Bootstrapping in Regression

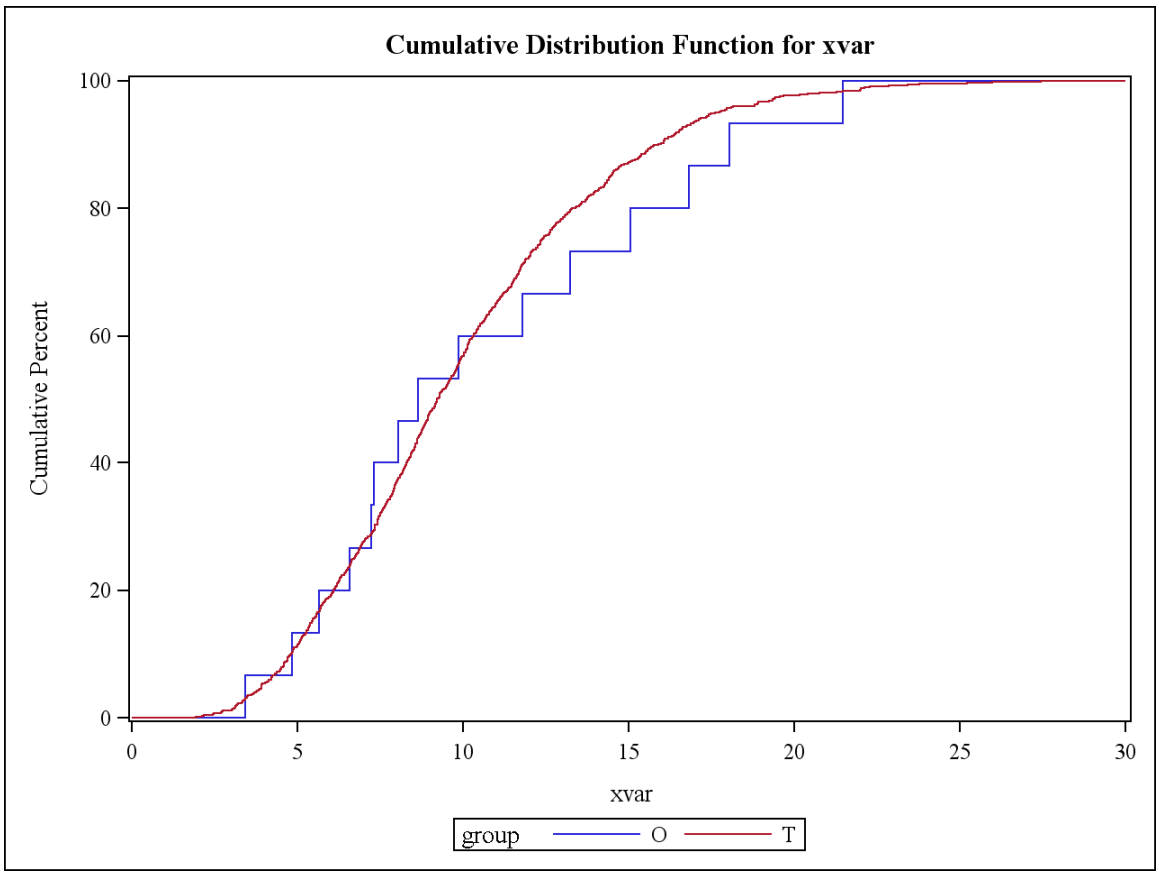
Bootstrapping combines Monte Carlo approach with the concept that Fhat is a good estimator for F.

That is, if $X_1, X_2, \ldots, X_n$ are sampled from F, then Fhat(t)= (#X's <= t) / n is a good estimator for F(t) = P(X <= t).

Example to show that Fhat is a Good Estimator for F = chi**^{2} with 10 df.

I sampled n=15 observations from a chi**^{2} with 10 df.

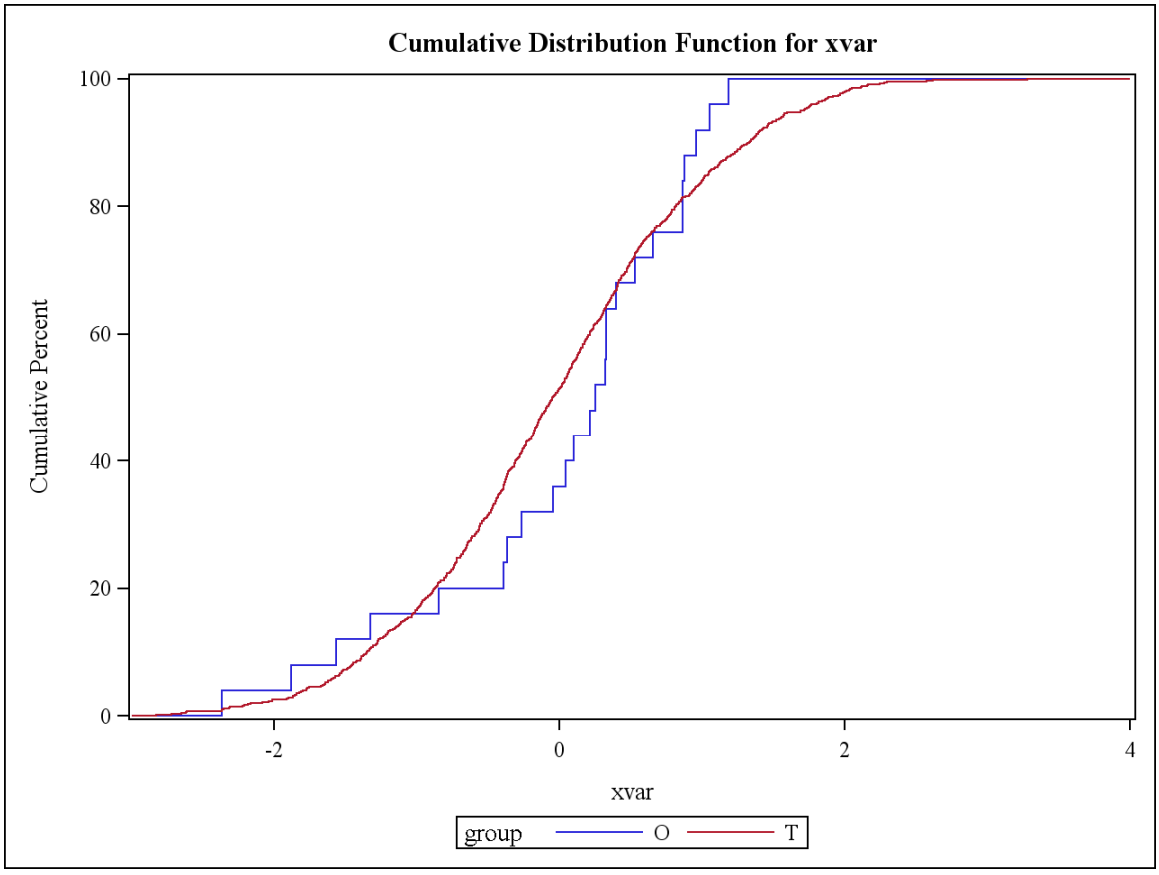
```
options pageno=1 orientation=landscape linesize=120
pagesize=40 nodate;
options nofmterr validvarname=v7;
OPTIONS FORMCHAR="|----|+|---+=|-/\<>*";

ods pdf file="FHAT1.pdf";

data chio;
do i=1 to 15;
xvar=rand('CHISQUARE',10);
group='O';
output;
end;
keep xvar group;
run;

data chit;
do i=1 to 1000;
xvar=rand('CHISQUARE',10);
group='T';
output;
end;
keep xvar group;
run;

data chisqsam;
set chit chio;
run;

ods graphics on;
proc capability data=chisqsam;
class group;
cdf xvar/overlay;
run;
ods graphics off;
run;
```



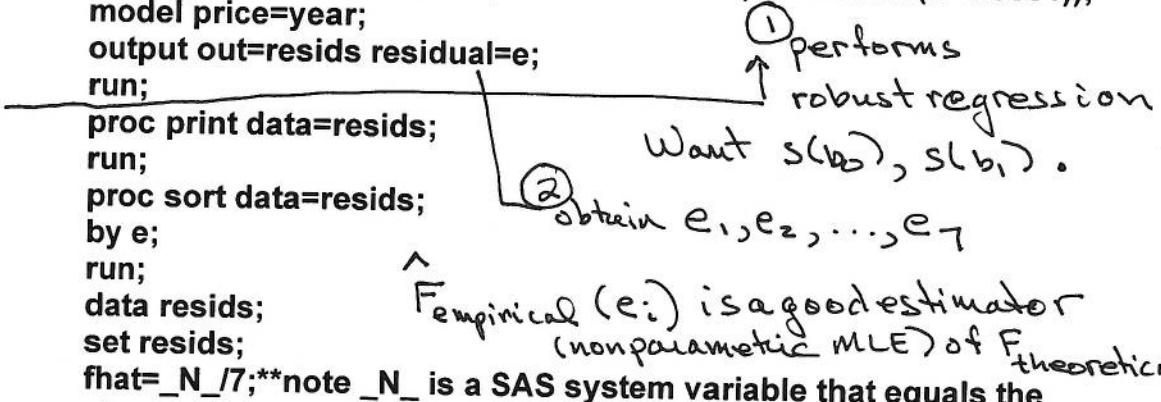
Another example: I sampled n=25 observations from a Normal (0,1) distribution.

```
options pageno=1 orientation=landscape linesize=120
pagesize=40 nodate;
options nofmterr validvarname=v7;
OPTIONS FORMCHAR="|----|+|---+=|-/\<>*";

ods pdf file="FHAT2.pdf";

data chio;
do i=1 to 25;
xvar=rand('NORMAL',0,1);
group='O';
output;
end;
keep xvar group;
run;

data chit;
do i=1 to 1000;
xvar=rand('NORMAL',0,1);
```

```
group='T';
output;
end;
keep xvar group;
run;

data normsam;
set chit chio;
run;

ods graphics on;
proc capability data=normsam;
class group;
cdf xvar/overlay;
run;
ods graphics off;
run;
```



```
options nonumber orientation=portrait linesize=90 pagesize=50
nodate;
options nofmterr validvarname=v7;
OPTIONS FORMCHAR="|----|+|---+=|-/\<>*";


/*example - growth of city prices in Taiwan 1940-1946*/;
/*interested in average rate of growth over these years*/;


title1 'Growth of City Prices in Taiwan 1940-1946';


data growprice;
input year price @@;
datalines;
40 1.62 41 1.63 42 1.90 43 2.64
44 2.05 45 2.13 46 1.94
;

proc robustreg data=growprice method=m(wf=median(c=0.0001));
model price=year;
output out=resids residual=e;
run;
proc print data=resids;
run;
proc sort data=resids;
by e;
run;
data resids;
set resids;
fhat=_N_/7;**note _N_ is a SAS system variable that equals the
observation number currently being processed;
run;

title 'Fhat (empirical distribution function) is a good estimator of the
theoretical F of the residuals';
proc gplot data=resids;
plot fhat*e='+'/haxis=-0.4 to 1 by .2 vaxis=0 to 1 by .1;
run;

/*Note the transpose procedure turns rows(observations) into
columns(variables)*/;
proc transpose data=resids out=boot prefix=e;
var e;
```



```
run;
proc print data=boot;
run;

data bootsam;
set boot;
x1=40;x2=41;x3=42;x4=43;x5=44;x6=45;x7=46;
b0=-2.4600;
b1=.1020;
array e(*) e1-e7;
array xvalue(7) x1-x7;
do samnum=1 to 300 by 1;*generate 300 bootstrap samples instead
of 200;
do j=1 to 7 by 1;
_i_=floor(ranuni(0)*7)+1;
estar=e(_i_);
year=xvalue(j);
pristar=b0+(b1*year) + estar;
output;
end;
end;
keep samnum pristar year estar;
run;
```

b_0, b_1 from LAR

③ resample $e_1^*, e_2^*, \ldots, e_7^*$ with replacement from $e_1, e_2, \ldots, e_7$

④ Compute $Y_i^* = b_0 + b_1 X_i + e_i^*$

Repeat a large number of times (300 in this case)

```
title1 'The results of the first 5 bootstrap samples are listed below';
proc print data=bootsam;
where samnum le 5;
run;

ods pdf close;

proc robustreg outest=bootest;
by samnum;
model pristar=year;
run;
```

⑤ Compute LAD to obtain b_0^*, b_1^* for each sample

```
title1 'We are interested in the standard deviation of the slope below';
proc means data=bootest n mean std;
var intercept year;
run;

ods pdf close;
run;
```

look at this output to obtain $s(b_0), s(b_1)$

⑥ 95% CI (bootstrapped) on B_1 is $b_1 \pm 1.96\, s(b_1)$

The ROBUSTREG Procedure

Model Information	
Data Set	WORK.GROWPRICE
Dependent Variable	price
Number of Independent Variables	1
Number of Observations	7
Method	M Estimation

Number of Observations Read	7
Number of Observations Used	7

Parameter Information	
Parameter	Effect
Intercept	Intercept
year	year

Summary Statistics						
Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
year	41.0000	43.0000	45.0000	43.0000	2.1602	2.9652
price	1.6300	1.9400	2.1300	1.9871	0.3471	0.2817

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-2.4600	.	.	.	.	.
year	1	0.1020	.	.	.	.	.
Scale	1	0.1127					

We can use bootstrapping to obtain s(b0) and s(b1).

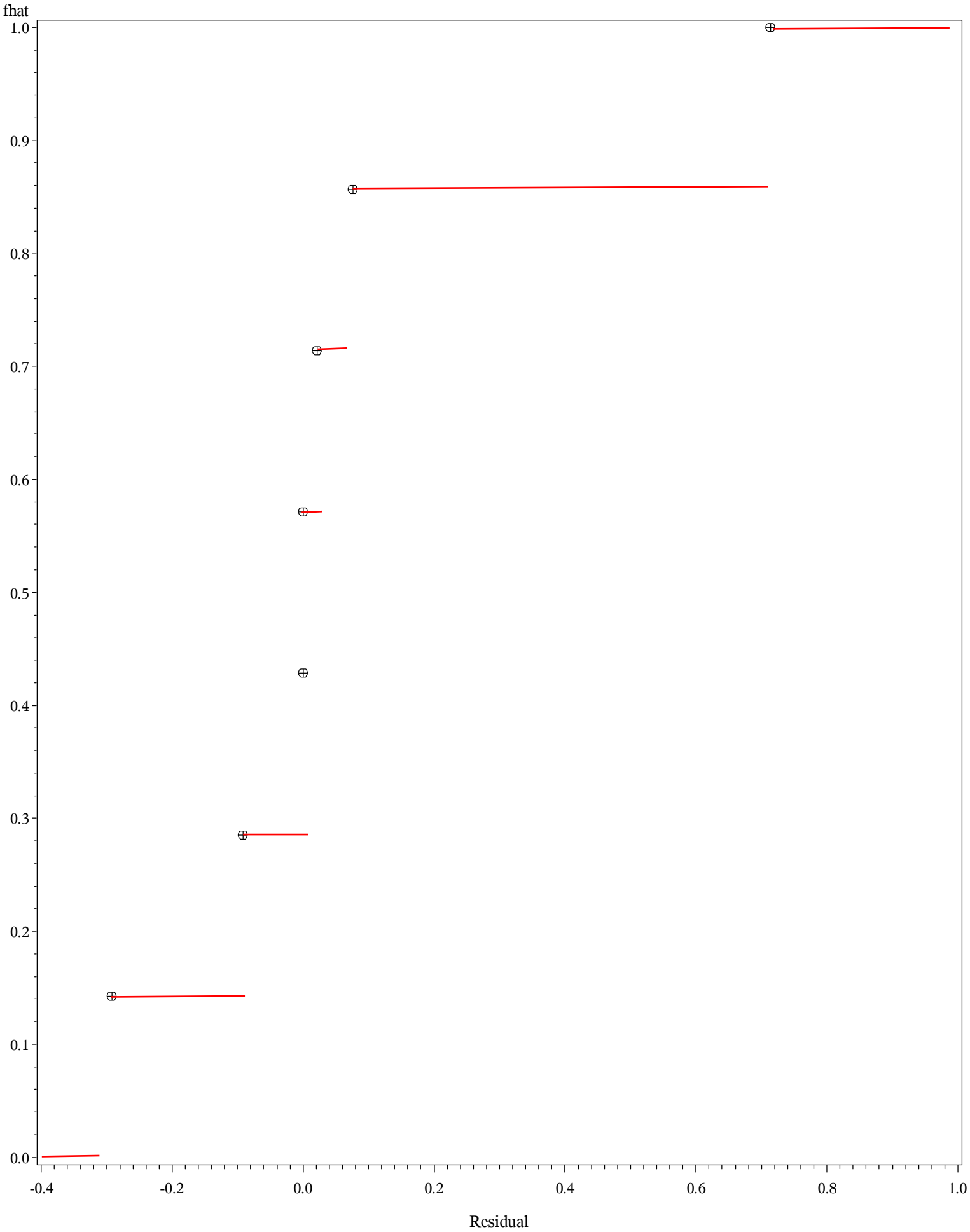
The ROBUSTREG Procedure

Diagnostics Summary		
Observation Type	**Proportion**	**Cutoff**
Outlier	0.1429	3.0000

Goodness-of-Fit	
Statistic	**Value**
R-Square	0.2838
AICR	.
BICR	25.1205
Deviance	0.2695

Obs	year	price	e
1	40	1.62	-0.00000
2	41	1.63	-0.09200
3	42	1.90	0.07600
4	43	2.64	0.71400
5	44	2.05	0.02200
6	45	2.13	-0.00000
7	46	1.94	-0.29200

Fhat (empirical distribution function) is a good estimator of the theoretical F of the residuals



Fhat (empirical distribution function) is a good estimator of the theoretical F of the residuals

Obs	_NAME_	_LABEL_	e1	e2	e3	e4	e5	e6	e7
1	e	Residual	-0.29200	-0.092000	-3.7956E-8	-1.4017E-15	0.022000	0.076000	0.71400

The results of the first 5 bootstrap samples are listed below

Obs	samnum	estar	year	pristar
1	1	0.71400	40	2.33400
2	1	-0.00000	41	1.72200
3	1	0.07600	42	1.90000
4	1	-0.09200	43	1.83400
5	1	0.71400	44	2.74200
6	1	0.02200	45	2.15200
7	1	0.71400	46	2.94600
8	2	0.02200	40	1.64200
9	2	0.07600	41	1.79800
10	2	0.07600	42	1.90000
11	2	0.02200	43	1.94800
12	2	0.02200	44	2.05000
13	2	-0.00000	45	2.13000
14	2	0.07600	46	2.30800
15	3	0.07600	40	1.69600
16	3	-0.00000	41	1.72200
17	3	0.07600	42	1.90000
18	3	0.71400	43	2.64000
19	3	-0.00000	44	2.02800
20	3	-0.09200	45	2.03800
21	3	-0.29200	46	1.94000
22	4	0.71400	40	2.33400
23	4	0.02200	41	1.74400
24	4	0.02200	42	1.84600
25	4	-0.29200	43	1.63400
26	4	-0.09200	44	1.93600
27	4	0.07600	45	2.20600
28	4	-0.09200	46	2.14000
29	5	0.71400	40	2.33400
30	5	-0.29200	41	1.43000
31	5	0.71400	42	2.53800
32	5	0.71400	43	2.64000
33	5	-0.29200	44	1.73600

The results of the first 5 bootstrap samples are listed below

Obs	samnum	estar	year	pristar
34	5	-0.00000	45	2.13000
35	5	0.02200	46	2.25400

We are interested in the standard deviation of the slope below

The MEANS Procedure

Variable	Label	N	Mean	Std Dev
Intercept	Intercept	297	-2.4833539	1.8906810
year		297	0.1028088	0.0442630

In 3 cases the IRLS algorithm did not converge so there are 297 bootstrapped estimates of b0 and b1 instead of 300.

95% CI on B1 is 0.1020 +/- 1.96 (.04426)
or 0.1020 +/- 0.0867 or (0.0153,0.1887)