

Getting Correct Results from PROC REG

Nate Derby

Stakana Analytics
Seattle, WA, USA

SUCCESS
3/12/15

Outline

- 1 PROC REG
 - Basics
 - Checking Assumptions
 - Understanding the Output

- 2 Conclusions

Basics

PROC REG = *Regression Analysis* done with SAS.

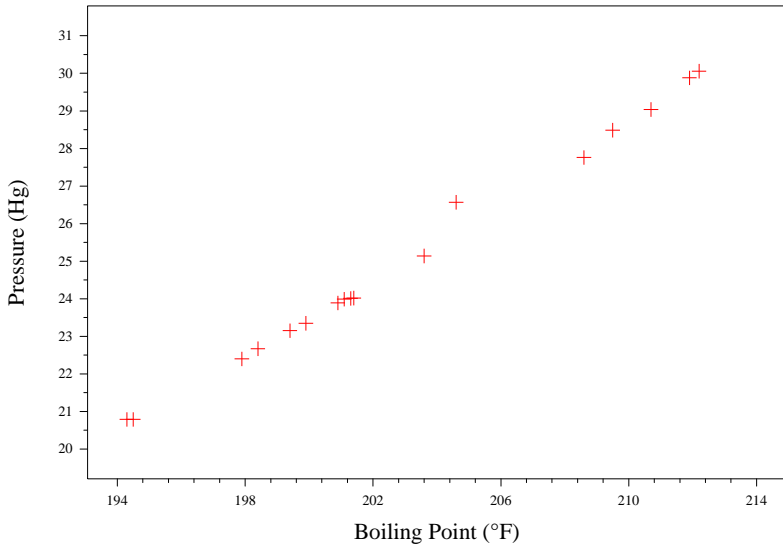
What is regression analysis?

- 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

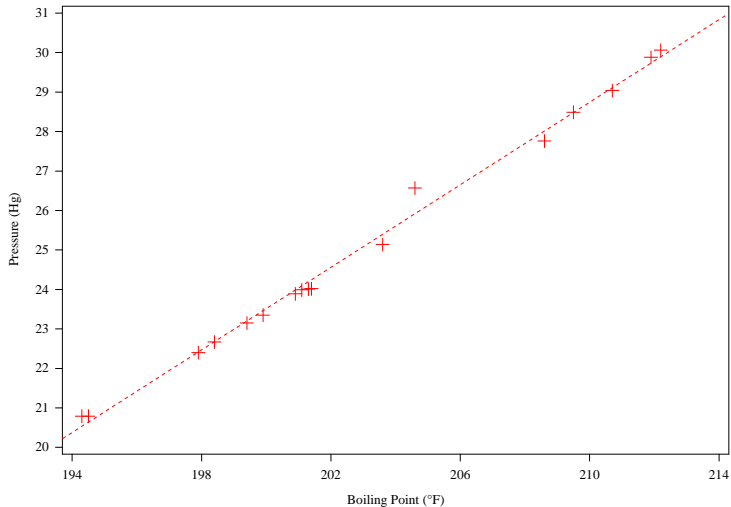
We want the line

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

SAS Code

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

Boiling Point vs Pressure



Checking Assumptions

Model must be appropriate for the data.

- Check mathematical assumptions of the model.
- Look at *residuals* = difference between a point and its fitted value (i.e., value on the line) [▶ Graph of Fitted Line](#)
 - **Do they form a pattern?** (Should be **NO**)
 - **Do they fit a normal distribution?** (Should be **YES**)
 - First one above more important than second.
- If assumptions above are violated, results could be false, **possibly to the point of being completely misleading.**

Checking for Residual Patterns

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

- Residual = What's left over after the modeled part.
 - ▶ Graph of Fitted Line
- We assume all patterns accounted for by the model.

Examples of patterns:

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

Checking for Residual Patterns: SAS Code

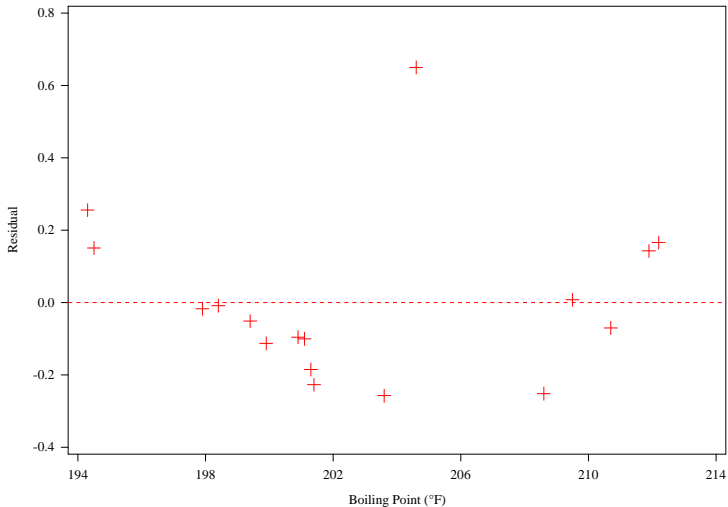
In General

```
proc reg data=blah;  
  model yyy = xxx;  
  plot residual.*xxx;  
  plot residual.*yyy;  
  plot residual.*predicted.;  
run;
```

Forbes' Data

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

Boiling Point vs Model 1 Residual



Trouble in Paradise

Pattern: Clusters of negative residuals.

⇒ Assumption violation!

Two options:

- **Modify the data:** Transform one of the variables in the model.
- **Modify the model:** Change the linear equation in the `model` statement.
 - Add/substitute some variables in the model.

Modifying the Data

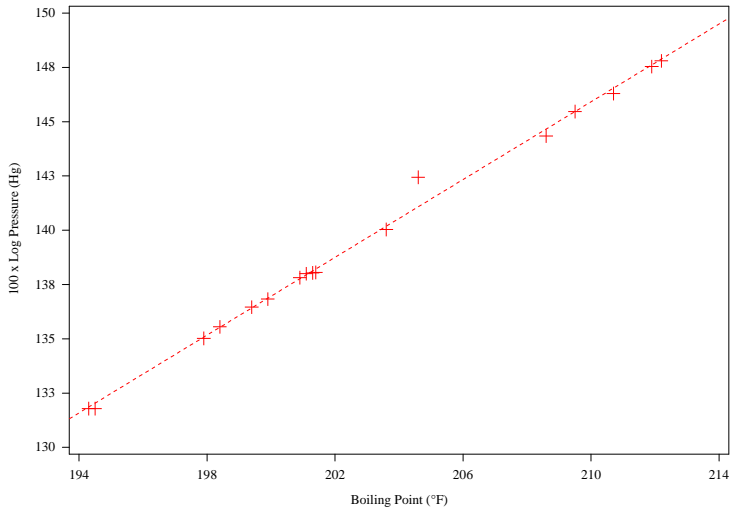
Pressure $\Rightarrow 100 \times \text{Log}(\text{Pressure})$:

$$100 \times \text{Log}(\text{ Pressure }) = \beta_0 + \beta_1 \text{Temperature} :$$

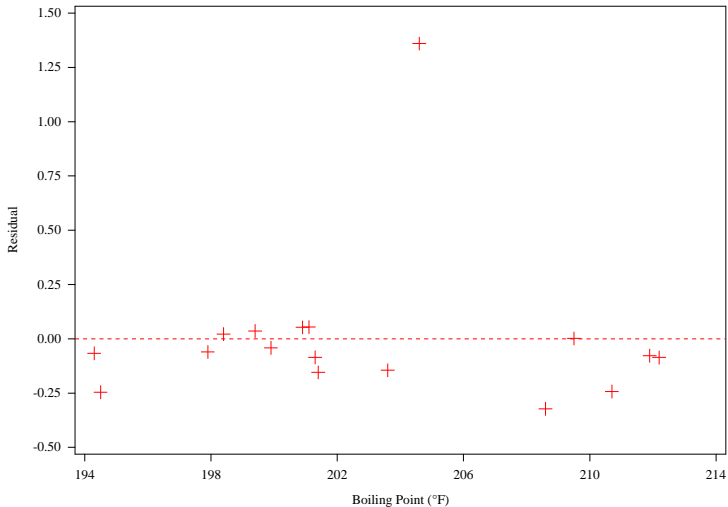
SAS Code

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

Boiling Point vs Log Pressure



Boiling Point vs Model 2 Residual



Checking for Residuals Fitting Normal Distribution

If residuals don't fit the *normal distribution* (bell curve), confidence intervals and hypothesis tests will be off.

- All other results (i.e., estimates) will be valid.

We check this via a *Quantile-Quantile Plot* (Q-Q Plot):

- Compares quantiles (percentiles) of residual distribution to those of standard normal distribution.
- **We want points to approximately fit a straight line.**

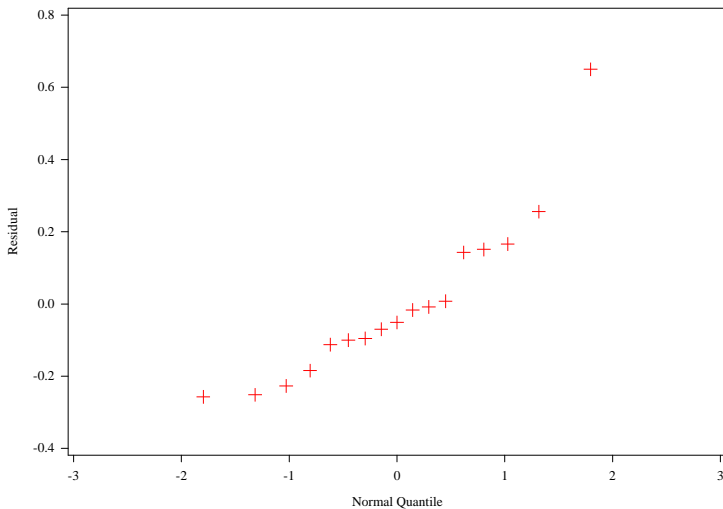
Checking for Residuals Fitting Normal Distribution

SAS Code

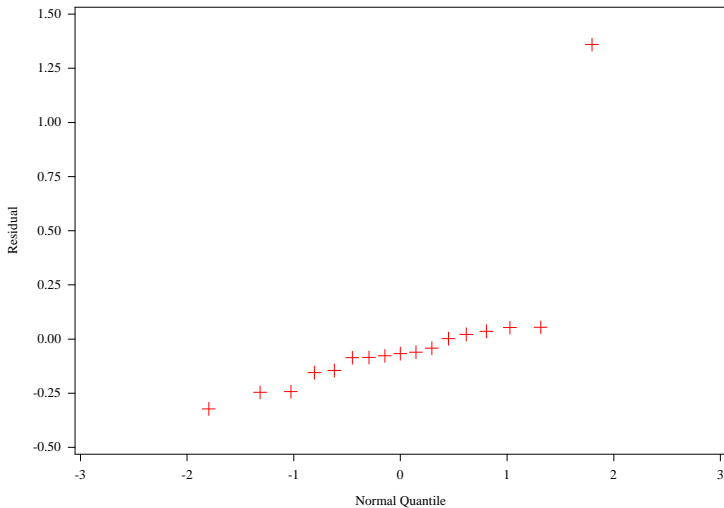
```
proc reg data=boiling noprint;  
  model press = temp;  
  plot residual.*nqq. / nostat nomodel noline;  
run;
```

```
proc reg data=boiling noprint;  
  model hlogpress = temp;  
  plot residual.*nqq. / nostat nomodel noline;  
run;
```


Model 1 Residuals vs Normal Quantiles



Model 2 Residuals vs Normal Quantiles



PROC REG Output: Forbes' Model 2

The REG Procedure
 Model: MODEL2
 Dependent Variable: hlogpress 100 x Log Pressure (Hg)

Number of Observations Read 17
 Number of Observations Used 17

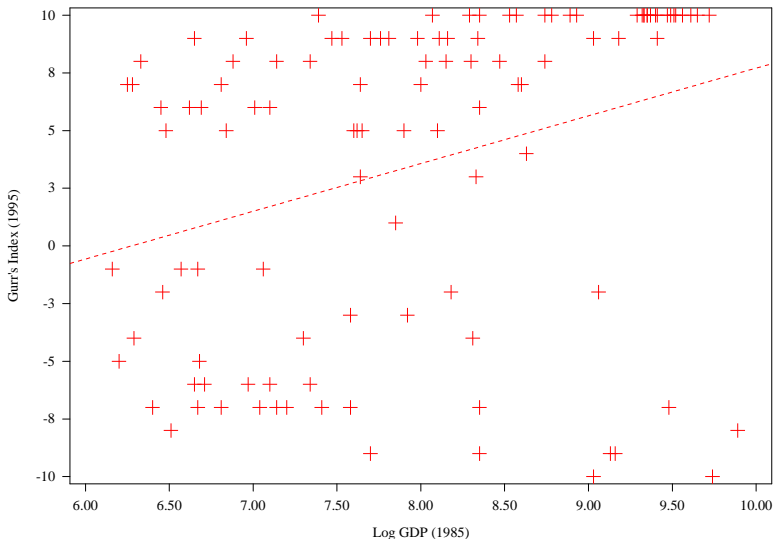
Analysis of Variance

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	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-42.13778	3.34020	-12.62	<.0001
temp	Boiling Point (F)	1	0.89549	0.01645	54.43	<.0001

log GDP vs Democracy Index



PROC REG Output: Democracy Index

The REG Procedure
 Model: MODEL1
 Dependent Variable: Gurr Index (1995)

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

Analysis of Variance

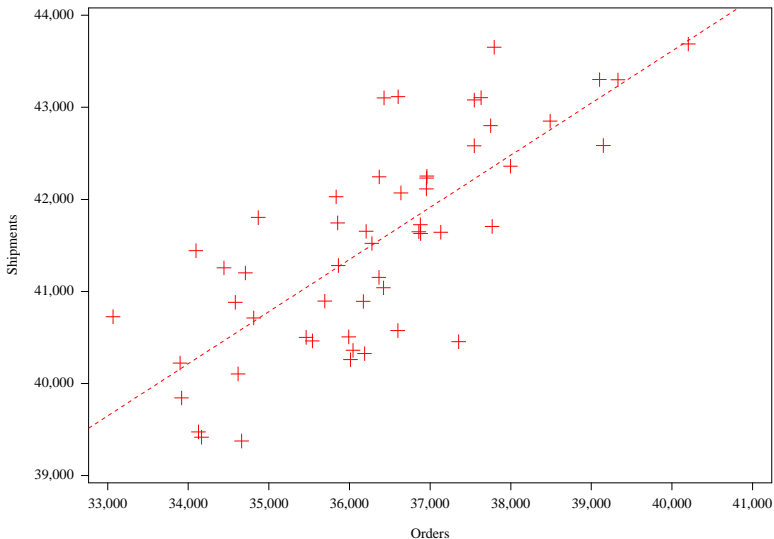
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	534.76792	534.76792	12.31	0.0007
Error	109	4734.97983	43.44018		
Corrected Total	110	5269.74775			

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

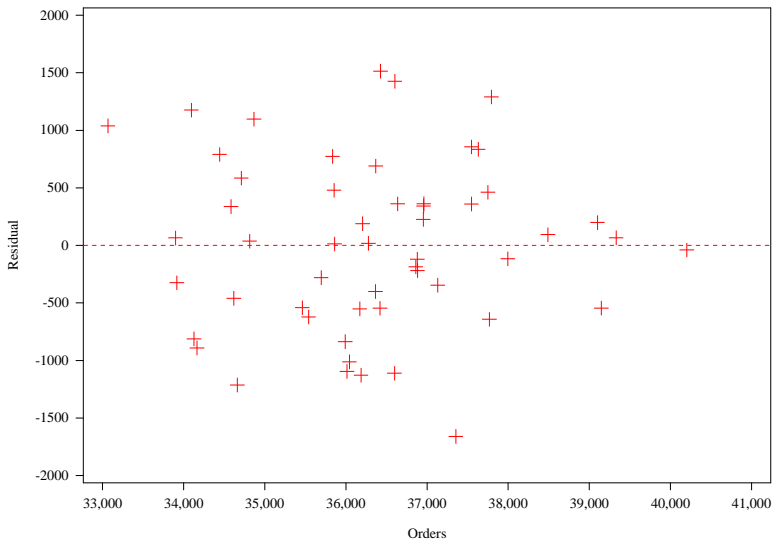
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-12.98347	4.74073	-2.74	0.0072
lgdp	Log GDP (1985)	1	2.06913	0.58973	3.51	0.0007

Valve Orders vs Shipments



Valve Orders vs Model 3 Residual



SAS Output

The REG Procedure
 Model: MODEL1
 Dependent Variable: shipments 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	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	20966	2456.79440	8.53	<.0001
orders	Orders	1	0.56613	0.06759	8.38	<.0001

Problems

Actually, **the conclusions are all false.**

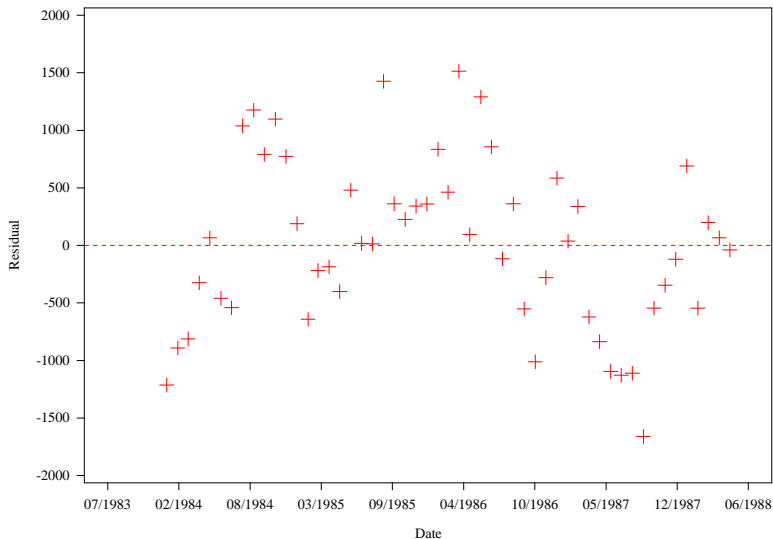
⇒ There is actually ***no* relationship between orders and shipments.**

Look at residuals another way:

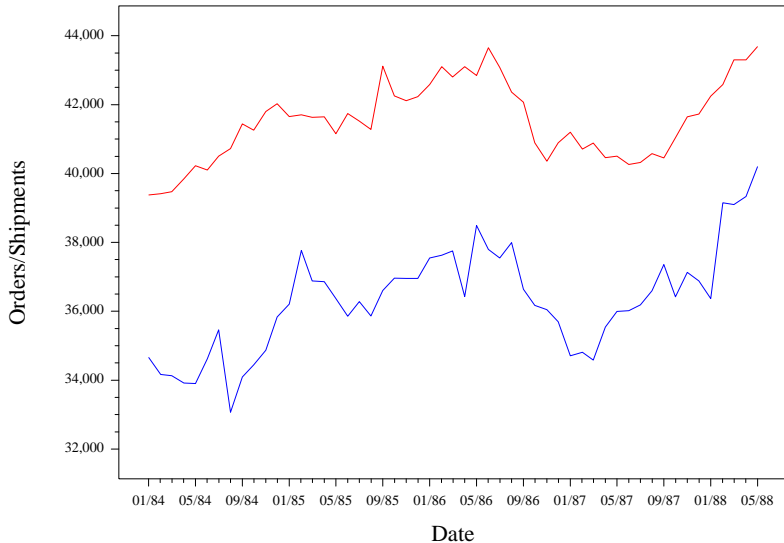
SAS Code

```
proc reg data=valves;  
  var date;  
  model shipments = orders;  
  plot residual.*date;  
run;
```

Date vs Model 3 Residual



Valve Orders vs Shipments



Conclusions

When fitting a model with PROC REG,

- Check the assumptions:
 - Is there a pattern with residuals vs other variables? **(NO)**
 - Do the residuals fit a bell curve? **(YES)**
 - For time series: Is there a pattern with residuals vs time? **(NO)**
- Look at results:
 - Is the R-squared value close to 1? **(YES)**
 - Are individual p -values less than 0.05? **(YES)**
 - Is the p -value for the analysis of variance less than 0.05?
(YES)

Further Resources



Sanford Weisberg.

Applied Linear Regression.

John Wiley and Sons, 1985.

UCLA Help:

`www.ats.ucla.edu/stat/sas/output/reg.htm`

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

`nderby@stakana.com`