International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

A Study Based on the Application of Bootstrap and Jackknife Methods in Simple Linear Regression Analysis

Tolga Zaman[a]*, Kamil Alakuş[b]

[a,b]*Department of Statistics, Faculty of Science and Arts, Ondokuz Mayıs University, Kurupelit, Turkey*

[a]*Email: tolga.zaman@omu.edu.tr*

[b]*Email: kamilal@omu.edu.tr*

Abstract

In the study, bootstrap and jackknife methods, which are used as a correction term when assumptions of the error in simple linear regressions are not met, are explored in detail. In the application, model parameters, coefficients of determination, standard errors, coefficients of correlation and %95 confidence intervals belonging to these methods are estimated with the help of a real data and the obtained results are interpreted.

Keywords: Bootstrap; Jackknife; Simple linear regression; Mean squared error; Coefficient of determination; Coefficient of correlation.

1. Introduction

Resampling methods used in applied statistics are also used in simple linear regression analysis. For the estimation in the least squares method to give good results, it needs to meet certain assumptions. Bootstrap method is used as an alternative approach when the assumptions are not met in regression analysis. In linear regression, using bootstrap and jackknife methods to estimate sampling distribution of coefficients of regression is firstly suggested by Efron in 1979 [1] and it is improved by Freedman in 1981 [2] and Wu in 1986 [3,4].

---

* Corresponding author.

In recent years, with the improvement of computer technology, it is possible to estimate more efficient and consistent parameters through the usage of methods known as resampling methods. In this regard, in the study, superiorities of bootstrap, jackknife and least squares methods in simple linear regression are explored according to the results obtained from the methods. In 2nd and 3rd Chapter, bootstrap and jackknife methods are going to be briefly explained. In 4th Chapter, stages of bootstrap and jackknife methods in simple linear regression are going to be explored with the help of a real data. And in the last part, results obtained in the study are going to be interpreted. The purpose of this study is to introduce and compare resampling methods in regression analysis.

2. Jackknife Method

Jackknife methods is suggested the first time by Quenoille in 1949 and in 1956 [5-6]. Tukey in 1958 [7] used it to calculate estimate and confidence intervals [8]. Jackknife method is based on excluding one observation value in each trial and each time, statistics for parameter, θ, in remaining observations are calculated.

Fundamental logic of the methods comes from calculating sampling statistics from remaining observations through excluding an observation one time in data set. Thus, only n different observations from n observations can be formed.

Let us have $X = (x_1, x_2, \dots, x_n)$ sample and $\hat{\theta} = s(X)$ be our estimator. According to jackknife methods, when $i.$ observation are excluded, new sample is;

$$x_{(i)} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \ ; \ i = 1,2,\dots,n \quad (1)$$

Because of this, its estimator is also;

$$\hat{\theta}_{(i)} = s\left(x_{(i)}\right) \quad (2)$$

Jacknife estimate of bias is defined as follows,

$$bias = \left(\hat{\theta}_{(.)} - \hat{\theta}\right) \quad (3)$$

Here, $\hat{\theta}_{(.)}$ is the estimate of θ and calculated through the equation, $\hat{\theta}_{(.)} = \frac{\sum_{i=1}^{n} \hat{\theta}_{(i)}}{n}$, Jackknife estimate of standard error is;

$$\widehat{se}_{Jackk} = \frac{\left\{\frac{\left[\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)}\right)^2\right]}{n-1}\right\}^{1/2}}{\sqrt{n}} \quad (4)$$

And to calculate pseudo values in jackknife methods, following equation [9-10] is used,

$$Psedovalue_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)} \quad (5)$$

The statistic to use here can be mean, median etc. Applying jackknife method in simple linear regression can be taken as the same logic. Through excluding one observation each time from the current dependent and independent variables, least squares method is applied. This process is repeated times of the sample size. Then, pseudo values belonging to estimate of the model parameters, model standard error, and coefficient of determination and correlation of the model are calculated. Regression model and statistics belonging to it are estimated through averaging pseudo values of calculated values.

3. Bootstrap Method

Bootstrap method, which came into the picture as a resampling method the first time, is suggested as an alternative to another resampling method, Jackknife method, by Efron in 1979. Bootstrap, which includes more calculations than distribution assumptions of traditional parametric result and mathematical analysis, is used as an approach to the statistical result [11]. Bootstrap is developed to calculate sample mean, standard error and to form confidence intervals [12].

Bootstrap method known as two different form. The first is parametric bootstrap and the second is nonparametric bootstrap method. Before using parametric bootstrap method, assumption is made for sample distribution. While, for example, two parameter is needed in normal distribution, one parameter is needed for Poisson distribution. And in nonparametric method, statistic is estimated by the usage of sampling with replacement and distribution of the statistic is tried to be determined [13].

This method, which is applied through repetition of error terms in regression analysis, is suggested by Bradley Efron in 1979 and is improved to obtain more efficient parameter estimations than classical least squares method. It is known as resampling of error term. Algorithm;

1) n sample is chosen from population depending on luck.

2) LSM is applied to the chosen sample.

3) e_i is calculated from this model.

4) n sized B bootstrap error subsamples are formed by giving $1/n$ probability to obtained e_i values.

So, experimental distribution function is obtained.

5) Bootstrap error of the mean values are calculated from experimental distribution function as follows;

$$\bar{\hat{\varepsilon}}_i^* = \frac{\sum_{b=1}^{B} \hat{\varepsilon}_{bi}}{B} \tag{6}$$

6) Obtained $\bar{\hat{\varepsilon}}_i^*$ values are put into place of e_i in the model, which is formed in 2nd step, and,

$$Y_i^* = X\underline{\hat{\beta}} + \bar{\hat{\varepsilon}}_i^* \tag{7}$$

So, bootstrap Y_i^* values are calculated as above.

7) β ′'s bootstrap estimator is calculated from Y_i^* and X with least squares method as following (Topuz, 2002),

$$\underline{\hat{\beta}}^{*k} = (X'X)^{-1}X'\underline{Y_i^*} \tag{8}$$

4. Real data application

Data set used in the study comes from Mikey and showed widely. Explanatory variable is the age of a child when he spoke his first word and dependent variable is Gesell adaptation value of it [15]. This data is for 21 children like in Table 1.

Table 1: First Word-Gesell Adaptation Score Data

Rank No	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Gesell Score	15	26	10	9	15	20	18	11	8	20	7	9	10	11	11	10	12	42	17	11	10
As month Age	95	71	83	91	102	87	93	100	104	94	113	96	83	84	102	100	105	57	121	86	100

Simple linear regression analysis of this data is analyzed in R programming language. Firstly, state of residual, which is obtained using estimation model, is interpreted in Figure 1.

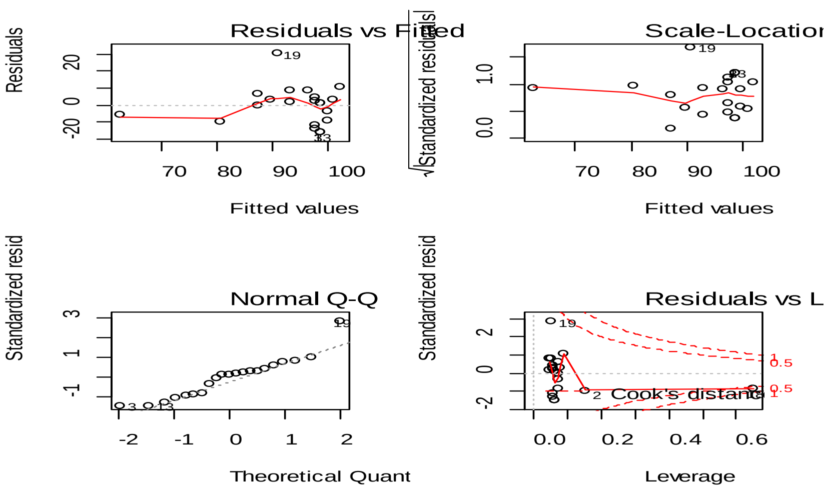

Figure 1: Graphs belonging to residual obtained using estimation model

If we look at Figure 1 carefully, we see that 19th observation is deviated value. In this regard, resampling methods can be used as a correction term. Thus, behaviors of bootstrap and jackknife methods in simple linear regression model are explored in the study.

Linear regression results of data set is shown in Table 2 and obtained results are indicated.

Table 2: Simple Linear Regression Results

Variable	Non-standardize Coefficients		t-test	Probability of Significance
	B	Std. Error		
Constant	109.874	5.068	21.681	0.000
X_1	-1.127	0.310	-3.633	0.002
$R^2 = 0.41; \hat{\sigma} = 11.023; r = -0.64$				
Significance test of the model, F=13.2002 and probability of significance is 0.002.				

When we look at Table 2, regression coefficients, which are estimated for the model, are found statistically significant ($p < 0.05$). Simple linear regression model, which is formed with the help of these variables, is also found significant ($p = 0.002 < 0.05$). It is calculated as $R^2 = 0.41$ for the model. In another words, rate for explanatory variable to explain variance of the model is 0.41. Standard error of the model is 11.023 and correlation between variables is -0.64.

And now, let us look into the estimation of model parameters in simple linear regression analysis with jackknife and bootstrap methods;

In Table 3, beta coefficients, coefficient of correlation, standard error of the model and coefficient of determination, which are obtained from applied regression analysis through excluding one observation one by one with jackknife method, is seen.

When one observation is deleted for each in turn.

$\hat{Y}_{-1} = a_{-1} + b_{-1}X = 109.7873 - 1.128X; r_{-1} = -0.64 ;\hat{\sigma}_{-1} = 11.31433 ; R^2_{-1} = 0.410$

$\hat{Y}_{-2} = a_{-2} + b_{-2}X = 108.9151 - 1.0228X; r_{-2} = -0.59 ;\hat{\sigma}_{-2} = 11.056 ; R^2_{-2} = 0.348$

...

$\hat{Y}_{-21} = a_{-21} + b_{-21}X = 109.7286 - 1.1218X; r_{-21} = -0.63 ;\hat{\sigma}_{-2} = 11.319 ; R^2_{-2} = 0.404$

Obtained values are shown in Table 3.

Table 3: Statistics obtained through Jackknife Method

	a_{-n}	b_{-n}	R^2_{-n}	$\hat{\sigma}_{-n}$	r_{-n}
None excluded	109.874	-1.127	0.41	11.02291	-0.64
1. excluded	109.7873	-1.128	0.410798	11.31433	-0.6409348
2. excluded	108.9151	-1.0228	0.347736	11.055957	-0.5896918
3. excluded	111.4973	-1.1847	0.45988	10.668705	-0.6781446
4. excluded	110.8967	-1.1670	0.429845	11.121977	-0.6556254
5. excluded	109,489	-1.1316	0.421076	11.11286	-0.648904
6. excluded	109.8679	-1.1253	0.40288	11.324667	-0.6347284
7. excluded	109.8506	-1.1373	0.41306	11.294609	-0.6426972
8. excluded	109.6435	-1.1198	0.405294	11.308398	-0.6366274
9. excluded	109.4631	-1.1097	0.395389	11.298614	-0.6287992
10. excluded	109.9915	-1.1589	0.4222	11.206822	-0.6497692
11. excluded	108.2788	-1.0561	0.382097	10.992783	-0.6181397
12. excluded	110.3109	-1.1440	0.412941	11.288168	-0.642605
13. excluded	111.4973	-1.1847	0.45988	10.668705	-0.6781446
14. excluded	111.1042	-1.1652	0.44527	10.842422	-0.6672854
15. excluded	109.4609	-1.1141	0.404415	11.271643	-0.635936
16. excluded	109.7286	-1.1218	0.404088	11.31986	-0.6356794
17. excluded	109.1919	-1.1097	0.409802	11.129664	-0.6401579
18. excluded	105.6299	-0.7792	0.112163	11.106756	-0.3349073
19. excluded	109.3047	-1.1933	0.571631	8.628196	-0.7560628
20. excluded	110.9216	-1.1595	0.436775	10.977128	-0.6608894
21. excluded	109.7286	-1.1218	0.404088	11.31986	-0.6356794

Pseudo values of the values in Table 3 are obtained with the help of following equation using these values and are in Table 4.

$$ps_i(X) = n * \emptyset_n(X_1, \ldots, X_n) - (n-1)\big(\emptyset_{n-1}(X_1, \ldots, X_n)\big)_{[i]}$$

For intercept estimates,

$$a_1^* = 21 * (109.874) - 20 * (109.7873) = 111.6$$

$$a_2^* = 21 * (109.874) - 20 * (108.9151) = 129.05$$

...

$$a_{21}^{*} = 21 * (109.874) - 20 * (109.7286) = 112.7$$

For slope estimates,

$$b_1^{*} = 21 * (-1.127) - 20 * (-1.128) = -1.1$$

$$b_2^{*} = 21 * (-1.127) - 20 * (-1.0228) = -3.2$$

...

$$a_{21}^{*} = 21 * (-1.127) - 20 * (-1.1218) = -1.2$$

In Table 4, pseudo values of regression coefficients, standard error of the model, coefficient of correlation and determination are given. With the help of this table, jackknife estimations are found. Let us show the estimation value, which is obtained using jackknife estimate, as $\hat{Y}_i^{*}$ and these values are obtained using following equation,

$$\hat{Y}_i^{*} = a^{*} + b^{*}X = 112.5322 - 1.347X$$

Moreover, $r_{\hat{Y}_i^{*},\hat{Y}_i} = 1$. This shows that the correlation between the original data and the estimation values is the same for both $\hat{Y}_i^{*}$ and the $\hat{Y}_i$. Jackknife estimations are calculated like Table 4.

Table 4: Pseudo Values

	a_n^{*}	b_n^{*}	R_n^{2*}	$\hat{\sigma}_n^{*}$	r_n^{*}
None excluded	109.874	-1,127	0.41	11.02291	-0.64
1. excluded	111.6047	-1.10608	0.3934469	5.194477	-0.62739
2. excluded	129.0488	-3.21011	1.65466794	10.361934	-1.65225
3. excluded	77.40537	0.028115	-0.5882058	18.106984	0.116803
4. excluded	89.41629	-0.32654	0.01250356	9.041542	-0.33358
5. excluded	117.5705	-1,034	0.18786752	9.223888	-0.46801
6. excluded	109.9917	-1.15904	0.55179401	4.987743	-0.75152
7. excluded	110.3382	-0.91941	0.34820159	5.588891	-0.59214
8. excluded	114.4804	-1.27019	0.50350789	5.313118	-0.71354
9. excluded	118.0882	-1.47205	0.70162701	5.508795	-0.87011
10. excluded	107.5214	-0.48739	0.16539742	7.344636	-0.45071
11. excluded	141.7749	-2.54299	0.96746245	11.625412	-1.0833
12. excluded	101.1318	-0.78494	0.35057232	5.717717	-0.59399
13. excluded	77.40537	0.028115	-0.5882058	18.106984	0.116803

14. excluded	85.26744	-0.36208	-0.2959997	14.632642	-0.10038
15. excluded	118.1321	-1.3837	0.52110385	6.048218	-0.72737
16. excluded	112.7787	-1.23033	0.52763134	5.083879	-0.7325
17. excluded	123.5129	-1.47106	0.41335457	8.887798	-0.64293
18. excluded	194.7533	-8.08235	6.36613795	9.345961	-6.74794
19. excluded	121.2571	0.199448	-2.8232241	58.91716	1.675167
20. excluded	88.91906	-0.47559	-0.1260996	11.938523	-0.2283
21. excluded	112.7787	-1.23033	0.52763134	5.083879	-0.7325
Mean	112.5322	-1.34726	0.4652	11.24096	-0.76856
Std. Error	5.487377	0.380179	0.34823524	2.545924206	0.328803

Mean jackknife parameter values related to explanatory coefficient of the model, standard error values related to pseudo jackknife values, "t" values related to jackknife parameter distribution and finally confidence intervals of parameter values are summarized in Table 5.

When we look into Table 5, for the jackknife, confidence intervals are computed. $\alpha = 0.05$ Critical value for a Student's t distribution for 19 degrees of freedom is equal to $t_t = 2.093$,

The confidence interval for the intercept;

$$a^* \mp t_t SE = 112.5322 \mp 2.093 * 5.487377 = (101.0471; 124.01728)$$

The confidence interval for the slope;

$$b^* \mp t_t SE = -1.3472 \mp 2.093 * 0.380179 = (-2.14291; -0.55148)$$

Table 5: %95 Confidence Intervals (CI) of calculated parameter estimates with Jackknife Method

	a_n^*	b_n^*	R_n^{2*}	$\hat{\sigma}_n^*$	r_n^*
Original Coefficient	109.874	-1.127	0.41	11.02291	-0.64
Mean	112.5322	-1.3472	0.4652	11.24096	-0.768
SE	5.487377	0.380	0.348	2.546	0.328
LowerBound	101.0471	-2.142	-0.2636	5.912	-1.454
UpperBound	124.01728	-0.551	1.194	16.56957	-0.077

Obtained results through bootstrap method is as following;

Table 6: Residuals Obtained From Regression Model

Rank No	e_i	$e_i/21$
1	2.0310	0.0967
2	-9.5721	-0.4558
3	-15.6040	-0.7431
4	-8.7309	-0.4158
5	9.0310	0.4300
6	-0.3341	-0.0159
7	3.4120	0.1625
8	2.5230	0.1201
9	3.1421	0.1496
10	6.6659	0.3174
11	11.0151	0.5245
12	-3.7309	-0.1777
13	-15.6040	-0.7431
14	-13.4770	-0.6418
15	4.5230	0.2154
16	1.3960	0.0665
17	8.6500	0.4119
18	-5.5403	-0.2638
19	30.2850	1.4421
20	-11.4770	-0.5465
21	1.3960	0.0665

If we look into Table 6, firstly, e_i residuals are computed with classical LSM. Then, to each obtained e_i value, $1/_{21}$ probability is given. Bootstrap method is applied to obtained 21 $^{e_i}/_{21}$ value.

Table 7: Bootstrap residuals and calculated estimates

Rank No	**1.Bootst Ex.**	**2.Bootst Ex.**	**…**	**999. Bootst Ex.**	**1000.Bootst Ex.**	$\bar{\hat{\varepsilon}}_i^*$	Y_{boot}^*
1	0.3174	0.4119	…	0.0665	0.0665	0.0177	92.9867
2	0.4119	0.2154		-0.4558	-0.7431	0.0032	80.5752
3	0.2154	0.0665		1.4421	0.4300	-0.0195	98.5845
4	-0.7431	-0.4558		-0.4158	0.0665	-0.0101	99.7209
5	0.3174	0.3174		-0.7431	-0.6418	0.0070	92,976
6	-0.7431	-0.0159		-0.7431	-0.6418	0.0208	87.3548
7	0.3174	-0.7431		-0.0159	-0.7431	0.0036	89.5916

8	0.0665	-0.2638		0.5245	0.1496	-0.01938	97.45762
9	0.0665	-0.6418		0.2154	0.4300	-0.00003	100,858
10	-0.6418	0.0665		-0.7431	-0.0159	-0.0013	87.3327
11	-0.7431	0.5245		0.0665	-0.7431	-0.0077	101.9773
12	-0.5465	1.4421		-0.4558	0.4119	-0.0191	99.7119
13	-0.7431	-0.4158		-0.5465	-0.6418	0.0164	98.6204
14	-0.7431	-0.6418		-0.1777	0.4119	0.0014	97.4784
15	-0.4158	-0.0159		0.4119	-0.2638	0.0241	97.5011
16	0.0665	0.0665		-0.4558	0.0967	-0.0075	98.5965
17	-0.1777	0.1201		0.1201	-0.6418	-0.0158	96.3342
18	0.0665	0.3174		0.5245	0.1201	-0.0012	62.5388
19	-0.1777	-0.7431		-0.0159	-0.4558	-0.0156	90.6994
20	0.0967	0.4300		0.1201	1.4421	-0.0100	97,467
21	0.1625	-0.4158	…	0.0967	-0.4558	0.0003	98.6043

If we look into Table 7, there are formed 1000 number of 25 sized bootstrap example belonging t this value. $\bar{\hat{\varepsilon}}_i^{\,*}$ value is calculated considering obtained bootstrap sample and with the help of this value, $\mathrm{Y}^*_{\mathrm{boot}}$ is calculated. $\bar{\hat{\varepsilon}}_i^*$ value is calculated by averaging row values in Table 7. In another words, it is the mean of the first values in each bootstrap sample.

Y^*_{boot}: Regression model, which is obtained from LSM, is expressed as $\bar{\hat{\varepsilon}}_i^*$. In another words,

$$Y^*_{boot} = 109.874 - 1.127x + \bar{\hat{\varepsilon}}_i^*$$

Y^*_{boot} value, which is obtained here, is based on resampling of error term.

If we summarize obtained values ,

Table 8: %95 Confidence Intervals (CI) of calculated parameter estimates with Bootstrap Method

	a	b
BootsMean	109.868	-1.127
BootsError	0.006	0.0003
Bottom Line % 95 CI	109.855	109.881
Top Line % 95 CI	-1.127	-1.126
$R^2 = 0.99$; $\hat{\sigma} = 0.0134$; r= -0.99		

When we look into Table 8, coefficient of determination of regression model, which is formed with the help of bootstrap method, is $R^2 = 0.99$ and rate for independent variable to explain the variance of dependent variable is observed as very high. Standard error of the formed regression model is 0.0134. This shows the success of the results obtained with the bootstrap method.

4. Conclusion

In the study, usage of bootstrap and jackknife methods in simple linear regression is explained in detail. Coefficients of determination, standard error of the model, coefficient of correlation and confidence intervals are calculated.

Pseudo standard error of the mean belonging to the model after calculations made considering jackknife method is 11.241 and, pseudo standard error of the standard error is 2.546. Also, jackknife mean of obtained coefficient of determination is 0.46. Estimate of standard error estimate belonging to the model with bootstrap method is 0.0134. Coefficient of determination of bootstrap method is calculated as 0.99.

As a result, in this study, bootstrap method, which is used as a correction method when the assumptions of error term are not met, is seen to give better results than jackknife and least squares method with the given data structure. In the application part, since the usage of bootstrap and jackknife methods in simple linear regression is given in detail, it can be used as a reference for similar studies.

References

[1] Efron. B. Bootstrap Method; Another Look at Jackknife. Annals of Statistics, Vol. 7, pp. 1-26, 1979.

[2] Freedman. D.A. Bootstrapping Regression Models, Annals of Statistics. Vol.1, No. 6, pp. 1218-1228, 1981.

[3] Wu, C. F. J. Jackknife, Bootstrap and other Resampling Methods in Regression Analysis, Annals of Statistics, Vol. 14, No. 4, pp. 1261-129, 1986.

[4] Algamal, Z. Y. And Rasheed, K . B. Re-sampling in Linear Regression Model Using Jackknife and Bootstrap , Iraqi Journal of Statistical Science. Vol. 18, pp. 59-73, 2010.

[5] Quenouille, M. H. Approximate tests of correlation in time series. Journal of the Royal Statistical Society, 11, 18-44, 1949.

[6] Quenouille, M. H. Notes on bias in estimation. Biometrika, 61, 353-360, 1956.

[7] Tukey, J. W. Bias and confidence in not-quite large sample. Ann Math Stat, 29, 614, 1958.

[8] Lohr, Sharon. Sampling: design and analysis. Nelson Education, 2009.

[9] Fenwick, I. Techniques in Market Measurement: The Jackknife. Journal of Marketing Research, 163,

410-414, 1979.

[10] Abdi, H. and Williams, J. L. Jackknife in Neil Salkind (Ed.). Eneyclopedia of Research Design. http://dx.doi.org/10.4135/9781412961288.n202, 2010

[11] Mooney Christopher Z.. Bootstrap Statistical Inference;Examples anad Evaluation for Political Science, American Journal of Political Science, Vol:40, No:2, Mayıs, 1996.

[12] Schenker, N. "Qualms about bootstrap confidence intervals." Journal of the American Statistical Association 80.390: 360-361, 1985.

[13] Avşar, P. E. Bootstrap Yönteminin Regresyon Analizinde Kullanımına İlişkin Olarak Türkiye İnşaat Sektöründe Bir Uygulama. Marmara Üniversitesi Ekonometri Anabilim Dalı Yüksek Lisans Tezi İstanbul, 2006.

[14] Topuz, D. Regresyonda Yeniden Örnekleme Yöntemlerinin Karsılastırmalı Olarak İncelenmesi. Yüksek Lisans Tezi, Nigde, 2002.

[15] Rousseeuw, P. J., & Leroy, A. M., Robust regression and outlier detection (Vol. 589). John Wiley & Sons. , 2005.