Hands on Workshop
Do You Know When Your Data Is Lying To You? The HOW of Regression Analysis with Quantitative And Qualitative Variables

Steven C. Myers, Ph.D.
Department of Economics, College of Business Administration
The University of Akron

ABSTRACT

A qualitative indicator such as a binary variable, D, may describe a population difference, such as 'male' and 'female,' or 'before' and 'after' some event. This workshop will walk you through the regression analysis of whether an outcome variable, Y, is influenced by a qualitative binary event, D. Using only 14 years of data on Y, you will learn that what seems a very simple actually takes 8 different regressions and many Wald statistical tests to reveal that the best conclusion requires complexity of model specification and a strategy of statistical inference. The take away is no matter how sophisticated the technique and how good the data, there is no substitute for thinking your way through a problem. Blindly following technique alone is a bad practice that leads you to make huge mistakes.

You will learn the value of articulating a problem, preparing data, exploring the data and the importance of the data generating process. You will experience interpreting the results and inferring the validity of the results and drawing a conclusion. You will learn a rather comprehensive set of techniques in a very simple example. Most importantly, you will learn that the roadmap for similar problem solutions is not guided by the techniques as much as the critical application of human thought.

BUSINESS PROBLEM

An entity, such as a business or government, has a metric of great interest to their operation. This success metric, denoted simply as Y, is tracked and recorded once during each period of time. Y could be output produced, number of people served, number of items sold, state domestic product, US gross domestic product, or any other success metric reasonably related to the entity's operation. This is the environment of our business problem.

The entity has implemented a policy change 7 periods ago and now wonders whether it "paid off" in higher values of the success metrics. To test this "pay off" we have 14 periods of data, seven years before and seven years after the policy change. These time periods, for example, can be months, quarters or years.

ARTICULATION OF THE PROBLEM

The first step in any applied analytic approach is to articulate the problem needing to be solved. Only by starting out with a clear statement of the problem can one hope to do all that is possible to solve the problem. The problem to be solved is this: Did a policy or procedure implemented during the seventh period cause a change in the outcome metric of interest? From this articulation various hypothesis will emerge to statistically test the problem.

DATA ACQUISITION

A data series of 14 periods of data are collected on a metric of interest. The measure of this variable is shown below and is called Y. We have 7 periods of data before and 7 periods of data after the change in the policy or procedure. A means procedure is shown to verify the data is entered correctly.

```
Data Y;
    input Y @@;
    datalines;
    12.35  13.71  16.00  17.94  20.76  21.11  24.63
    27.56  32.88  35.16  39.26  44.28  47.27  51.55;
    run;
title 'The correct mean of y is 28.890 and the Std Dev is 12.9719';
proc means data=y mean std maxdec=4;
  run;
title;
```

PREPARING THE DATA FOR ANALYSIS

DATA CLEANING

Preparation of the data should involve much work exploring and cleaning the data, but in this illustration we accept the data as high quality and accurate. In practice one should never skip this step. Doing so will lead to peril.

DATA TRANSFORMATIONS FOR ANALYSIS

Our business problem suggests a course of action since the hypothesis The first question to explore is what is the pattern of the data and how does it trend over time? Are there any obvious characteristics to that data and its trend? Are their significant perturbations in the data or is it fairly smooth? These and other questions will be addressed below, but it become obvious that some variables need to be created or transformed as shown in the next code block.

```
Data      trdata;
    /* Problem is to explain the trend in variable Y.*/
    /*    H0: An intervention that begins in T=8 has no effect on the trend line.*/
    /*    H1: An intervention at T=8 changes the trend line.*/
    /* Alternative problem: */
    /*    The actual equation is simply nonlinear in variables such as y = T TSQ.*/

    set Y;
    T=_N_;                    /* 1. create time variable. */
    TSQ = T*T;                /* 2. and time-squared value. */
    D=0; if T>=8 then D=1;    /* 3. Create binary variable for the intervention. */
    DT = D*T;                 /* 4. create interaction of D and T. */
    run;
```

Four variables are created to aid the analysis, T which is the linear time measure, TSQ which will allow for a quadratic bend, D which is the intervention or treatment variable and DT which allows that the influence of time, T, may vary with D.

MODEL SELECTION AND SPECIFICATION

We must resolve how we will solve the problem. Much of this design must stem from the thought exercise of how to solve the problem and some of this designed is informed as one is in process of looking for the truth. In this paper we will see that journey influenced by data that at once appears to support a structural break with little curvature and a quadratic form with little evidence of a structural break. This paper is designed to walk one through that exact landscape with applicable and general lessons of structural integrity.

Some analytic approaches start with the idea that the truth is in the data and what is revealed is true, but the premise of this paper is that data can and does “lie” or greatly mislead. As pointed out by Nobel Lauriate Ronald Coase: “If you torture the data long enough, it will confess.” Of course the tormentor will stop the beating when their biases are confirmed. We have all seen the drama of a torturer beating the innocent person because the (wrongly) expected truth fails to emerge. If all you want to do is to confirm your expected solution, then why do the data work at all? And if you are willing to let the data speak for itself without much or any human thought intervention then why think at all. The lesson of this paper takes on both fallacies.

WE WILL SELECT BOTH QUANTATIVE AND QUALITATIVE VARIABLES FOR THIS EXERCISE

The values Y and T and TSQ are quantitative. The binary class variable, D, is Qualitative with value of zero for the first seven periods (called ‘before’) and a value of 1 for the last seven periods (called ‘after’). While D is certainly qualitative, because of the 0,1 coding it has desirable quantitative properties, chiefly among these is the mean is the portion of observations ‘after’ when compared to the total sample. As important is the value of reducing a model specification into conditional expectations of E(Y|D=0) and E(Y|D=1), that is what do we expect the values of our metric to be ‘before’ as opposed to ‘after’ the intervention. Helping in that will be the interaction variable, DT, allowing for the effect of the intervention, D, to vary by time. TSQ is a similar interaction variable allowing for the effect of time

to also vary with time with a direct application to forming a quadratic in time to explain the path of our metric, Y. Because all models will be linear in parameters, non-parametric regression is used as a check on that assumption and to offer model free insight to our exploration.

I. A VISUAL EXPLORATORY APPROACH USING PROC SGPLOT

Regression is a statistical procedure that can be a highly visual one in the few parameter case. This is especially true in this paper since we are using a small dataset and few explanatory variables. It is a small data set with much meaning. We suspect that an intervention may have had an effect and are interested in examining whether that effect has meaning or is just an artifact of the particular sample. In part one, we will take a visual approach and in part two we will apply a statistical inference approach to solve our problem whether the intervention has a statistical effect.

Make sure you include this code before continuing:

```
ods graphics on / noborder width=5in;
%let xref = %str(xaxis values=(1 to 14 by 1); refline 7.5 /
axis=x label="<-- Policy change" labelloc=inside labelpos=min ;);
```

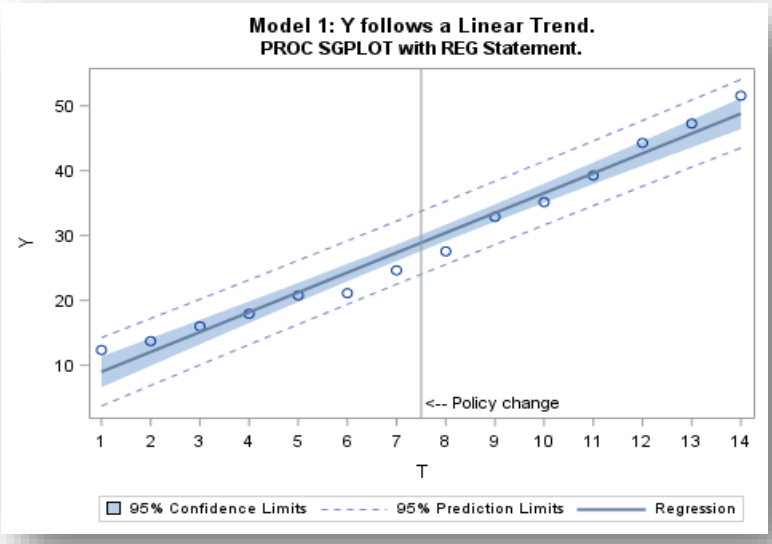
The code assures graphics are on and the token &xref is defined as a common setting for all X axis in the following graphs.

FIRST LOOK: IS IT LINEAR?

So what does the trend look like? Is it linear before and after the intervention? Figure 1 suggests that the regression line is possibly linear, but that at least five residuals are large enough to be outside the 95% confidence interval. The points from the scatter also look like they may be quadratic.

```
title1 'Model 1: Y follows a Linear Trend.';
title2 'PROC SGPLOT with REG Statement.';
PROC SGPLOT data=trdata ;
    reg x=T y=Y / CLM CLI ;
    &xref;
    run;
```

Figure 1: Explore the trend of the metric Y



The pattern of residuals (vertical differences between the actual observations and the line plotted in Figure 1) suggest a “U” or “V” pattern as at low and high values of T the residuals are positive while in the middle of the series, the residuals are negative.

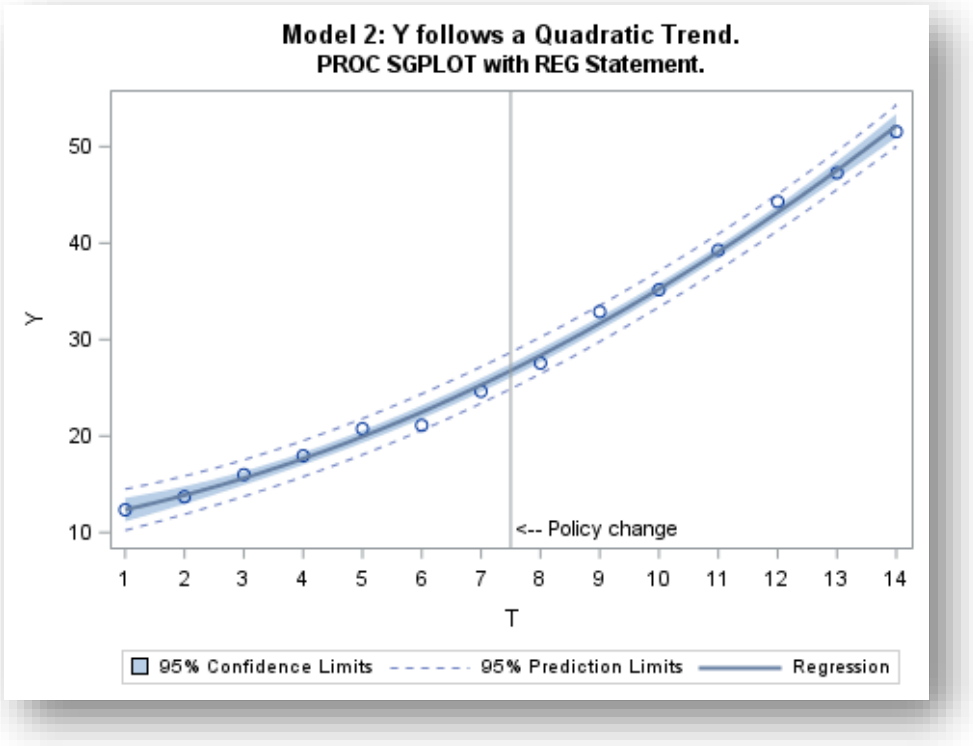
FRIST LOOK: IS IT QUADRATIC?

Change the last code to add degree=2 as an option to the reg command. The default is degree=1 and plots the regression of Y on X, while a change to degree=2 plots Y as a quadratic in T. The code is:

```
title1 'Model 2: Y follows a Quadratic Trend.';
title2 'PROC SGPLOT with REG Statement.';
PROC SGPLOT data=trdata ;
    reg x=T y=Y / degree=2 CLM CLI ;
    &xref;
    run;
```

The result of this change is shown in Figure 2 and rather clearly suggests that the trend in Y may be quadratic and have little to do with the intervention, D.

Figure 2: Y as a quadratic function of T



FIRST LOOK: A NON-PARAMETRIC LOOK AT THE TREND.

What if we do not impose either a linear or quadratic form on the trend in Y, but rather use a non-parametric technique invoked by the plot command loess to trace out the most obvious pattern. Loess calculates a large number of regressions by using a neighborhood sample of the points at each observation to calculate a weighted linear or cubic regression of degree 1 or 2 on the neighborhood where the weights are greatest for the nearest neighbors.[1]

To examine the exact path we again use PROC SGPLOT to trace out the points, but this time with a LOESS statement. The LOESS is a non-parametric regression called local regression that will trace out the scatter points by

[1] A high-level look at Loess is found at https://en.wikipedia.org/wiki/Local_regression. See recommended readings for more.

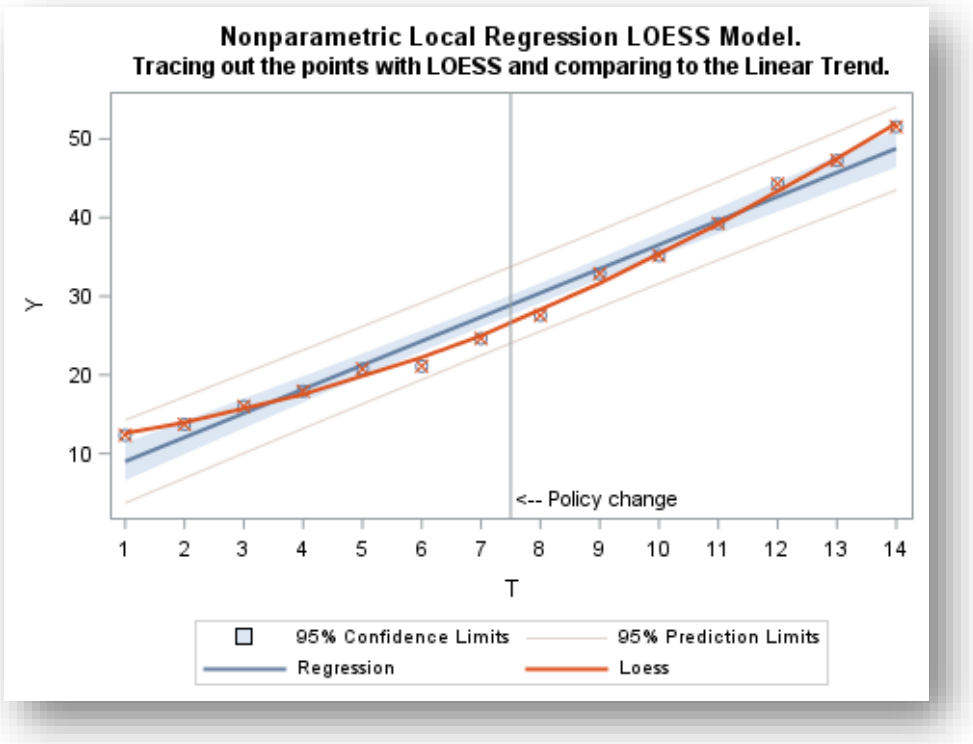
running a regression only among the nearest neighbors to each point. That way the many regressions run are not sensitive to points far away from the point of interest. In the loess command we can choose cubic or linear and each with a degree of 1 or two. You can run all 4 combinations and see that in this case the lines look the same for each combination.

On the next graph, Figure 3, the linear reg plot (changing the code from last time to degree=1 and cause it to be somewhat transparent) is left for reference. The loess plot is listed next and hence is drawn over the reg plot. The code for this investigation is:

```
title1 'Nonparametric Local Regression LOESS Model.';
title2 'Tracing out the points with LOESS and comparing to the Linear Trend.';
PROC SGPLOT data=trData;
    reg x=T y=Y / degree=1 CLM CLI CLMTRANSPARENCY=.5;
    loess x=T y=Y /interpolation=linear degree=2;
    &xref;
   run;
```

Figure 3 shows the loess plot seems to trace out a fairly definite quadratic-looking relationship and in this case the scatter seems to show points closer to the quadratic than the linear. This is the first lie in this data, but we have to finish the analysis to know that!

Figure 3: What does a non-parametric trend reveal?

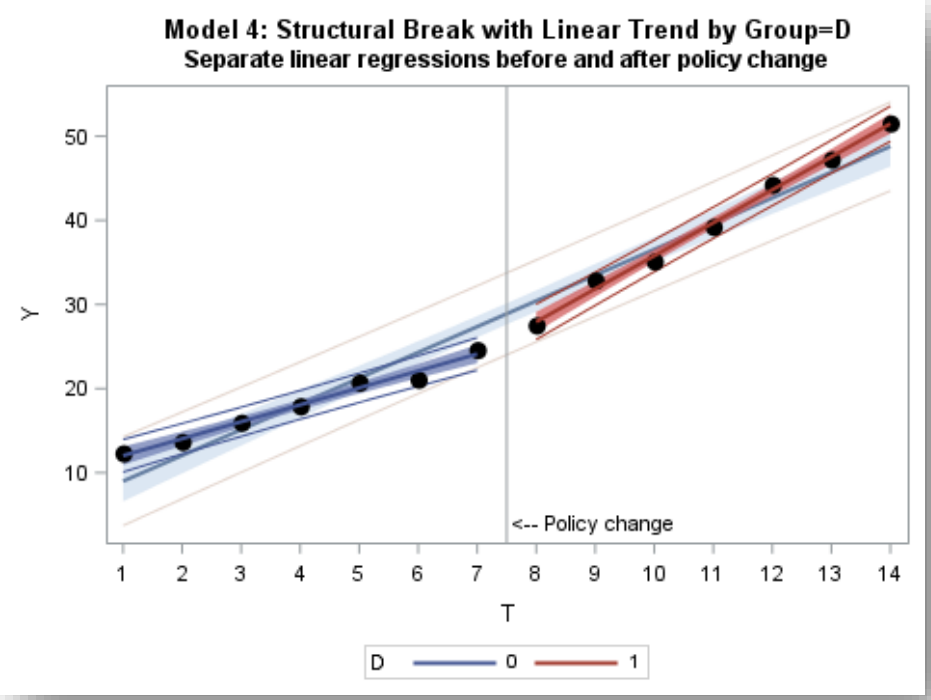


HOW DOES THE TREND LOOK SEPARATELY BY GROUP?

Since our interest is before and after the midpoint of the data, that is did the intervention, D, effect the trend, we turn again to linear regression and apply it separately by group, before (D=0) and after (D=1). For this visual we modify the code to overlay the before and after regressions on top of the full-sample linear regression. The group=D option on the second reg plot instructs the plot to plot first the observations with D=0 and then the last observations with D=1. The code now looks like:

```
title1 'Model 4: Structural Break with Linear Trend by Group=D';
title2 'Separate linear regressions before and after policy change';
PROC SGPLOT data=trdata;
    reg x=T y=Y / CLM CLI CLMTRANSPARENCY=.5;
    reg x=T y=Y / CLM CLI CLMTRANSPARENCY=.25 group=D
                markerattrs=(symbol=circlefilled color=black size=10px);
    &xref;
run;
```

Figure 4: What does the regression look like before and after the intervention?



It seems obvious from Figure 4 that there is a different trend line before (blue line) and after (red line). This suggests that there is strong evidence that the trend is consistent with a structural break and that D has a strong influence.

DOES THE TREND APPEAR QUADRATIC, BEFORE AND AFTER?

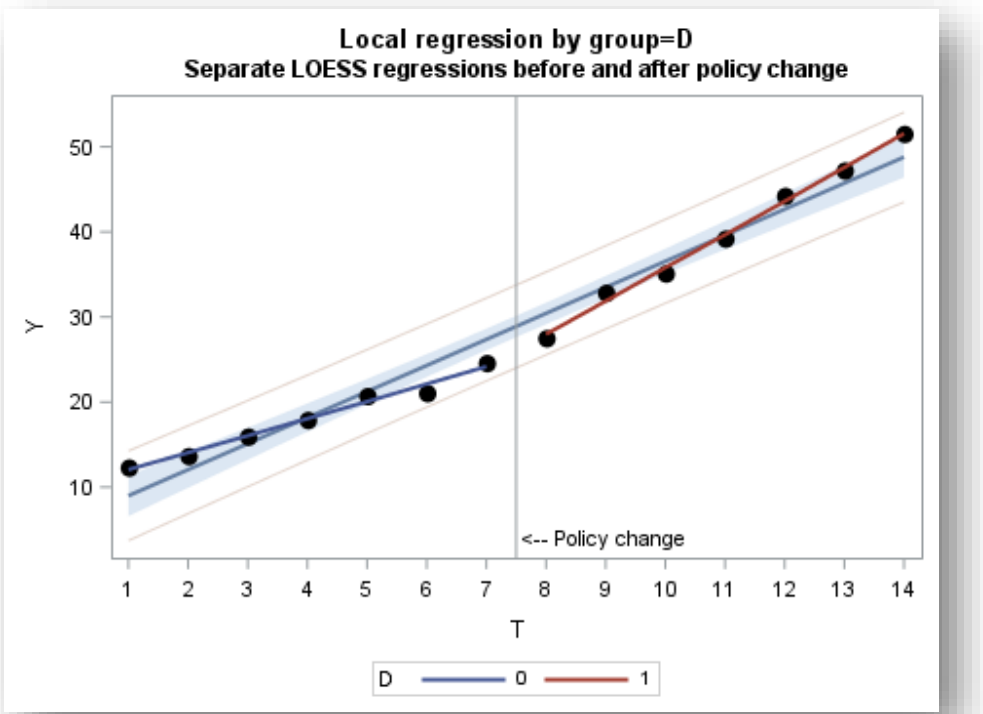
The local regression in Figure 3 traced out a strong looking quadratic relationship and that was said to be the first lie in the data. To see why we can visualize the local regression of the last 7 periods separately from the first 7 periods to get a feel if the separate periods (groups) are more linear such as in Figure 4 or more likely due to a natural quadratic-like relationship as in Figure 2. The following code will overlay the full-sample linear trend with the local regressions in the before and after time-slice of the full sample:

```
title1 'Local regression by group=D';
title2 'Separate LOESS regressions before and after policy change';
PROC SGPLOT data=trData;
    reg x=T y=Y / CLM CLI CLMTRANSPARENCY=.5;
    loess x=T y=Y / group=D interpolation=linear degree=1
                    markerattrs=(symbol=circlefilled color=black size=10px)
                    CLMTRANSPARENCY=.25;
    &xref;
    run;
```

Our result in

Figure 5 is that the quadratic relationship visualized over the entire sample, seemingly vanishes when each group is analyzed separately. But of course we have only beat the data a little, so far, and this confession seems pretty weak. Like much analysis we have dug into the data, but have not yet confirmed our assertions using statistical inference. We turn to that next.

Figure 5: What do the separate trends by D look like?



SUMMARY OF THE VISUAL ANALYSIS

To summarize this section, we wondered whether there was a break in the trend of Y due to an intervention indicated by D. Because of large residuals in Figure 1 and the strong looking fit in Figure 2 we could have thought that the trend was natural following a quadratic pattern. However, when regressing separately by

group (in Figure 4) we see the possibility of completely separate linear trends before and after the intervention.

Figure 5 suggests that the separate trends are actually linear and not just artifacts of an overall quadratic trend since the quadratic pattern fails to visually emerge when separated by group. Our data so far seems to suggest that the intervention did have an effect. The effect illustrated is called a structural break. Now to quantify that.

II. AN APPLIED ECONOMETRIC/STATISTICAL APPROACH USING PROC REG

HASTY REGRESSION

I use the term hasty regression to describe when a researcher quickly runs a regression with minimal or no articulation of the problem, virtually no data cleaning and preparation, a lack of understanding of the data generating process (DGP) and no critical thought on or empirical investigation into the appropriate model specification. Who would do this you might ask? Nearly all of us. The temptation is just too great. We become anxious about what may be learned, but we can never unseen that initial and hasty regression which can bias our approach whether the binary variable is significant or not.

Hasty regression is running a model pulled out of the air without regard to the theoretical problem under study and without rigor in the model specification stage, all the while giving no regard to whether the data is clean and ready for analysis.[2]

Mistakes of commission and omission, made in the articulation of the problem and the specification and selection of the model may prove fatal to your analysis, but the results do not identify as fatal, indeed they look the same as better well-conceived results. The poorly conceived problem and the undoubtedly misspecified model can lead us to what can only be described as a failure.

Researchers legitimately use dummy variables to ascertain whether a point or many such points are significant deviations from the overall trend in a linear regression model. Our dummy variable is D and defined as D=0 for before the intervention and D=1 after the intervention. And we ask our first statistical inference question: Is D significant in a linear regression of Y on T? To test that end we run two models.

```
ods graphics on;
proc       reg data=work.trdata;
           model_1: model y = t;
           model_3: model y = t d;
           title2 'Full Sample, T=1,..., 14';
           run;
```

Model_1 is illustrated in

Figure 6 and Table 1 while the results of model_3 are given in Figure 7 and Table 2. Model_3 is our suggested hasty regression based on “just throw in a dummy variable to see if it has an effect.” In good applied econometrics this same model may be a first step, but not a last step, if critical human thought is to be applied.

[2] The data in this paper are assumed clean and if they were not then a whole another layer of concern emerges. Fortunately, that worry is outside the purpose of this paper.

IS IT LINEAR?

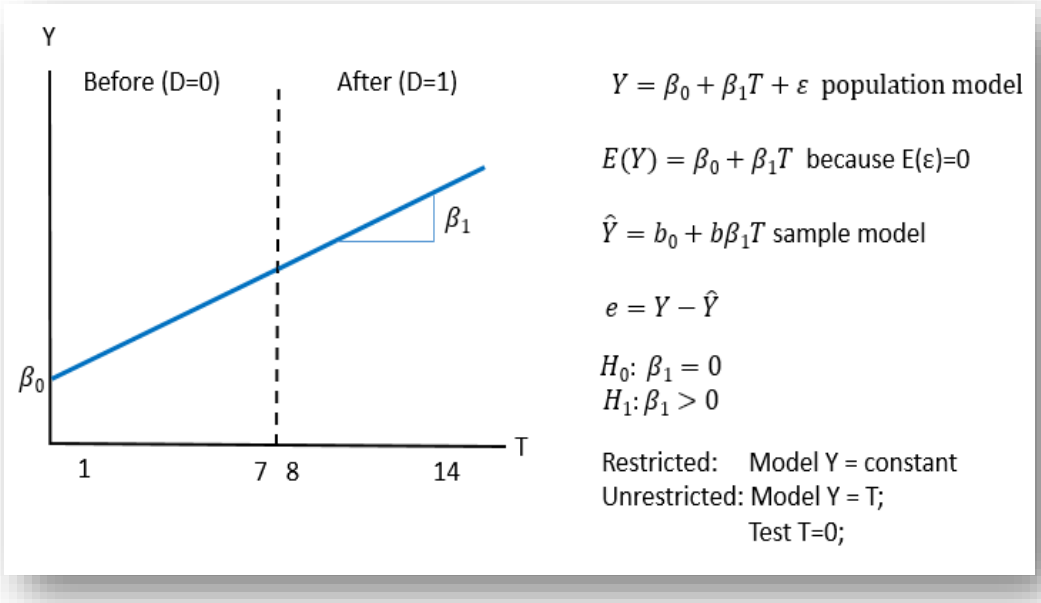
Table 1 shows the parameters estimated for Model_1. Using labels for models and tests in a PROC REG specifications helps avoid confusion when sifting through the copious output. Model_1 has an adjusted R-square of 0.972 and a RMSE (root mean Squared Error) of 2.16 showing that this seems to be a pretty good representation of the trend of Y. One might be tempted to say that with such low p-values that this 'proves' that the regression is linear,

Table 1: Model 1 Linear results

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.93242	1.21836	4.87	0.0004
T	1	3.06101	0.14309	21.39	<.0001

but this proves nothing. The t-values and the F statistic for the model show that the variable T is statistically significance based on a null hypothesis that Y is equal to random error. This is hardly confirmation where the 'bar' is T is literally better than nothing. **Figure 6 shows the restricted and unrestricted specifications that lead to this hypothesis.** If you were to think that these results prove anything, then that would be the second lie in this data.

Figure 6: The linear model without regard to the intervention



The Results of our Hasty Regression: Does D matter?

Model_3 can tell us something about D. The hypothesis test of D is illustrated in Figure 7 and is whether the unrestricted model (model Y = T D;) offers significantly more explanatory power than the restricted model (model Y = T);

Figure 7: Test of an intercept difference holding slopes the same (Model 3)

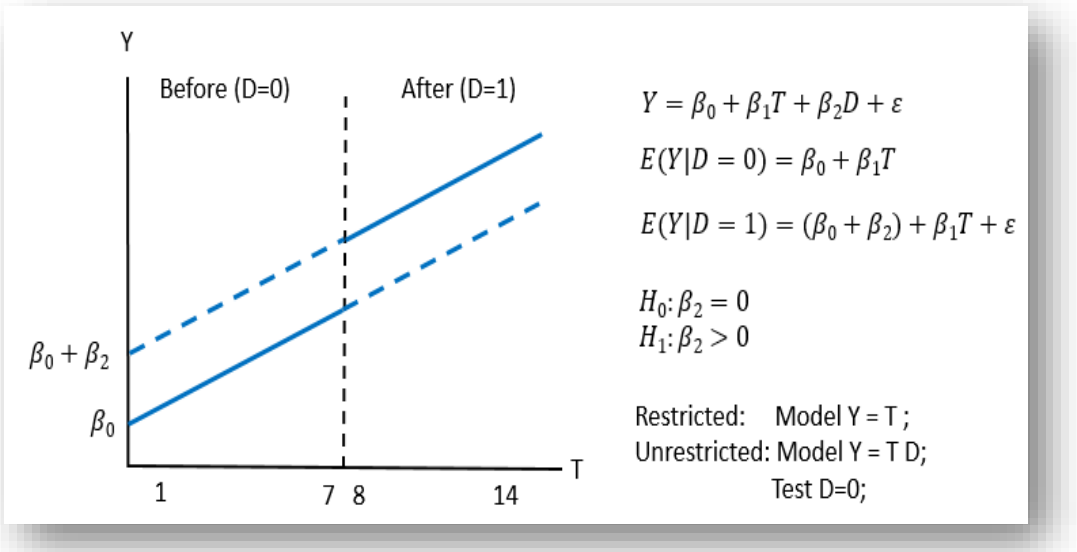


Table 2 shows the parameters estimates of Model_3 with an adjusted R-square of 0.970 (worse than the previous model) and a RMSE of 2.24 (worse than the previous model). Further the parameter estimate of D is found to be not statistically different from zero in this model. That is, the value of 0.85339 cannot be differentiated from zero.

Table 2: Model 3 linear and dummy variable.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.19500	1.46741	4.22	0.0014
T	1	2.96911	0.29953	9.91	<.0001
D	1	0.85339	2.41493	0.35	0.7305

So at this point many researchers may conclude that D has no effect on Y and stop their questioning. This is hasty regression at its finest (read sarcasm).

By saying that D is unimportant in explaining Y as a general statement then this is the next lie from our data. For as we will show, D is far more important than this would imply. Our hasty researcher is on to other tasks, while our better researcher is just getting started.

WHY ISN'T THE HASTY REGRESSION THE LAST STEP?

The statistician's dilemma is he or she can never know truth. What is known is an estimate based on an estimator (formula) and given a specific sample data set. The statistician wants to test a hypothesis on an unknown population estimator, by using a known sample estimate. Type I errors occur when we reject the null hypothesis but in reality the

hull hypothesis is false while type II errors have us failing to reject the null hypothesis when indeed the null hypothesis is true.

We can never prove, we can only falsify on the bases of this estimator and this dataset. The best we can do is use the most powerful test based on our best estimators from our best specified model. In a controlled experiment we can replicate the experiment and test over and over, but in economics and in the testing of observational events that have historically occurred, there is often if not always no replication possible, the history we have is all we have.

The statistical community has called into disuse the declaration of models being significant or not significant based on a crossed threshold by our test. The tables below do show a breakdown by threshold, but in this case this is for convenience and the exact p-values in this example are universally extremely small (very close to 0.00) or quite large (over 0.30).

DEALING WITH INDECISION

Evidence! I need more evidence! I am a fan of mystery and thrillers and enjoy the twists and turns of a problem where typically the protagonist is up to their ears in a mystery and every step frustrates more, because it seems that they will never get to the end. But of course they do and we will as well. In the words of Fenton Hardy, the fictional father of the Hardy Boys, a series I read religiously a half century ago, we must “leave no stone unturned.”

So what to do? Obviously time to overturn some stones. That is to dig deeper, to go beyond the obvious as in the hasty regression above.

Our grand hypothesis is about D, or rather about what D is measuring. Did the intervention have an impact on the outcome variable as measured by the Y variable? Two points: First it is easy to think we are seeking truth from data, but the manner as I write this paper is that data will lie to you. We need to first think of and seek an overall explanation of why D would affect Y and then measure every effect we can think of. Falling short of that means the data may still lead us astray. Second point: tests of statistical hypothesis have three parts, (1) the null hypothesis which we try to reject, (2) the alternative hypothesis and (3) the maintained hypothesis that is not subject to test. The latter is indicative that what other variables are in a model and equally important what is not in a model affects the restricted/unrestricted tests. Consider models A and B below, each with the same test of the effect of D. While the null and alternative tests are the same, the maintained hypotheses are different, meaning we may find a failure to reject Test A and an ability to reject Test B, that is a rejection of Model_A and an “acceptance” of Model_B.

```
Model_A: Model Y = T D;
Test A: Test D=0;
Model_B: Model Y = T D DT;
Test B: Test D=0;
```

Analysts need not only to develop a modeling and estimation strategy, but must have a testing strategy to answer the overall question. In this simple case of this paper I show that it takes 8 separate tests to answer the one grand hypothesis. No single equation answers the question *until we have rejected all other competitors*.

STATISTICAL SIGNIFICANCE AND ECONOMIC SIGNIFICANCE

Two questions to further contemplate:

(1) have you found statistical significance or sufficient evidence to draw an empirical conclusion.

(2) Do the findings make economic or theoretical or financial, or common sense.

Something can be statistically significant, but have little or no economic significance. Something can be economically significant without being statistically significant.

REGRESSION MODELING: DID A STRUCTURAL CHANGE OCCUR OR IS IT A NATURAL QUADRATIC PROGRESSION?

In the visual inspection of the data in part 1, two possible models stood out, one called a structural break where D is quite important (see Figure 8) and a quadratic model where D is likely unimportant(see Figure 9).

Figure 8: Test of changing intercept and changing slope by the intervention, D.

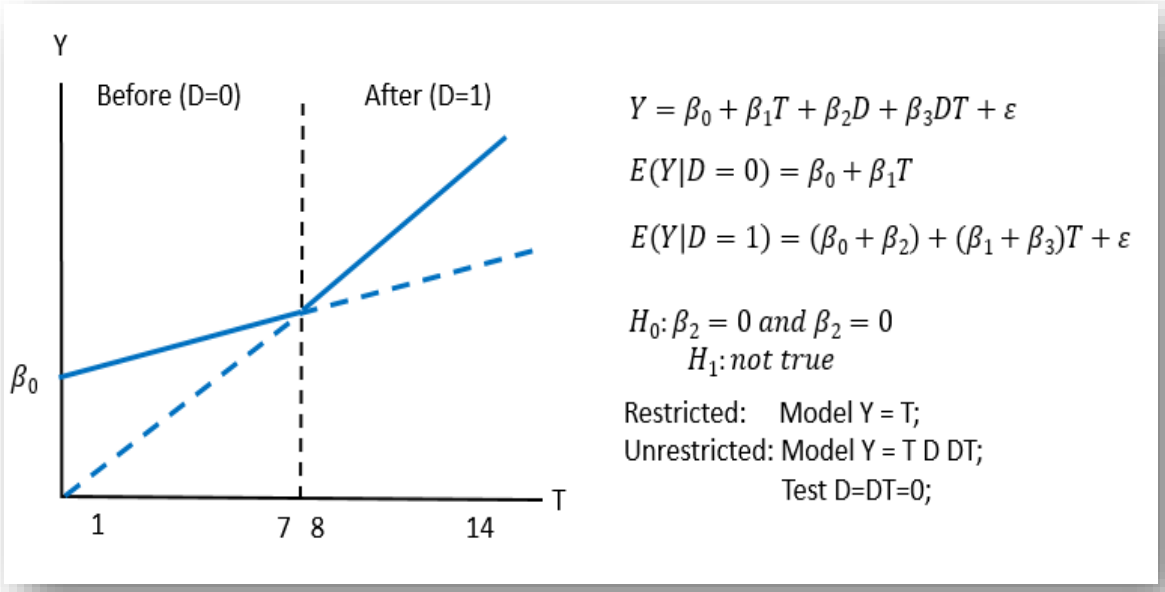
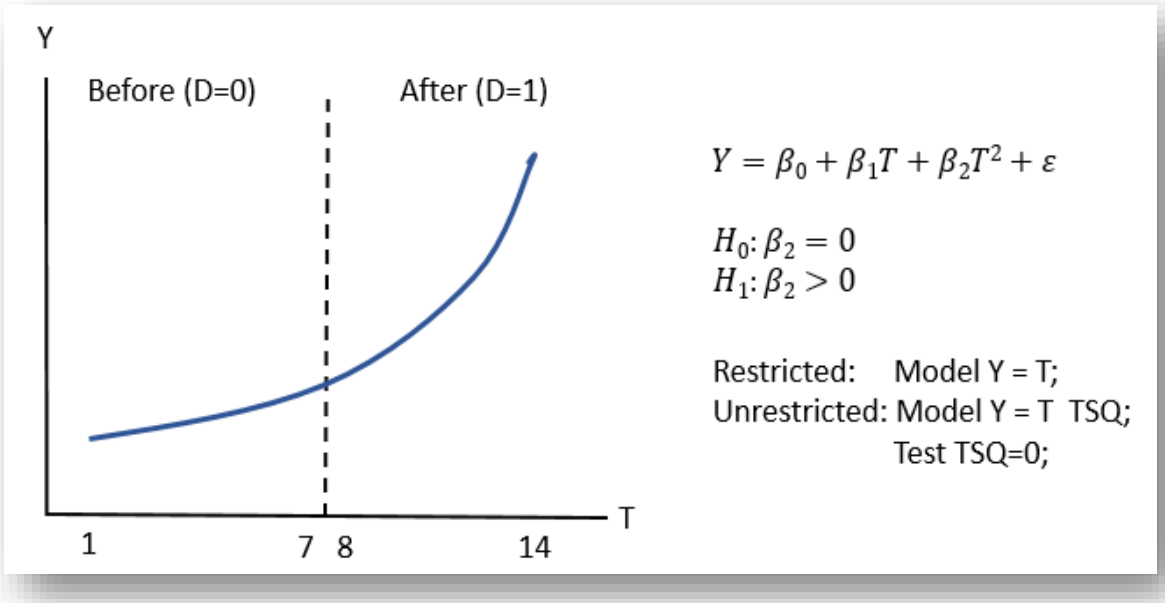


Figure 9: Alternate model, test of a quadratic form



The following code block sets up the full set of regressions and tests needed for this investigation in a linear regression format. The results of the first PROC REG are shown in Table 4 and the results of the second set or PROC REGS based on the WHERE command of D=0 and D=1 are shown in Table 5. A summary of all eight models is shown in Table 3.

```
Title1 'Statistical models';
```

```
proc       reg data=work.trdata;
           model_1: model Y = T      ;
           model_2: model Y = T TSQ  ;
           model_3: model Y = T D    ;
           model_4: model Y = T D DT ;
           title2 'Full Sample, T=1,..., 14';
           run;

proc       reg data=work.trdata;
           model_5: model Y = T      ;
           model_6: model Y = T TSQ  ;
           title2 'Partial Sample, T=1,...,7 Before';
           where D=0;;
           run;

proc       reg data=work.trdata;
           model_7: model Y = T      ;
           model_8: model Y = T TSQ  ;
           title2 'Partial Sample, T=8,..., 14 After';
           where D=1;
           run;
quit;
```

Table 3: Summary of all eight statistical models

Model	Sample	Model name	What we learn by each model	
1	Full n=14	Linear	Y is trending upward linearly.	
2	Full n=14	Quadratic	Y is better described as trending upward quadratically.	
3	Full n=14	Linear intervention	Based on the linear model, D has no apparent effect.	
4	Full n=14	Structural break	Based on the linear model, D (through D and DT) has a large effect.	The Structural break model seems better than the quadratic
5	Before n=7	Linear for D=0	Before, Linear model is trending upward at about 2 points a period.	
6	Before n=7	Quadratic for D=0	The linear model of Model 5 is not rejected in favor of the quadratic model in the before period.	Before the trend is linear not quadratic
7	Before n=7	Linear for D=1	After, Linear model is trending upward at almost 4 points a period.	
8	Before n=7	Quadratic for D=1	The linear model of Model 5 is not rejected in favor of the quadratic model in the after period.	After the trend is linear not quadratic

Table 4 shows every adjusted R-square term is very high and close to 1 and RMSE are all low, but the quadratic and structural break models stand out the most with the structural break model edging out the quadratic model with adjusted R^2 of .998 > .996 for the quadratic model. Also, the structural break model has a lower RMSE of 0.65 <0.80 for the quadratic model. Nevertheless, both models appear quite strong and while both reject the linear model, there is at this point no test of difference between the two.

Table 4: Full Sample Statistical Models

	Sample defined as years 1 to 14							
	(1)		**(2)**		**(3)**		**(4)**	
constant	5.93	***	11.12	***	6.2	***	10.01	***
	(4.87)		(14.96)		(4.22)		(18.25)	

T	3.06	***	1.12	***	2.97	***	2.01	***
	(21.39)		(4.9)		(9.91)		(16.42)	
TSQ			0.13	***				
			(8.77)					
D					0.85		-13.47	***
					(0.35)		(-9.12)	
DT							1.91	***
							(11.01)	
n	14		14		14		14	
adj R sq	0.972		0.996		0.970		0.998	
F	457.6		1713.4		212.2		1717.1	
root MSE	2.16		0.80		2.24		0.65	
DW	0.44		2.07		0.46		3.30	

note: All regressions estimated with OLS using the SAS REG procedure.
t-stats in parentheses.
*** significant at the .01 level
** significant at the .05 level
* significant at the .10 level

Table 5 shows a way to "test" between the two models. If the quadratic model holds in the full sample, it must likewise hold in the before and after samples. Models 5 and 7 show strong linear models which have higher adjusted R^2s and lower RMSE scores than the quadratic models 6 and 8. Indeed the addition of TSQ to the linear model is not significantly different from zero. Hence the linear model superiority on the full model is confirmed in each of the segments of time, before and after. Nevertheless, this still is not a best test between the two.

Table 5: Partial Sample statistical models: Before and After Regressions

	Sample year 1 to 7				Sample year 8 to 14			
	(5)		(6)		(7)		(8)	
constant	10.01	***	10.42	***	-3.45	**	-3.19	
	(17.13)		(9.87)		(-2.42)		(-0.03)	
T	2.01	***	1.74	**	3.92	***	3.87	**
	(19.04)		(2.88)		(30.76)		(2.13)	
TSQ			0.034				0.002	
			(0.46)				(0.03)	
n	7		7		7		7	
adj R sq	0.980		0.976		0.994		0.992	
F	293.5		123.7		946.1		378.5	
root MSE	0.62		0.68		0.68		0.75	

note: All regressions estimated with OLS using the SAS REG procedure.
t-stats in parentheses.
*** significant at the .01 level
** significant at the .05 level

*	significant at the .10 level

Before we go on to a test of the structural break model versus the quadratic model directly in part III, lets consider an alternative manner of selecting the models. Table 6 shows the results of using the powerful features in PROC REG to automatically choose variables for the model. The choice is based on criteria explained in the SAS Documentation for PROC REG and in some cases betrayed by the name of the selection option. The data of this paper were subjected to all 7 methods of automatic selection and the results show the structural break model in three cases, the quadratic model in 1 case and all of the variables in the data 3 times. That is, the revealed model by the "black box" of selection is totally dependent on which selection method used. Obviously, this is not a superior solution to reveal the truth of the data as we know it.

Table 6: Use of automatic model selection in PROC REG

Regression selection process	winning model
Selection=adjRsq	T D DT
Selection=Stepwise	T TSQ
Selection=Forward	T TSQ D DT
Selection=Backward	T D DT
Selection=maxR	T TSQ D DT
Selection=minR	T TSQ D DT
Selection=CP	T D DT

Another approach and highly recommended is to subject our models to a Ramsey Specification Test (called RESET).[3] This tests whether there is any additional information in higher order polynomials of the fitted Y values when added to the model being tested for misspecification. The test is whether there are any nonlinear terms not picked up in the linear expression. The SAS code follows and it is worth noting that this is available only in the SAS/ETS PROC Autoreg.

```
PROC autoreg data=trdata;
     model_1: model y = T  / reset;
     run;
PROC autoreg data=trdata;
     model_2: model y = T TSQ/ reset;
     run;
PROC autoreg data=trdata;
    model_3: model y = T D / reset;
    run;
PROC autoreg data=trdata;
    model_4: model y = T D DT/ reset ;
    run;
```

The reset option produces estimates of these higher order polynomials (power 2=squared, 3=cubic, 4=quadratic) and tests whether they are significant. The results for the first model are in Table 7 and show a rejection of the linear specification. The results for all of the models are shown in Table 8 and show that both the structural and quadratic models seem to be successful.

[3] Ramsey (1969).

Table 7: Ramsey Specification Test of Model_1

Ramsey's RESET Test		
Power	RESET	Pr > F
2	79.3697	<.0001
3	35.9250	<.0001
4	28.9662	0.0001

Table 8: Ramsey Specification Tests Results (SAS/ETS Proc Autoreg)

Model	P=2	P=3	P=4
1 Y=T Linear	Reject	Reject	Reject
2 Y=T TSQ Quadratic	Fail to reject	Fail to reject	Fail to reject
3 Y=T D Hasty Regression	Reject	Reject	Reject
4 Y=T D DT Structural break	Fail to reject	Fail to reject	Fail to reject

NON-NESTED HYPOTHESIS TESTING: QUADRATIC VERSUS STRUCTURAL BREAK MODEL

In the above visual analysis and in the regression-statistical-inference analysis lead to the conclusion that the model of a structural break and the implication that D has a large influence is a better model if only slightly better than the quadratic model and the implication that D has no effect.

The models of structural break and quadratic have not been tested directly, but mostly which specification is better than the linear. Looking beyond the structural break "win," we turn to test whether the two models are actually statistically different from each other. To do so requires a test called a non-nested hypothesis test, because the models are different and one does not fit inside (or nest within) the other. That is, there is no model we can run (either Model_2 or Model_4) that will allow linear restrictions on the parameters of one to reveal the other. If you notice every test above is a nested test.

Two tests can be run on the two alternative models, the variance encompassing J-test and the mean encompassing F-test as described in Kennedy (2008) with appropriate references.[4] Together the variance and mean encompassing tests make up the complete encompassing test and has 16 possible results (as shown in Table 9). Each test, the J-test and the F-test, have 4 possibilities: Neither model is acceptable, Both models are acceptable and model A is acceptable and B is not, or Model B is acceptable and A is not.

H_A: $Y = \beta_0 + \beta_1 T + \beta_2 D + \beta_3 DT + \varepsilon$ → a structural break

versus H_B: $Y = \beta_0 + \beta_1 T + \beta_2 T^2 + \varepsilon$ → a quadratic model

[4] The J-test as implemented here is found in the ETS documentation. Accessed at https://support.sas.com/rnd/app/ets/examples/spec/index.htm. Kennedy (2008, pp. 87-88, gives a clear description). Mizon and Richard (1986) discuss the encompassing tests in detail.

Table 9: Possible outcomes from the complete encompassing test for Non-Nested Hypotheses (with results of this paper highlighted)

J-test		F-test	
Quadratic Model	**Structural Break**	**Quadratic Model**	**Structural Break**
"Acceptable"	"Acceptable"	"Acceptable"	"Acceptable"
		"Acceptable"	"not Acceptable"
		"not Acceptable"	"Acceptable"
		"not Acceptable"	"not Acceptable"
"Acceptable"	"not Acceptable"	"Acceptable"	"Acceptable"
		"Acceptable"	"not Acceptable"
		"not Acceptable"	"Acceptable"
		"not Acceptable"	"not Acceptable"
"not Acceptable"	"Acceptable"	"Acceptable"	"Acceptable"
		"Acceptable"	"not Acceptable"
		"not Acceptable"	"Acceptable"
		"not Acceptable"	"not Acceptable"
"not Acceptable"	"not Acceptable"	"Acceptable"	"Acceptable"
		"Acceptable"	"not Acceptable"
		"not Acceptable"	"Acceptable"
		"not Acceptable"	"not Acceptable"

The following code block shows the non-nested tests, first for the J-test and then followed by the non-nested F-test. The J-test starts with estimating the quadratic and the structural break model and outputing the predicted values of each model. Then the prediction of one model is added as an explanatory variable of the second model. If that new variable, the prediction of the other model, is significant then the model being estimated is rejected.

In the case of the F-test a super model of all the variables is formed with wald type TEST statements to reduce the super model to the model being tested. If the Wald test is significant then the model is rejected.

```
Title1 'Non-nested hypothesis - J-test';
Proc reg data=trdata;
    model_2: model Y = T TSQ;
    output out=Mquad p=Yquadhat;
    run;
Proc reg data=trdata;
    model_4: model Y = T D DT;
    output out=Minter p=Yinterhat;
    run;
Proc reg data=mquad;
```

```
    model_4A: model Y = T D DT Yquadhat;
    run;
Proc reg data=Minter;
    model_2A: model Y = T TSQ Yinterhat;
    run;
Title1 'Non-nested hypothesis test - super model, F-test';
Proc reg data=trdata;
    Super: model Y = T TSQ D DT ;
    quad: test tsq = 0;
    interactive: test d =dt=0;
   run;
```

Table 10: Non-nested Hypotheses Testing

	Variance encompassing test				mean encompasing test
	Quadratic	J-TEST	Interactive	J-TEST	F-TEST
constant	11.12 ***	1.26	10.01 ***	8.67 ***	10.23214 ***
	(14.96)	(-0.33)	(18.25)	(2.22)	12.01
T	1.12 ***	0.11	2.01 ***	1.71 ***	1.86905 ***
	(4.9)	(0.25)	(16.42)	(1.94)	4.26
TSQ	0.13 ***	0.02			0.01815
	(8.77)	(0.33)			0.35
D			-13.47 ***	-11.56 ***	-11.5627 **
			(-9.12)	(-2.02)	-2.02
DT			1.91 ***	1.66 ***	1.65548
			(11.01)	(2.19)	-2.02
prediction tests		**0.89 ** (2.61)**		**0.14 (0.35)**	**quad: test tsq = 0; F= 0.12** / **interactive: test d =dt=0; F=3.07 ***
n	14	14	14	14	14
adj R sq	0.972	0.998	0.970	0.997	0.997
F	457.6	1746.3	212.2	1181.4	1181.4
root MSE	2.16	0.06	2.24	0.68	0.68

note: All regressions estimated with OLS using the SAS REG procedure.
T-stats in parentheses.
*** significant at the .01 level
** significant at the .05 level
* significant at the .10 level

J-TEST – THE VARIANCE ENCOMPASSING TEST

Table 10 shows that the predictions from the structural break model (column 2 j-test highlighted) are significant and thus contribute additional explanatory power to the quadratic model and therefore rejects the quadratic model. Also shown is the predictions from the quadratic model (column 4 j-test highlighted) fail to reach significance and thus do

not contribute additional explanatory power to the structural break model and therefore fails to reject (finds "acceptable") the structural break model.

F-TEST – THE MEAN ENCOMPASSING TEST

When a super model of all variable from the structural break model and the quadratic model combined the variable unique to the quadratic model fails to reject the null hypothesis of zero effect. So the quadratic model is not acceptable. However the unique variables to the structural break model do contribute to the explanatory power over and above the quadratic model meaning the structural break model is not rejected (found "acceptable").

J-TEST AND F-TEST TOGETHER – THE COMPLETE ENCOMPASSING TEST

As highlighted in Table 9, the J and F non-nested hypothesis tests both agree that the only model "acceptable" is the structural break model therefore the intervention as measured by D did have an effect.

CONCLUSION

Visually, we liked two models on the entire sample and only the linear on the sub-samples.

We showed that a Hasty Regression led to a false conclusion, namely that D did not matter. This was the biggest lie in the data. Not only did D have no effect in Model Y = T D; it also has no effect in Model Y = T TSQ D; either. Hasty regression lies whether you assume a base linear or nonlinear model in time.

Eight regressions were necessary to complete a testing strategy that convinced us that the structural break model was a better representation of the data.

A Ramsey test for misspecification was run on all models and found the structural break and quadratic models both "acceptable."

The 'quadratic model' was tested against the 'structural break' model directly by the use of non-nested hypotheses: four tests were run and in each case the 'quadratic model' was found lacking.

Finally, automatic processes may arrive at the same conclusion, but that is not at all clear. Human critical thinking processes are critical for making sense of this data.

REFERENCES

https://www.lexjansen.com/wuss/2006/posters/POS-Calise.pdf Detecting Structural Change Using SAS®/ETS Procedures Archie J. Calise, Queensborough College of the City University of New York Joseph Earley, Loyola Marymount University, Los Angeles

https://support.sas.com/resources/papers/proceedings09/306-2009.pdf Paper 306-2009 Structural Analysis of Time Series Using the SAS/ETS® UCM Procedure Rajesh Selukar, SAS Institute Inc., Cary, NC

Davidson and Mackinnon (1981). "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica*, 49, 781-793.

Kennedy, Peter. (2008), *A Guide to Econometrics*, 6th edition, Blackwell Publishing.

Mizon, G. and Richard, J. F. (1986). The encompassing principle and its applications to testing nonnested hypothesis. Econometrica 3, 657–78

Ramsey (1969) "Tests for Specification Errors in Classical Linear Least Squares Analysis." *Journal of the Royal Statistical Association*, Series B, 71, 350–371.

ACKNOWLEDGMENTS

I wish to thank the section chair, Kirk Paul Lafler, for inviting this paper for a HOW session at the 2019 SCSUG Educational Forum and for his patience in my completion of it. And my thanks to Joni Shreve, the Academic Chair, for her support as well. I am grateful to Kirk Paul Lafler and Josh Horstman for their encouragement to begin presenting at SAS conferences in general.

RECOMMENDED READING

Calise, Archie J. and Joseph Earley (2006) Detecting Structural Change Using SAS®/ETS Procedures. Poster Session, WUSS. Accessed at https://www.lexjansen.com/wuss/2006/posters/POS-Calise.pdf.

Fan, J. and Gijbels, I. (1996) Local Polynomial Modelling and Its Applications. Chapman and Hall, London.

Horstman, Joshua M. (2018) Getting Started with the SGPLOT Procedure, SCSUG. Accessed at https://www.lexjansen.com/scsug/2018/Horstman_SCSUG2018_Getting_Started_With_SGPLOT.pdf.

Horstman, Joshua M. (2018) Doing More with the SGPLOT Procedure, SESUG Paper 205-2018 Accessed at https://www.lexjansen.com/sesug/2018/SESUG2018_Paper-205_Final_PDF.pdf.

Selukar, Rajesh (2017) Detecting and Adjusting Structural Breaks in Time Series and Panel Data Using the SSM Procedure, Paper SAS456-2017, SAS Global Forum. Accessed at https://www.lexjansen.com/wuss/2006/posters/POS-Calise.pdf.

Bilenas, Jonas V. (2014) Scatter Plot smoothing using PROC LOESS and Restricted Cubic Splines, Paper 1503-2014, SAS Global Forum, Accessed at https://support.sas.com/resources/papers/proceedings14/1503-2014.pdf.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Steven C. Myers
myers@uakron.edu
https://www.linkedin.com/in/stevencmyers/
https://econdatascience.com