Title: Bootstrapping Inference in Regreesion
Date: 2015-03-08
Category: Blog

Summary

The purpose of this notebook is illustrate how apply bootstrap a confidence interval for the slope of the `MATH` independent variable in a multivariable regression environment using the `GENMOD` procedure in SAS.

This technique is useful in an inference problem where the interpretation of the Beta parameter is critical for the business problem and we are not really sure if our model follows all the GLM assumptions.

This technique only is practicable with models that fit very fast.

Introduction

Apply bootstrapping to check the inference (p-value) of the coefficient estimates of the one or all explanatory variables. In this expample only the `MATH` Beta parameter. The aim of this consists in determine if the beta estimator of the one or all explanatory variables are statistically significant. Therefore, if the assumptions of our GLM model are correct the p-value we get in the output of the GLM model should be correct and our work finish here. But, if one or several assumptions of the GLM are not correct the inference is not valid. The bootstrap approach does not rely on any of these assumptions, and so it is likely giving a more accurate estimate of the coefficient estimates (and of the standard errors). I think it worth checking if the bootstrap confidence interval are the same of the GLM Wald confidence interval.

Let's build up a SAS datasets to illustrate the bootstrapping technique:

```
options nosource;
```

```
504  ods listing close;ods html5 file=stdout options(bitmap_mode='inline') device=png;
NOTE: Writing HTML5 Body file: STDOUT
505
506  options nosource;
```

```
data HSB2 ;
        infile datalines dsd
        ;
input
        id
        female
        race
        ses
        schtyp
        prog
        read
        write
        math
        science
        socst
;
datalines;
70,0,4,1,1,1,57,52,41,47,57
121,1,4,2,1,3,68,59,53,63,61
86,0,4,3,1,1,44,33,54,58,31
141,0,4,3,1,3,63,44,47,53,56
172,0,4,2,1,2,47,52,57,53,61
113,0,4,2,1,2,44,52,51,63,61
50,0,3,2,1,1,50,59,42,53,61
11,0,1,2,1,2,34,46,45,39,36
84,0,4,2,1,1,63,57,54,58,51
48,0,3,2,1,2,57,55,52,50,51
75,0,4,2,1,3,60,46,51,53,61
60,0,4,2,1,2,57,65,51,63,61
95,0,4,3,1,2,73,60,71,61,71
104,0,4,3,1,2,54,63,57,55,46
38,0,3,1,1,2,45,57,50,31,56
115,0,4,1,1,1,42,49,43,50,56
76,0,4,3,1,2,47,52,51,50,56
195,0,4,2,2,1,57,57,60,58,56
114,0,4,3,1,2,68,65,62,55,61
85,0,4,2,1,1,55,39,57,53,46
167,0,4,2,1,1,63,49,35,66,41
143,0,4,2,1,3,63,63,75,72,66
41,0,3,2,1,2,50,40,45,55,56
20,0,1,3,1,2,60,52,57,61,61
12,0,1,2,1,3,37,44,45,39,46
53,0,3,2,1,3,34,37,46,39,31
154,0,4,3,1,2,65,65,66,61,66
178,0,4,2,2,3,47,57,57,58,46
196,0,4,3,2,2,44,38,49,39,46
29,0,2,1,1,1,52,44,49,55,41
126,0,4,2,1,1,42,31,57,47,51
103,0,4,3,1,2,76,52,64,64,61
192,0,4,3,2,2,65,67,63,66,71
150,0,4,2,1,3,42,41,57,72,31
```

```
199,0,4,3,2,2,52,59,50,61,61
144,0,4,3,1,1,60,65,58,61,66
200,0,4,2,2,2,68,54,75,66,66
80,0,4,3,1,2,65,62,68,66,66
16,0,1,1,1,3,47,31,44,36,36
153,0,4,2,1,3,39,31,40,39,51
176,0,4,2,2,2,47,47,41,42,51
177,0,4,2,2,2,55,59,62,58,51
168,0,4,2,1,2,52,54,57,55,51
40,0,3,1,1,1,42,41,43,50,41
62,0,4,3,1,1,65,65,48,63,66
169,0,4,1,1,1,55,59,63,69,46
49,0,3,3,1,3,50,40,39,49,47
136,0,4,2,1,2,65,59,70,63,51
189,0,4,2,2,2,47,59,63,53,46
7,0,1,2,1,2,57,54,59,47,51
27,0,2,2,1,2,53,61,61,57,56
128,0,4,3,1,2,39,33,38,47,41
21,0,1,2,1,1,44,44,61,50,46
183,0,4,2,2,2,63,59,49,55,71
132,0,4,2,1,2,73,62,73,69,66
15,0,1,3,1,3,39,39,44,26,42
67,0,4,1,1,3,37,37,42,33,32
22,0,1,2,1,3,42,39,39,56,46
185,0,4,2,2,2,63,57,55,58,41
9,0,1,2,1,3,48,49,52,44,51
181,0,4,2,2,2,50,46,45,58,61
170,0,4,3,1,2,47,62,61,69,66
134,0,4,1,1,1,44,44,39,34,46
108,0,4,2,1,1,34,33,41,36,36
197,0,4,3,2,2,50,42,50,36,61
140,0,4,2,1,3,44,41,40,50,26
171,0,4,2,1,2,60,54,60,55,66
107,0,4,1,1,3,47,39,47,42,26
81,0,4,1,1,2,63,43,59,65,44
18,0,1,2,1,3,50,33,49,44,36
155,0,4,2,1,1,44,44,46,39,51
97,0,4,3,1,2,60,54,58,58,61
68,0,4,2,1,2,73,67,71,63,66
157,0,4,2,1,1,68,59,58,74,66
56,0,4,2,1,3,55,45,46,58,51
5,0,1,1,1,2,47,40,43,45,31
159,0,4,3,1,2,55,61,54,49,61
123,0,4,3,1,1,68,59,56,63,66
164,0,4,2,1,3,31,36,46,39,46
14,0,1,3,1,2,47,41,54,42,56
127,0,4,3,1,2,63,59,57,55,56
165,0,4,1,1,3,36,49,54,61,36
174,0,4,2,2,2,68,59,71,66,56
3,0,1,1,1,2,63,65,48,63,56
58,0,4,2,1,3,55,41,40,44,41
```

```
146,0,4,3,1,2,55,62,64,63,66
102,0,4,3,1,2,52,41,51,53,56
117,0,4,3,1,3,34,49,39,42,56
133,0,4,2,1,3,50,31,40,34,31
94,0,4,3,1,2,55,49,61,61,56
24,0,2,2,1,2,52,62,66,47,46
149,0,4,1,1,1,63,49,49,66,46
82,1,4,3,1,2,68,62,65,69,61
8,1,1,1,1,2,39,44,52,44,48
129,1,4,1,1,1,44,44,46,47,51
173,1,4,1,1,1,50,62,61,63,51
57,1,4,2,1,2,71,65,72,66,56
100,1,4,3,1,2,63,65,71,69,71
1,1,1,1,1,3,34,44,40,39,41
194,1,4,3,2,2,63,63,69,61,61
88,1,4,3,1,2,68,60,64,69,66
99,1,4,3,1,1,47,59,56,66,61
47,1,3,1,1,2,47,46,49,33,41
120,1,4,3,1,2,63,52,54,50,51
166,1,4,2,1,2,52,59,53,61,51
65,1,4,2,1,2,55,54,66,42,56
101,1,4,3,1,2,60,62,67,50,56
89,1,4,1,1,3,35,35,40,51,33
54,1,3,1,2,1,47,54,46,50,56
180,1,4,3,2,2,71,65,69,58,71
162,1,4,2,1,3,57,52,40,61,56
4,1,1,1,1,2,44,50,41,39,51
131,1,4,3,1,2,65,59,57,46,66
125,1,4,1,1,2,68,65,58,59,56
34,1,1,3,2,2,73,61,57,55,66
106,1,4,2,1,3,36,44,37,42,41
130,1,4,3,1,1,43,54,55,55,46
93,1,4,3,1,2,73,67,62,58,66
163,1,4,1,1,2,52,57,64,58,56
37,1,3,1,1,3,41,47,40,39,51
35,1,1,1,2,1,60,54,50,50,51
87,1,4,2,1,1,50,52,46,50,56
73,1,4,2,1,2,50,52,53,39,56
151,1,4,2,1,3,47,46,52,48,46
44,1,3,1,1,3,47,62,45,34,46
152,1,4,3,1,2,55,57,56,58,61
105,1,4,2,1,2,50,41,45,44,56
28,1,2,2,1,1,39,53,54,50,41
91,1,4,3,1,3,50,49,56,47,46
45,1,3,1,1,3,34,35,41,29,26
116,1,4,2,1,2,57,59,54,50,56
33,1,2,1,1,2,57,65,72,54,56
66,1,4,2,1,3,68,62,56,50,51
72,1,4,2,1,3,42,54,47,47,46
77,1,4,1,1,2,61,59,49,44,66
61,1,4,3,1,2,76,63,60,67,66
```

```
190,1,4,2,2,2,47,59,54,58,46
42,1,3,2,1,3,46,52,55,44,56
2,1,1,2,1,3,39,41,33,42,41
55,1,3,2,2,2,52,49,49,44,61
19,1,1,1,1,1,28,46,43,44,51
90,1,4,3,1,2,42,54,50,50,52
142,1,4,2,1,3,47,42,52,39,51
17,1,1,2,1,2,47,57,48,44,41
122,1,4,2,1,2,52,59,58,53,66
191,1,4,3,2,2,47,52,43,48,61
83,1,4,2,1,3,50,62,41,55,31
182,1,4,2,2,2,44,52,43,44,51
6,1,1,1,1,2,47,41,46,40,41
46,1,3,1,1,2,45,55,44,34,41
43,1,3,1,1,2,47,37,43,42,46
96,1,4,3,1,2,65,54,61,58,56
138,1,4,2,1,3,43,57,40,50,51
10,1,1,2,1,1,47,54,49,53,61
71,1,4,2,1,1,57,62,56,58,66
139,1,4,2,1,2,68,59,61,55,71
110,1,4,2,1,3,52,55,50,54,61
148,1,4,2,1,3,42,57,51,47,61
109,1,4,2,1,1,42,39,42,42,41
39,1,3,3,1,2,66,67,67,61,66
147,1,4,1,1,2,47,62,53,53,61
74,1,4,2,1,2,57,50,50,51,58
198,1,4,3,2,2,47,61,51,63,31
161,1,4,1,1,2,57,62,72,61,61
112,1,4,2,1,2,52,59,48,55,61
69,1,4,1,1,3,44,44,40,40,31
156,1,4,2,1,2,50,59,53,61,61
111,1,4,1,1,1,39,54,39,47,36
186,1,4,2,2,2,57,62,63,55,41
98,1,4,1,1,3,57,60,51,53,37
119,1,4,1,1,1,42,57,45,50,43
13,1,1,2,1,3,47,46,39,47,61
51,1,3,3,1,1,42,36,42,31,39
26,1,2,3,1,2,60,59,62,61,51
36,1,3,1,1,1,44,49,44,35,51
135,1,4,1,1,2,63,60,65,54,66
59,1,4,2,1,2,65,67,63,55,71
78,1,4,2,1,2,39,54,54,53,41
64,1,4,3,1,3,50,52,45,58,36
63,1,4,1,1,1,52,65,60,56,51
79,1,4,2,1,2,60,62,49,50,51
193,1,4,2,2,2,44,49,48,39,51
92,1,4,3,1,1,52,67,57,63,61
160,1,4,2,1,2,55,65,55,50,61
32,1,2,3,1,3,50,67,66,66,56
23,1,2,1,1,2,65,65,64,58,71
158,1,4,2,1,1,52,54,55,53,51
```

```
25,1,2,2,1,1,47,44,42,42,36
188,1,4,3,2,2,63,62,56,55,61
52,1,3,1,1,2,50,46,53,53,66
124,1,4,1,1,3,42,54,41,42,41
175,1,4,3,2,1,36,57,42,50,41
;
```

```
NOTE: Writing HTML5 Body file: STDOUT
NOTE: The data set WORK.HSB2 has 192 observations and 11 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      cpu time            0.00 seconds
```

We want to check if the coefficient estimate of the `MATH` explanatory variable really doesn't contains the 0 value; that is if the `MATH` coefficient is statistically significant (p-value less than .05).

How can I bootstrap estimates in SAS?

Bootstrapping allows for estimation of statistics through the repeated resampling of data. In this page, we will demonstrate several methods of bootstrapping a confidence interval about the slope of the `MATH` explanatory variable in SAS. We will be using the `hsb2` dataset that can be found here. We will begin by running an OLS regression, predicting `read` with `female`, `math`, `write`, and `ses`, and saving the slope of `math` value in a dataset called `t0`. The estimated `MATH` paramter value in this regression is 0.4333.

```
proc genmod data = HSB2 ;
  model read = female math write ses;
  ods output parameterestimates = t0;
run;
```

The SAS System

The GENMOD Procedure

Model Information	
Data Set	WORK.HSB2
Distribution	Normal
Link Function	Identity

Model Information	
Dependent Variable	read

Number of Observations Read	192
Number of Observations Used	192

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	187	9749.9643	52.1388
Scaled Deviance	187	192.0000	1.0267
Pearson Chi-Square	187	9749.9643	52.1388
Scaled Pearson X2	187	192.0000	1.0267
Log Likelihood		-649.4785	
Full Log Likelihood		-649.4785	
AIC (smaller is better)		1310.9569	
AICC (smaller is better)		1311.4110	
BIC (smaller is better)		1330.5019	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	6.3343	3.2981	-0.1299	12.7985	3.69	0.0548
female	1	-2.4704	1.1133	-4.6525	-0.2883	4.92	0.0265

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
math	1	0.4333	0.0734	0.2894	0.5773	34.82	<.0001
write	1	0.4106	0.0744	0.2647	0.5564	30.43	<.0001
ses	1	1.3787	0.7423	-0.0761	2.8336	3.45	0.0633
Scale	1	7.1261	0.3637	6.4478	7.8757		

Note: The scale parameter was estimated by maximum likelihood.

Store the estimated `MATH` parameter:

```
data _null_;
    set t0;
    if parameter =  "math" then call symput('est_bar', estimate);
run;

%put &est_bar;
```

```
NOTE: Writing HTML5 Body file: STDOUT
NOTE: Numeric values have been converted to character values at the places given by: (L
      738:56
NOTE: There were 6 observations read from the data set WORK.T0.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

0.4333340545
```

To bootstrap a confidence interval about this `MATH` beta value, we will first need to resample. This step involves sampling with replacement from our original dataset to generate a new dataset the same size as our original dataset. For each of these samples, we will be running the same regression as above and saving the `MATH` beta parameter value. `proc surveyselect` allows us to do this resampling in one step.

Before carrying out this step, let's outline the assumptions we are making about our data when we use this method. We are assuming that the observations in our dataset are independent. We are also assuming that the statistic we are estimating is asymptotically normally distributed.

We indicate an output dataset, a seed, a sampling method, and the number of replicates. The sampling method indicated, `urs` , is unrestricted random sampling, or sampling with replacement. The samprate indicates how large each sample should be relative to the input dataset. A `samprate` of one means that the sampled datasets should be of the same size as the input dataset. So in this example, we will generate 500 datasets of 200, so our output dataset `bootsample` will have 100,000 observations.

```
%let rep = 100000;
proc surveyselect data= HSB2 out=bootsample
     seed = 1347 method = urs
         samprate = 1 outhits rep = &rep;
run;
```

The SAS System

The SURVEYSELECT Procedure

Selection Method	Unrestricted Random Sampling

Input Data Set	HSB2
Random Number Seed	1347
Sampling Rate	1
Sample Size	192
Expected Number of Hits	1
Sampling Weight	1
Number of Replicates	100000
Total Sample Size	19200000
Output Data Set	BOOTSAMPLE

With this dataset, we will now run our regression model, specifying by replicate so that the model will be run separately for each of the 100000 sample datasets. After that, we use a data step to convert the `MATH` beta paramter values to numeric.

```
ods select none;
ods output parameterestimates = tdata (where = (parameter =  "math"));
proc genmod data = bootsample;
  by replicate;
  model read = female math write ses;
  ods output parameterestimates = t (where = (parameter =  "math"));
run;
quit;
```

NOTE: The `ods select none` suppresses displayed output in SAS `GENMOD` . The `noprint` options deosn't work in this procedure. More about the use of the `ODS` system in this example see this

```
ods select all;
```

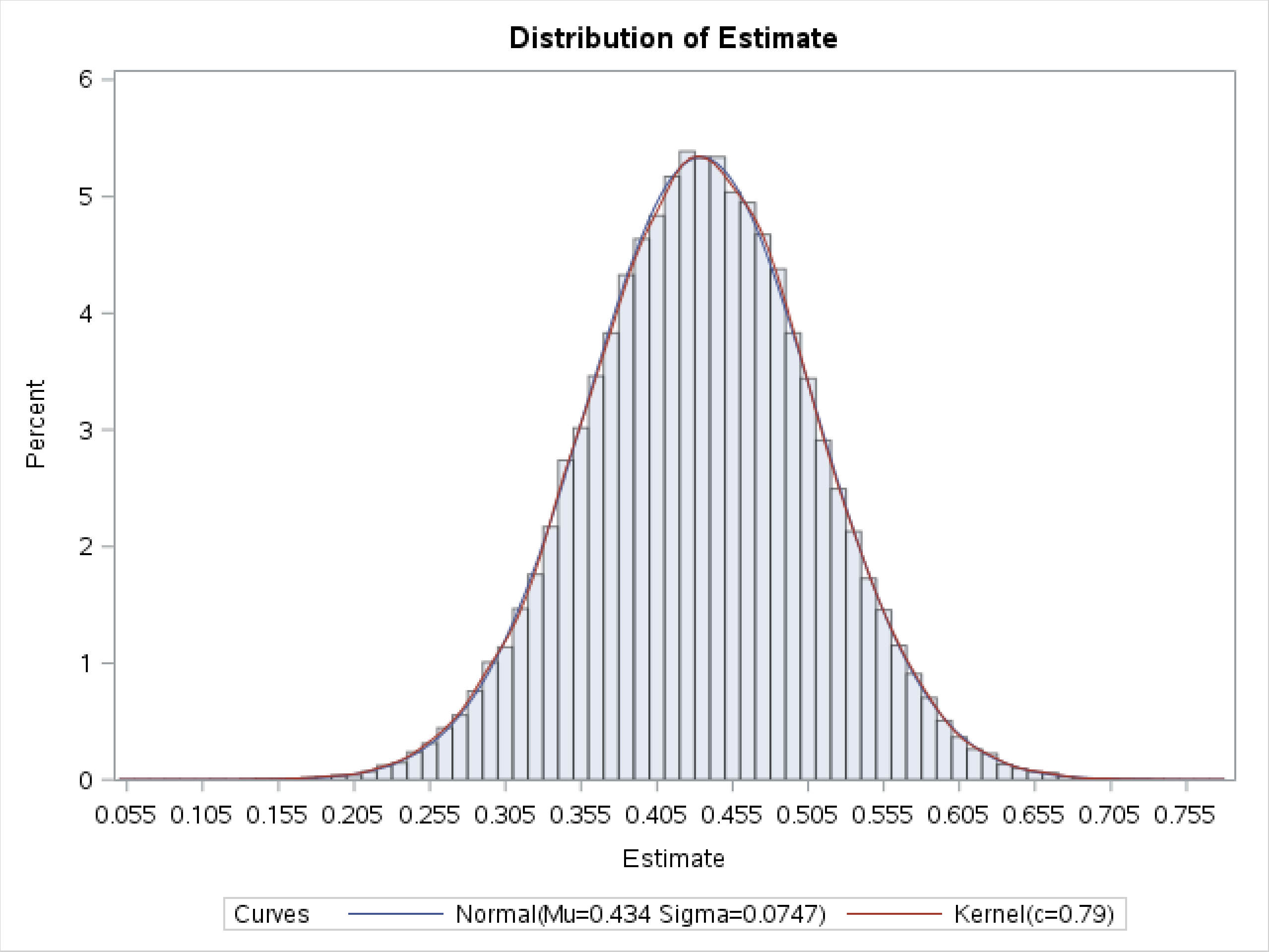
```
NOTE: Writing HTML5 Body file: STDOUT
```

The following histogram shows the distributions of the bootstrapped `MATH` beta estimate parameter for the 100000 samples follows a Normal distribution:

```
proc univariate data=tdata noprint;
    histogram Estimate / kernel (noprint) normal (noprint);
run;
```

The SAS System

The UNIVARIATE Procedure



Method 1: Normal Distribution Confidence Interval

We will first create a confidence interval using the normal distribution theory. This assumes that the Beta values follow a t distribution, so we can generate a 95% confidence interval by about the mean of the Betea values based on quantiles from a t-distribution with 99000 degrees of freedom. We find the critical t values for our confidence interval and multiply these by the standard deviation of the Beta values that arose in our 100000 replications. Our confidence interval using this method is symmetric about the Beta value we saw in our original regression. We can see that the 95% confidence interval using this method is (0.286893, 0.579775). We have also calculated the bias in our original value of Beta as the difference between that value and the mean of the 100000 Beatas in our bootstrap sample.

```
%let alphalev = .05;
ods listing;
proc sql;
  select  &est_bar as estimate,
          mean(estimate) - &est_bar as bias,
                  std(estimate) as std_err,
          &est_bar - tinv(1-&alphalev/2, &rep-1)*std(estimate) as lb,
          &est_bar + tinv(1-&alphalev/2, &rep-1)*std(estimate) as hb
  from t;
quit;
```

The SAS System

estimate	bias	std_err	lb	hb
0.433334	0.000637	0.074715	0.286893	0.579775

Method 2: Percentile Confidence Interval

Another way to generate a bootstrap 95% confidence interval from the sample of 100000 Betas for `MATH` values is to look at the 2.5th and 97.5th percentiles in this distribution. This approach to the confidence interval has some advantages over the normal approximation used above. This interval is not symmetric about the original estimate of the Beta and this method is unaffected by monotonic transformations on the estimated statistic. The first advantage is relevant because our original estimate is subject to bias. The second advantage is less relevant in this example than in an instance where the estimate might be subject to a transformation. The bootstrap estimates that form the bounds of the interval can be transformed in the same way to create the bootstrap interval of the transformed estimate.

We can easily generate a percentile confidence interval in SAS using `PROC UNIVARIATE` after creating some macro variables for the percentiles of interest and using them in the output statement. We can see that the confidence interval from this method is (0.28637, 0.57973). Since we have put the information of interest into a new dataset, pmethod, we have omitted the standard output from the proc univariate.

```
%let alphalev = .05;
%let a1 = %sysevalf(&alphalev/2*100);
%let a2 = %sysevalf((1 - &alphalev/2)*100);
* creating confidence interval, percentile method;
proc univariate data = t alpha = .05 noprint;
  var estimate;
  output out=pmethod mean = estimate_hat pctlpts=&a1 &a2 pctlpre = p pctlname = _lb _ub
run;

data t2;
  set pmethod;
  bias = estimate_hat - &est_bar;
  estimate = &est_bar;
run;
ods listing;
proc print data  = t2;
  var estimate bias p_lb p_ub;
run;
```

The SAS System

Obs	estimate	bias	p_lb	p_ub
1	0.43333	.000636566	0.28637	0.57973

Method 3: Bias-Corrected Confidence Interval

We can also correct for bias in calculating our confidence interval. We have calculated bias in the previous method as the difference between the Beta we observed in our initial regression and the mean of the 100000 Beta values from the bootstrap samples. The Beta estimate from the initial regression is assumed to be an unbiased estimate of the true parameter. If we wish to correct for the bias in calculating our confidence interval, we can go through the steps below. These are described by Cameron and Trivedi in Microeconomics Using Stata.

We first calculate the proportion of the bootstrap Beta that are less than our original value. We will adjust the percentiles used to define our confidence interval based on how this proportion differs from 0.5. We then find the probit of this proportion (z0) and the proportion associated with our alpha level (zalpha). Next, we calculate the two percentiles that will be used to find our confidence interval, p1 and p2, from these values. We then calculate our interval with proc univariate. From this method, our interval is (0.27954, 0.57413).

```
%let alphalev = .05;
%let alpha1 = %sysevalf(1 - &alphalev/2);
%put &alpha1;
proc sql;
  select sum(estimate<=&est_bar)/count(estimate) into :z0bar
  from t;
quit;

data _null_;
  z0 = probit(&z0bar);
  zalpha = probit(&alpha1);
  p1 = put(probnorm(2*z0 - zalpha)*100, 3.0);
  p2 = put(probnorm(2*z0 + zalpha)*100, 3.0);
  output;
  call symput('a1', p1);
  call symput('a2', p2);
run;
```

The SAS System

0.4952

Creating confidence interval, bias-corrected method

```
proc univariate data = t alpha = .05 noprint;
  var estimate;
  output out=pmethod mean = estimate_hat pctlpts=&a1 &a2 pctlpre = p pctlname = _lb _ub
run;

data t2;
  set pmethod;
  bias = estimate_hat - &est_bar;
  estimate = &est_bar;
run;

ods listing;

proc print data  = t2;
  var estimate bias p_lb p_ub;
run;
```

The SAS System

Obs	estimate	bias	p_lb	p_ub
1	0.43333	.000636566	0.27954	0.57413

Conclusion

Because the bootstrapping confident interval doesn’t contains the zero and is very similar to the confident interval of the `GENMOD` we conclude that the estimated slope of 0.4333 for the `MATH` explanatory variable is statistically significant and that our assumption in the GLM are correct.

The SAS code can be useful also to bootstrap the standard errors. Another good example consists in check the FEMALE variable that has a p-value close to 0.05.

Reference:

The code is based in the article titled "How can I bootstrap estimates in SAS?" where I changed the R-squared statistic by the slope of the `MATH` explanatory variable. There are also some other slightly differences.