

PROC REG ... and Its Discontents

Nathaniel Derby

Statis Pro Data Analytics
Seattle, WA, USA

Vancouver SAS Users Group, 10/8/2008

Outline

- 1 PROC REG
 - Basics
 - Checking Assumptions
 - Options
- 2 Discontents
 - Overview
 - Time Series Data
- 3 Conclusions

Basics

PROC REG = *Regression Analysis* done with SAS.

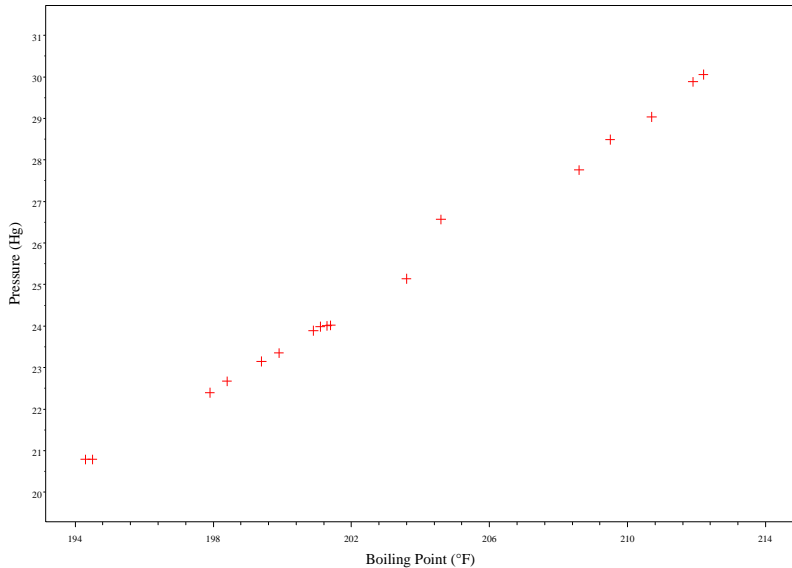
What is regression analysis?

- Fitting the best-fit straight line through the data.
- Some assumptions required ...

Start with a *scatterplot*:

- Data: James Forbes, 1857.
- Boiling point vs air pressure.
- `work.boiling`.
- Does it fit a straight line?

Boiling Point vs Pressure



Fitting a Line

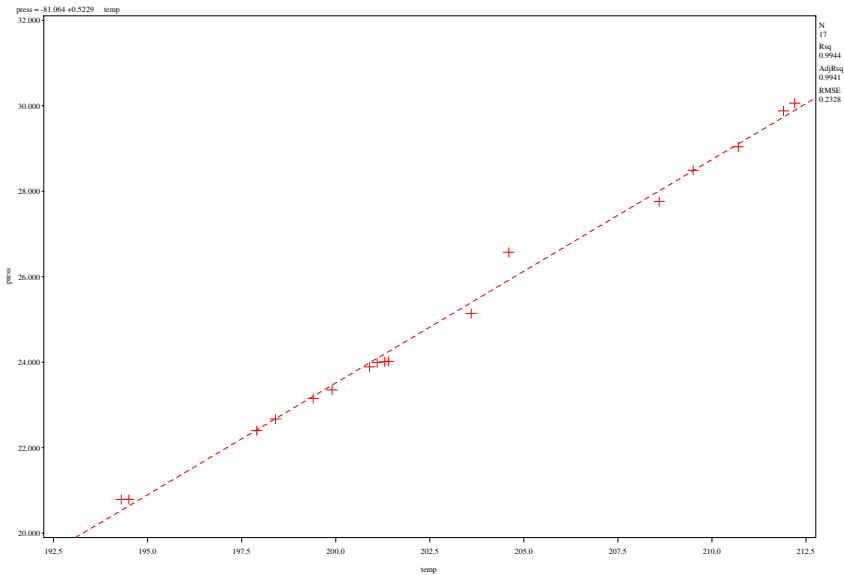
We want the line

$$\text{Pressure} = \beta_0 + \beta_1 \text{Temp} :$$

SAS Code

```
proc reg data=boiling;  
  model press = temp;  
  plot press*temp;  
run;
```

Boiling Point vs Pressure



Trouble in Paradise

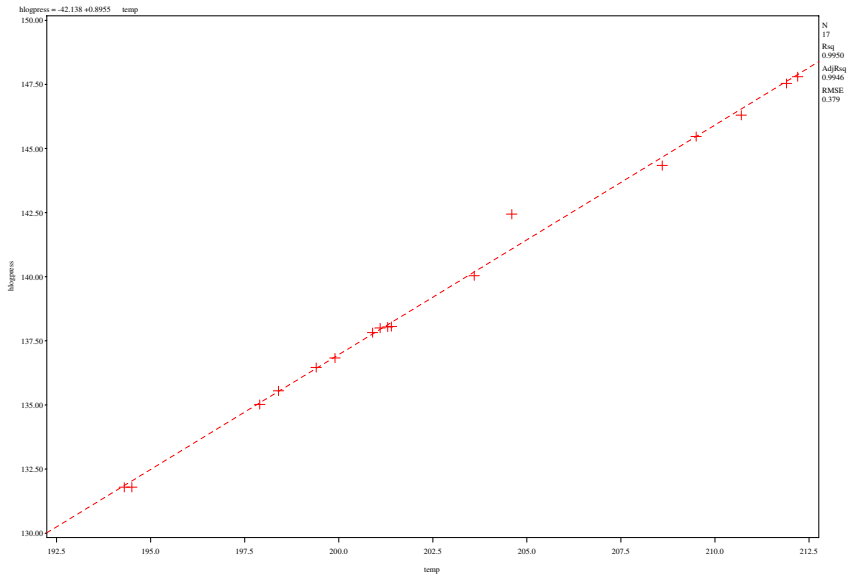
Doesn't work if underlying data not on a (straight) line!

- Try a transformation!
- Pressure $\Rightarrow 100 \times \text{Log}(\text{Pressure})$.

New SAS Code

```
proc reg data=boiling;  
  model hlogpress = temp;  
  plot hlogpress*temp;  
run;
```

Boiling Point vs Log Pressure



We're Done, Right?

The REG Procedure
Model: MODEL1
Dependent Variable: hlogpress

Number of Observations Read 17
Number of Observations Used 17

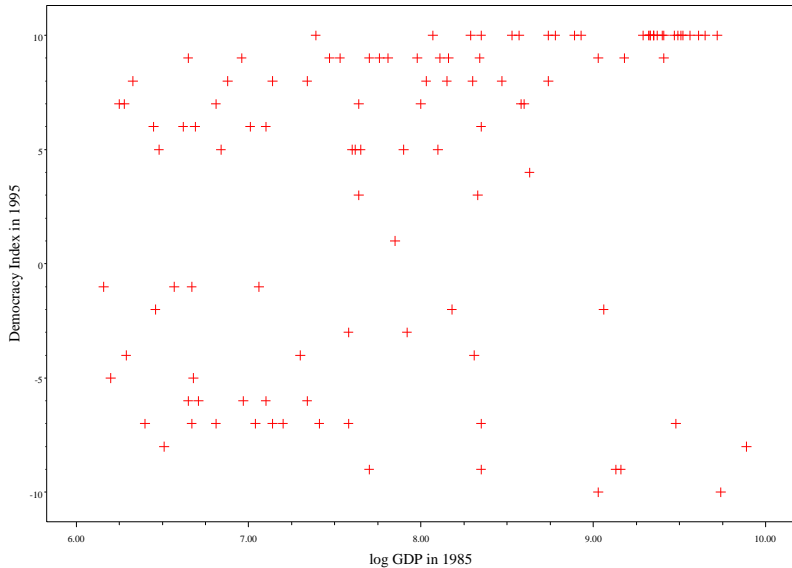
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	425.63910	425.63910	2962.79	<.0001
Error	15	2.15493	0.14366		
Corrected Total	16	427.79402			

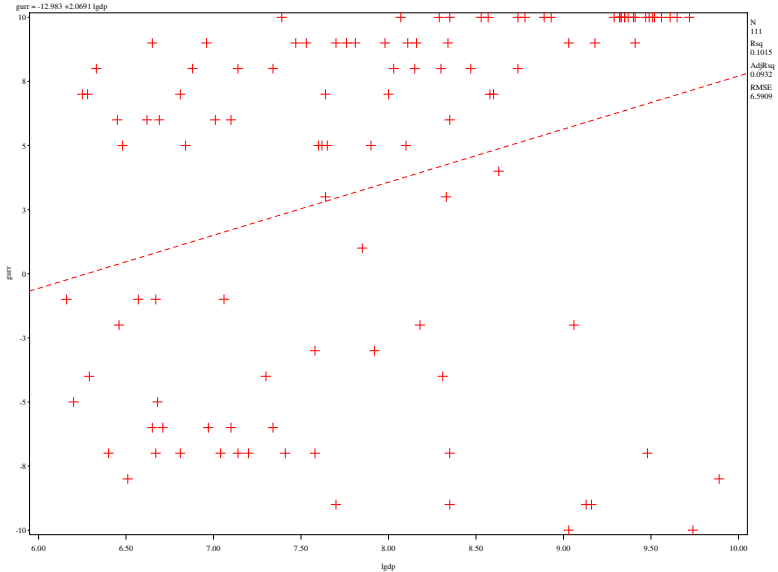
Root MSE		0.37903	R-Square	0.9950
Dependent Mean		139.60529	Adj R-Sq	0.9946
Coeff Var		0.27150		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-42.13778	3.34020	-12.62	<.0001
temp	1	0.89549	0.01645	54.43	<.0001

log GDP vs Democracy Index

log GDP vs Democracy Index



SAS Output

The REG Procedure
Model: MODEL1
Dependent Variable: gurr

Number of Observations Read	112
Number of Observations Used	111
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	534.76792	534.76792	12.31	0.0007
Error	109	4734.97983	43.44018		
Corrected Total	110	5269.74775			

Root MSE	6.59092	R-Square	0.1015
Dependent Mean	3.50450	Adj R-Sq	0.0932
Coeff Var	188.06986		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-12.98347	4.74073	-2.74	0.0072
lgdp	1	2.06913	0.58973	3.51	0.0007

Checking Assumptions

We need to be sure model is appropriate for the data.

- Checking mathematical assumptions of data distribution.
- Look at *residuals vs predicted values*:
 - *Predicted value*: The y value on the line.
 - *Residual*: What's left over.

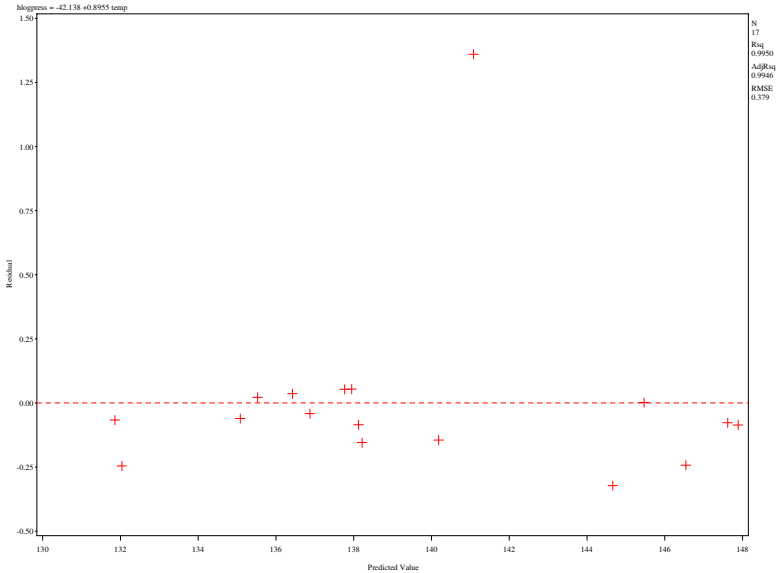
▶ Example 1

▶ Example 2

SAS Code

```
proc reg data=boiling;  
  model press = temp;  
  plot press*temp;  
  plot residual.*predicted.;  
run;
```

Boiling Point vs Log Pressure



What Are We Looking For?

Goal: We want residuals to *have no pattern whatsoever*.

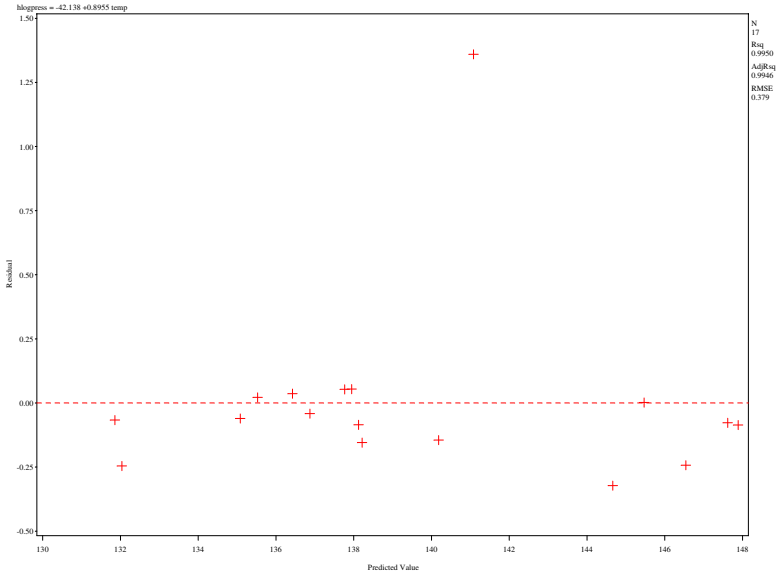
The following are bad:

- Grouped together into “clumps”.
- All of one part of range above/below line.
- Some parts farther away from line than others.
- Outliers (sometimes, sometimes not).

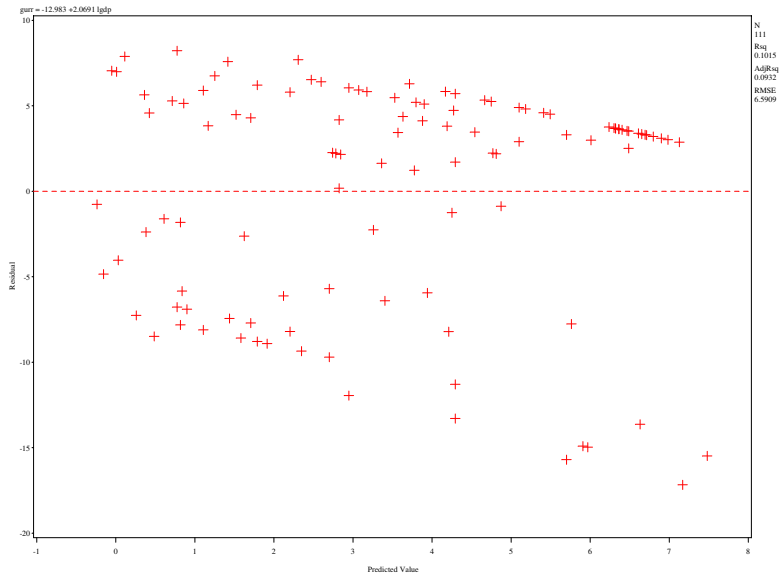
What does “bad” mean?

- Results we just saw not exactly valid.
- Stay tuned for an example!

Boiling Point vs Log Pressure



log GDP vs Democracy Index



Options

SAS Code

```
proc reg data=boiling noprint/simple/all covout  
      outest=SASdataset outscp=SASdataset;  
  model press = temp;  
  plot press*temp;  
run;
```

- **noprint/simple/all**: How much output do you want?
- **covout**: Outputs *covariance matrix*.
- **outest/outscp**: Outputting parameter estimates/covariance matrix.

Options

SAS Code

```
proc reg data=boiling;  
  model press = temp / options;  
  plot press*temp;  
run;
```

Many options for statistical output. Very useful:

- **clm**: Prints *95% confidence limits* for expected value of mean of each observation.
- **cli**: Prints *95% confidence limits* for a predicted value..
- **r**: Analyses residuals.

Options

SAS Code

```
proc reg data=boiling;  
  model press = temp;  
  weight xxx;  
  plot press*temp;  
run;
```

Use if **xxx** is a variable of *weights* of each observation.

- If Bill Gates were an observation.
- Countries weighted by population.

Options

SAS Code

```
proc reg data=boiling;  
  model press = temp;  
  output out=SASData keyword=variable;  
  plot press*temp;  
run;
```

Some (of *many*) keywords:

- **predicted**: Predicted values.
- **residuals**: Residual values.
- **l95**: Lower 95% bound for predicted value.
- **stdp**: Standard error of mean predicted value.

Options

SAS Code

```
proc reg data=boiling;  
  model press = temp;  
  by xxx;  
  plot press*temp;  
run;
```

Gives a different analysis for each value of variable **xxx**.

Problems

What happens when the assumptions fail?

⇒ Proposed model is wrong!

Solution: Change the model!

Questions:

- Is there a *linear* relationship?
- Do we have the right variables?
- Are the errors *uncorrelated*? (Time series!)

Often Hidden Problems with Time Series Data!

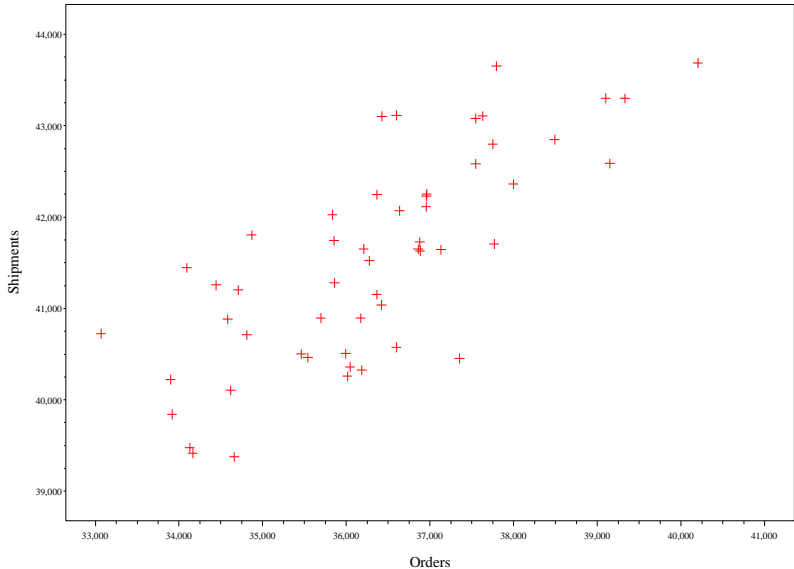
Time Series Data

Valve shipments and orders (Pankratz, 1991):

	date	shipments	orders
1	01/1984	39377	34662
2	02/1984	39417	34165
3	03/1984	39475	34127
4	04/1984	39843	33917
5	05/1984	40223	33900
6	06/1984	40105	34618
7	07/1984	40502	35463
8	08/1984	40726	33067
9	09/1984	41444	34095
10	10/1984	41256	34443
11	11/1984	41803	34868

Suppose we ignore the time variable (**date**).

Valve Orders vs Shipments

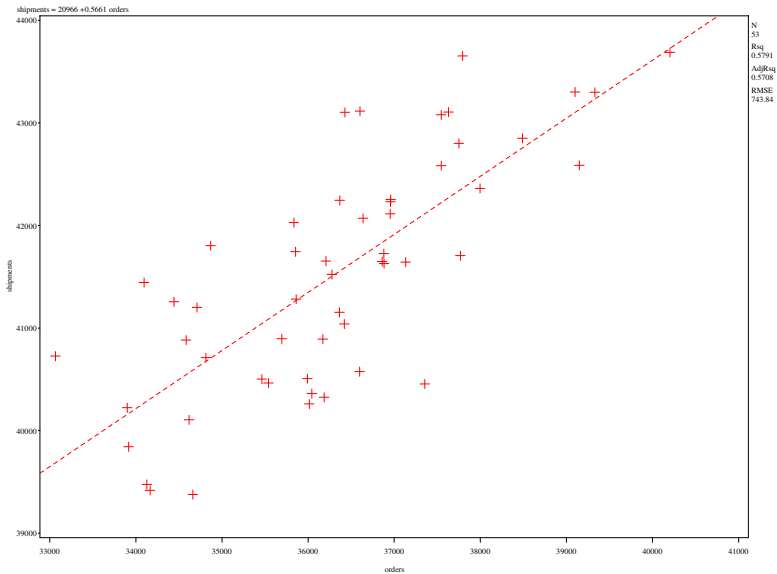


Doing the Usual

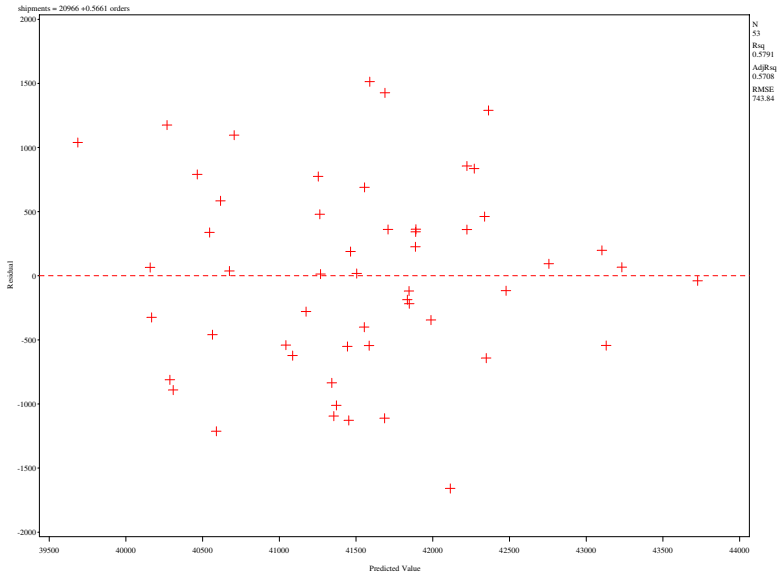
SAS Code

```
proc reg data=boiling;  
  model press = temp;  
  plot press*temp;  
  plot residual.*predicted.;  
run;
```

Valve Orders vs Shipments



Valve Orders vs Shipments



SAS Output

The REG Procedure
Model: MODEL1
Dependent Variable: shipments

Number of Observations Read	54
Number of Observations Used	53
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	38818277	38818277	70.16	<.0001
Error	51	28218196	553298		
Corrected Total	52	67036473			

Root MSE	743.84001	R-Square	0.5791
Dependent Mean	41527	Adj R-Sq	0.5708
Coeff Var	1.79124		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	20966	2456.79440	8.53	<.0001
orders	1	0.56613	0.06759	8.38	<.0001

Problems

Actually, the **SAS output is all false.**

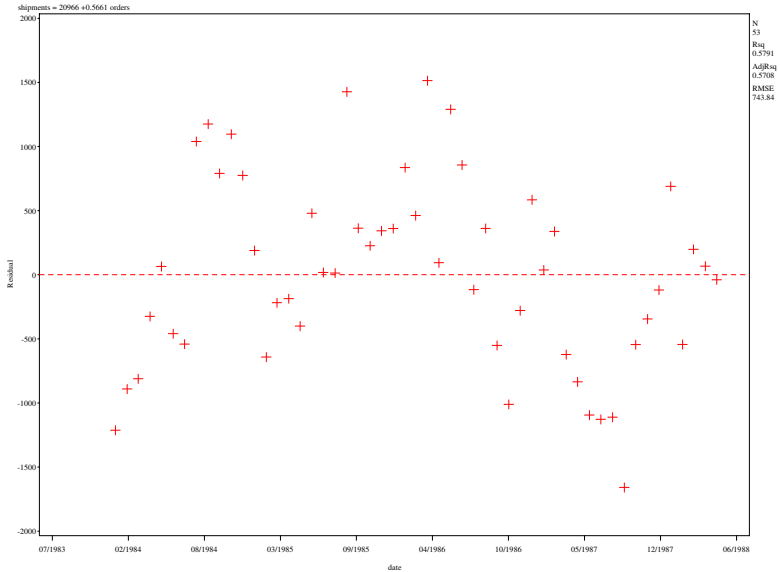
⇒ There is actually ***no* relationship between orders and shipments.**

Look at residuals another way:

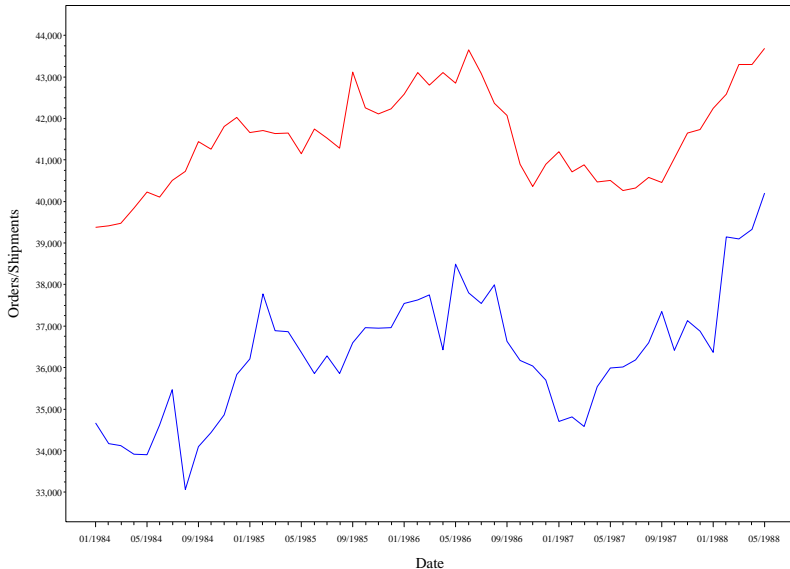
SAS Code

```
proc reg data=boiling;  
  var date;  
  model press = temp;  
  plot press*temp;  
  plot residual.*date;  
run;
```

Valve Orders vs Shipments



Valve Orders vs Shipments



Conclusions

- PROC REG is very useful and powerful.
- Don't lose sight of what you're actually doing.
- Check your assumptions.
- Look at time variables in the data set.

Further Resources



Sanford Weisberg.

Applied Linear Regression.

John Wiley and Sons, 1985.

UCLA Help:

`www.ats.ucla.edu/stat/sas/library/
SASReg_mf.htm`

Nate Derby: `http://nderby.org`

`nderby@sprodata.com`