Paper 1579-2015

The bookBot Ultimatum!

David A. Dickey, NC State University, John Vickery NC State University

ABSTRACT

The bookBot Identity: January 2013. With no memory of it from the past, students and faculty at NC State awake to find the Hunt Library just opened, and inside it, the mysterious and powerful bookBot. A true physical search engine, the bookBot, without thinking, relentlessly pursues, captures, and delivers to the patron any requested book (those things with paper pages—remember?) from the Hunt Library. The bookBot Supremacy: Some books were moved from the central campus library to the new Hunt Library. Did this decrease overall campus circulation or did the Hunt Library and its bookBot reign "supreme" in increasing circulation? The bookBot Ultimatum: To find out if the opening of the Hunt Library decreased or increased overall circulation. To address the bookBot Ultimatum, the Circulation Statistics Investigation (CSI) team uses the power of SAS® analytics to model library circulation before and after the opening of the Hunt Library. The bookBot Legacy: Join us for the adventure-filled story. Filled with excitement and mystery, this talk is bound to draw a much bigger crowd than had it been more honestly titled "Intervention Analysis for Library Data." Tools used are PROC ARIMA, PROC REG, and PROC SGPLOT.

INTRODUCTION

This paper takes the reader through a sequence of analyses of library circulation data before and after the opening of a new library at NC State University. This paper's name derives from a robotic book retrieval system that delivers the majority of volumes. A sequence of increasingly sophisticated analyses shows the effect of the library's opening on circulation. Three circulation data sets, total circulation and two subsets based on whether or not books were destined to be moved to the new James B. Hunt Jr. library on NC State's Centennial Campus or were destined to remain in the historic D. H. Hill library on the central campus. The "bookBot" and other futuristic features of the library have been nationally recognized and are drawing visitors to the new library, possibly inducing increased circulation of books. In addition to containing a tongue in cheek reference to a recently popular movie, the title refers to a request (not really an ultimatum) by library staff to statistically analyze the effect, if any, of the opening of this new library on circulation. Ultimately a time series intervention model is found to provide a good fit to the data.

MOTIVATION FOR THE STUDY

In January 2013, NC State opened the signature James B. Hunt Jr. Library shown in Figure 1.

Figure 1: NCSU Hunt Library, exterior view

The Hunt Library serves as a second main library with collections focusing on engineering, computer science, textiles, and related interdisciplinary research. NCSU's original main library, the D. H. Hill Library, houses collections across the sciences, humanities, and social science disciplines. The NCSU Libraries also include three branch libraries for veterinary medicine, design, and natural resources.

A defining feature of the new Hunt Library is the robotic book delivery system called the bookBot, shown in Figure 2. The bookBot requires substantially less space than traditional library shelving allowing for many more collaborative learning spaces for the NCSU community. Books in the bookBot are barcoded and automatically stored in more than 18,000 bins. While the majority of the collections at the Hunt Library are in the bookBot stacks, recent publications are on open shelving for browsing. The Hunt Library is located on NC State's Centennial Campus, a satellite campus, while the historic D.H. Hill is located on the central campus.

Figure 2: NCSU Hunt Library: yellow bookBot carrying a bin of books in the stacks. Each handle is connected to one of the over 18,000 bins.

With the novel approach to storing books in the bookBot stacks as well as the new reality of having much of the libraries' collections in two locations, librarians at NCSU wanted to investigate what, if any, impact the opening of the Hunt Library and bookBot has had on the use of circulating collections. By the summer of 2014, the library had enough post-Hunt circulation data for a time series intervention analysis.

Would curiosity and/or the appeal of this modern space cause a lift in circulation or would the fact that some volumes of interest had been moved to the other campus dissuade potential borrowers on the central campus and reduce circulation? Books can be ordered from either library to be delivered to the other one but then must be checked out rather than just browsed on a visit and left in the library. Might that increase circulation? Might the effect differ for books destined to move into the new Hunt Library versus those destined to stay at D. H. Hill?

DATA COLLECTION

Data required for this analysis brought up several challenges and highlighted the effectiveness of SAS. Data needed to be collected from several different sources and systems. These included a MySQL data warehouse, irregularly formatted text log files, SAS datasets and delimited data files.

The MySQL data warehouse contains transaction data from August 2012 to July 2014. Log files from the libraries' integrated library system were used to get data for Jan. 2011 through July 2012. The analysis also required that items be identifiable as either a Hunt Library / bookBot item or an item from the D.H. Hill library or one of the other branches. The additional data were added from previously generated SAS datasets and delimited data files of monthly snapshots from the integrated library system.

In order to import data from the irregularly formatted log files, perl regular expressions were used to extract the relevant transactions. In particular, PRXPARSE and CALL PRXSUBSTR, were used. For example, the following PRXPARSE function compiles an expression to match one or more A-Z characters following the data code of "^FE": `libid = prxparse("/\^FE[A-Z]+?\^/");`.

Each library item has a unique barcode that was used as the key variable for merges and SQL joins. The ABS function in combination with PROC SQL joins were used to match a particular item with its most likely item type (i.e. book, non-circulating journals, etc.) as of the date of the transaction. The ABS function was used to calculate the smallest date difference between an item's transaction and the corresponding snapshot dataset.

PROC FREQ was used to aggregate the combined and cleaned transactional data into a dataset with monthly circulation counts from January 2011 through June 2014. Finally, the necessary dummy variables for analysis with PROC ARIMA were created with the DATA step. The data set of combined campus wide circulation is divided into two parts by the books' ultimate destinations. Complete data on total circulation are available monthly from January 2011 through June 2014.

NAIVE ANALYSIS I

A basic staple of any beginning statistics class is the two sample t test. It assumes two normal populations with the same variance but possibly different means and the idea is to test to see if those two means are the same. Two independent random samples, one from each population, are assumed. A careless practitioner or a person with limited statistical tools at their disposal might apply the test whenever two samples are encountered without carefully checking assumptions. If the sample sizes are reasonably large (we have 24 observations before and 18 after the new library opened) then failure of the distributions to be normal is not much of a problem as a result of the central limit theorem. If the variances are different, then an approximation to the distribution of the t test is available. It uses Satterthwaite's formula. SAS PROC TTEST has a test of the null hypothesis that the variances are the same. This can be used as a guide for selecting between the usual ("pooled sample") t test and the approximation that is standard when there is evidence that the populations variances differ.

Data should not be analyzed without plots. Figure 3 shows the two distributions, one for the circulation data before the intervention (the name for an important event) in January 2013 and one for the data after the intervention. The data in Program 1 are only those books destined to be stored in the new library. The distributions look reasonably normal. Below those are the corresponding box plots. The program is set up to analyze any of the three data sets by changing the macro variable &Y.

Program 1: T-test procedure with ODS graphics:

```
ods html close;
ods listing;
ods listing gpath="%sysfunc(pathname(work))";
ods html gpath="%sysfunc(pathname(work))";
proc format;
  value change 0="BEFORE" 1="AFTER";
proc ttest data=all; var &y; class step;
  format step change.; run;
```

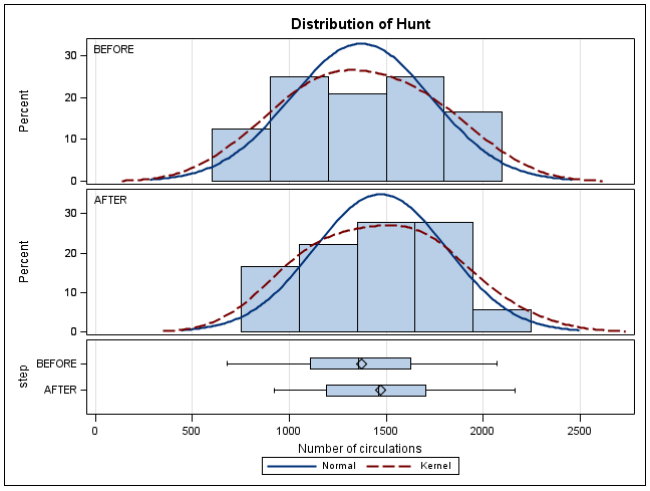

Figure 3: Distributional Analysis from PROC TTEST (for Hunt books only).

This graph is part of the automatically generated graphics from PROC TTEST. Such graphs are by default permanently stored on your computer. If you do a lot of statistical analysis in SAS, the accumulation of all graphs on your computer may be undesirable. Graphs can be sent to your WORK directory using an ODS GPATH command so that they will not be permanently stored. The first three lines of the Program 1 code turn off the HTML destination, turn on the LISTING destination, and send graphics to the work directory. For completeness, a GPATH statement is shown for each output destination. The macro variable &Y is set to Hunt, Hill, or Total depending on which data are of interest.

Both the before (Hunt destined books in DH Hill) and after data (same books now moved to Hunt) appear to be approximately normal. The automatic graphics also include a quantile-quantile plot which can be understood as a plot of what you got, versus what you would have gotten if the data were normal (centered and scaled to mean 0 variance 1) so if the plot looks like a straight line, then the assumption of normality is not drastically violated – you got, more or less, what you would expect under normality.

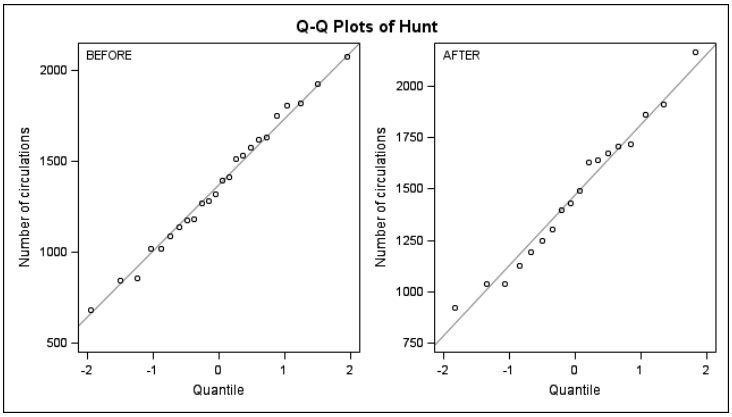

Figure 4: Q-Q plots for Hunt destined books before and after Hunt Library opened.

The Q-Q plots in Figure 4, before on the left and after on the right, agree with Figure 3 in showing that the data seem close to normal. Plots, not shown, for total circulation and for the books remaining at D.H. Hill do not look quite as good. In the data step there is a variable STEP which is 0 before the intervention and 1 from the intervention date, Jan 1, 2013, onwards. See the lower portion of Figure 3. The variable

STEP is known in statistics as a “dummy” or “indicator” variable. Output from PROC TTEST shows that there is no difference in the two means regardless of whether the two variances are assumed equal.

Program 2: T-test computations:

```
proc format;
  value change 0="before" 1="after";
proc ttest data=all;
  var &y; class step;
  format step change.; run;
```

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	40	-0.90	0.3749
Satterthwaite	Unequal	37.897	-0.91	0.3710

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	23	17	1.13	0.8091

Following those t tests is a test of the hypothesis that the two variances are the same. It is labelled as a “folded F test,” a name derived from the nature of the test. The alternative hypothesis is that the variances are unequal (a two sided alternative) but the test divides the larger mean square by the smaller and compares to the right tail of F resulting in what is essentially a test for the two sided alternative that uses only one tail of the F distribution. With p-value 0.8091, the test is consistent with the hypothesis of equal population variances and allows the use of the pooled t test for the two means.

As a practical point, some experts suggest that when using a preliminary test like this to decide which of two subsequent test methods to use for the hypothesis test of main interest, one should do the preliminary test at the 0.25 rather than 0.05 level.

The t test (t=-0.90, p=0.3749) shows no evidence of a change in mean circulation of this subset of books before versus after they were moved to the Hunt library. The pooled variance two sample t test results can alternatively be obtained by regression of circulation on the dummy variable STEP. This will be especially helpful here because the technique extends nicely to more complicated models.

Program 3: T-test obtained with PROC REG:

```
proc reg data=all;
model hunt=step; run;
```

Here is part of the output:

Variable	Label	Parameter DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1371.12500	72.45203	18.92	<.0001
step		1	99.31944	110.67231	0.90	0.3749

The equation predictions are obtained by adding 99 times the dummy variable STEP to 1371. Because STEP is 0 before 2013 and 1 after that, the two means are 1371 and 1371+99=1470. Therefore 99 is the difference in two means. Its t test is 0.90, and its p-value is 0.3749, which exactly matches that of the

equal variances t test from PROC TTEST. The "root MSE," not shown, is 355 and is an estimate of the standard deviation of the residuals.

The naïve analyst might stop at this point for the Hunt data, however, despite the seemingly normal distributions and equal variances, the t test is not appropriate here. The problem is that these two samples are not random samples but are really the first and last parts of a time series. There are two samples of course, but this is not a case where treatment units are randomly assigned to two treatments. You would still be OK *if* the two means model were appropriate and the deviations from the two means were serially independent but with time series like this, it is unsafe to assume this without checking. Furthermore, the same books are in the pre and post intervention subsets. The two subsets are not independent which violates an underlying assumption for PROC TTEST.

Before using a slightly better approach, one more comment on the change in default output destinations might be of interest. With the ODS LISTING command, it is possible to go to the KEYS window and define a key (for example F12) as follows:

```
SUBMIT "QUIT"; CLEAR OUTPUT; CLEAR LOG; SUBMIT;
```

Using F12, rather than the running person icon, to submit your SAS program causes the LOG and OUTPUT windows to first be cleared. There is no more searching through an accumulation of LOG and OUTPUT content to find that of your most recent submission.

A SLIGHTLY LESS NAÏVE ANALYSIS

A fundamental mistake was made in the previous analysis. When data are taken over time, a time plot of the data should be made. Again using the Hunt destined book data, Figure 5 shows two such plots with the intervention date indicated.

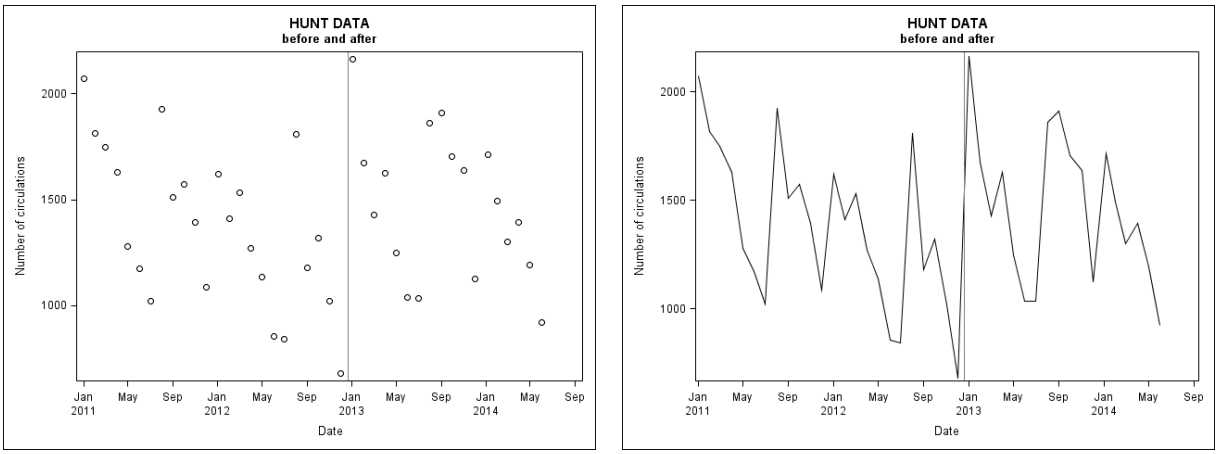

Figure 5: Results of scatter (left) and series (right) statements in PROC SGPLOT.

The left plot suggests a random scatter around a mean. This is not the type of plot that should be made for time series data. The plot on the right reveals a somewhat obvious repeating seasonal pattern of duration about 5 months, corresponding to academic semesters, that is not at all clear in the scatter plot on the left.

Just as a dummy variable allowed a shift in mean when the Hunt library opened, a January dummy variable, MON1 for example, can measure a repeating January effect. MON1 is created using 1 in

January and 0 elsewhere. It is used to shift the January mean away from the mean of all other months. With a January and a February variable, you have shifts of the means of each of those two months from the mean of the remaining months. Up to 11 monthly shifts can be accommodated with the intercept representing the mean of the remaining month(s). There must be at least one month without a dummy variable when an intercept is used. Note that the same principle was in force when you used one dummy variable to compare two means, one before and one after the intervention. Here is a regression that includes dummy variables for all months except December. The monthly dummy variable names are MON1 through MON11. The MON1 (January) variable's coefficient 913.82 is estimating the January minus December effect difference, the MON2 coefficient estimates the February minus December difference, etc.

We can combine the level shift dummy variable for the intervention with the intercept and seasonal dummy variables to get a model that captures the seasonality and any level shift in circulation for the books destined for the new Hunt library.

Program 4: Accounting for seasonality and the January 2013 intervention.

```
proc reg data=all;
  model hunt=step mon1 - mon11;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	934.13333	119.00485	7.85	<.0001
STEP	1	88.60000	64.12171	1.38	0.1776
MON1	1	913.81667	155.23693	5.89	<.0001
MON2	1	619.81667	155.23693	3.99	0.0004
MON3	1	524.56667	155.23693	3.38	0.0021
MON4	1	502.31667	155.23693	3.24	0.0030
MON5	1	235.56667	155.23693	1.52	0.1400
MON6	1	18.81667	155.23693	0.12	0.9044
MON7	1	2.66667	165.56154	0.02	0.9873
MON8	1	900.66667	165.56154	5.44	<.0001
MON9	1	569.33333	165.56154	3.44	0.0018
MON10	1	569.33333	165.56154	3.44	0.0018
MON11	1	387.00000	165.56154	2.34	0.0265

The relevant part of the output is the coefficient 88.6 of STEP indicating an increase in circulation for these books but it is small and statistically not significantly different from 0. The addition of the seasonal dummy variables reduced the root MSE substantially: from 355 to 203. The same analyses for the books destined to stay in the D. H. Hill library and for the total circulation are shown below.

The Hill data show a statistically significant decrease in circulation of 931 volumes per month (p=0.0001) after the intervention when the model with seasonal dummy variables and the STEP variable is used.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5158.72222	394.92809	13.06	<.0001
STEP	**1**	**-931.16667**	**212.79354**	**-4.38**	**0.000**1
MON1	1	3411.61111	515.16744	6.62	<.0001

MON2	1	3648.86111	515.16744	7.08	<.0001
MON3	1	3169.11111	515.16744	6.15	<.0001
MON4	1	4595.61111	515.16744	8.92	<.0001
MON5	1	268.11111	515.16744	0.52	0.6067
MON6	1	-399.63889	515.16744	-0.78	0.4442
MON7	1	-202.66667	549.43055	-0.37	0.7149
MON8	1	2620.33333	549.43055	4.77	<.0001
MON9	1	3579.66667	549.43055	6.52	<.0001
MON10	1	4338.33333	549.43055	7.90	<.0001
MON11	1	4357.00000	549.43055	7.93	<.0001

Not surprisingly, the total circulation shows a significant drop in circulation as well. The drop -842.6 is the small increase 88.6 in circulation for the Hunt destined books minus the large decrease 931.2 in the Hill book circulation.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6092.85556	492.08632	12.38	<.0001
STEP	**1**	**-842.56667**	**265.14393**	**-3.18**	**0.0035**
MON1	1	4325.42778	641.90635	6.74	<.0001
MON2	1	4268.67778	641.90635	6.65	<.0001
MON3	1	3693.67778	641.90635	5.75	<.0001
MON4	1	5097.92778	641.90635	7.94	<.0001
MON5	1	503.67778	641.90635	0.78	0.4390
MON6	1	-380.82222	641.90635	-0.59	0.5576
MON7	1	-200.00000	684.59870	-0.29	0.7723
MON8	1	3521.00000	684.59870	5.14	<.0001
MON9	1	4149.00000	684.59870	6.06	<.0001
MON10	1	4907.66667	684.59870	7.17	<.0001
MON11	1	4744.00000	684.59870	6.93	<.0001

The diagnostics for these models seem reasonable and again, you might be content to report the results of these analyses but you can still do better. Before leaving this section, there is a twist on the seasonal dummy variable approach that is worth mentioning. As it stands, the intercept represents the December circulation level with each dummy variable coefficient representing a difference between the associated month and December. For the total circulation data above, the estimated long run average December circulation was 6093 before the intervention and 6093-843 after. Horizontal lines at these two heights could be plotted over time to show the effect of the new library, but why should it be located at the December level? Why not at the average seasonal level? In fact this can be accomplished.

Recall that the deviations of any set of numbers from their mean will sum to 0. If you can reparameterize your model so that the 12 monthly effects sum to 0 then the intercept will represent the (pre intervention) level for the average monthly effect. Adding the step coefficient will give the corresponding centered mean after the intervention. For 12 numbers that sum to 0, if you have the first 11 numbers then setting the 12^{th} to the negative of the sum of these other 11 will give an overall sum 0. This can easily be accomplished by adjusting your 11 monthly dummy variables. If you adjust MON1 through MON11 by entering -1 instead of 0 in December, then for each December, the effect will be modeled not as 0 but rather as -1 times the sum of the coefficients for the other 11 months. A MON1 coefficient 1439.575, for example, tells us that the January circulation is 1439.575 more than the average monthly circulation. . With those new variables MON1-MON11 you get this output from PROC REG for the total circulation data:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8978.70833	171.14967	52.46	<.0001
STEP	1	-842.56667	265.14393	-3.18	0.0035

MON1	1	1439.57500	405.01405	3.55	0.0013
MON2	1	1382.82500	405.01405	3.41	0.0019
MON3	1	807.82500	405.01405	1.99	0.0556
MON4	1	2212.07500	405.01405	5.46	<.0001
MON5	1	-2382.17500	405.01405	-5.88	<.0001
MON6	1	-3266.67500	405.01405	-8.07	<.0001
MON7	1	-3085.85278	461.36399	-6.69	<.0001
MON8	1	635.14722	461.36399	1.38	0.1791
MON9	1	1263.14722	461.36399	2.74	0.0105
MON10	1	2021.81389	461.36399	4.38	0.0001
MON11	1	1858.14722	461.36399	4.03	0.0004

The intercept now has changed from 6093 to 8979 and represents the (pre intervention) mean circulation over all months. The large increase is not surprising in light of the Christmas break at the university. December has much lower circulation than the average month. A nice advantage of this approach is that a graph of Y=8979 – 843*STEP overlaid on a plot of the data and predicted values will show the two nicely centered mean circulation levels. Figure 6 shows such a graph for the total circulation and Figure 7 is the graph for the two component data sets. Future values of the step and seasonal dummy variables in the data set cause predictions for two years ahead to appear in the graphs.

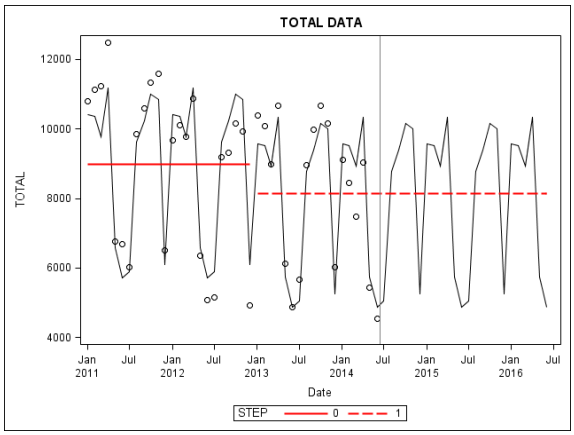

Figure 6: Data and predictions for total circulation before and after the intervention.

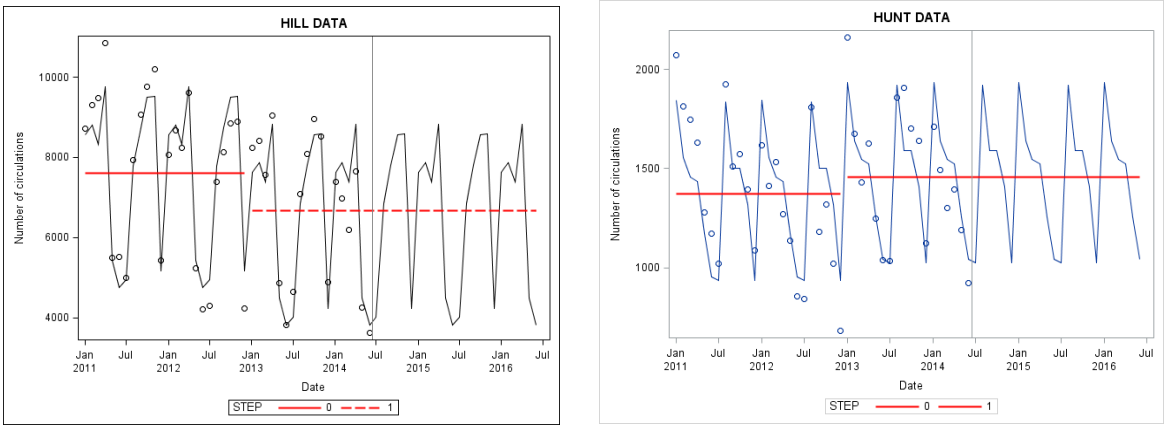

Figure 7: Data and predictions for Hill (left) and Hunt (right) data subsets.

It is worth noting that these results are disappointing. It is disappointing that the overall effect of this beautiful new library is a decrease in circulation overall – or is it? Perhaps we're not done yet.

For time series, plotting the residuals over time is generally a good idea. Figure 8 shows residual plots for the Hill and Hunt circulation data. A vertical reference line at January 1, 2013 marks the date the new Hunt library opened and to the left of that line you see a fairly steady linear decrease in circulation. One possible reason for this is the increasing ease with which electronic resources can be accessed on line from anywhere on campus. After the intervention for the Hunt data and possibly for the Hill and total data, (the total data plot resembles that of the Hill data) you see a somewhat linear decline with slope visually near that of the data before the intervention.

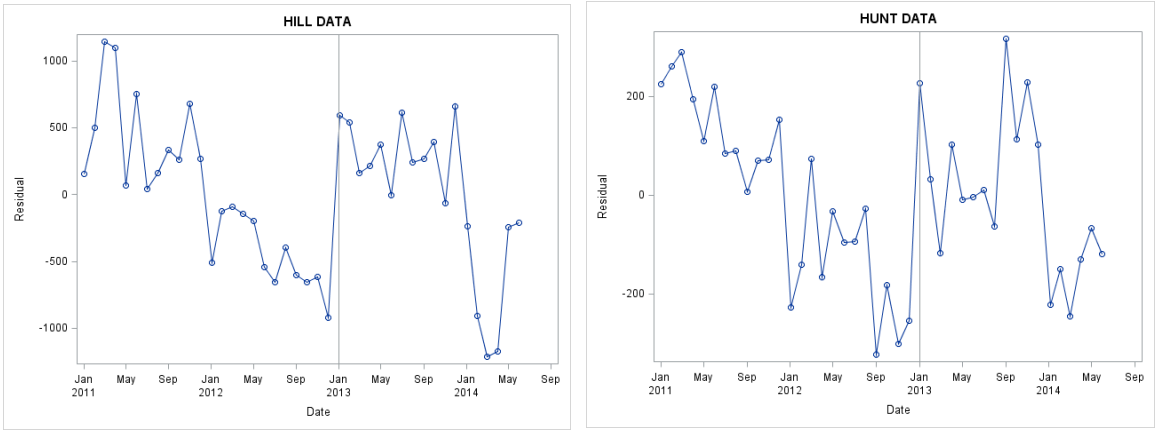

Figure 8: Residuals from step function plus seasonal dummy models for subcollections.

A TIME SERIES INTERVENTION APPROACH

Based on the results thus far you anticipate a model with the STEP intervention variable, seasonal dummy variables, and now possibly an additional overall linear trend. In what follows, the error term from this model at time t is denoted W_t. It is likely that these W_t error terms are not independent as is needed for valid inference in least squares regression. If the error terms are autocorrelated, then none of the standard errors and resulting p-values that have been seen thus far is valid. This is true with any form of model misspecification. With time series like these, the dummy variables will capture the overall long term seasonality but there are often local effects such as extra predictability coming from last month's circulation and/or the circulation from the same month last year.

To capture this autocorrelation you can model the current error term W_t. Imagine predicting this month's error W_t as a multiple of last month's error αW_{t-1}. If positive, the coefficient α would represent a proportion of last month's error that carries over into this month and if negative it would indicate a tendency for positive errors to be followed by negative and vice versa. Had you done the same thing last year, your prediction error would have been $W_{t-12} - \alpha W_{t-13}$. Perhaps the amount by which you missed the target last year is predictive of this year's error $W_t - \alpha W_{t-1}$ in which case some proportion γ of last year's error could be included along with αW_{t-1} to improve the forecast. Putting this all together suggests that the deviation W_t of the t^{th} observation from the deterministic part of the model might itself be modelled as $W_t = \alpha W_{t-1} + \gamma W_{t-12} - \alpha\gamma W_{t-13} + e_t$ where e_t would be interpreted as the ultimate error term in the model and would be assumed to have the standard regression error properties of independence and constant variance. Error terms with this property are referred to collectively as "white noise," and a good time series model should leave only white noise as the part it does not predict. The model for W is called the seasonal multiplicative model or sometimes the "airline model" because of its use in an early edition of the seminal book of Box and Jenkins (1976) to model an international airline passenger time series. The

backshift operator $B^j(W_t)=W_{t-j}$ is commonly used to describe models for autocorrelated errors which in this case gives $(1-\alpha B)(1-\gamma B^{12})W_t = e_t$. The airline model is very well known among time series practitioners. It is a good idea to always try it on seasonal data as an initial start in modelling, perhaps to be modified later.

In summary, you are led to try a model with the trend, seasonal, and step intervention variables and errors modelled using the airline model. For such a model it is necessary to move into the SAS/ETS procedures with PROC ARIMA being an obvious candidate for estimating the model parameters. The code for the model, using circcount as the name for total circulation, is

Program 5: PROC ARIMA takes account of autocorrelation:

```
proc arima data=all;
  identify var=circcount crosscor = (mon1 mon2 mon3 mon4 mon5 mon6 mon7
    mon8 mon9 mon10 mon11 mon12 step);
  estimate input = (step mon1 mon2 mon3 mon4 mon5 mon6 mon7 mon8 mon9
           mon10 mon11)
  p=(1)(12) ml; run;
```

Note that this model contains deterministic seasonal effects, a step function, and seasonal autoregressive adjustments but no deterministic trend. Here is the most interesting part of the output from PROC ARIMA:

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	Approx t Value	Pr > |t|	Lag	Variable	Shift
MU	5887.6	203.44308	28.94	<.0001	0	circcount	0
AR1,1	0.43966	0.16427	2.68	0.0074	1	circcount	0
AR2,1	-0.71717	0.14674	-4.89	<.0001	12	circcount	0
NUM1	-1094.5	226.03054	-4.84	<.0001	0	step	0
NUM2	4667.9	227.82388	20.49	<.0001	0	mon1	0
NUM3	4639.8	257.56021	18.01	<.0001	0	mon2	0
NUM4	4028.2	269.92921	14.92	<.0001	0	mon3	0

(Program 5 output continued)

NUM5	5430.3	274.05698	19.81	<.0001	0	mon4	0
NUM6	852.65835	273.23354	3.12	0.0018	0	mon5	0
NUM7	-134.87683	273.55329	-0.49	0.6220	0	mon6	0
NUM8	-69.18609	283.41235	-0.24	0.8071	0	mon7	0
NUM9	3712.0	282.93997	13.12	<.0001	0	mon8	0
NUM10	4240.1	275.71015	15.38	<.0001	0	mon9	0
NUM11	5010.7	259.23604	19.33	<.0001	0	mon10	0
NUM12	4820.7	217.06428	22.21	<.0001	0	mon11	0

June and July do not differ significantly from December, June being about 135 lower and July 69 lower, but typically you would not leave out seasonal dummy variables you put in the model unless they can all be left out. After all, their estimates and significance are an artifact of which of the 12 months was used as a baseline (December here, for example). The estimated autoregressive coefficients α and γ are estimated as .44 and -.72, both being significant. The -1094.5 indicates a loss in circulation. It is statistically significant and an expert would have to judge its practical significance. Taken as an estimate of the true long term loss, you can add and subtract 1.96 standard errors 1.96(226)=443 to get a number that can be added and subtracted from the estimate to get a 95% confidence interval.

Is this model sufficient? Note that you have ignored the seemingly downward trends exhibited in some of your plots. A good time series model should result in e_t being an uncorrelated sequence (i.e. white noise). After all, you are using correlation to help you predict, and if the errors e_t are still correlated, then

why did you not use that information? A test with null hypothesis that e_t is an uncorrelated sequence is based on the estimated correlations in e_t. In the output below, these are 0.047, -0.119, 0.056, …, -0.202. This is the white noise test. Various types of estimated residual correlations are also plotted. Here are some results:

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	1.40	4	0.8450	0.047	-0.119	0.056	0.007	0.100	0.002
12	4.18	10	0.9388	-0.134	0.107	-0.022	-0.087	0.084	-0.073
18	6.21	16	0.9857	-0.107	-0.055	-0.084	0.047	0.041	-0.066
24	13.90	22	0.9051	0.074	-0.025	0.057	0.169	-0.028	-0.202

This printed output sums the squares of the first 6, the first 12, the first 18, then the first 24 residual correlations, each time multiplying by n(n-j)/(n-1) (roughly n when n is large relative to j) where j is the lag number of the correlation and n is the sample size. This test, labelled Chi-square is the Ljung-Box test described in Ljung and Box (1979). The authors prove the distributional results for any arbitrary number k of summands that the user chooses. In PROC ARIMA, the k values are multiples of 6, perhaps a throwback to ASCII output where these fit nicely on single lines of output. The important things to see are the p-values, none of which supply evidence to refute the hypothesis of uncorrelated e_t. There is no evidence of a problem with this model. SAS will also give p-values for this test for all k in a range, not just in sets of 6, now that graphical output is available. The white noise tests have p-values where small p-values indicate problems with the model. In graphical output, large rather than small objects catch the viewer's attention so bars representing small p-values would not be very obvious. In the lower right corner of Figure 6, bars representing those p-values are plotted on an inverted logarithmic scale, that is, smaller p-values have large bars and larger p-values have small bars. Small bars are insignificant which in this case means no significant evidence against the current model.

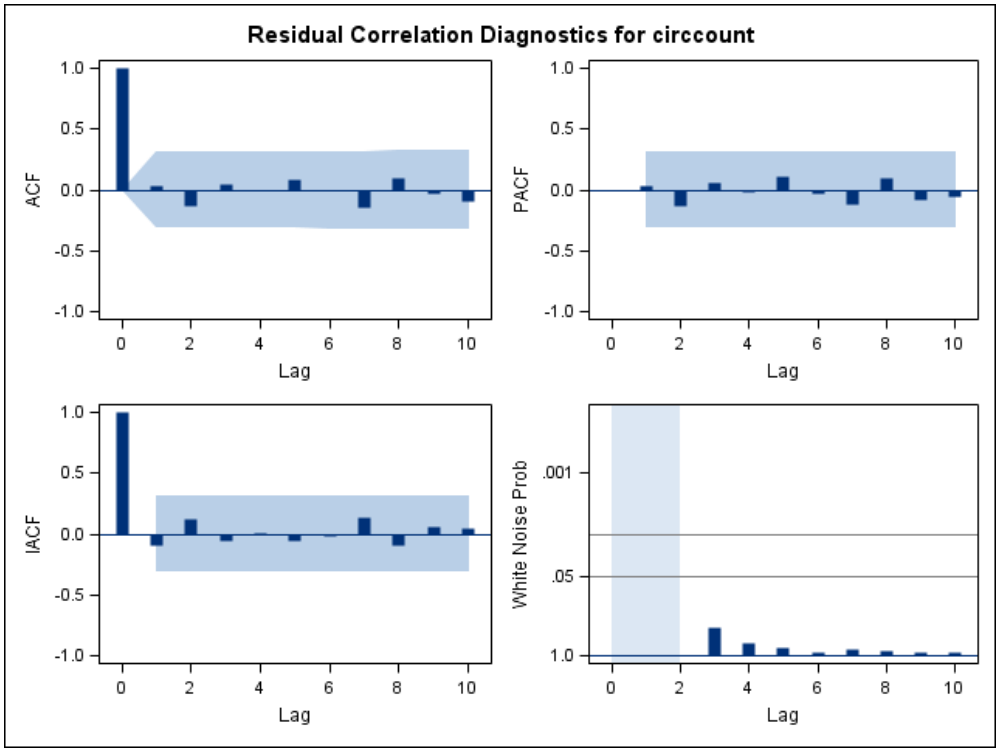

Figure 9: Residual diagnostics for total circulation model without trend.

All but the lower right plot are various correlation plots: autocorrelation, partial autocorrelation, and inverse autocorrelation. Besides the ignorable bars at 0 (everything is perfectly correlated with itself – its lag 0 correlation) there are no bars penetrating the two standard error bands around 0 - no significant correlation at any lag. Your model t tests thus indicate that (almost) everything you put in the model is significant and the diagnostics indicate that you need nothing more in your model – almost a textbook case!

Figure 10 shows the data (circles) the one step ahead forecasts and one step ahead forecast error bands, forecasts one year ahead, and the two different means pegged to the average seasonal factor. Vertical reference lines appear at the intervention date (thin, solid blue) and at the end of the data (dashed green).

Other considerations: As stated before, this model is adequate, not necessarily the best or only acceptable model. A model with a trend term might show a significant downward trend. The plot in Figure 10 does seem to have a slight downward trend in the pre and post Hunt periods.

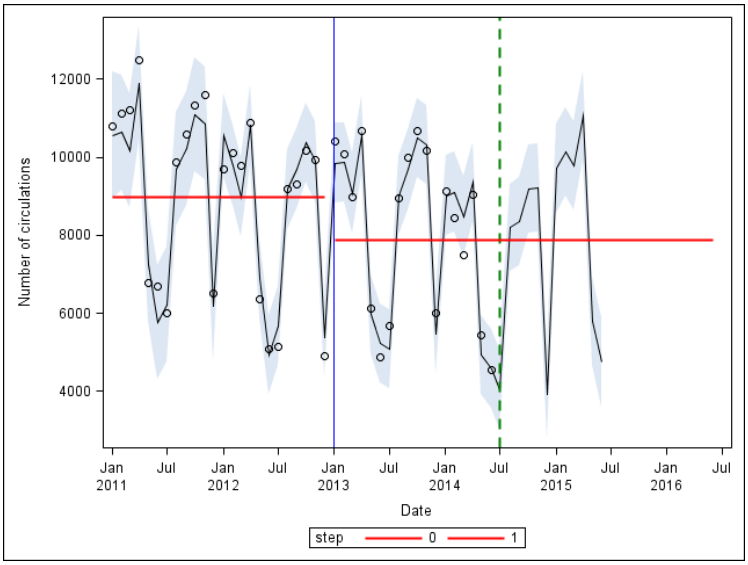

Figure 10: Combined data. Total circulation model with 2 means and seasonal dummy variables.

SUBCOLLECTIONS

Program 5 was modified to capture the intervention effect and its p-value as macro variables. These were then added to the graph titles. Here is the Hunt data result.

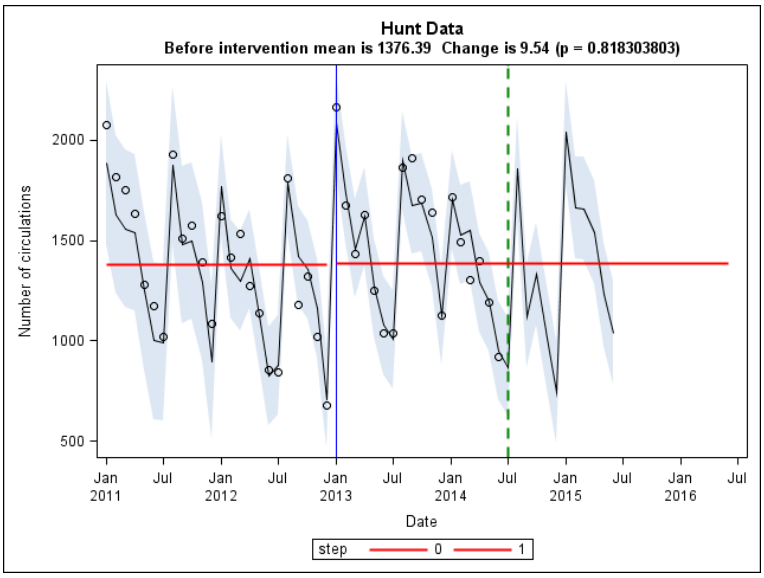

Figure 11: Hunt data, 2 means and seasonal dummy variables in the model.

For the Hunt data you have an increase of only 9.54 after the intervention which is nowhere near significance. Applying the same model to the Hill data you see a more substantial drop of 1113.4 which is strongly significant, however this graph prior to the forecast appears to have a fairly consistent downward trend, as did the combined data. It seems prudent to try a trend model.

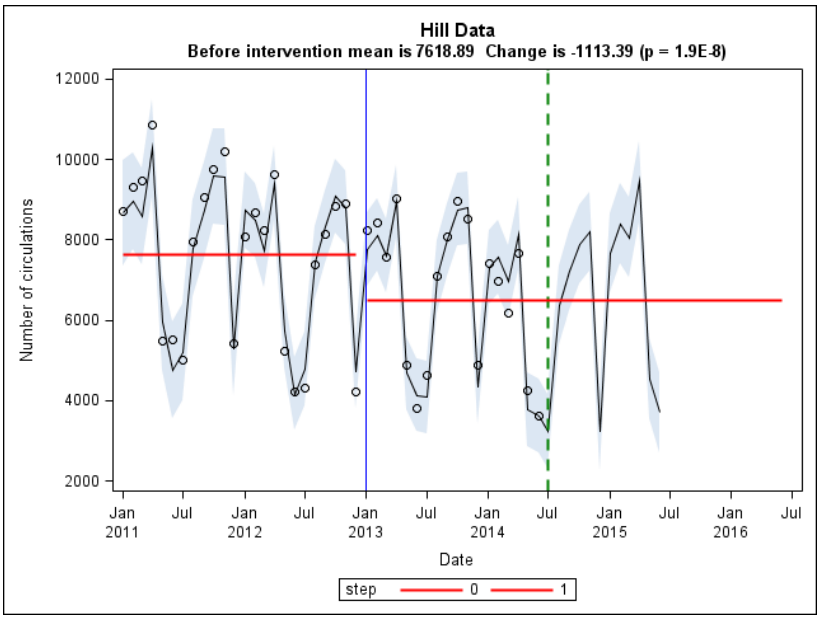

Figure 12: Hill data, 2 means and seasonal dummy variables in the model.

ADDING A TREND

Adding a trend is a simple modification to the program. From the graphs it seems possible that a single trend line will suffice with no intervention effect but it is prudent to leave the step variable in the model and base its possible removal on the statistics. Here variable t, which starts at 1 and increases by 1 for each

observation, is added to the data set and to the inputs shown in Program 5 above. You see a monthly decline 106 in circulation for the combined data. See also Figure 13.

Parameter	Variable	Estimate	Probt
MU	Combined	10304.8208	<.0001
AR1,1	Combined	0.13841323	0.5039
AR2,1	Combined	-0.0544686	0.8480
NUM1	step	1476.40278	<.0001
NUM2	t	-106.6494	<.0001
NUM3	mon1	981.440709	<.0001
NUM4	mon2	1034.22572	<.0001
NUM5	mon3	562.207672	0.0073
NUM6	mon4	2072.88743	<.0001
NUM7	mon5	-2413.0437	<.0001
NUM8	mon6	-3201.2287	<.0001
NUM9	mon7	-3139.5847	<.0001
NUM10	mon8	677.827681	0.0032
NUM11	mon9	1401.26414	<.0001
NUM12	mon10	2267.56915	<.0001
NUM13	mon11	2206.9914	<.0001

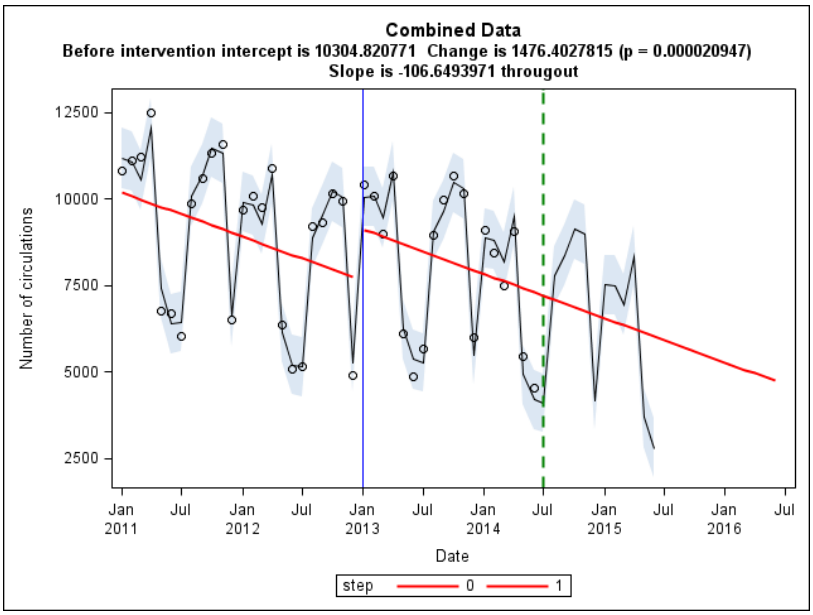

Figure 13: Combined data, model with parallel trend lines and seasonal dummy variables.

In the printout above, rather than a December baseline number, the seasonal dummy variables are set up so that the seasonal factors sum to 0 placing the red line at the average seasonal as has been done in previous plots. The December seasonal factor is the negative of the sum of the other 11, namely -(981.44+ 1034.23+…+2206.99). Here are the same analyses for the subcollections.

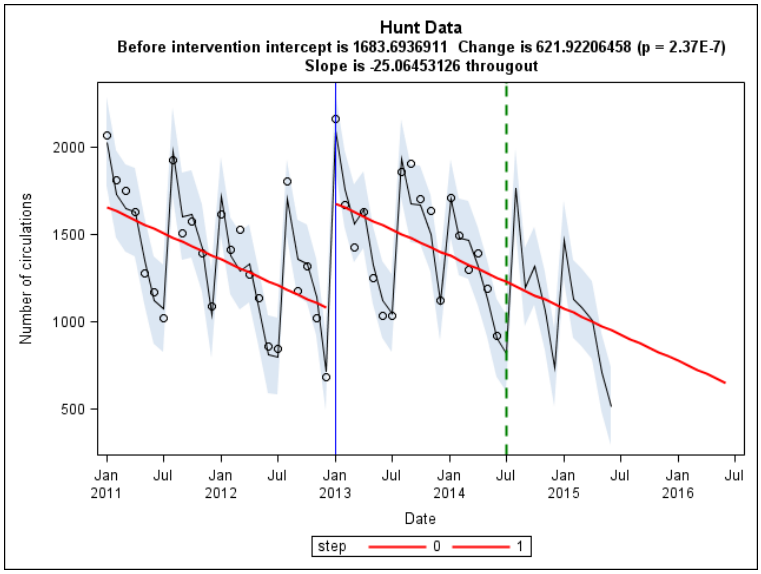

Figure 14: Hunt data, model with parallel trend lines and seasonal dummy variables.

Parameter	Variable	Estimate	Probt
MU	Hunt	1683.69369	<.0001
AR1,1	Hunt	0.10766467	0.5750
AR2,1	Hunt	-0.4639023	0.0545
NUM1	step	621.922065	<.0001
NUM2	t	-25.064531	<.0001

(Hunt analysis continued)

NUM3	mon1	373.674585	<.0001
NUM4	mon2	94.4781372	0.0315
NUM5	mon3	30.5498612	0.4684
NUM6	mon4	31.5356305	0.4373
NUM7	mon5	-208.21496	<.0001
NUM8	mon6	-405.31898	<.0001
NUM9	mon7	-436.48798	<.0001
NUM10	mon8	494.915632	<.0001
NUM11	mon9	148.864926	0.0020
NUM12	mon10	192.791268	<.0001
NUM13	mon11	19.596356	0.6962

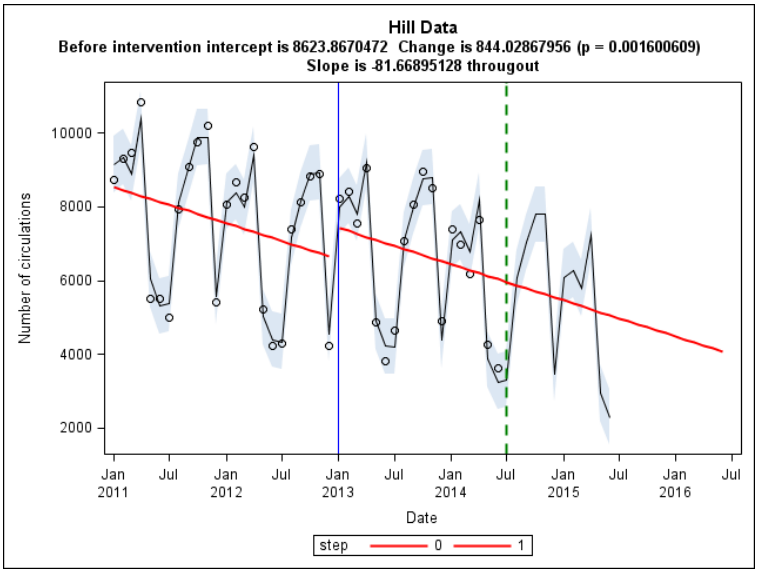

Figure 15: Hill data, model with parallel trend lines and seasonal dummy variables.

Parameter	Variable	Estimate	Probt
MU	Hill	8623.86705	<.0001
AR1,1	Hill	0.07761512	0.7142
AR2,1	Hill	0.05962787	0.8507
NUM1	step	844.02868	0.0016
NUM2	t	-81.668951	<.0001
NUM3	mon1	609.240538	0.0025
NUM4	mon2	923.152124	<.0001
NUM5	mon3	530.360878	0.0086
NUM6	mon4	2038.48253	<.0001
NUM7	mon5	-2208.9977	<.0001
NUM8	mon6	-2783.9681	<.0001
NUM9	mon7	-2696.2815	<.0001
NUM10	mon8	193.67804	0.3754
NUM11	mon9	1238.10868	<.0001
NUM12	mon10	2079.37124	<.0001
NUM13	mon11	2178.07992	<.0001

Because the autoregressive terms are insignificant, you could omit them and fit the resulting model with just regression or refit it in PROC ARIMA with no autoregressive terms.

WHAT HAPPENED TO THE AUTOREGRESSIVE TERMS?

The models with two means and no trends showed seasonal autoregressive residuals despite the deterministic seasonal dummy variables that you included. The p-values for the two autoregressive terms were below 0.01, fairly strong evidence of autocorrelation. It appeared that the seasonal dummy variables were insufficient to explain the seasonality in the plot. Once the trend term was introduced, with no change to the seasonal dummy variables, there was no evidence of autoregressive errors even at the 70% level. How could that have happened? All of a sudden the seasonal dummies now seem sufficient to handle the seasonality. How could the addition of a trend affect the seasonality in the model?

Imagine data with a downward trend similar to what we've seen and with an exactly repeating 12 month seasonal pattern around it, a sine wave for example. Suppose only a constant mean and seasonal dummy variables are used to model the data. Now in the data, the sequence of January numbers would be decreasing because of the trend. The model, not having a trend, would leave a sequence of positive

January deviations at first and negative ones at the end of the data as in Figure 16. In that figure, the data trend line is 100 – 0.5t, the seasonality is 3cos(2πt/12), the data are the circles and the predictions from a trendless model are the dots. Residuals for the Januaries are shown as vertical lines.

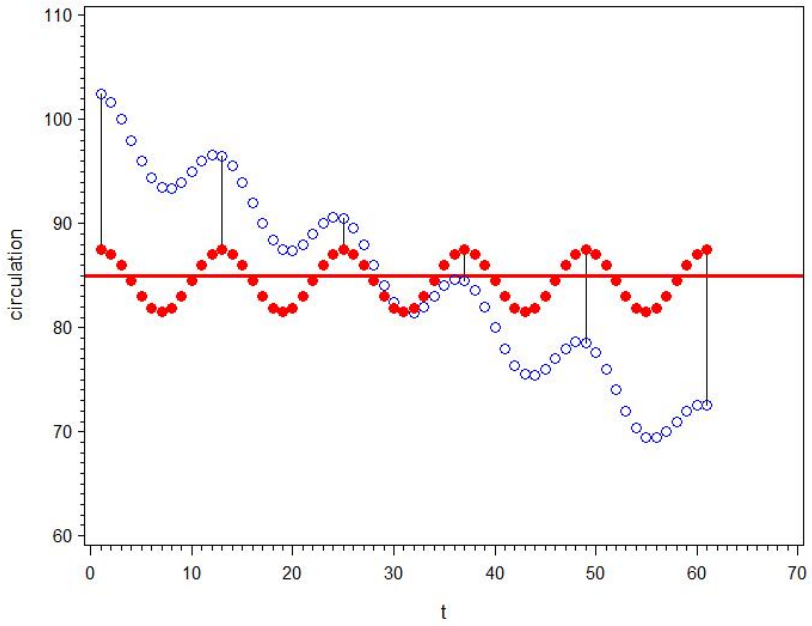

Figure 16: Trend in the seasonal data (circles) but not in the fitted model (dots).

Each January residual is exactly 6 less than its predecessor 12 months ago, namely 15, 9, 3, -3, -9, -15 because of the -0.5t in the data that is not captured by the model. The same association holds for February and all the other months. In summary, all of the residuals r_t after the first 12 satisfy $r_t = r_{t-12} - 6$, a perfect lag 12 correlation for this simple errorless example. Knowing the residual 12 months ago would definitely help predict the current one. Had a model with trend been used for prediction, its slope would have been -0.5 and the predictions (dots) would exactly match the data (circles). In the same way, adding the trend to the library data left no evidence of correlation in the residuals.

This is a general lesson. Evidence of autocorrelation, especially exceptionally strong autocorrelation can come from true autocorrelation of course, but could also be a signal of model misspecification.

CONCLUSION

Data on library circulation are analyzed with increasingly sophisticated models. The data fall into before and after groups. Treating two samples with a two sample t test, while a common approach, is seen to be inappropriate here because of failure of the two groups to satisfy the assumptions. Including seasonal dummy variables did not seem to handle all of the seasonality until a linear trend was added. This trend was not obvious in the original data but was revealed somewhat more obviously in residual plots. Once the trend was added no longer was any residual autocorrelation suggested. Its presence prior to adding the trend might have simply been a consequence of model misspecification, a lesson that is good to keep in mind whenever dealing with time series data.

REFERENCES

Box, G.E.P. and G.M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.

Ljung, G. M. and G.E.P. Box (1979). “On a measure of lack of fit in time series models,” *Biometrika*, 65, 297-303.

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

David A. Dickey
Department of Statistics
5230 SAS Hall
NC State University
dickey@stat.ncsu.edu