Paper SP05 - 2015

Using ANCOVA to Assess Regression to the Mean

Kathryn Schurr, M.S., Spectrum Health – Healthier Communities, Grand Rapids, MI

ABSTRACT

Regression to the mean (RTM) is a statistical phenomenon in which results appear to be statistically significant, but, in fact, they are spurious. There are many documented occurrences of this effect throughout history where researchers have been led to erroneous conclusions. Generally, these conclusions are based on a statistically significant change having been observed in a population or sample, but the change is in whole or in part due to chance. By implementing SAS and PROC GLM, a user can conduct an Analysis of Covariance (ANCOVA) to determine whether or not RTM is present. This will help prevent the user from making a presumptuous and erroneous statement regarding an observed change within their study.

INTRODUCTION

Regression to the Mean (RTM) is a statistical artifact which occurs when a sample from a population is selected on the basis of extreme values on a measure of disease or abnormality and then a subsequent observation of the same measure of that sample progresses toward the population mean. Historically, RTM has been seen in various fields such as genetics, biology, and revenue analyses, an example of which will be discussed later. In genetics, a simplified example shows the RTM phenomenon. Suppose a researcher selected a sample from the lowest 10% of the adult population in regards to weight and where these people were otherwise free of disease. If a researcher were to follow these individuals over time and evaluate the weight of their offspring, one would find that once the offspring reached adulthood, they would not represent the next generation's lowest 10% in terms of weight. Although there is variability in terms of lifestyle and other factors, researchers will find that the offspring's weight will tend towards the population's mean weight – even though their parents were of the lowest 10%. This natural occurrence of RTM is a highly simplified example.

Within Spectrum Health's Healthier Communities Department, programs are designed that specifically target diseased or abnormal populations. One program in particular evaluates individuals with Heart Failure (HF) and helps them manage their disease by using a heart monitor that communicates wirelessly back to the hospital. This program called Telehealth (TH) is a digital monitoring system which helps individuals track their disease state and evaluate their health continuously in real-time. This system alerts patients of biometric readings that are out of the ordinary or are potentially problematic. TH allows patients to seek appropriate medical care before they have even recognized their abnormal symptoms. RTM appears when comparing the costs of inpatient admissions of the TH patients before using TH monitoring with their costs while they were being monitored. Based on the cost data from this program, it was found that there was a definite drop in their costs between the 'Before' period and the 'During' period. This example of RTM will be used to explain how SAS and PROC GLM can be used to identify possible RTM.

METHODOLOGY

Individuals enrolled in TH are in the program typically for six months unless if they choose to opt out during the program. For the purpose of this paper, the individuals included in the RTM analysis are those who were on the TH monitor for some six months and who had at least one inpatient visit in the same length time period before their TH enrollment. The "mirrored" time frame is exactly the length of time they were enrolled in TH, only directly prior to enrolling in TH.

Once the group of patients were selected that would be included in the TH sample, control groups were then constructed based on certain criteria. The individuals who were chosen as controls had to have had an inpatient hospital stay with a diagnosis of HF during the same time as the TH group. A small subset of those patients were then randomly selected based on the last digit of their medical record number. To simulate the same type of 'Before' and 'During' time period, the initial inpatient stay that triggered the individuals as a possible control was designated the 'index' visit for a simulated 'During' time period. The patients' inpatient stays were then recorded for approximately six months so that the control groups and TH patients had similar observation periods. Similarly, the 'Before' period was approximately six months prior to the simulated index visit.

Once the control groups were established, an analysis of average costs incurred per patient were evaluated. These costs were looked at for both the 'Before' and 'During' periods. Table 1 compares some summary statistics regarding the distribution of costs for 'Before' and 'During' TH monitoring. These results of a simple PROC MEANS evaluate the initial output for the original study group and the two control groups.

TABLE 1: Comparative Costs (in dollars) of Hospital Admissions Before and During Telehealth Monitoring

Group	Variable	N	Mean	Std Dev	Minimum	10th Pctl	25th Pctl	Median	75th Pctl	Maximum
20	Before	153	32137.90	40189.10	0.00	7240.70	10507.90	17760.89	35314.69	276288.27
	During	153	14014.52	26179.52	0.00	0.00	0.00	0.00	17931.31	189555.91
21	Before	149	32251.65	41251.25	5316.08	8316.00	13131.36	18809.71	36707.75	377129.18
	During	149	14875.13	38749.13	0.00	0.00	0.00	0.00	13238.72	340497.57
22	Before	134	40405.93	64337.62	4651.87	9839.21	12399.32	21859.40	38564.83	481504.42
	During	134	9105.62	22576.79	0.00	0.00	0.00	0.00	10050.56	149466.73

Group 20 is the group of TH patients. By looking at the Mean and the change that was seen between the 'During' and 'Before' periods it appears that TH was effective in reducing inpatient admission costs. By having the two control groups present, group 21 and group 22, it can be seen that similar changes occurred. By looking further at the means table, it can be seen that the 'During' period percentiles contain numerous zero-costs data. This indicates that patients who incurred costs in the 'Before' period did not have any hospital admissions in the 'During' period.

To test if the charges differed between time periods, many researchers may take the simple and straightforward route and conduct a paired t-test. By doing this, the researchers will more than likely fall prey to what RTM is. Tables 2 through 4 show what the output of doing a simple paired t-test on the TH patient group would produce.

TABLE 2: Difference in Costs of Hospital Admissions Before and During Telehealth Monitoring

N	Mean Difference	Std Dev of Difference	Std Err	Minimum	Maximum
153	18123.4	45027.3	3640.2	-189556	276288

TABLE 3: 95% confidence interval estimates of the difference in hospitalization costs

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
18123.4	10931.4	25315.4	45027.3	40484.5	50727.6

TABLE 4: Paired t-test and significance level

DF	t Value	Pr > |t|
152	4.98	<.0001

By looking at the results in Table 4 we see that the paired t-test for the study group shows a significant difference between the two time periods. Without the control groups being analyzed, a researcher may conclude that there is a statistically significant difference when in fact the difference can be attributed to RTM. To test whether or not there is a significant difference between the changes seen in the study group and the changes incurred in the control groups, an ANCOVA will be run using PROC GLM in SAS.

UTILIZING PROC GLM IN SAS

Now that the similarities between the TH group and the control groups have been highlighted, it is time to conduct an ANCOVA to determine how likely it is that there is a true difference between the two control group changes and the change observed in the TH group. An ANCOVA was chosen as the analytical test due to its ability to reduce within-group error variance. In a typical ANOVA, the effect of an experiment is assessed by comparing the amount of variability in the data that can be explained by the treatment contrasted with the variability that cannot be explained. By allowing the ANOCOVA to take into effect and thus control the effects of unknown covariates, the ability to assess the treatment is enhanced. Another reason the ANCOVA is an appropriate method of assessing RTM is because it eliminates some of the unknown confounding variables. This allows some of the unknown or unintended bias to be removed.

To conduct an ANCOVA in SAS, PROC GLM is used. The code is given below.

```
PROC GLM DATA = Average_Charges2;
      CLASS Group;
      MODEL During = Group Before / SOLUTION;
      LSMEANS Group / STDERR PDIFF COV OUT = AdjMeans;
RUN;

PROC PRINT DATA = AdjMeans;
RUN;
```

The DATA = portions specifies the input dataset. Using the CLASS statement allows the user to identify which variable represents the group. The MODEL statement in PROC GLM allows the user to specify the dependent variable, 'During', which is the average costs incurred in the 'During' time period. The independent variable is placed on the left of the equal sign, the average costs incurred in the 'Before' period, and the inclusion of the grouping variable, 'Group.' Including the option SOLUTION allows the user to indicate the desire to see parameter estimates for the variables included in the model.

The LSMEANS statement in PROC GLM tells SAS that the user is interested in looking at multiple comparisons on interactions as well as main effects of the classification variable. The options for this statement that were used are STDERR PDIFF COV OUT = ADJMEANS. STDERR produces the standard error of the Least Squares Means and the probability level for the hypothesis H_0 : LSMEANS = 0. . PDIFF requests that p-values for the differences of the Least Squares Means be produced. COV is used in conjunction with the OUT = ADJMEANS options to include variances and covariances. By outputting this to a dataset ADJMENS, the user can view the covariance matrix for the Least Squares Means themselves. This can only be used if there is only one effect listed in the LSMEANS statement.

The output produced from PROC GLM is given below. Table 5 gives the overall ANOVA. This shows whether or not the test for differences between groups and within groups is significant. Based on the p-value of 0.0796, the null hypothesis is not rejected.

TABLE 5

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6165364542	2055121514	2.27	0.0796
Error	432	390719610956	904443543.88		
Corrected Total	435	396884975498			

Table 6 gives the R-Square value and the overall average for the 'During' period.

TABLE 6

R-Square	Coeff Var	Root MSE	During Mean
0.015534	234.9541	30073.97	12799.93

TABLE 7

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Group	2	2696196894	1348098447	1.49	0.2264
Before	1	3469167648	3469167648	3.84	0.0508

Table 7 above shows the Type I SS for Group which represents the sums of squares that are obtained for the ANCOVA model Before = Group. This measures the difference between arithmetic means of the 'Before' costs for the three different groups. The p-value here suggests that neither Source is significant. Table 8 below shows the TYPE III SS for Group adjusted for the covariate. This measures the differences between the Least Squares Means, controlling for the covariate. The p-value here also suggests that neither Source is significant.

TABLE 8

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	2	3165364471	1582682236	1.75	0.1750
Before	1	3469167648	3469167648	3.84	0.0508

Table 9 shows the Least Squares Means that were specified on the LSMEANS statement. This shows all the probability values for the hypothesis to be displayed. Table 10 shows the covariance matrix that shows which comparisons are statistically different. Although the within-group differences are statistically different for all groups, there is no evidence to suggest that the between group differences are statistically different.

TABLE 9

Group	During LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
20	14162.9708	2432.5174	<.0001	1
21	15017.0353	2464.8206	<.0001	2
22	8778.3397	2603.3640	0.0008	3

TABLE 10

Least Squares Means for effect group Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: During			
i/j	**1**	**2**	**3**
1		0.8052	0.1318
2	0.8052		0.0828
3	0.1318	0.0828	

By specifying the COV option and the OUT = option, SAS produces a data set of the estimates, their standard errors, and the variances and covariances of the Least Squares Means. These are shown in Table 11.

TABLE 11

Obs	_NAME_	Group	LSMEAN	STDERR	NUMBER	COV1	COV2	COV3
1	During	20	14162.97	2432.52	1	5917140.81	5491.79	-12666.25
2	During	21	15017.04	2464.82	2	5491.79	6075340.55	-12107.78
3	During	22	8778.34	2603.36	3	-12666.25	-12107.78	6777504.01

The ODS graphics portion of SAS produces Figure 1 and Figure 2. Figure 1 shows the ANCOVA plot of the 'During' costs by Group and 'Before' costs. By looking at Figure 1, it is seen that the TH group, Group 20, is in between the two control groups on the graph and almost overlays the regression line for control Group 21. This suggests that there is no difference between groups, which is a confirmation of what was previously presented in the tables. Figure 2 shows a plot of differences between Group and the Least Squares Means for 'During' costs. This graphic allows the user to readily distinguish groups that are significantly different from the rest.

FIGURE 1

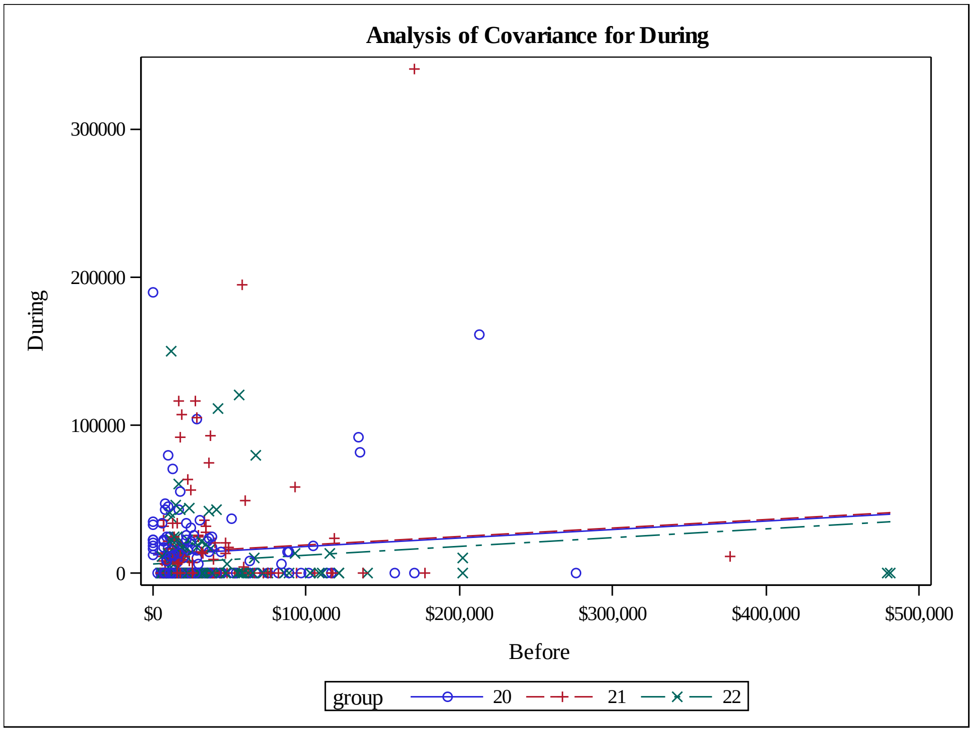
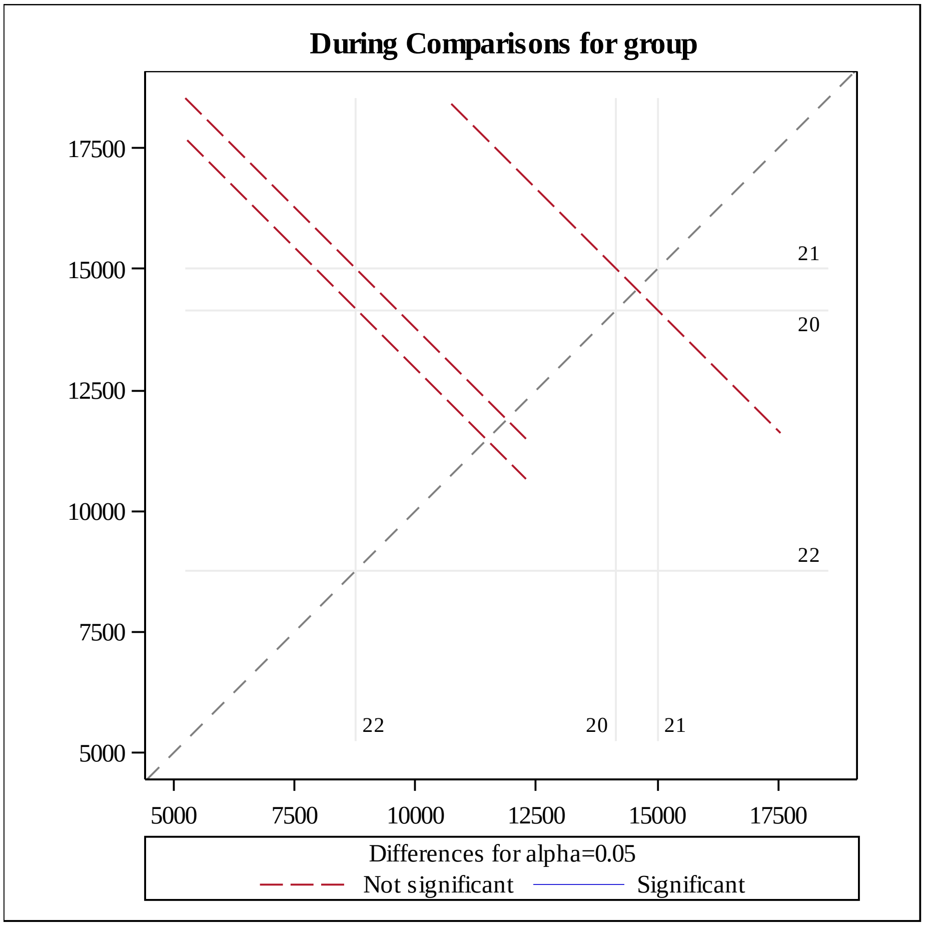


FIGURE 2



By looking at the output produced by PROC GLM and by evaluating the ANCOVA results, it is apparent that although there originally appeared to be differences between costs for the TH group and control groups, the ANCOVA has found it not statistically significant. It is in these cases that individuals who were to simply run a paired t-test on the TH group would have developed an erroneous result.

CONCLUSION

Regression to the Mean is a statistical phenomenon seen in a variety of different fields and applications. It is with thorough knowledge of solid statistical practice and methodology that analysts are able to distinguish RTM from a real significant change within their data. The best way to avoid RTM is by using randomized controlled experiments so that the study group can be evaluated against an experimental control. By having the controls in the example used for this paper, the researchers were able to determine there was not a statistical difference in the costs incurred by the TH patients. The control groups were able to show a similar change in costs which proves RTM was present in the all three of the groups.

REFERENCES

"Analysis of Covariance Example." *Analysis of Covariance*. SAS Institute, 1 Jan. 2015. Web. 12 Mar. 2015. <www.sas.com>.

Barnett, Adrian G., Jolieke C. Van Der Pols, and Annette J. Dobson. "Regression to the Mean: What it is and How to Deal With It." *International Journal of Epidemiology* 34.1 (2004): 215-20. Web. 12 Mar. 2015.

Field, Andy. "Analysis of Covariance (ANCOVA)." *Analysis of Covariance (ANCOVA)*. Discovering Statistics, 2012. Web. 12 Mar. 2015. <www.discoveringstatistics.com>.

Trochim, William. "Regression to the Mean." *Research Methods Knowledge Base*. N.p., 20 Oct. 2006. Web. 12 Mar. 2015.

CONTACT INFORMATION

Your comments and questions are much appreciated. Contact the author at:

Kathryn Schurr, M.S.
Statistical Database Analyst
Spectrum Health - Healthier Communities
665 Seward Avenue NW, Suite 110
Grand Rapids, MI 49504
Work Phone: 616.391.2983
Work E-Mail: kathryn.schurr@spectrumhealth.org