CHAPTER 26

Resampling Procedures (Bootstrapping)

The ability to do a lot of computation extremely fast has led to the use of techniques that provide "new" sets of data by resampling numbers generated from a single data set. These methods are used in many different statistical situations. In this brief introduction, resampling methods are illustrated specifically in the context of linear regression. Via resampling methods, we can reexamine a regression analysis already made, by comparing it with a population of results that might have been obtained under certain assumed circumstances. Often, an important point of bootstrapping is not just to evaluate estimates of the parameters, but also to obtain good estimates of standard errors from the distributions generated by the parameter estimates in bootstrapped iterations. This would be especially valuable in fitting situations where standard errors were *not* directly derivable from theory, for example, in nonlinear estimation situations, or situations where least squares is inappropriate, such as for generalized linear models. In our linear model examples, we compare resampling parameter estimates and standard errors with the corresponding least squares values.

26.1. RESAMPLING PROCEDURES FOR REGRESSION MODELS

Two resampling procedures that can be used in the regression context are:

(a) Fit the linear model and obtain the n residuals. Choose a sample of size n from the residuals, generated with probability $1/n$ for each residual, and sampling with replacement. Attach these sampled values to the n predicted $\hat{Y}_i$ to give a resampled set of Y's. Thus if the model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and $\hat{\mathbf{Y}} = \mathbf{Xb}$, the new $\mathbf{Y}$-values are

$$\mathbf{Y}^* = \mathbf{Xb} + \mathbf{e}^*, \tag{26.1.1}$$

where $\mathbf{e}^*$ is a resampled set from the vector $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. Least squares regression is now performed on the model

$$\mathbf{Y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{26.1.2}$$

to obtain an estimate $\mathbf{b}^*$ (say). As many iterations as desired can be performed, and the usual sample mean and sample standard deviation of each of the elements of those vector estimates can be found. In our tables, these are called the "bootstrap averages" and the "bootstrap standard errors," respectively.

(b) The resampling could also be carried out on the pairs $(Y_i, \mathbf{x}_i')$, where Y_i is the ith observation and $\mathbf{x}_i'$ is the ith row of the $\mathbf{X}$ matrix. The resampling involves selecting a set of n of the $(Y_i, \mathbf{x}_i')$, each selected with probability 1/n, and sampling with replacement, to obtain (say) $\mathbf{Y}^{**}$ and $\mathbf{X}^{**}$. The regression model

$$\mathbf{Y}^{**} = \mathbf{X}^{**}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{26.1.3}$$

is now fitted by least squares. Again, the properties of the corresponding $\mathbf{b}^{**}$ values can be examined after any desired number of iterations.

Note that, after either resampling (a) or (b) has been carried out, the model could also be fitted robustly, rather than by least squares. Numerous variations are possible.

Both procedures (a) and (b), and variations of them, are called *bootstrapping procedures.* In (a) we have "bootstrapped the residuals"; in (b) we have "bootstrapped pairs."

26.2. EXAMPLE: STRAIGHT LINE FIT

We make use of the steel employment data of Table 25.2, but with the 1992 Germany value of 132 adjusted to 104 (to remove the effect of the former East Germany; see Section 25.3). This gives the data X, Y of Table 26.1. The predicted values and the residuals are those from the fitted least squares equation $\hat{Y} = 2.803 + 0.3337X$ and are rounded to one decimal place.

(a) A MINITAB program (see Appendix 26A) was used to bootstrap residuals. Five example iterations gave the results of the upper portion of Table 26.2. We see that such a small data set can lead to some peculiarities in the samples, for example, iteration 2 used the same residual (No. 2, 15.2) five times out of 10. We see, however, that after 100 iterations, the averages of the estimates are only slightly different from the least squares values and the standard errors are of comparable size. Another anomaly that can arise is an occasional negative Y^* value, which is of course not physically possible. Such observations could be adjusted to zero, or discarded and replaced.

(b) A second MINITAB program (see Appendix 26B) was used to bootstrap pairs (Y, X) from the original data set. Table 26.3 shows results in a format similar to that of Table 26.2. The bootstrapped "averaged fits" provide lines quite close to the least squares values.

TABLE 26.1. Steel Employment by Country in Europe in Thousands, in 1974 (X) and 1992 (Y) Together with Predicted Values and Residuals from the Least Squares Fit: Y_1 Has Been Adjusted to 104 from 132

Observation Number	1974 (X)	1992 (Y)	$\hat{Y}$	Residuals e_i
1	232	104	80.2	23.8
2	96	50	34.8	15.2
3	158	43	55.5	−12.5
4	194	41	67.5	−26.5
5	89	33	32.5	0.5
6	64	25	24.2	0.8
7	25	16	11.1	4.9
8	23	8	10.5	−2.5
9	4	3	4.1	−1.1
10	2	1	3.5	−2.5

TABLE 26.2. Five Sample Iterations of Bootstrapping Residuals with Results of 100 Iterations

Iteration	Residual Numbers Used		b_0^*	b_1^*
1	9, 6, 8, 9, 1, 2, 2, 8, 2, 7		14.479	0.2784
2	1, 9, 8, 7, 2, 2, 5, 2, 2, 2		13.317	0.3294
3	10, 4, 8, 7, 8, 8, 8, 2, 7, 1		8.878	0.2762
4	4, 10, 8, 5, 7, 2, 7, 9, 6, 3		6.119	0.2750
5	7, 3, 10, 4, 4, 1, 7, 6, 5, 9		4.180	0.2794
	Five iterations (residuals)	Bootstrap averages	9.395	0.2877
		Bootstrap standard errors	4.456	0.0234
	Least squares values		2.803	0.3337
	Least squares standard errors		6.992	0.0593
	100 iterations (residuals)	Bootstrap averages	3.523	0.3296
		Bootstrap standard errors	6.032	0.0530

Using the Original Data

We reinstated the original first reading of 132 (which was replaced by 104 in the calculations above) and performed 100 simulations on bootstrapped residuals, and another 100 on bootstrapped pairs. The results were (standard errors in parentheses):

Least squares values	$b_0 = -0.314(9.997)$,	$b_1 = 0.4004(0.0849)$
Bootstrapping residuals	$b_0^* = -0.469(8.626)$,	$b_1^* = 0.4078(0.0712)$
Bootstrapping pairs	$b_0^{**} = -0.004(7.258)$,	$b_1^{**} = 0.3927(0.1258)$

We see that the least squares values are approximately reproduced and that the first observation (which has the largest internally studentized residual of 2.53 in the least squares fit) is not "flagged" in any way by the bootstrap procedures. Bootstrapping procedures will not do anything for us in this regard because such residuals and/or the corresponding pairs $(Y_i, \mathbf{x}_i')$ have their share of appearances in the random choices of "new" data.

TABLE 26.3. Five Sample Iterations of Bootstrapping Pairs and Results of 100 Iterations

Iteration	Numbers of (Y, X) Pairs Used		b_0^{**}	b_1^{**}
1	5, 7, 5, 9, 9, 8, 3, 10, 9, 5		3.378	0.2920
2	4, 5, 7, 4, 4, 1, 5, 6, 4, 1		1.910	0.3051
3	7, 3, 3, 7, 5, 3, 2, 8, 9, 1		2.718	0.3428
4	7, 3, 9, 10, 7, 1, 4, 1, 10, 6		0.918	0.3676
5	2, 3, 5, 6, 9, 7, 1, 5, 3, 9		1.664	0.3660
	Five iterations (pairs)	Bootstrap averages	2.119	0.3347
		Bootstrap standard errors	0.953	0.0347
	Least squares values		2.803	0.3337
	Least squares standard errors		6.992	0.0593
	100 iterations (pairs)	Bootstrap averages	3.810	0.3242
		Bootstrap standard errors	3.995	0.0841

TABLE 26.4. **Five Hundred Bootstrappings on the Steam Data (Y, X_5, X_6, X_8)**

b_0	b_5	b_6	b_8	Method Used
−2.968	0.402	0.199	−0.074	Least
4.833	0.157	0.041	0.0072	squares
−2.724	0.393	0.200	−0.074	Bootstrapping
4.356	0.141	0.038	0.0068	residuals
−4.047	0.410	0.238	−0.074	Bootstrapping
6.880	0.234	0.107	0.0074	pairs

26.3. EXAMPLE: PLANAR FIT, THREE PREDICTORS

We next look briefly at a larger example, using the steam data (see Appendix 1A) to fit a plane

$$Y = \beta_0 + \beta_5 X_5 + \beta_6 X_6 + \beta_8 X_8 + \epsilon. \qquad (26.3.1)$$

Table 26.4 gives results of 500 "connected" sets of sampled data. Averages of the 500 sets of bootstrapped parameter estimates are in the top row of each pair of lines and the corresponding bootstrapped standard errors are in the bottom row of each pair of lines. The two sets of iterations (a) on the residuals e_i and (b) on the $(Y_i, \mathbf{x}_i')$ pairs were done using the same choices of random numbers for the observations selected. That is, 500 sets of random choices were made, and each set was used on (a) and (b). This was done as a matter of convenience. We see that the results obtained by bootstrapping residuals are, on the whole, closer to the least squares results, and that bootstrapping pairs appears to provide larger variance estimates then does least squares.

26.4 REFERENCE BOOKS

Edgington, E. (1980). *Randomization Tests*, 2nd ed. New York: Marcel Dekker.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Procedures*. Philadelphia: SIAM Publications.

Efron, B., and R. Tibshirani (1993). *Introduction to the Bootstrap*. London: Chapman and Hall.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer-Verlag.

Hall, P. (1995). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Hjorth, J. S. U. (1994). *Computer Intensive Statistical Methods*. London: Chapman and Hall.

Manley, B. (1992). *Randomization and Monte Carlo Methods in Biology*. London: Chapman and Hall.

Maritz, J. S. (1994). *Distribution Free Statistical Methods*. London: Chapman and Hall.

Noreen, E. (1989). *Computer-Intensive Methods for Testing Hypotheses*. New York: Wiley.

Westfall, P., and S. Yound (1992). *Resampling-Based Multiple Testing*. New York: Wiley.

For other useful information, see the review by J. Albert and M. Berliner of the "Resampling Stats" computing package in *American Statistician*, **48**, 1994, 129–131 and "Bootstrap: more than a stab in the dark?" by G. A. Yound, *Statistical Science*, **9**, 1994, 382–415, which contains extensive discussion.

Acknowledgment. We are grateful to Ying Kuen (Ken) Cheung who provided computations and commentary while pursuing "directed study" credits at the University of Wisconsin–Madison.

APPENDIX 26A. SAMPLE MINITAB PROGRAMS TO BOOTSTRAP RESIDUALS FOR A SPECIFIC EXAMPLE

```
#ken.r  MAIN PROGRAM
oh=0
set cl
104 50 43 41 33 25 16 8 3 1
set c2
232 96 158 194 89 64 25 23 4 2
name cl 'Y'
name c2 'X'
name c4 'fitted'
name c5 'resids'
let k1=1
let k2=10                 # number of data points
let k10=100               # number of iterations
let k20=1
let c9=0
let c10=0
name c7 'Y*'
noecho
print k2
print k10
regress c1 1 c2 c22 c4;
residuals c5.
name c5 'resids'
name c4 'fitted'
print c2 c1 c4 c5
plot c1 c2
exec 'kena.r' k10      #call subprogram 1
let c11=c9/k10
let c12=(c10-k10*cll**2)/(k10-1)
let c13=sqrt(c12)
print c11-c13
end
stop
```

```
#kena.r    SUBPROGRAM 1
random k2 c6;
integer k1 k2.
print k20
print c6
let k3=1
exec 'kenab.r' k2      #call subprogram 2
noecho
print 'Y*' 'X'
regress 'Y*' 1 'X';
coeff c8.
let c9=c9+c8
let c10=c10+c8**2
let k20=k20+1
end
```

```
#kenab.r    SUBPROGRAM 2
let k4=c6(k3)
let c7(k3)=c4(k3)+c5(k4)
let k3=k3+1
end
```

APPENDIX 26B. SAMPLE MINITAB PROGRAMS TO BOOTSTRAP PAIRS FOR A SPECIFIC EXAMPLE

```
#ken.p  MAIN PROGRAM
oh=0
set c1
104 50 43 41 33 25 16 8 3 1
set c2
232 96 158 194 89 64 25 23 4 2
name c2 'X'
let k1=1
let k2=10                #number of data points
let k10=100              #number of iterations
let k20=1
let c9=0
let c10=0
name c4 'Y*'
name c5 'X*'
noecho
print k2
print k10
regress c1 1 c2
print c1 c2
plot c1 c2

exec 'kena.p' k10      #call subprogram 1
let c11=c9/k10
let c12=(c10-k10*c11**2)/(k10-1)
let c13=sqrt(c12)
print c11-c13
end
stop
```

```
#kena.p    SUBPROGRAM 1
random k2 c6;
integer k1 k2.
print k20
print c6
let k3=1

exec 'kenab.p' k2      #call subprogram 2
noecho
print 'Y*' 'X*'
regress 'Y*' 1 'X*';
coeff c8.
let c9=c9+c8
let c10=c10+c8**2
let k20=k20+1
end
```

```
#kenab.p    SUBPROGRAM 2
let k4=c6(k3)
let c4(k3)=c1(k4)
let c5(k3)=c2(k4)
let k3=k3+1
end
```

ADDITIONAL COMMENTS

Bootstrapping residuals can be thought of as working with a fixed **X** matrix; bootstrapping pairs corresponds to using a random **X**. Both procedures above can be described as nonparametric procedures. If (as a third alternative) the residuals were independently sampled from a $N(0, s^2)$ distribution where s^2 was the least squares estimate of σ^2 from the original data, we would have a *parametric* bootstrap procedure. For discussion, see the references on page 588.

EXERCISES FOR CHAPTER 26

A. Consider any of the data sets in this book and use the least squares fit as a basis for bootstrapping the residuals as described in the text. Start by using 100 iterations and compare the least squares fit with the bootstrap results. Then do more iterations on the same data to see what effect that has on your conclusions.

B. Consider any of the data sets in this book and bootstrap pairs as described in the text. Begin with 100 iterations and then do more to see the effect on your results. Compare your results with the least squares fit to the original data and to your bootstrap results in Exercise A.