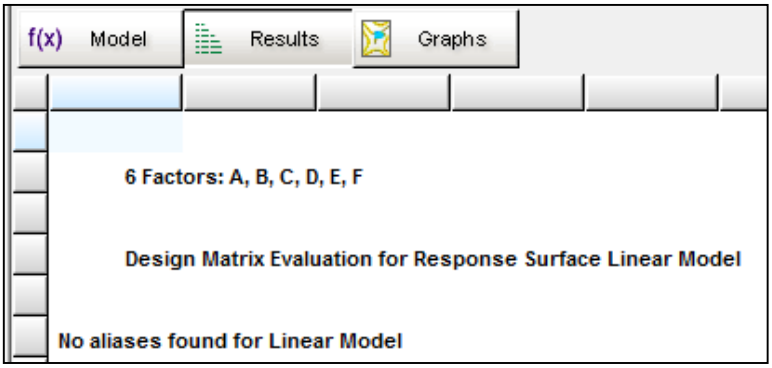
Results of evaluation for quadratic polynomial

This model is badly aliased. For example, the effect of A is confounded with -24.5 CD, etc. Go back to ***Model*** and reduce the **Order** to **Linear**.



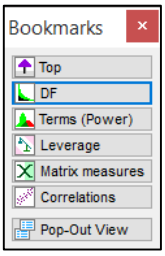
Re-setting order to linear

Press ***Results*** again and note "No aliases found…" Much better!



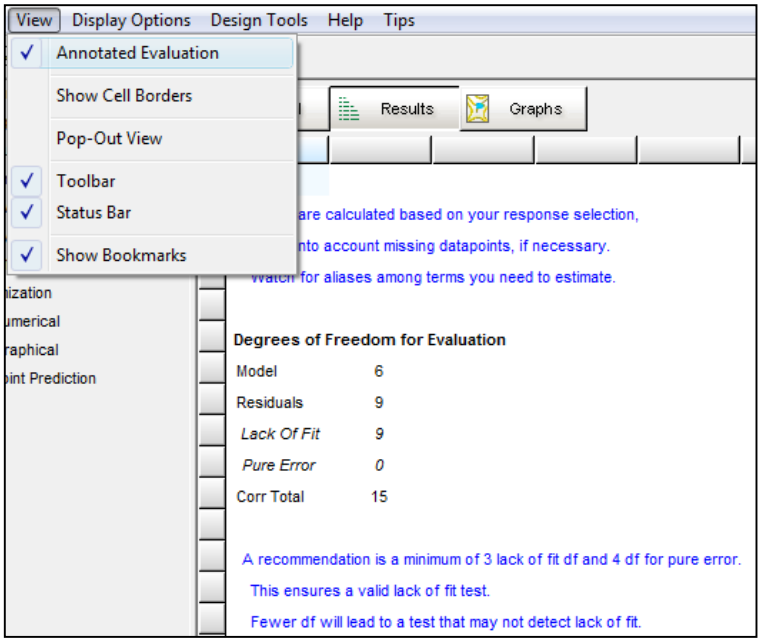
Results of evaluation for linear model

On **Bookmarks** click the **DF** option to bring up the accounting for degrees of freedom.



Bookmarking to evaluate degrees of freedom (DF)

Looking over the annotations provided by the software (activated via **View**, **Annotated Evaluation**), notice this design flunks the recommendation for pure error df. Of course this really is not a designed experiment, but rather historical data collected at happenstance.



Annotations for degrees of freedom

Study the next section of the evaluation by Design-Expert. Do any of the statistics pass the tests suggested for a good design? No!

				Power at 5 % alpha level to detect signal/noise ratios of		
Term	StdErr[1]	VIF	Ri-Squared	0.5 Std. Dev.	1 Std. Dev.	2 Std. Dev.
A	4.72	135.53	0.9926	5.0 %	5.1 %	5.4 %
B	17.61	1788.51	0.9994	5.0 %	5.0 %	5.0 %
C	2.35	33.62	0.9703	5.1 %	5.4 %	6.7 %
D	0.75	3.59	0.7214	6.0 %	9.2 %	22.2 %
E	8.33	399.15	0.9975	5.0 %	5.0 %	5.1 %
F	11.21	758.98	0.9987	5.0 %	5.0 %	5.1 %

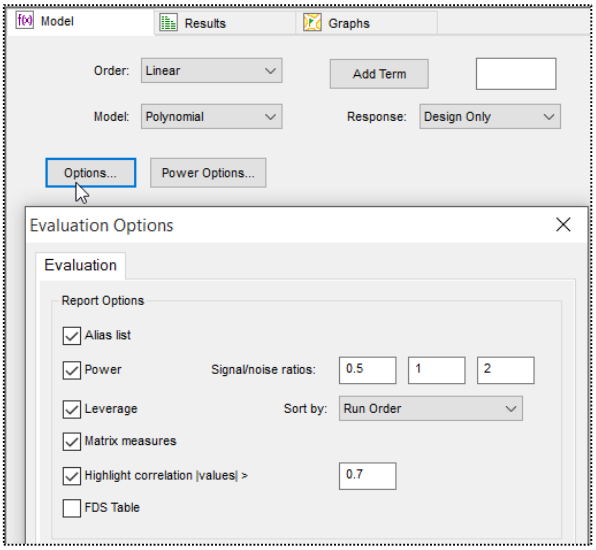
[1] Basis Std. Dev. = 1.0

Power is evaluated over the -1 to +1 coded factor space.

Details on model terms, including power

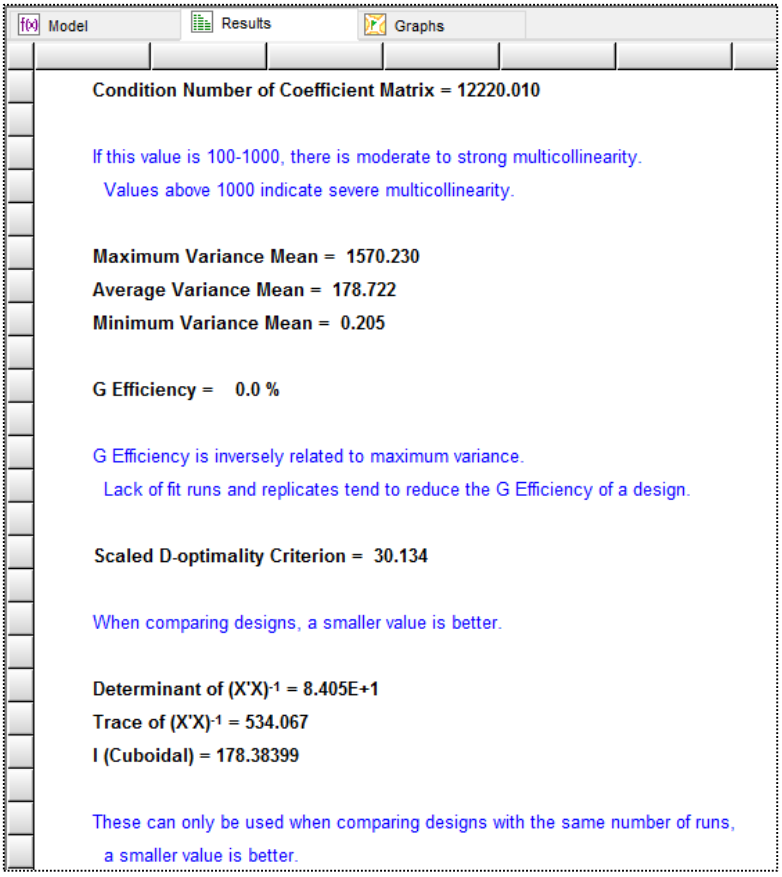
Scroll down or bookmark to the leverage report. These statistics come out surprisingly good – none exceeds twice the average.

More statistics are available by going back to ***Model***, selecting **Options,** and turning on (checkmarks) **Matrix Measure** and **Highlight Correlation Values**.



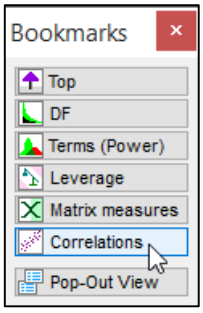
Turning on more options for report

Click **OK** and view the ***Results***. On **Bookmarks** choose **Matrix** to see new statistics.



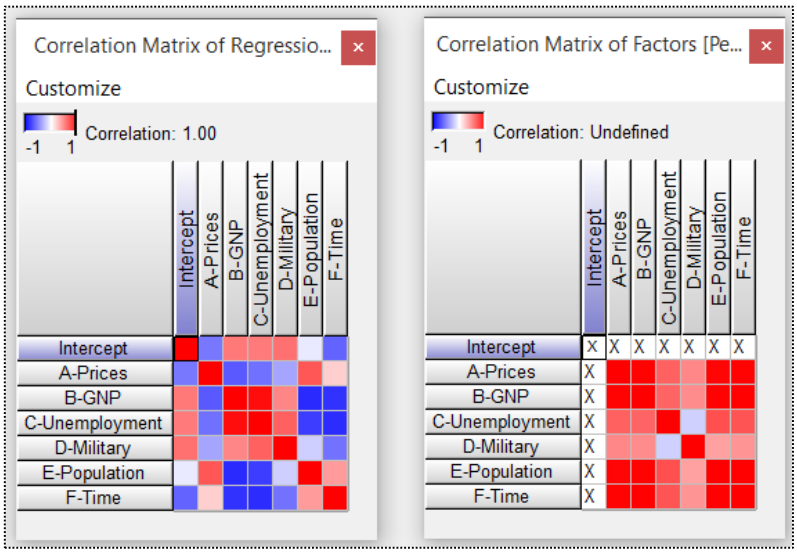
Matrix measures for design evaluation

Notice the condition number (12,220) far exceeds the level considered to represent severe multicollinearity for a design matrix (1000 or fewer). Viewing specific correlations by clicking on the **Correlations** bookmark reveals why.



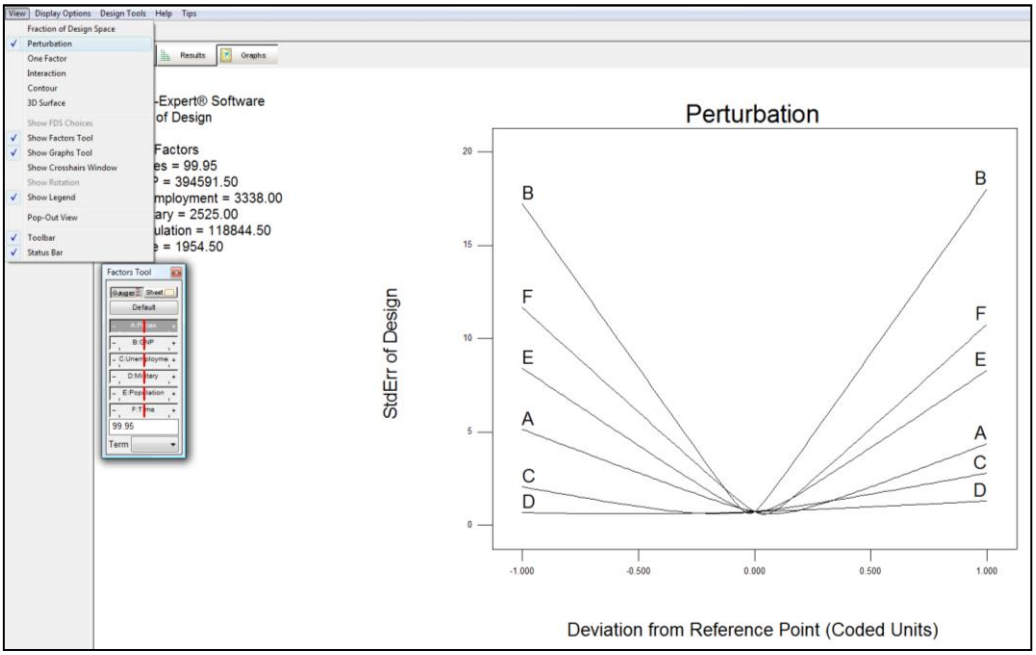
Bookmarking to see Correlation Matrices

You are presented with two Correlation matrices in new windows. They are shown below. The Correlation Matrix of Regression Coefficients shows how the factors are correlated with one another on a scale of -1 (perfect negative correlation) to +1 (perfect positive correlation). These correlations are shown in a grid form and color coded to see at a glance where there may be issues. Remember, we don't want our factors to be correlated. We want independent estimates of how they affect the responses. Therefore, white boxes on the grid are good. By just glancing at this grid, you can see there are a lot of correlations among factors (dark blue and red colors). It's no wonder Longley picked this data set to test regression software! The matrix grid on the right shows Pearson's correlation coefficients. It's just a different way of calculating correlation. You can learn more about that by clicking on the tips (light bulb) icon.



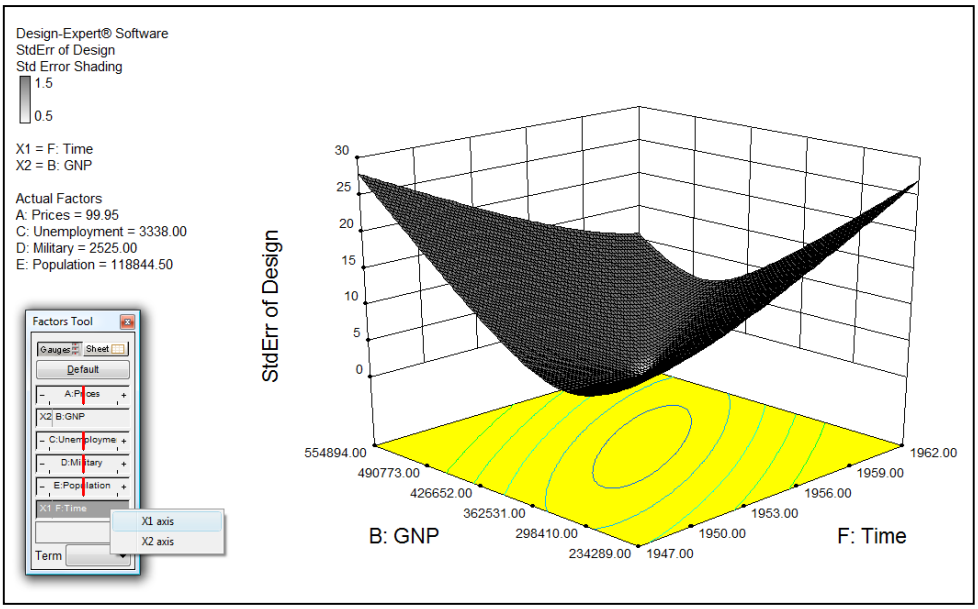
Correlation matrices

Now, just for fun, press the ***Graphs*** button and select **View**, **Perturbation** (or press this option on the floating Graphs Tool).



Perturbation plot for standard error

Notice factors B and F exhibit the most dramatic tracks for standard error. On the floating **Graphs Tool** select **3D Surface.** On the **Factors Tool,** right-click factor **F:Time** and change it to **X1 axis**.



3D view of standard error for factors B and F

There's no sense doing anything more. By now it's clear that this 'design' fails all the tests for a good experiment, but that's generally the nature of the beast for happenstance data.