
VIRTUAL
SAS® GLOBAL FORUM 2021
#SASGF

Haste Makes Waste: Don't Ruin Your Reputation with Hasty Regression

Dr. Steven C. Myers, The University of Akron

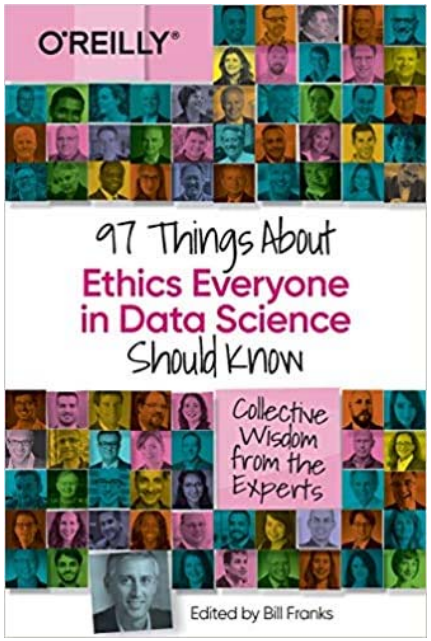
Steven Myers is an economist, educator in applied econometrics and an evangelist for economists in data science. He is the 2019 SAS Distinguished Educator. He specializes in combining economic and business acumen with rigorous statistical and programming techniques to solve problems and relate solutions to business leaders. He taught and mentored SAS programming, data handling, statistics and applied econometrics for over 40 years. He is a recovering CIO and is Associate Professor Emeritus at The University of Akron serving on their adjust faculty.

How do you know if your data is lying to you?

The answer lies in-following ethical applied econometric rules and being aware of and avoiding pitfalls in regression practice.

These essential skills are not typically taught in single-semester econometric (or regression) course.

What are the Ethical Process rules for Applied Econometrics?



Rule 1:	Use common sense and economic reasoning
Rule 2:	Avoid Type III errors
Rule 3:	Know the context
Rule 4:	Inspect the data
Rule 5:	Keep it sensibly simple
Rule 6:	Use the interocular trauma test
Rule 7:	Understand the costs and benefits of data mining.
Rule 8:	Be prepared to compromise
Rule 9:	Do not confuse statistical significance with meaningful magnitude
Rule 10:	Report a sensitive analysis

Steven Myers. "Ethics Rules in Applied Econometrics and Data Science," Essay 85, Policy Implications, in Bill Franks (ed.). 97 Things About Ethics Everyone in Data Science Should Know. O'Reilly, 2020, pp. 231-233.

What are the Pitfalls to Regression Practice?

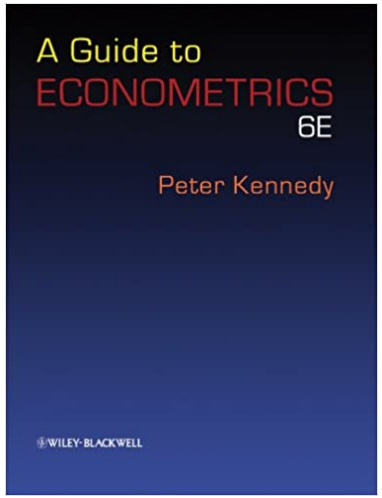
This presentation is part of a larger effort I call

Avoiding the Pitfalls in Regression Analysis

EconDataScience.com

1. Failure to understand why you are running the regression.
2. Failure to be a data skeptic and ignoring the data generating process.
3. Failure to examine your data before you regress.
4. Failure to examine your data after you regress.
5. Failure to understand how to interpret regression results.
6. Failure to model both theory and data anomalies, and to know the difference.
7. Failure to be ethical.
8. Failure to provide proper statistical testing
9. Failure to properly consider causal calculus
10. Failure to meet the assumptions of the classical linear model.

Pitfall #7. Failure to Be ethical



"… my opinion is that regardless of teachability, we have a moral obligation to inform students of these rules, and , through suitable assignments, socialize them to incorporate them into the standard operating procedures they follow when doing empirical work…. (I) believe that these rules are far more important than instructors believe and that students at all levels do not accord them the respect they deserve."

Kennedy, Peter E. (2002) "Sinning in the Basement: What are the rules? The Ten Commandments of Applied Econometrics." Applied Econometrics, Blackwell Publishers Ltd, pp.571-2.

"To be an effective and ETHICAL data scientist, you MUST understand mathematics and statistics. Many data scientists make bad decisions - with ethical implications - not because they are intentionally trying to do harm, but because they do not have an understanding of how the algorithms they are taking responsibility for actually work. "

Jennifer Priestley. Ph.D.
Associate Dean, The Graduate College at Kennesaw State University

Pitfall #6. Failure to model theory and data anomalies and know the difference.

Business problem:

What was the value of a policy change, D, on our outcome metric, tax revenue, Y?

Articulate the problem:

Modeling theory:

Is the trend of our metric, Y, changed positively or negatively by our measure of the policy/intervention, D?

H_0: An intervention that begins in T=8 has no effect on the trend line of Y.
H_1: An intervention at T=8 changes the trend line of Y.

Data Anomalies:

The actual equation is nonlinear in variables and the intervention has no effect. That is, the natural nonlinear progression is why it may appear that the intervention D has an effect.

Set up the Data

(1) Read the 14 observations from work.Y,
(2) The policy variable, D, taking values of
0 for time periods 1-7 and values of
1 for time periods 8 to 14.
There is no other data available,
(3) Create and transform variables for testing, and
(4) in this example only assume the data is clean.

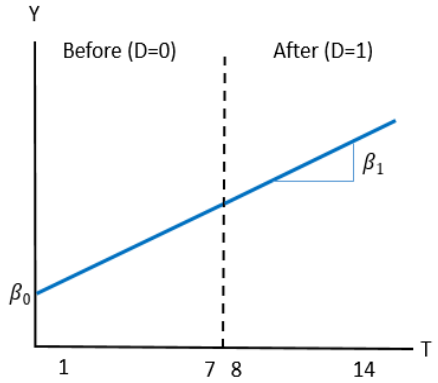
```
DATA Y;
   input Y @@;
   datalines;
   12.35  13.71  16.00  17.94  20.76  21.11  24.63
   27.56  32.88  35.16  39.26  44.28  47.27  51.55;
   run;
```

```
Data trdata;

      set Y;

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

First regression: Establish the trend in the base model.



$Y = \beta_0 + \beta_1 T + \varepsilon$ population model

$E(Y) = \beta_0 + \beta_1 T$ because E(ε)=0

$\hat{Y} = b_0 + b\beta_1 T$ sample model

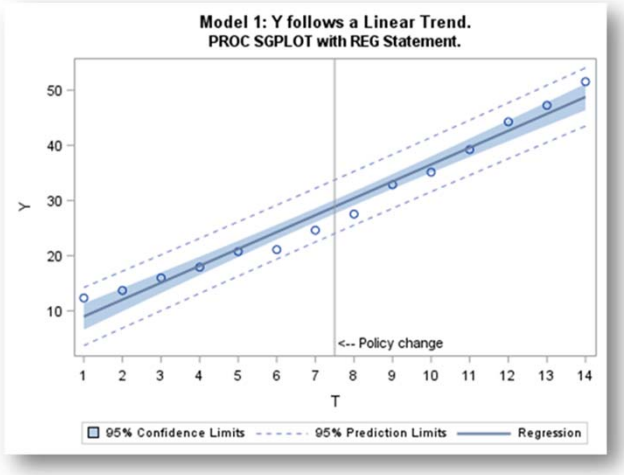
$e = Y - \hat{Y}$

H_0: $\beta_1 = 0$
H_1: $\beta_1 > 0$

Restricted: Model Y = constant
Unrestricted: Model Y = T;
Test T=0;

$\overline{R}^2$ = 0.972

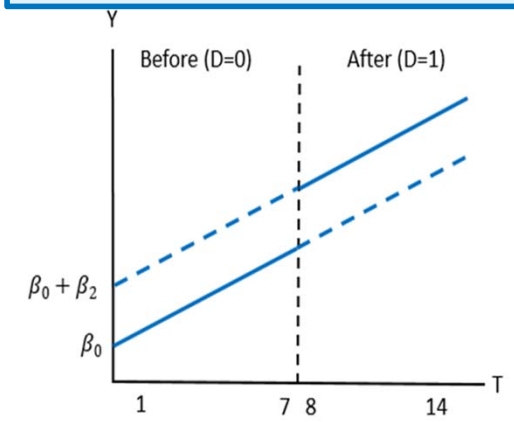
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.93242	1.21836	4.87	0.0004
T	1	3.06101	0.14309	21.39	<.0001




```
ods graphics on;
Title1 'Regression Specifications -
full sample' ;
PROC   REG data=trdata;
       var T TSQ D DT;
       model_1: model Y = T;
       run;
```

Hasty Regression: Just throw in a dummy variable.

```
title2 'Just throw in a dummy variable';
proc reg;
   model_3:  model y = t d;
   run;
```



$$Y = \beta_0 + \beta_1 T + \beta_2 D + \varepsilon$$

$$E(Y|D=0) = \beta_0 + \beta_1 T$$

$$E(Y|D=1) = (\beta_0 + \beta_2) + \beta_1 T + \varepsilon$$

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 > 0$$

Restricted: Model Y = T ;
Unrestricted: Model Y = T D;
Test D=0;

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

$\overline{R}^2$ = 0.970
< .972 for linear model_1

The intervention apparently does not affect the outcome.

Y = 6.19 + 2.97T + e_1
Y = 6.19 + 2.97T + 0.853D + e_2

Effect of D is the difference in the expected values

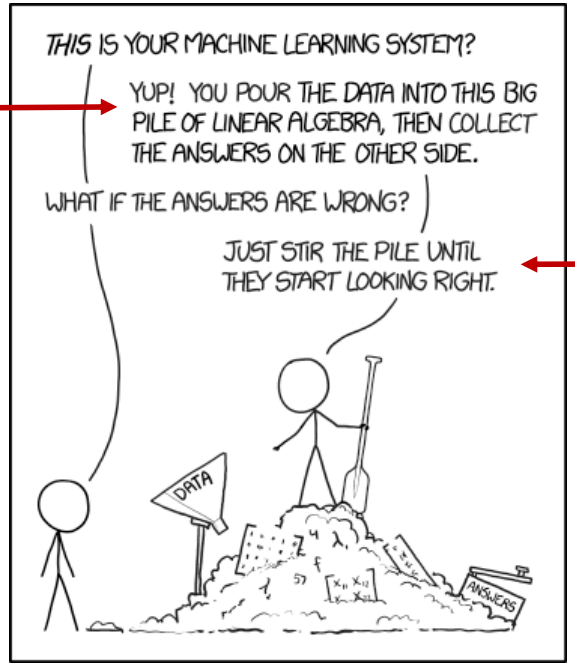
$$E(Y|T, D=1) - E(Y|T, D=0)$$
= (6.19 + 2.97T)
- (6.19 + 2.97T – 0.853)
= 0.853

Effect of Time is the partial derivative of Y with respect to T,

$$\frac{\partial Y}{\partial T} = \beta_1$$ holding D constant

Why isn't hasty regression enough?

Garbage in, Garbage out.



https://xkcd.com/1838/

If you torture the data long enough, it will confess.
-Ronald Coase

So why not use automatic model selection?

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

How about use some thought?
Economics and common sense
Know the DGP and context
Clean the data before
Inspect the data after
Validate your specification and testing.

Lack of Proof, its all about rejection, not acceptance

H_0: $\beta_2=0$ The intervention had no effect
H_1: $\beta_2\neq0$ The intervention had an effect

The problem with this is it is sensitive to model specification

	The statistician can never "know" truth	
The statistician's decision	H_0 is true	H_0 is false
Rejects H_0	$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$ *A wrong decision* *Convicting the innocent* Type I Error	$1-\beta = P(\text{reject } H_0 \mid H_0 \text{ is false})$ *A correct decision* *Convicting the guilty* Power of the test
Fails to Reject H_0	$1-\alpha = P(\text{fail to reject } H_0 \mid H_0 \text{ is true})$ *A correct decision* *Letting the innocent go free* Confidence in the test	$\beta = P(\text{fail to reject } H_0 \mid H_0 \text{ is false})$ *A wrong decision* *Letting the guilty go free* Type II Error

The most damning are **Type III errors** which occur when you get the right answer to the wrong question (Kennedy, 2008).

Look at the residuals in the baseline regression

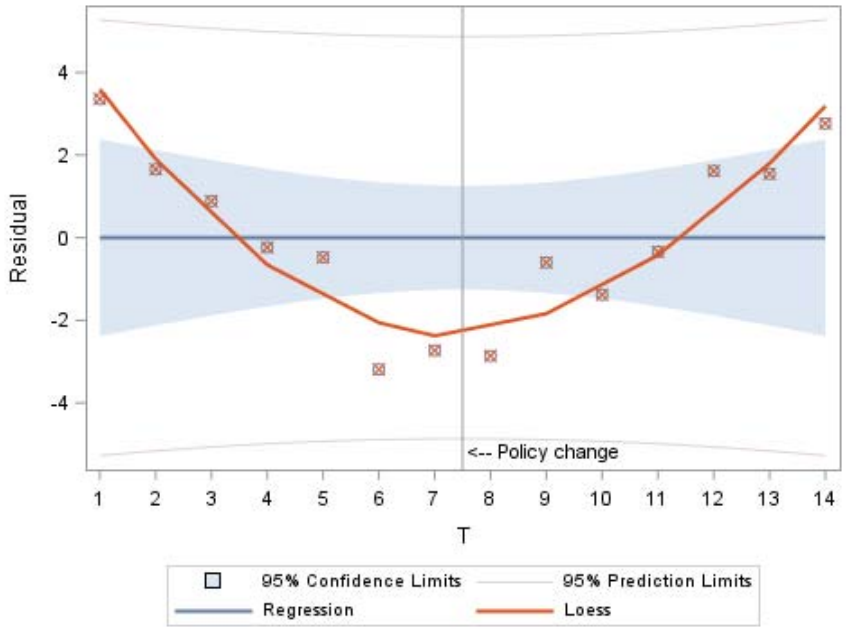
$$Y = \beta_0 + \beta_1 T + \varepsilon$$

$$\hat{Y} = b_0 + b_1 T + e$$

$$e = Y - \hat{Y}$$

```
Title2 'Examining the residuals to the linear trend model.';
PROC REG data=trdata;
            model_1R: model Y = T;
            output out=TR  r=residual;
            run;

PROC SGPLOT data=TR;
            reg    x=T y=residual / degree=1 CLM CLI
                                    CLMTRANSPARENCY=.5;
            loess x=T y=residual / interpolation=linear degree=2;
            &xref;
           run;
```

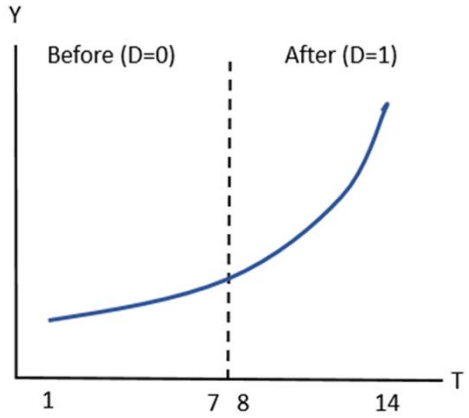


Two models are suggested by analysis

```
title2 'Quadratic Model';
model_2:   model Y = T  TSQ;
Quadratic:   test tsq=0;
run;
```

Effect of Time in Model 2 is the partial derivative of Y with respect to T,

$$\frac{\partial Y}{\partial T} = \beta_1 + 2\,\beta_2 T$$



$$Y = \beta_0 + \beta_1 T + \beta_2 T^2 + \varepsilon$$

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 > 0$$

Restricted: Model Y = T;
Unrestricted: Model Y = T TSQ;
Test TSQ=0;

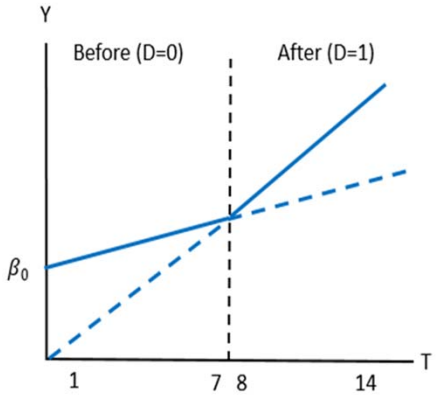
```
title2 'Structural break model';
model_4:  model Y = T  D  DT;
Structural_break:  test D=DT=0;
run;
```

Effect of Time in Model 4 is the partial derivative of Y with respect to T,

$$\frac{\partial Y}{\partial T} = \beta_1 + \beta_3 D$$

Effect of D in Model 4 is the differences in expectations

$$E(Y|T, D = 1) - E(Y|T, D = 0) = \beta_2 + \beta_3 T$$



$$Y = \beta_0 + \beta_1 T + \beta_2 D + \beta_3 DT + \varepsilon$$

$$E(Y|D = 0) = \beta_0 + \beta_1 T$$

$$E(Y|D = 1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) T + \varepsilon$$

$$H_0: \beta_2 = 0 \text{ and } \beta_2 = 0$$
$$H_1: \text{not true}$$

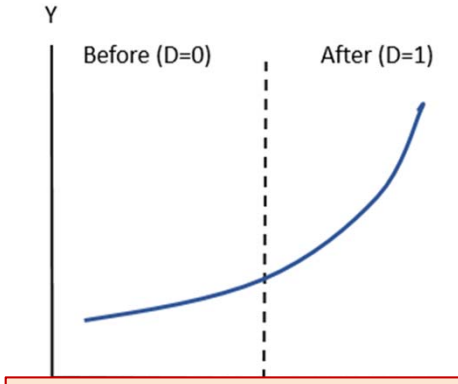
Restricted: Model Y = T;
Unrestricted: Model Y = T D DT;
Test D=DT=0;

Testing both models against the base model

```
title2 'Quadratic Model';
model_2:    model Y = T  TSQ;
Quadratic:  test tsq=0;
run;
```

	F Value	P-value	$\bar{R}^2$	Root MSE
Quadratic: test tsq=0;	76.84	<.0001	0.996	0.80
Structural_break: test D=DT=0;	61.32	<.0001	0.998	0.65

```
title2 'Structural break model';
model_4:            model Y = T  D  DT;
Structural_break:   test D=DT=0;
run;
```

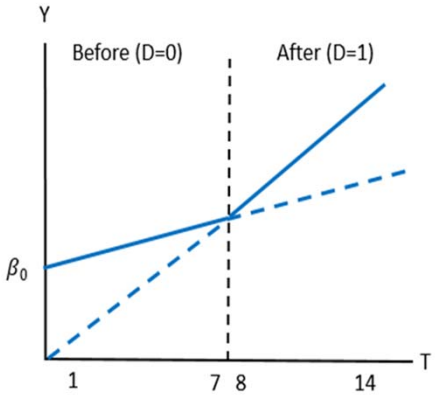


$$Y = \beta_0 + \beta_1 T + \beta_2 T^2 + \varepsilon$$

$H_0: \beta_2 = 0$
$H_1: \beta_2 > 0$

Restricted: Model Y = T;
Unrestricted: Model Y = T TSQ;
Test TSQ=0;

Arghhhhh: Both models are valid.



$$Y = \beta_0 + \beta_1 T + \beta_2 D + \beta_3 DT + \varepsilon$$

$$E(Y|D=0) = \beta_0 + \beta_1 T$$

$$E(Y|D=1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) T + \varepsilon$$

$H_0: \beta_2 = 0 \text{ and } \beta_2 = 0$
$H_1: \text{not true}$

Restricted: Model Y = T;
Unrestricted: Model Y = T D DT;
Test D=DT=0;

Both models perform very well.

Both models 2 and 4 reject the null hypothesis of a linear trend.

Both the quadratic model and structural break model are "acceptable."

	F Value	P-value
Quadratic: test tsq=0;	76.84	<.0001
Structural_break: test D=DT=0;	61.32	<.0001

	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
$\bar{R}^2$	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

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

These are called nested hypothesis. The null hypothesis Is a subset of the model being tested.

Model 1 is the restricted model compared to the unrestricted models 2 and 4.

How do the models fit in the before and after period?

```
title2 'Before Sample, T=1,...,7';

PROC   reg data=work.trdata;
       model_5: model Y = T      ;
       model_6: model Y = T TSQ  ;
       where D=0;
       run;

title2 'After Sample, T=8,..., 14';

PROC   reg data=work.trdata;
       model_7: model Y = T      ;
       model_8: model Y = T TSQ  ;
       where D=1;
       run;
```

If the quadratic model ruled, then it would exist in the sub samples and it does not! There was a structural break!

	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
$\bar{R}^2$	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

Non-nested hypothesis testing

Are we so confident that there is a structural break and not a quadratic model?

After all, the tests used were completely nested:

$H_0: Y = \beta_0 + \beta_1 T + \varepsilon$

versus $H_1: Y = \beta_0 + \beta_1 T + \beta_2 D + \beta_3 DT + \varepsilon$ → a structural break (model_2)

&

$H_0: Y = \beta_0 + \beta_1 T + \varepsilon$

versus $H_1: Y = \beta_0 + \beta_1 T + \beta_2 T^2 + \varepsilon$ → a quadratic model (model_4)

Shouldn't we be testing one model against the other?:

$H_0: Y = \beta_0 + \beta_1 T + \beta_2 D + \beta_3 DT + \varepsilon$ → a structural break

versus $H_1: Y = \beta_0 + \beta_1 T + \beta_2 T^2 + \varepsilon$ → a quadratic model

These are non-nested models and require a different testing procedure.

Non-nested hypothesis testing code

```
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=Mbreak p=Ybreakhat;
          run;

Proc reg data=Mbreak;
          model_2A: model Y = T TSQ Ybreakhat;
          run;

Proc reg data=Mquad;
          model_4A: model Y = T D DT Yquadhat;
          run;
```

```
Title1 'Non-nested hypothesis test, super model, F-test';


Proc reg data=trdata;

          model_4A: model Y = T TSQ D DT;
          quad: test TSQ = 0;
          structural_break: test D=DT=0;
          run;
```

Four tests all reject ‘quadratic’ in favor of the ‘structural break’ model.

The ‘break’ model contributes significantly to the explanation of the ‘quadratic’ model.

Conclusion: reject the ‘quadratic’ model

The ‘quadratic’ model contributes nothing to the explanation of the ‘break model.

Conclusion: fail to reject the ‘’break’ model

	Variance encompassing test				Mean encompassing test
	Quadratic	J-TEST	Break	J-TEST	F-TEST
constant	11.12***	1.26	10.01***	8.67***	10.23***
	(14.96)	(-0.33)	(18.25)	(2.22)	(12.01)
T	1.12***	0.11	2.01***	1.71***	1.87***
	(4.9)	(0.25)	(16.42)	(1.94)	(4.26)
TSQ	0.13***	0.02			0.02
	(8.77)	(0.33)			(0.35)
D			-13.47***	-11.56***	-11.56**
			(-9.12)	(-2.02)	(-2.02)
DT			1.91***	1.66***	1.66
			(11.01)	(2.19)	(2.19)
Y-prediction from other model		0.89** (2.61)		0.14 (0.35)	
quad: test TSQ = 0;					F = 0.12
break: 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

Parameters unique to the ‘quadratic’ model are not rejected from zero.

Conclusion: the ‘break’ model is not rejected.

Parameters unique to the ‘break’ model are rejected from zero.

Conclusion: the ‘quadratic’ model is rejected.

Complete encompassing reconciliation of the

Mean Encompassing tests (J-test)

and the

Variance Encompassing Tests (F-test)

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"

Ramsey misspecification tests (RESET) from SAS/ETS

```
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;
```

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

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

Automatic selection leads to the

Structural break model 3 times
Quadratic model 1 time, and the
Super model 3 times.

Which to believe?

Conclusion:

There is strong evidence of a structural break in the metric Y because of the intervention, D, as studied.

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

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

3. 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.]

4. 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.

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

6. Finally, automatic processes do not arrive at the same conclusion. Human critical thinking processes are critical for making sense of this data.

Thank you for watching!

Contact me at

Dr. Steven Myers
myers@uakron.edu
LinkedIn: stevencmyers
Website: EconDataScience.com


VIRTUAL
SAS® GLOBAL FORUM 2021
#SASGF