Paper 270-2010

Getting Correct Results from PROC REG

Nathaniel Derby, Stakana Analytics, Seattle, WA

ABSTRACT

PROC REG, SAS®'s implementation of linear regression, is often used to fit a line without checking the underlying assumptions of the model or understanding the output. As a result, we can sometimes fit a line that is not appropriate for the data and get erroneous results. This paper gives a brief introduction to fitting a line with PROC REG, including assessing model assumptions and output to tell us if our results are valid. We then illustrate how one kind of data (time series data) can sometimes give us misleading results even when these model diagnostics appear to indicate that the results are correct. A simple method is proposed to avoid this.

Examples and SAS code are provided. The SAS/GRAPH® package is used in this paper but not required to use these methods, although SAS/STAT® is required to use PROC REG.

KEYWORDS: SAS, PROC REG, assumptions, residuals, time series.

All data sets and SAS code used in this paper are downloadable from `http://nderby.org/docs/SGF270-2010.zip`.

INTRODUCTION: PROC REG BASICS

PROC REG is SAS's implementation of *linear regression*, which is simply a mathematical algorithm to express a variable Y as a linear function of other variables $X_1, X_2, \ldots, X_n$. That is, if we know the variables $X_1, \ldots, X_n$, we want to estimate the variable Y as a linear function of those variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n.$$

which means we essentially have to estimate the values of the quantities (called *parameters*) $\beta_0, \beta_1, \ldots, \beta_n$. For simplicity, let's suppose that n=1, so that Y is estimated as a linear function of just one variable:

$$Y = \beta_0 + \beta_1 X_1.$$

Thus, if we have X_1, our estimate of Y is $\beta_0 + \beta_1 X_1$. β_0 and β_1 are the intercept and slope of the line. We determine the values β_0 and β_1 via a method called *least squares estimation*, which roughly means that we minimize the squared distance between each point and the line. For more details, see Weisberg (2005).

As an example, let's turn to data from a 19th century Scottish physicist, James D. Forbes, as explained in Forbes (1857) and Weisberg (2005, pp. 4-6). Forbes wanted to estimate altitude above sea level from the boiling point of water. Thus, he wanted to establish a relationship between the boiling point and air pressure (in Hg). The data are given in Figure 1. The question for linear regression to answer is,

What is the equation of the line that best fits the given data points?

Note that linear regression (and thus, PROC REG) is used only to establish a *linear* relationship. Since the data in Figure 1 look like they follow a line, a linear relationship is appropriate. However, since the points do not lie exactly on a line, it is impossible to put a straight line through all the data points. How do we even define a "best fit line"?

Via PROC REG, SAS computes these values for us, and can even graph the resulting line. Thus, for our example, we would like the equation

$$\text{Pressure} = \beta_0 + \beta_1 \times \text{Temperature} \tag{1}$$

The SAS code for this:

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

This gives us the output in Figure 2(a). Here we see the original data, plus the fitted line. That is, this is the line that best fits the data points that we have. However, there is a problem with this line, which can lead to false results.

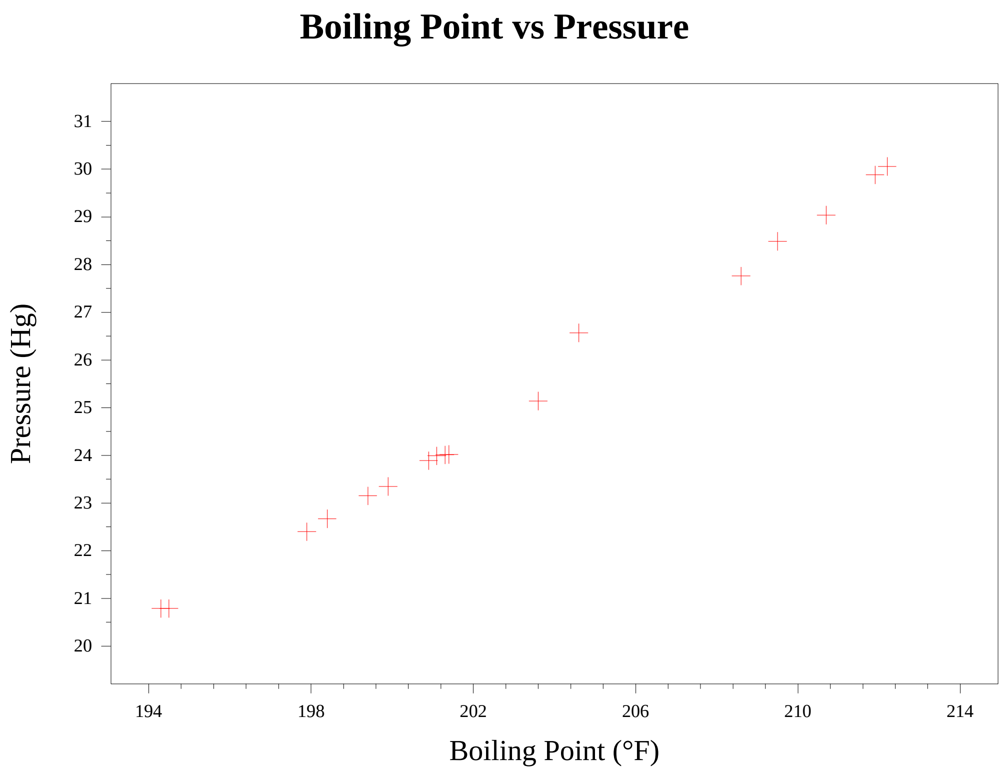


Figure 1: Scatterplot of Forbes' Data, originally from Forbes (1857) and described in detail in Weisberg (2005, pp. 4-6). This graph is generated from `PROC GPLOT`.

CHECKING ASSUMPTIONS

A *residual* is the difference between the point and its fitted value (i.e., its value on the line). A chief mathematical assumption of the estimation method for creating a line in linear regression and `PROC REG` is that these residuals are *completely random*. Therefore, the residuals should have *no pattern whatsoever*. If there is a pattern in the residuals, we have violated one of the central assumptions of the mathematical algorithm, and our results (which we shall see a little later) can be false – **possibly to the point of being completely misleading**.

A secondary assumption of our estimation method is that the residuals fit a normal distribution (the "bell curve"). This assumption isn't as necessary as the first one, about being completely random. If the residuals are random but do not fit a normal distribution, then some but not all our results will be invalid.

In summary, whenever we fit a model with `PROC REG`, there are two assumptions we must check:

- **Do the residuals form any kind of pattern whatsoever?** There a different patterns we should check.
- **Do the residuals fit a *normal distribution*?** In other words, when we put the residuals together into a data set, do they fit the standard bell curve?

Fortunately, both sets of assumptions can easily be checked via `PROC REG`.

CHECKING FOR RESIDUAL PATTERNS

When fitting a line, `PROC REG` creates some additional variables, which end with a period. They include `residual.` (containing the residuals) and `predicted.` (the fitted or predicted values). We basically want to look at plots of residual values versus various other values to look for patterns, which would indicate a lack of randomness. The main three variables that the residuals should be checked against are the *x* variable, the *y* variable, and the fitted value `predicted.`):

(a)

(b)

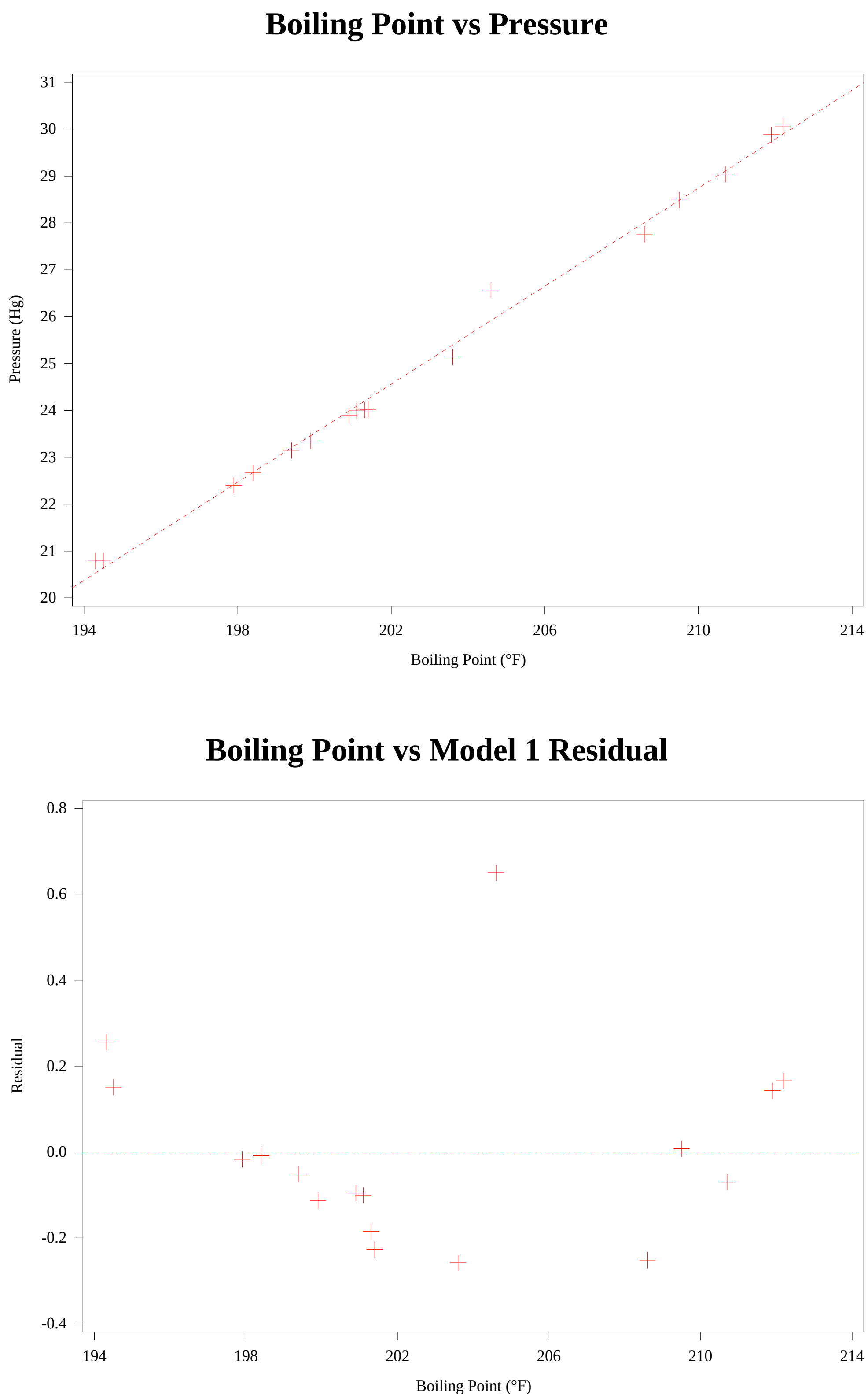


Figure 2: Scatterplot and regression line (a) and residual plot (b) of Forbes' Data, originally from Forbes (1857) and described in detail in Weisberg (2005, pp. 4-6).

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

Looking at these plots for any pattern whatsoever can provide valuable insights into whether the output is accurate or not. As an example, let's look at Forbes' temperature data against the residuals from the `PROC REG` model shown in Figure 2(a):

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

This gives us the output in Figure 2(b). Here we see a pattern: there are clusters of data points with negative residuals. In fact, the residuals form a rough concave curve, which is definitely a pattern. When an assumption like this fails, it is a sign that a straight line is an inappropriate model for the data. To deal with this situation, we must modify either the data or the model (i.e., the line we are trying to fit):

- **Modifying the data** entails transforming one or more of the variables and then using it in `PROC REG` in place of the original data.
- **Modifying the model** entails changing the linear equation, which means the `model` statement in `PROC REG`. That is, we add or substitute some variables in the model.

A proper discussion of how to modify the data or the model in different situations is outside the scope of this paper. In our case here, we shall substitute the pressure variable with the natural logarithm of the pressure variable, which is a common remedy for concave residuals:

```
proc reg data=boiling noprint;
  format hlogpress temp 4.;
  model hlogpress = temp;
  plot hlogpress*temp / haxis=( 194 to 214 by 4 ) nostat nomodel;
run;
```

This transformation gives us Figures 3(a)-(b). Here we see that the residuals vacillate between positive and negative values, which is indicative of random residuals. Therefore, this gives us a model that fits our data better. Note that our model is no longer that of equation (1). Instead, it is

$$\text{Log Pressure} = \beta_0 + \beta_1 \times \text{Temperature}. \quad (2)$$

CHECKING FOR RESIDUALS FITTING THE NORMAL DISTRIBUTION

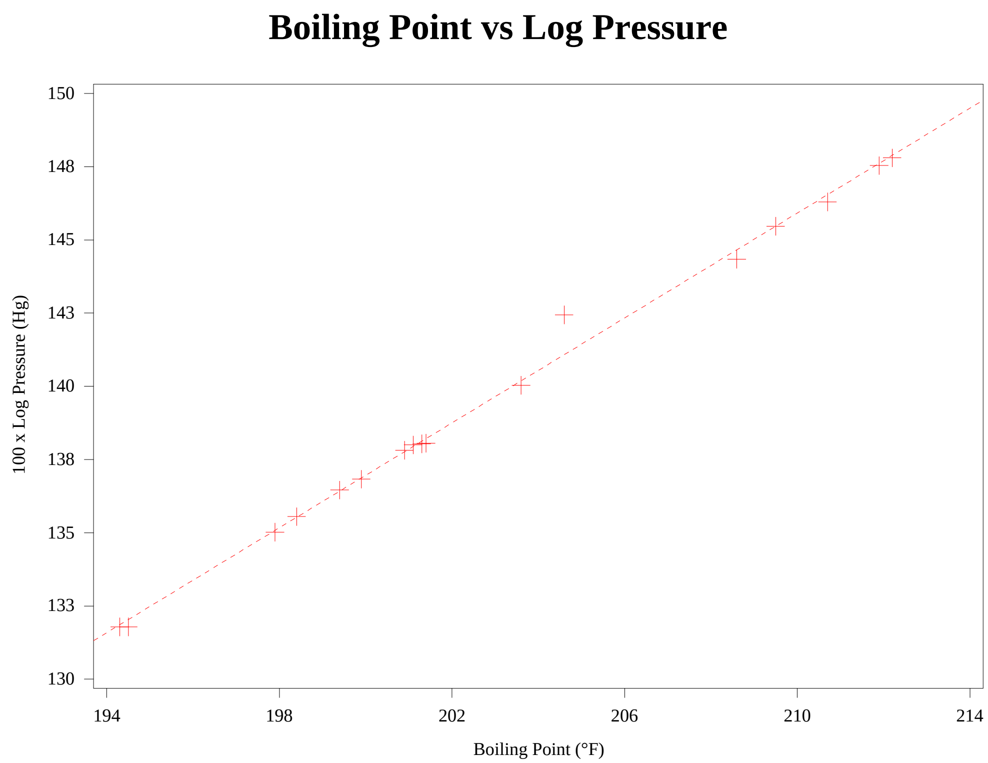
Here we want to test the assumption that the distribution of the residuals roughly matches the distribution of a normal distribution. We can do this via a *quantile-quantile plot*, or Q-Q plot.[1] We can generate these for each of our models above via the following code:

```
proc reg data=boiling noprint;
  format press temp 4.;
  model press = temp;
  plot residual.*nqq. / nostat nomodel noline;
run;

proc reg data=boiling noprint;
  format hlogpress temp 4.;
  model hlogpress = temp;
  plot residual.*nqq. / nostat nomodel noline;
run;
```

The output of these are given in Figure 4. Here the objective is to get the data points to line up in a straight line. This would indicate that the quantiles of the residuals are a linear function of the quantiles of a standard normal distribution, which is what we want. Q-Q plots can also be generated from `PROC UNIVARIATE`, which would provide a line to compare the points against – see UCLA (2009b) for an example. In Figure 4, we see that while roughly both models show a linear relationship, the Q-Q plot for model 2 (boiling point vs log pressure) has a stronger linear relationship.

[1] Strictly speaking, we can also do this via a probability-probability plot, or P-P plot, which compares empirical distribution functions. This is also available from `PROC REG` via the `npp.` variable, but Q-Q plots are generally more widely used.

(a)



(b)

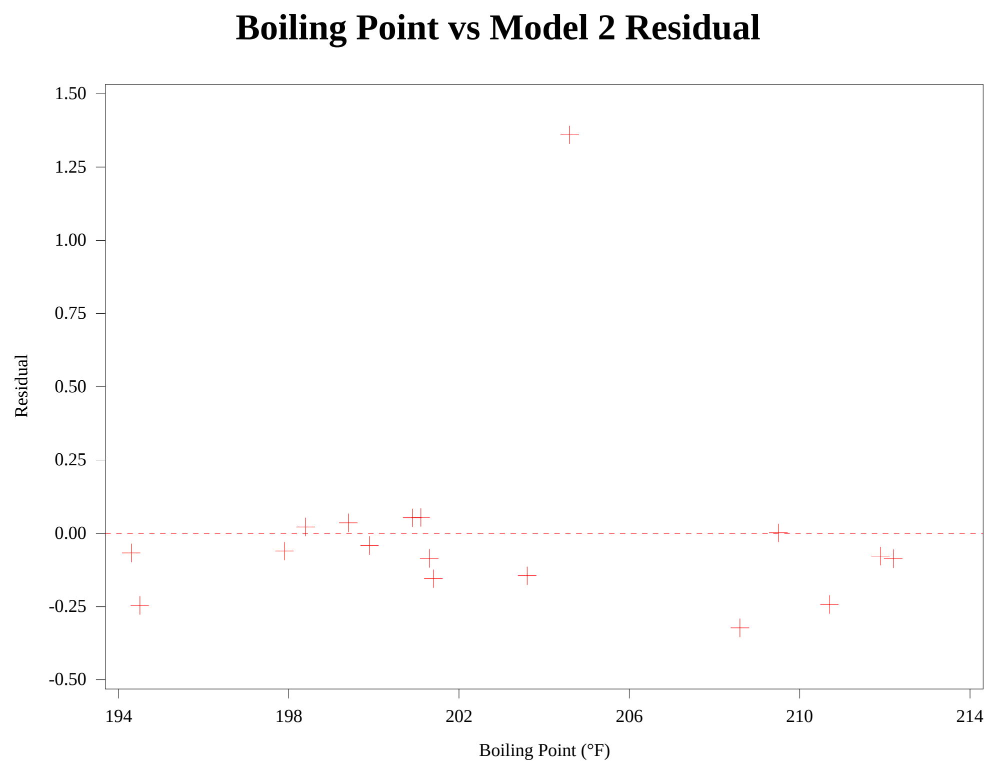
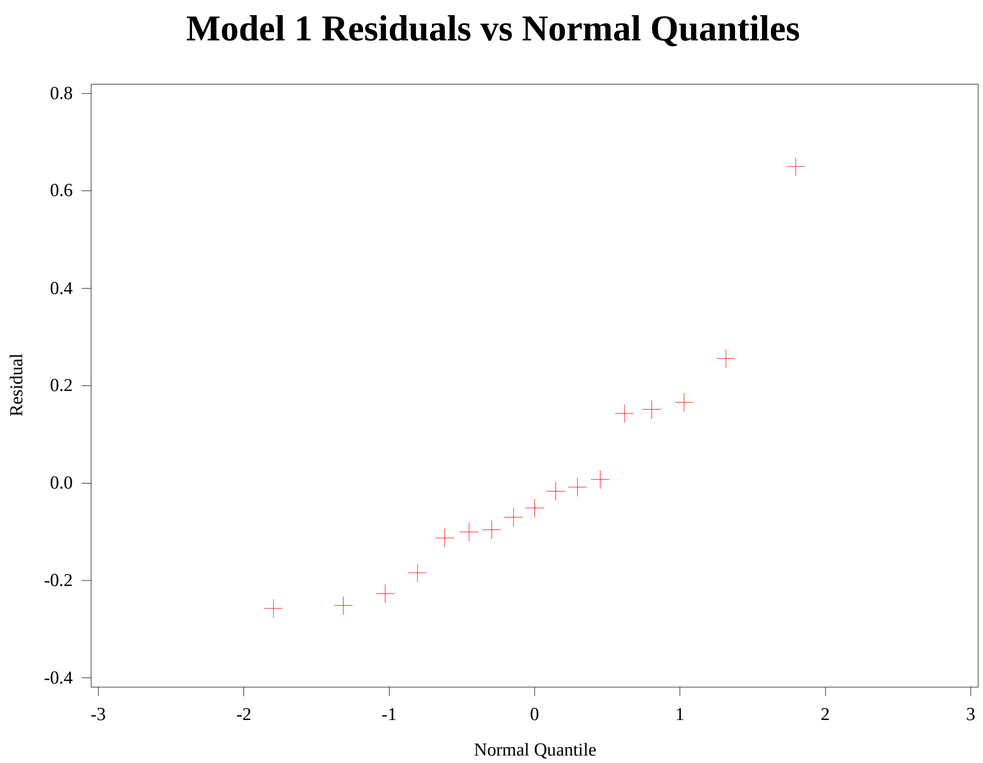


Figure 3: Scatterplot and regression line (a) and residual plot (b) of model 2 of Forbes' Data, originally from Forbes (1857) and described in detail in Weisberg (2005, pp. 4-6).

(a)



(b)

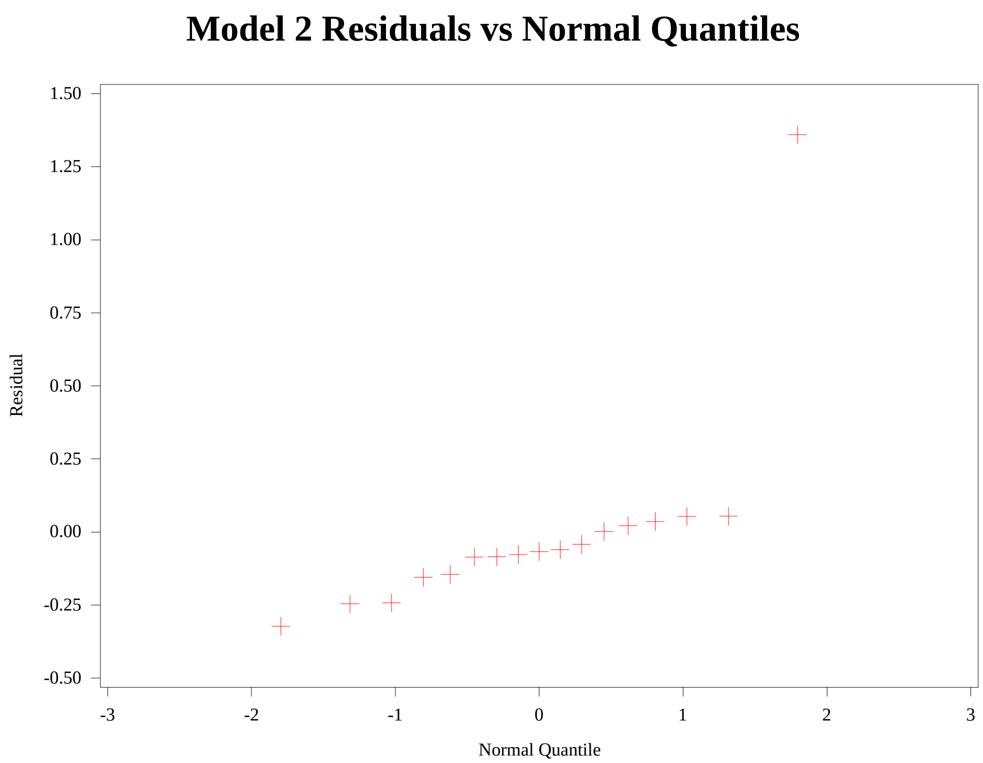


Figure 4: Q-Q Plots for model 1 (a: boiling point vs pressure) and model 2 (b: boiling point vs log pressure) of Forbes' data.

```
                              Boiling Point vs Log Pressure

                                   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

                                          Sum of           Mean
         Source                   DF     Squares         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

                                              Parameter       Standard
Variable      Label                     DF     Estimate          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
```

Figure 5: `PROC REG` Output from Model 2 of Forbes' data (boiling point vs log pressure).

PROC REG OUTPUT

The output from fitting model 2 of Forbes' data (boiling point vs log pressure) is shown in Figure 5. Note that **much of this output depends on the assumptions we checked earlier, and can be invalid if these assumptions are violated**. Here we first see that the title matches that of the graphs. The other items of this output are described as follows:

- `Number of Items Read`: The number of observations in our input data set.
- `Number of Items Used`: The number of observations used in fitting our model. Clearly missing values will not be used in the model. This basically tells us the number of observations with no missing values in any of the variables.
- `Parameter Estimates`: This gives the *parameters* of our model, which are the estimates of the values of β_0 and β_1, our coefficients in equation (2).
  - `Variable`: The name of the variable.
  - `Label`: The label of the variable.
  - `DF`: The *degrees of freedom*, which is a an internal variable usually of interest only to statisticians.
  - `Parameter Estimate`: Our estimate of the coefficient β_i.
  - `Standard Error`: Our estimate of how volatile our estimate of β_i is. The larger the standard error, the less reliable our `Parameter Estimate` is. In practice, this number becomes smaller as the data are less scattered, as we shall see in a further example.
  - `t Value`: Our test statistic for a *t-test*. This tests the hypothesis that our parameter is actually equal to zero.
  - `Pr > |t|`: Our *p-value*, which can be interpreted as the estimated probability that the parameter is actually equal to zero or further in the opposite direction from the estimate. For instance, if our estimate is positive, this is the estimated probability that our parameter is actually less than or equal to zero. If this number is below 5% (0.05), we usually consider it to be sufficiently different from zero.

- `Root MSE`: The *root mean squared error*, which is the square root of the average squared distance of a data point from the fitted line:

$$\texttt{Root MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

  where y_i is the i^{th} data point out of a total of n data points. This gives a measure of fit of the regression line to the data. We want this number to be small.

- `Dependent Mean`: This is simply the mean of the dependent variable, which in our model (2) is log pressure.

- `Coeff Var`: The *coefficient of variation*, which is simply the ratio of the root MSE to the mean of the dependent variable:

$$\texttt{Coeff Var} = \frac{\texttt{Root MSE}}{\texttt{Dependent Mean}}.$$

  It is used as a unitless measure of the variation of the data, which can be useful. For example, a mean variation of 100 is large if the mean of the dependent variable is 1, but very small if the mean is 10,000.

- `R-Square`: The *R-squared value* addresses the question: *What percentage of the variation in the data is due to the independent variable?* We want this number to be as close to 1 as possible.

- `Adj R-Sq`: The *adjusted R-squared value* has the same interpretation of the R-squared value, except that it adjusts for how many independent variables are included in the model. This is useful only when we have more than one independent variable.

- `Analysis of Variance`: All output in this section tests the hypothesis that *none* of the independent variables have an effect on the dependent variable. If we have only one independent variable, it is equivalent to `Pr > |t|` shown in the `Parameter Estimates` section. There will always be three (and only three) observations in this table.
  - `Source`: The source of variation of the data. Is it from the model (`Model`), random variations (`Error`), or total (`Corrected Total`)?
  - `DF`: The degrees of freedom – again, only of interest to a statistician.
  - `Sum of Squares`: An intermediate calculation only used in later columns.
  - `Mean Square`: An intermediate calculation equal to `Sum of Squares`/`DF`.
  - `F Value`: A calculation equal to the mean square of the model divided by the mean square of the error. This gives us our *test statistic*, which we shall test against the F-distribution.
  - `Pr > F` roughly gives us the probability that the coefficients are all equal to zero. We want this number to be very small (generally, below 5%).

Further descriptions and examples of these terms can be found in UCLA (2009a). Generally, when looking at the output, we look only at the following output:

- `R-Square` or `Adj R-Sq`: Are they close to 1? It is not a good model if this number is below 50%.
- `Parameter Estimates`: `Pr > |t|`: We want this column to be less than 5% for each variable.
- `Analysis of Variance`: `Pr > F`: We want this to be less than 5%. However, note that this variable is not as useful as it may seem – if we have a large number here, it means that *at least one* of the independent variables is not significant, but it doesn't tell us which one. Hence, the parameter estimates above is generally more useful.

Keep in mind, however, that **all output listed here can be misleading if our assumptions fail**. For an illustration of this, we can compare this output to that of another data set that is (much) different from Forbes'. Figure 6 shows a fitted regression line for a much different data set. Here, the X variable is log GDP (gross domestic product) of a country in 1985, and the Y variable is Gurr's democracy index in 1995. A linear regression can help us investigate whether the GDP affects how democratic a given country is. From this scatterplot and regression line, we see that the data are much more spread out than for Forbes' data, and that while a line has been fit to the data (indicating that as GDP increases, so does the democratic index), it appears that there actually is not a linear relationship between the two. Simply fitting a regression line to the data does not mean that a linear relationship is present.

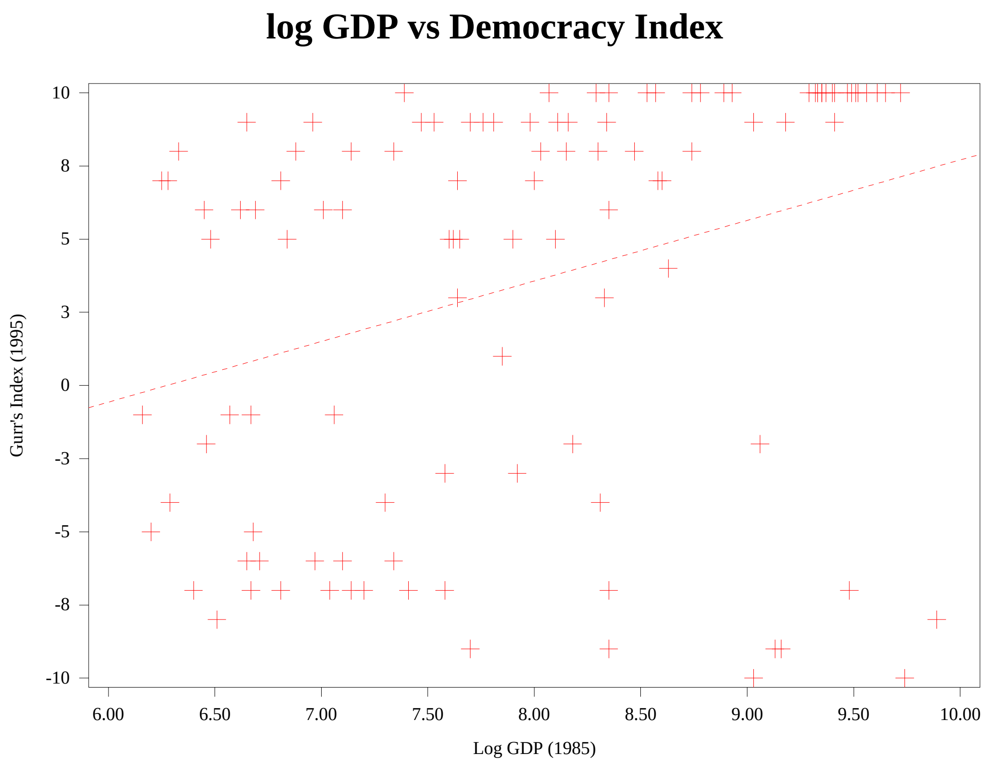


Figure 6: A fitted regression line for log GDP (gross domestic product) of a country in 1985 vs. Gurr's democracy index in 1995, from a data set provided to the author from private correspondence. Higher numbers of Gurr's index implies a higher level of democracy in a given country.

Figure 7 shows the `PROC REG` output for the democracy data. Comparing this to the output from Forbes' data in Figure 5, what can we deduce? We can simply compare the three main measures discussed earlier:

- `R-Square`: 99.5% of the variability of Forbes' data is explained by the independent variable, whereas only 10.15% of the variability of the democracy data is. This is a *major* different, and it definitely suggests that either the democracy data set is nonlinear, or (more likely) we will need more variables in our model. This is a reflection of how much the data are spread out from the line.
- `Parameter Estimates`: `Pr > |t|`: For the intercept term, this is very small for Forbes' data ($<$ 0.0001), but a little larger (0.0072, or 0.72%) for the democracy data set. This indicates that the line might go through the point (0,0). This is not a major concern. For the coefficient of the independent variable, Forbes' data give us $<$ 0.0001 again, whereas the democracy data has 0.0007. Again, this is not a large number, but certainly larger than for Forbes' data.
- `Analysis of Variance`: `Pr > F`: For Forbes' data, we have $<$ 0.0001, whereas the democracy data give us 0.0007. This is the same as the p-value for the independent variable as mentioned above, and illustrates a fact we saw earlier: These two values are the same for a model with only one independent variable. Once again, this is a reflection of how spread out the data are.

However, all of this output may be suspect, because checking the assumptions (not shown here) would reveal that there are doubts about the distribution of the residuals. For a better illustration of how we can get misleading results, we turn to an example with time-series data.

```
                          log GDP vs Democracy Index

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

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


                           Analysis of Variance

                                      Sum of          Mean
    Source                   DF      Squares        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

                                        Parameter       Standard
Variable     Label               DF      Estimate          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
```

Figure 7: `PROC REG` Output from the democracy data.

SPECIAL PROBLEMS WITH TIME-SERIES DATA

Generally, we never want to use `PROC REG` on time series data because there is a trend component that is not part of the model. We illustrate this with the data shown in Figure 8, as described in Pankratz (1991, pp. 2-3). This data is indexed by time, but for illustration we ignore the time component and act as though each data point is independent other ones. As shown in the output in Figures 8 and 9, everything looks fairly normal here in terms of assumptions (the Q-Q plot is not shown).[2]

In the `PROC REG` output in Figure 9, we see that there are no large *p*-values for either the analysis of variance or the parameter estimates, and we have an R^2 value of over 50%. We would normally accept this model, and its accompanying conclusion that valve orders are positively correlated with valve shipments (since the parameter for `Orders` is greater than zero).

In fact, **this conclusion is false, and valve orders are shipments are actually not correlated at all**. How is this possible? Our assumptions are that the residuals are *completely* random, which means that there is no pattern in them whatsoever. Indeed, as shown in the code on page 4, to be completely random, we want to look at graphs of different variables versus the residuals. One commonly overlooked variable to check against residuals is time. If we take a look at the residuals versus the time variable (date), as shown in Figure 10(a), we see a fairly obvious pattern. Indeed, the residuals are all negative for earlier and later dates, while they tend to be positive for dates in the middle of the range. This is indeed a pattern, and missing it can lead to the dramatically misleading results shown in Figure 9.

What is actually the case is that *both* the valve orders and shipments are driven by a trend, as shown in Figure 10(b). Any attempt to model the relationship between valve orders and shipments must also include a component to model the time trend. As shown in Pankratz (1991, pp. 191-192), once the trend is modeled and its effect removed from the valve orders and shipments, there is no direct relationship between the orders and shipments. Unfortunately, this cannot be done by `PROC REG`.[3]

[2]Technically we could be concerned that the variation in the higher end of the orders (the *x*-axis) is less than at the lower end, but in practice this is minor.

[3]It can be done in `PROC ARIMA`, which requires the ETS package and some skill at modeling the residuals. See Pankratz (1983), Pankratz (1991) or Brocklebank and Dickey (2003) for details.

In practice, this situation is typical of time-series data. That is, omitting a time component typically leads to misleading results such as in this example. Modeling time-series data properly requires incorporating a trend component into the model using a time series technique such as ARIMA (`PROC ARIMA`) or a state-space model.

CONCLUSIONS

In this paper we have covered the main ideas behind getting accurate results from `PROC REG`. When fitting a model with it,

- First check the assumptions:
  - Make a histogram of the residual values. Does it look like they fit a bell curve? (They should)
  - Make several plots of the residuals verses other quantities. Is there a pattern? (There shouldn't be)
  - If there is a time component, make a plot of residuals versus that time value. Is there a pattern? (There shouldn't be)
- Then take a look at the results:
  - Is the R-squared (or adjusted R-squared) value close to 1.00? (It should be)
  - Are the individual *p*-values less than 0.05? (They should be)
  - Is the *p*-value for the analysis of variance less than 0.05? (It should be)

Furthermore, it is generally ill advised to model time-series data with `PROC REG`, as it ignores the time component. However, it might be possible – just be sure to check the residuals against time to make sure that there is no discernible pattern.

REFERENCES

Brocklebank, J. C. and Dickey, D. A. (2003), *SAS for Forecasting Time Series*, second edn, SAS Institute, Inc., Cary, NC.

Forbes, J. D. (1857), Further experiments and remarks on the measurement of heights by the boiling point of water, *Transactions of the Royal Society of Edinburgh*, **21**, 135–143.

Pankratz, A. (1983), *Forecasting with Univariate Box-Jenkins Models*, John Wiley and Sons, Inc., New York.

Pankratz, A. (1991), *Forecasting with Dynamic Regression Models*, John Wiley and Sons, Inc., New York.

UCLA (2009a), Introduction to SAS: Annotated output of regression analysis, Academic Technology Services: Statistical Consulting Group.
`http://www.ats.ucla.edu/stat/sas/output/reg.htm`

UCLA (2009b), Regression with graphics by Lawrence Hamilton, Chapter 2: Bivariate regression analysis, Academic Technology Services: Statistical Consulting Group.
`http://www.ats.ucla.edu/stat/sas/examples/rwg/rwgsas2.htm`

Weisberg, S. (2005), *Applied Linear Regression*, third edn, John Wiley and Sons, Inc., New York.

ACKNOWLEDGMENTS

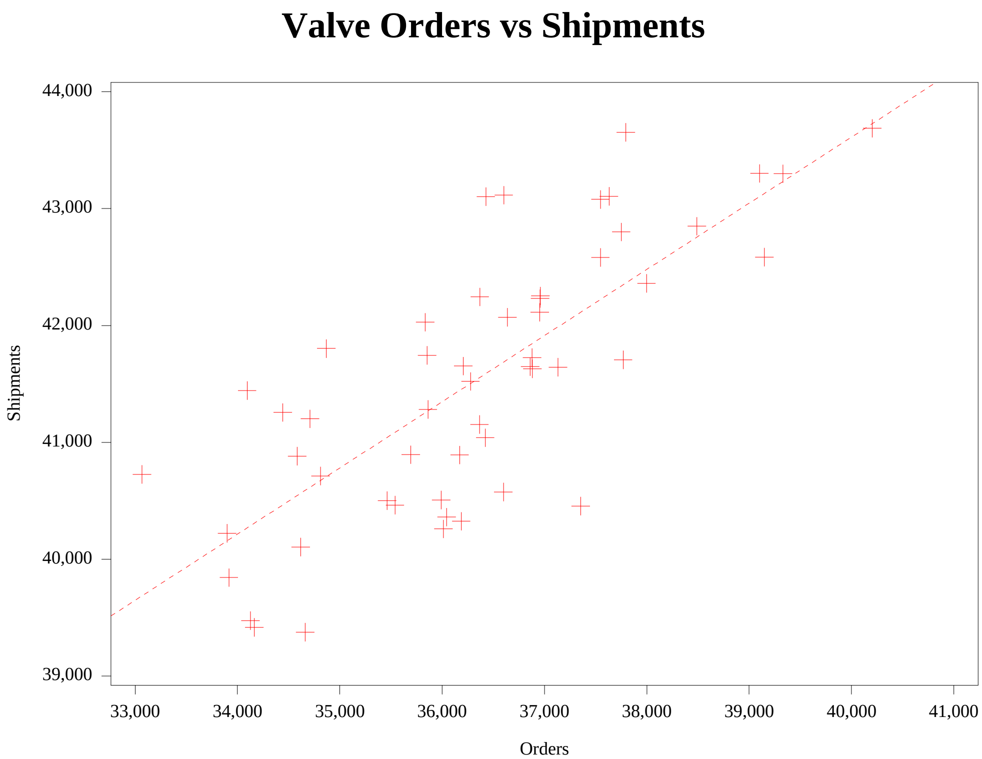
I thank Colleen McGahan and the rest of the executive committee of the Vancouver SAS Users Group for giving me the idea to present this topic in the fall of 2008. I thank Adrian Raftery of the University of Washington for giving me the democracy data set. Lastly, and most importantly, I thank Charles for his patience and support.

CONTACT INFORMATION

Comments and questions are valued and encouraged. Contact the author:

Nathaniel Derby
Stakana Analytics
815 First Ave., Suite 287
Seattle, WA 98104-1404
206-973-2403
`nderby@stakana.com`
`http://nderby.org`

(a)



(b)

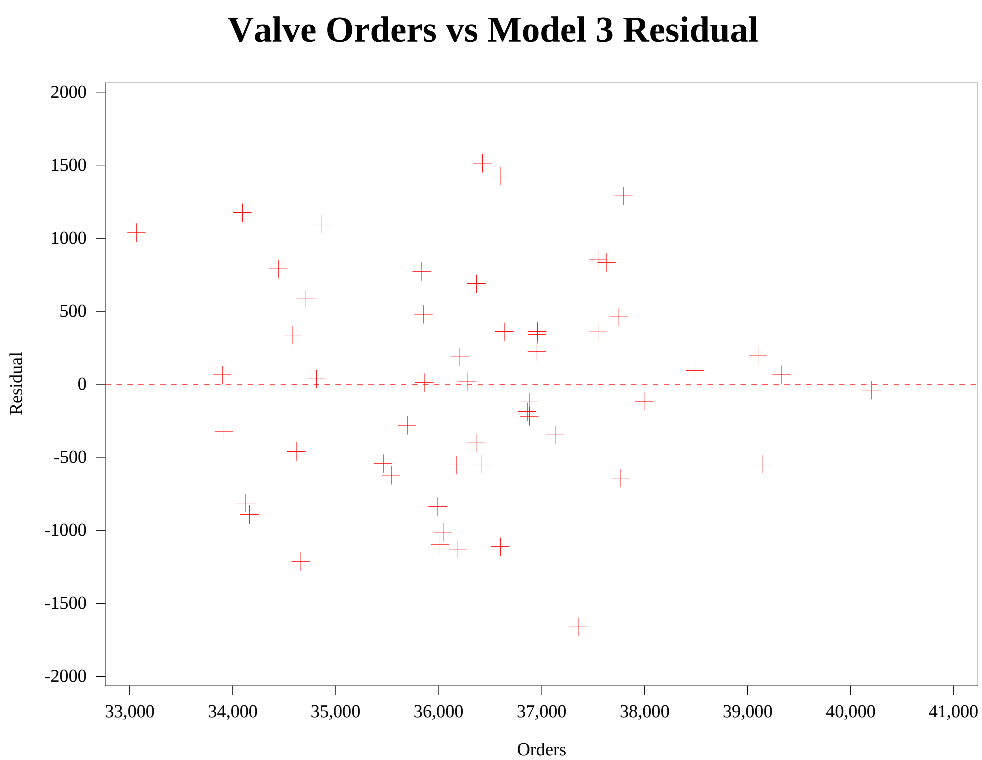
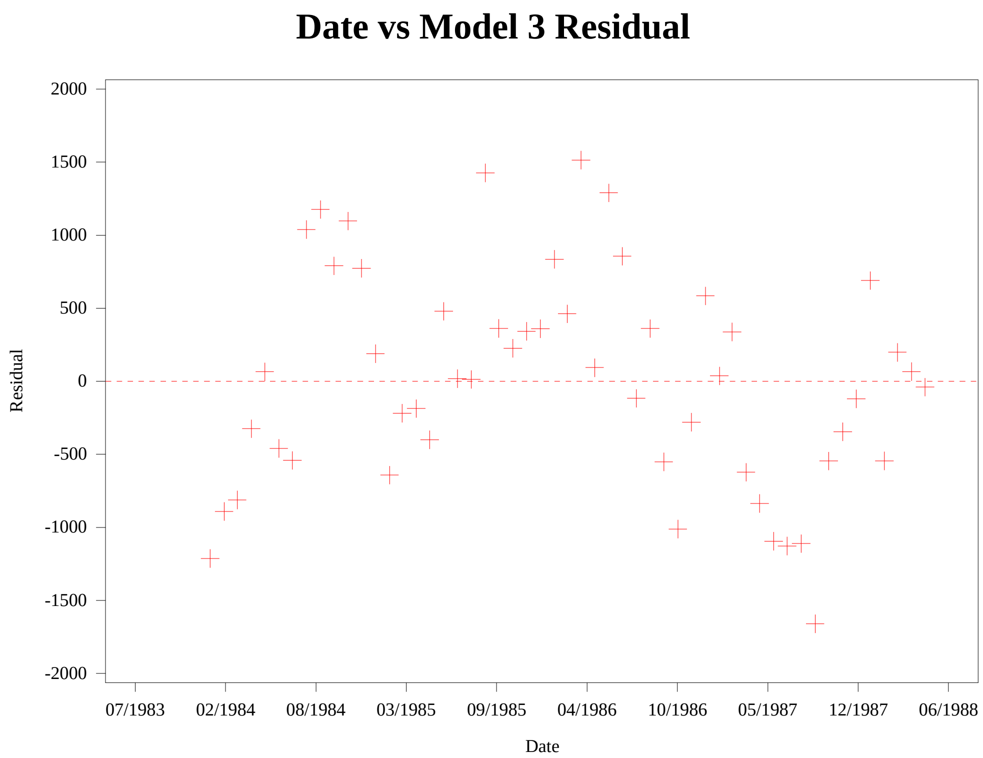


Figure 8: Data of value orders vs shipments, as described in Pankratz (1991, pp. 2-3). (a) is the data and fitted regression line, while (b) is a plot of the orders versus model residuals.

```
                           Valve Orders vs Shipments

                              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

                                    Sum of           Mean
  Source                  DF       Squares         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

                                    Parameter       Standard
Variable      Label          DF      Estimate          Error    t Value    Pr > |t|

Intercept     Intercept       1         20966     2456.79440       8.53      <.0001
orders        Orders          1       0.56613        0.06759       8.38      <.0001
```

Figure 9: `PROC REG` output from the valve order and shipment data.

(a)



(b)

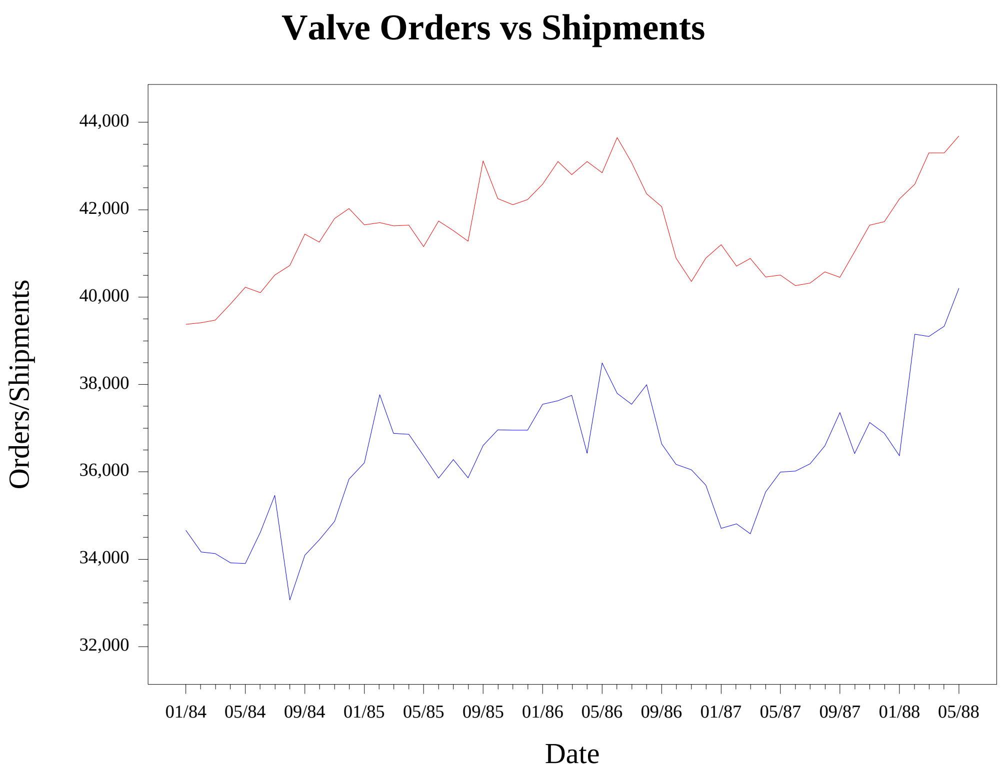


Figure 10: Plot of the data date vs the residuals (a) and of date vs both the orders and shipments (b) for the valve data.