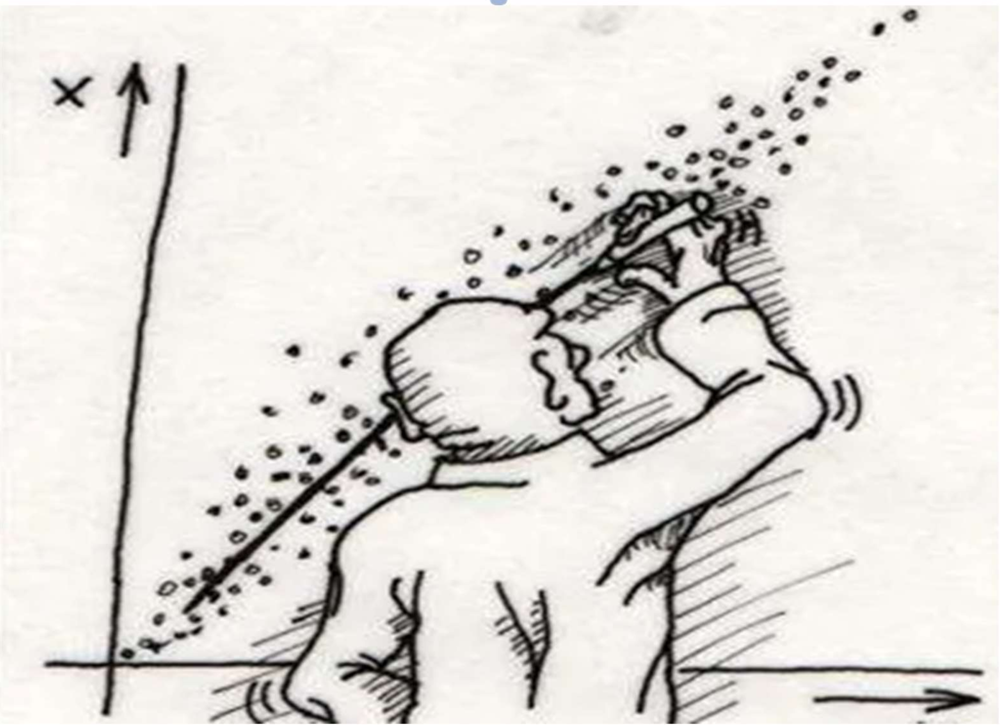
REPORT
LINEAR REGRESSION STUDY

A Simulated framework for the comparison of confidence intervals of $\beta_i{}'s$ and predictive accuracy using lm model in R vs Bootstrap method of computing coefficient estimates.



KANAV MALIK

Contents

1. Introduction

Regression is a very simple yet very powerful tool which can be used to solve very complex problems. We use linear regression for prediction and inference of events. However often we do not realize the impact of violation of linear regression assumptions on the predictive accuracy of linear regression model. The motivation behind the project is to study the impact of violation of assumptions on the prediction accuracy of the linear regression model and infer whether bootstrapping can help in such cases.

When the assumptions are violated and the linear regression model is fitted, the coefficient estimates might not be accurate. The idea is to simulate and check whether bootstrapping might help us in such cases to more accurately predict the coefficient estimates so that the predictive power of linear regression can be improved when the assumptions are not completely violated.

It is expected that when the data conforms to the assumptions of linear regression, the coefficient estimates of linear model should fit as good as the bootstrapped linear model coefficient estimates. When the assumptions of linear regression are slightly or fairly violated, the bootstrap coefficient estimates might give better results. In cases where the assumptions are completely violated, both estimates are expected to perform poor.

The purpose of this study is to simulate different datasets which either conforms to all assumptions of a linear regression model, to some of the assumptions or violate all the assumptions; and compare the confidence interval of $\beta_i's$ using lm function in R with the bootstrap intervals. The comparison will be made for different scenarios where the data for each scenario will either agree with the assumptions for linear regression or violate some/all assumptions. Based on the train set, the bootstrap confidence intervals for $\beta_i's$ will be computed and compared with the former coefficient estimates. Test MSE's will be computed and compared for each model.

2. Background: Linear Regression

The linear regression model is represented as :

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_i\,(j) + \varepsilon_i$$

where
β_0 *is the intercept*, β_j *are slopes*, $x_i(j)$ *are independant variables*, y_i *is depandant variable*.

The method of Ordinary Least Squares is used to find the intercept and slope coefficients. The linear regression has the following assumptions :

- **Linear relationship :** There exists a linear relationship between the independent variables, x_i, and the dependent variable, y_i.
- **Independence :** The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
- **No Multicollinearity :** Independent variables do not have a strong correlation among themselves.
- **Homoscedasticity :** The residuals have constant variance at every level of x.
- **Normality :** The residuals of the model are normally distributed

The last 2 assumptions means that $\varepsilon_i \sim N(0, \sigma^2)$ and so subsequently $E[\varepsilon_i] = 0$.

The performance of the linear regression model can be tested using two measures :

- **R^2 (Coefficient of determination)**
  It is a measure of how well the regression predictions approximate the real data points(actual response).

  $$\mathrm{R}^2 = \frac{\sum(\check{y}_i - \bar{y})}{\sum(y_i - \bar{y})}$$

- **Mean Squared Error**.
  It is a measure of square of errors which is helpful for determining the predictive power of our coefficient estimates.

  $$\mathrm{MSE} = \frac{1}{n}\sum(\check{y}_i - y_i)$$

3. Dataset

The following datasets of 1 million records each were simulated for the purpose of study :

Data Set	Assumption Violation	$x_i(1)$	$x_i(2)$	$x_i(3)$	$x_i(4)$	$x_i(5)$	$x_i(6)$	ε_i	y_i
1	None	$N(1, 0.5^2)$	bin(100,0.6)	exp (0.3)	weib(2,1)	$N(5, 2^2)$	pois(0.3)	0	$\beta_0 + \sum_{j=1}^{n} \beta_j x_i(j)$
2	None	$N(1, 0.5^2)$	bin(100,0.6)	exp (0.3)	weib(2,1)	$N(5, 2^2)$	pois(0.3)	$N(0, 10^2)$	$\beta_0 + \sum_{j=1}^{n} \beta_j x_i(j) + \varepsilon_i$
3	Linear relationship	$N(1, 0.5^2)$	bin(100,0.6)	exp (0.3)	weib(2,1)	$N(5, 2^2)$	pois(0.3)	$N(0, 10^2)$	$\beta_0\, x_i(5) \times \prod_{j=1}^{4,6} x_i(j)^{\beta_j} + \varepsilon_i$
4.	Independence	$N(1, 0.5^2)$	bin(100,0.6)	exp (0.3)	weib(2,1)	$N(5, 2^2)$	pois(0.3)	Gamma Markov chain	$\beta_0 + \sum_{j=1}^{n} \beta_j x_i(j) + \varepsilon_i$
5	Homoscedasticity	$N(1, 0.5^2)$	bin(100,0.6)	exp (0.3)	weib(2,1)	$N(5, 2^2)$	pois(0.3)	$N(0, \sigma_{i*}{}^2)$	$\beta_0 + \sum_{j=1}^{n} \beta_j x_i(j) + \varepsilon_i$
6	No multicollinearity	$N(1, 0.5^2)$	bin(100,0.6)	exp (0.3)	$N(50, 1^2)$ $r_{4,5} = 0.98$ $r_{4,6} = 0.98$ $r_{4,\varepsilon} = 0.98$	$N(60, 1^2)$ $r_{5,4} = 0.98$ $r_{5,6} = 0.98$ $r_{5,\varepsilon} = 0.98$	$N(70, 1^2)$ $r_{6,4} = 0.98$ $r_{6,5} = 0.98$ $r_{6,\varepsilon} = 0.98$	$N(30, 1^2)$ $r_{\varepsilon,4} = 0.98$ $r_{\varepsilon,5} = 0.98$ $r_{\varepsilon,6} = 0.98$	$\beta_0 + \sum_{j=1}^{n} \beta_j x_i(j) + \varepsilon_i$
7	Normality	$N(1, 0.5^2)$	bin(100,0.6)	exp (0.3)	weib(2,1)	$N(5, 2^2)$	pois(0.3)	$\text{unif}(-10,10)$	$\beta_0 + \sum_{j=1}^{n} \beta_j x_i(j) + \varepsilon_i$
8	Linear relationship, Normality, No collinearity, Homoscedasticity	$N(1, 0.5^2)$	bin(100,0.6)	pois(0.3)	$N(50, 1^2)$ $r_{4,5} = 0.8$ $r_{4,6} = 0.8$	$N(60, 1^2)$ $r_{5,4} = 0.8$ $r_{5,6} = 0.8$	$N(70, 1^2)$ $r_{6,4} = 0.8$ $r_{6,5} = 0.8$	$N(\mu_{i*}, \sigma_{i*}{}^2)$	$\beta_0\, x_i(5) \times \prod_{j=1}^{4,6} x_i(j)^{\beta_j} + \varepsilon_i$

* For computations, refer Appendix 1

where $\beta_0 = 4, \beta_1 = 2, \beta_2 = 0.5, \beta_3 = 3, \beta_4 = 7, \beta_5 = 0.1, \beta_6 = 1$. These are the actual(population) coefficient given by us. These will be compared against the LM coefficients and Bootstrapped LM coefficients.

$\mu_{i*}, \sigma_{i*} = \{5,10,15,20\}$ for values in $\{1\text{st}, 2\text{nd}, 3\text{rd}, 4\text{th}\}$ quartile of y

Once each data is generated, the following plots are generated to check some assumptions as follows:

- **Scatterplot of errors** – Check whether the errors are randomly distributed.
- **Autocorrelation plot of errors** – Check whether the errors are independently distributed .i.e. each error is not dependent on any previous error.
- **Histogram of errors** – Check normality of errors.
- **Correlation plot** – Check whether any correlation exists between independent variables, or between independent variables and error.

Careful consideration of the plots shows that each data simulated violates the assumptions we want to violate.(Refer Appendix 1 for the plots for each dataset.)

Further, for Validation of our conclusions regarding the coefficients estimates and their CI, each dataset is split into a train/test sample in 80/20 ratio. The reason is that the Train/Test R^2 and MSE can be compared to see they are consistent or inconsistent(underfit or overfit model). In the latter case, we cannot rely on how close our coefficient estimates are to population coefficients.

4. Methodology

Linear Model coefficients
A linear model is fit on each Training dataset to get the regressed coefficients and their 95% confidence interval.

For each model, the fitted-residual plot is generated to check the assumption of homoscedasticity of residuals and the Normal Q-Q plot is generated to check whether the assumption of normality of residuals.

Further the Train/Test R^2 and Mean Squared Error are computed.

Bootstrapped Linear Model coefficients
A function is created which takes the Train data, number of bootstrap simulations to perform, Test data as the input to give a list of output as follows : Bootstrapped coefficient estimates, their 95% CI, the Train/Test R^2 and MSE, table of coefficient estimates and 95% CI for each simulation.

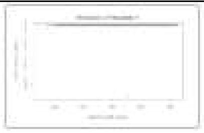
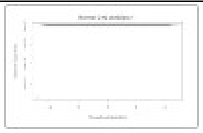
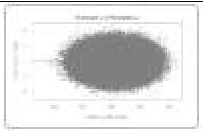
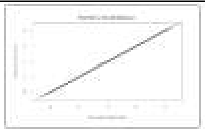
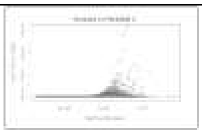
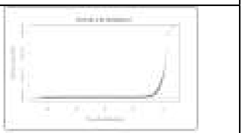
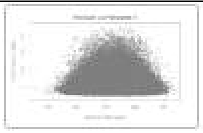
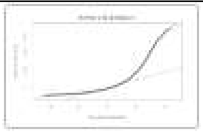
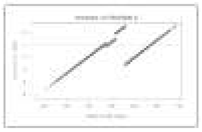
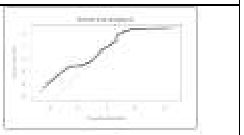
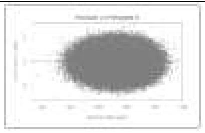
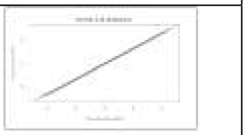
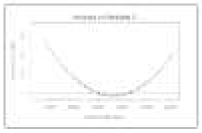
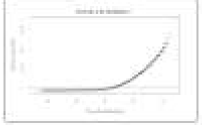
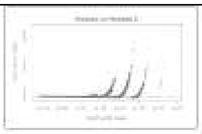
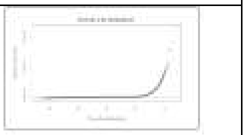
For each simulation, an 80% random sample is created from the training data and the coefficient estimates and their 95% CI is stored in a data frame. This is repeated n times to get n coefficient estimates. Finally the average of all these estimates give us the bootstrapped coefficient estimates and their 95% CI. Based on these bootstrapped coefficient estimates, the bootstrapped linear model is fit to the train data and predictions are made on the train, test data to get the Train/Test R^2 and MSE.

** For computations, refer Appendix 2.*

5. Results

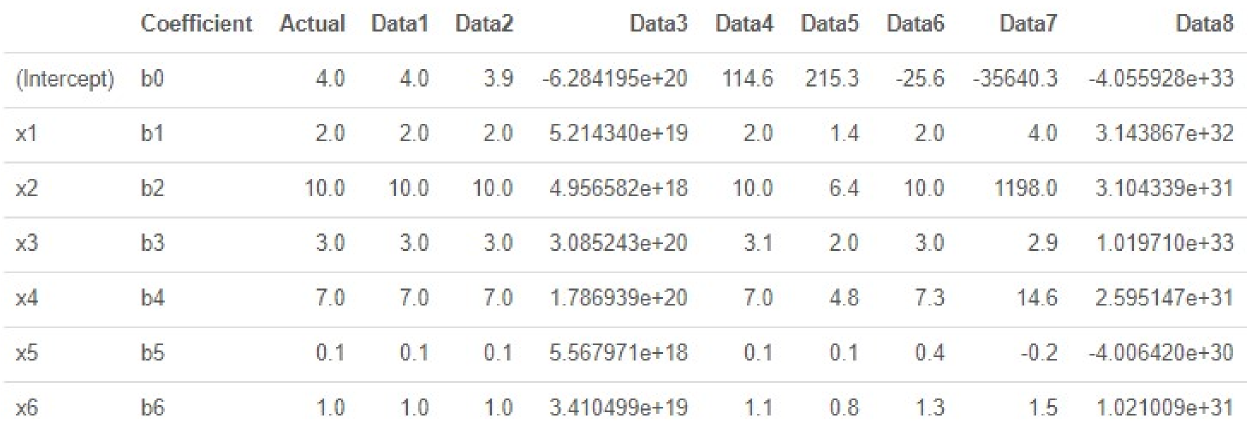
**Refer Appendix 3 for the results computation and Appendix 4 for complete set of displayed results.*

Here is a high-level screenshot of the Residual-Fitted Plot and Normal Q-Q plot for each train dataset.

Data Set	Assumption Violation	Residual-Fitted Plot	Normal Q-Q Plot
1	None		
2	None		
3	Linear relationship		
4.	Independence		
5	Homoscedasticity		
6	No multicollinearity		
7	Normality		
8	Linear relationship, Normality, No collinearity, Homoscedasticity		

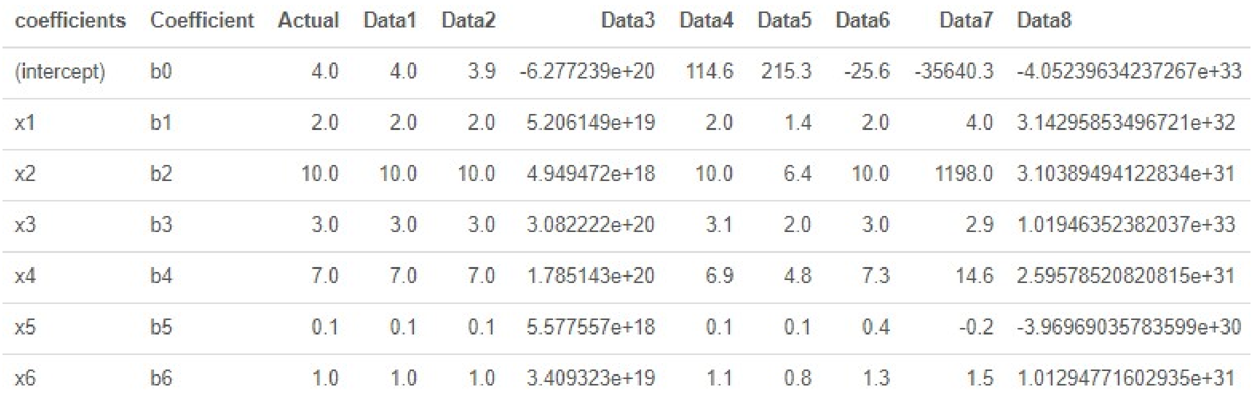
Coefficient Estimates of Linear Model and Bootstrapped Linear Models

Table 1: Linear Model coefficients



	Coefficient	Actual	Data1	Data2	Data3	Data4	Data5	Data6	Data7	Data8
(Intercept)	b0	4.0	4.0	3.9	-6.284195e+20	114.6	215.3	-25.6	-35640.3	-4.055928e+33
x1	b1	2.0	2.0	2.0	5.214340e+19	2.0	1.4	2.0	4.0	3.143867e+32
x2	b2	10.0	10.0	10.0	4.956582e+18	10.0	6.4	10.0	1198.0	3.104339e+31
x3	b3	3.0	3.0	3.0	3.085243e+20	3.1	2.0	3.0	2.9	1.019710e+33
x4	b4	7.0	7.0	7.0	1.786939e+20	7.0	4.8	7.3	14.6	2.595147e+31
x5	b5	0.1	0.1	0.1	5.567971e+18	0.1	0.1	0.4	-0.2	-4.006420e+30
x6	b6	1.0	1.0	1.0	3.410499e+19	1.1	0.8	1.3	1.5	1.021009e+31

Table 4: Boostrap coefficient Estimates



coefficients	Coefficient	Actual	Data1	Data2	Data3	Data4	Data5	Data6	Data7	Data8
(intercept)	b0	4.0	4.0	3.9	-6.277239e+20	114.6	215.3	-25.6	-35640.3	-4.05239634237267e+33
x1	b1	2.0	2.0	2.0	5.206149e+19	2.0	1.4	2.0	4.0	3.14295853496721e+32
x2	b2	10.0	10.0	10.0	4.949472e+18	10.0	6.4	10.0	1198.0	3.10389494122834e+31
x3	b3	3.0	3.0	3.0	3.082222e+20	3.1	2.0	3.0	2.9	1.01946352382037e+33
x4	b4	7.0	7.0	7.0	1.785143e+20	6.9	4.8	7.3	14.6	2.59578520820815e+31
x5	b5	0.1	0.1	0.1	5.577557e+18	0.1	0.1	0.4	-0.2	-3.96969035783599e+30
x6	b6	1.0	1.0	1.0	3.409323e+19	1.1	0.8	1.3	1.5	1.01294771602935e+31

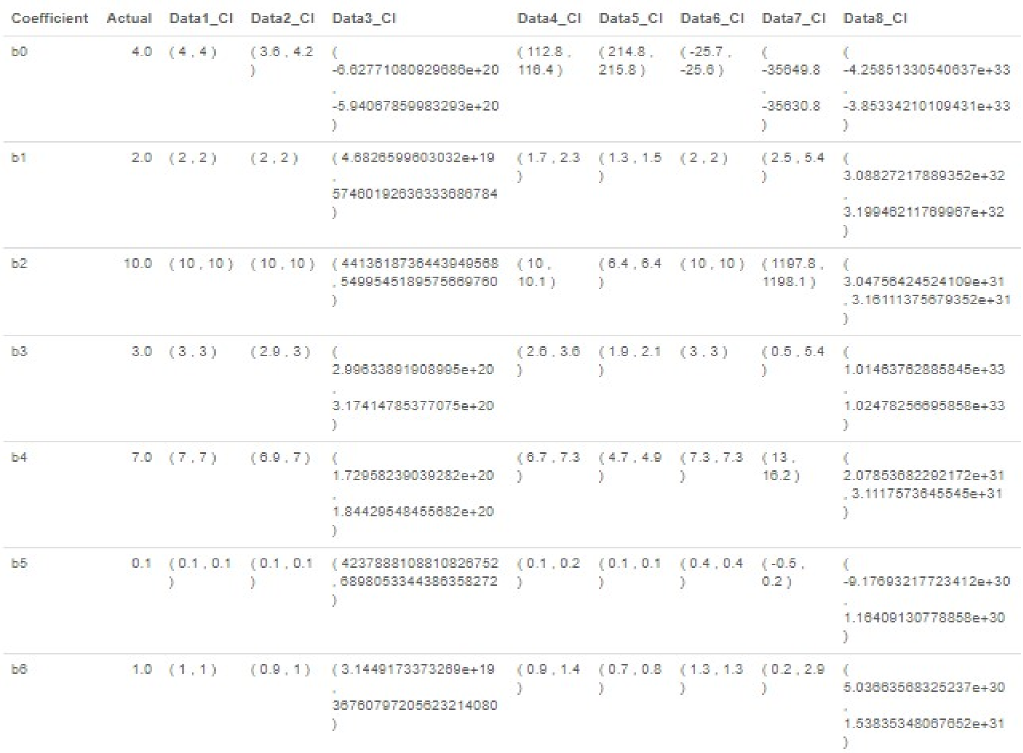
It can be noticed that the coefficient estimates for Data 1, Data 2 are estimated perfectly by Linear model and Bootstrapped Linear model.

The coefficient estimates for Data 5,6 and Data 7 are estimated the same by both methods. It can be noticed that both methods are not able to estimate the intercept correctly in Data 5(heteroscedastic residuals) and Data 6(multicollinearity among independent variables and error)

Other datasets 3,4,8 have slightly different coefficient estimates using each method but none of them are close to the actual coefficients.

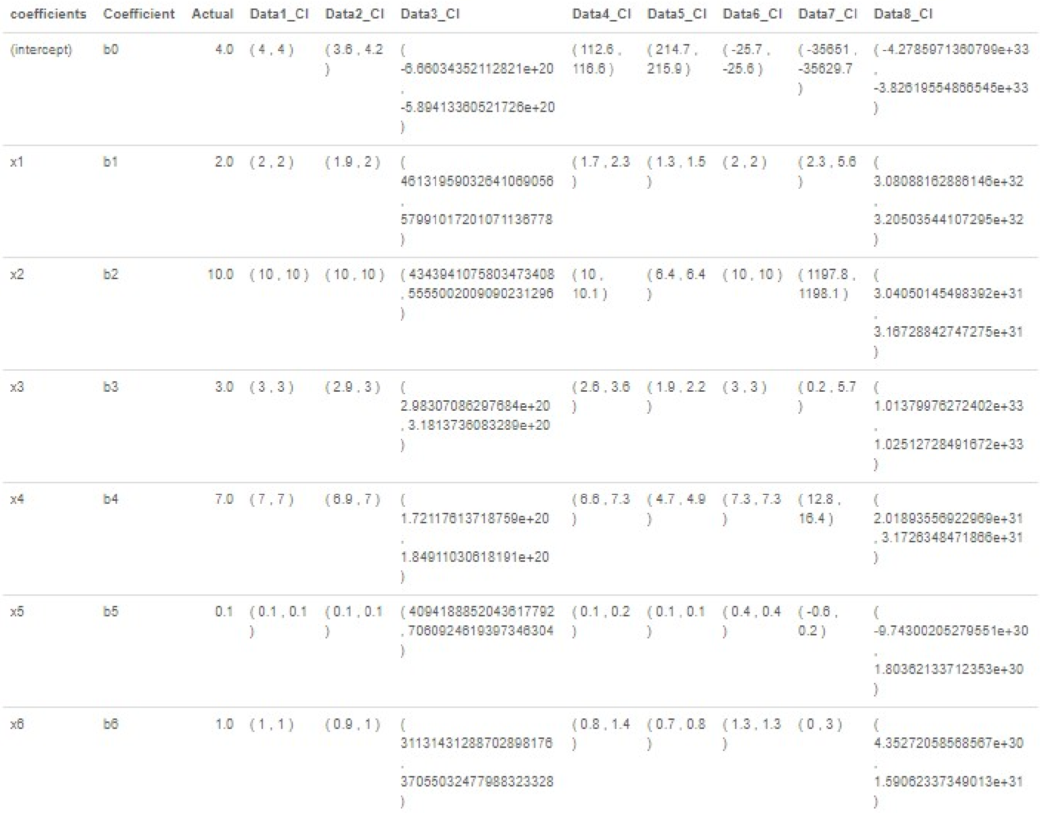
95% CI for coefficient estimates of Linear Model and Bootstrapped Linear Models

Table 2: Linear Model coefficients 95% CI



Coefficient	Actual	Data1_CI	Data2_CI	Data3_CI	Data4_CI	Data5_CI	Data6_CI	Data7_CI	Data8_CI
b0	4.0	(4 , 4)	(3.6 , 4.2)	(-6.62771080929686e+20 , -5.94067859983293e+20)	(112.8 , 116.4)	(214.8 , 215.8)	(-25.7 , -25.6)	(-35649.8 , -35630.8)	(-4.25851330540637e+33 , -3.85334210109431e+33)
b1	2.0	(2 , 2)	(2 , 2)	(4.6826599603032e+19 , 57460192636333686784)	(1.7 , 2.3)	(1.3 , 1.5)	(2 , 2)	(2.5 , 5.4)	(3.08827217889352e+32 , 3.19946211789967e+32)
b2	10.0	(10 , 10)	(10 , 10)	(441361873644394968 , 5499545189575669760)	(10 , 10.1)	(6.4 , 6.4)	(10 , 10)	(1197.8 , 1198.1)	(3.04756424524109e+31 , 3.16111375679352e+31)
b3	3.0	(3 , 3)	(2.9 , 3)	(2.99633891908995e+20 , 3.17414785377075e+20)	(2.6 , 3.6)	(1.9 , 2.1)	(3 , 3)	(0.5 , 5.4)	(1.01463762885845e+33 , 1.02478256695858e+33)
b4	7.0	(7 , 7)	(6.9 , 7)	(1.72958239039282e+20 , 1.84429548455682e+20)	(6.7 , 7.3)	(4.7 , 4.9)	(7.3 , 7.3)	(13 , 16.2)	(2.07853682292172e+31 , 3.1117573645545e+31)
b5	0.1	(0.1 , 0.1)	(0.1 , 0.1)	(4237888108810826752 , 6898053344386358272)	(0.1 , 0.2)	(0.1 , 0.1)	(0.4 , 0.4)	(-0.5 , 0.2)	(-9.17693217723412e+30 , 1.16409130778858e+30)
b6	1.0	(1 , 1)	(0.9 , 1)	(3.1449173373269e+19 , 36760797205623214080)	(0.9 , 1.4)	(0.7 , 0.8)	(1.3 , 1.3)	(0.2 , 2.9)	(5.03663568325237e+30 , 1.53835348067652e+31)

Table 5: Boostrap coeffficient Estimates 95% CI



coefficients	Coefficient	Actual	Data1_CI	Data2_CI	Data3_CI	Data4_CI	Data5_CI	Data6_CI	Data7_CI	Data8_CI
(intercept)	b0	4.0	(4 , 4)	(3.6 , 4.2)	(-6.66034352112821e+20 , -5.89413360521726e+20)	(112.6 , 116.6)	(214.7 , 215.9)	(-25.7 , -25.6)	(-35651 , -35629.7)	(-4.2785971360799e+33 , -3.82619554866545e+33)
x1	b1	2.0	(2 , 2)	(1.9 , 2)	(46131959032641069056 , 57991017201071136778)	(1.7 , 2.3)	(1.3 , 1.5)	(2 , 2)	(2.3 , 5.6)	(3.08088162888146e+32 , 3.20503544107295e+32)
x2	b2	10.0	(10 , 10)	(10 , 10)	(4343941075803473408 , 5555002009090231296)	(10 , 10.1)	(6.4 , 6.4)	(10 , 10)	(1197.8 , 1198.1)	(3.04050145498392e+31 , 3.16728842747275e+31)
x3	b3	3.0	(3 , 3)	(2.9 , 3)	(2.98307086297684e+20 , 3.1813736083289e+20)	(2.6 , 3.6)	(1.9 , 2.2)	(3 , 3)	(0.2 , 5.7)	(1.01379976272402e+33 , 1.02512728491672e+33)
x4	b4	7.0	(7 , 7)	(6.9 , 7)	(1.72117613718759e+20 , 1.84911030618191e+20)	(6.6 , 7.3)	(4.7 , 4.9)	(7.3 , 7.3)	(12.8 , 16.4)	(2.01893556922969e+31 , 3.1726348471866e+31)
x5	b5	0.1	(0.1 , 0.1)	(0.1 , 0.1)	(4094188852043617792 , 7060924619397346304)	(0.1 , 0.2)	(0.1 , 0.1)	(0.4 , 0.4)	(-0.6 , 0.2)	(-9.74300205279551e+30 , 1.80362133712353e+30)
x6	b6	1.0	(1 , 1)	(0.9 , 1)	(31131431288702898176 , 37055032477988323328)	(0.8 , 1.4)	(0.7 , 0.8)	(1.3 , 1.3)	(0 , 3)	(4.35272058568567e+30 , 1.59062337349013e+31)

It is observed the CI and their width is the same for Linear model and Bootstrapped Linear model when the assumptions of data are not violated.

However in cases where the assumptions of the data are violated, bootstrapping leads to wider confidence intervals.

Test and Train $\mathbf{R}^2$ and MSE of Linear Model and Bootstrapped Linear Models

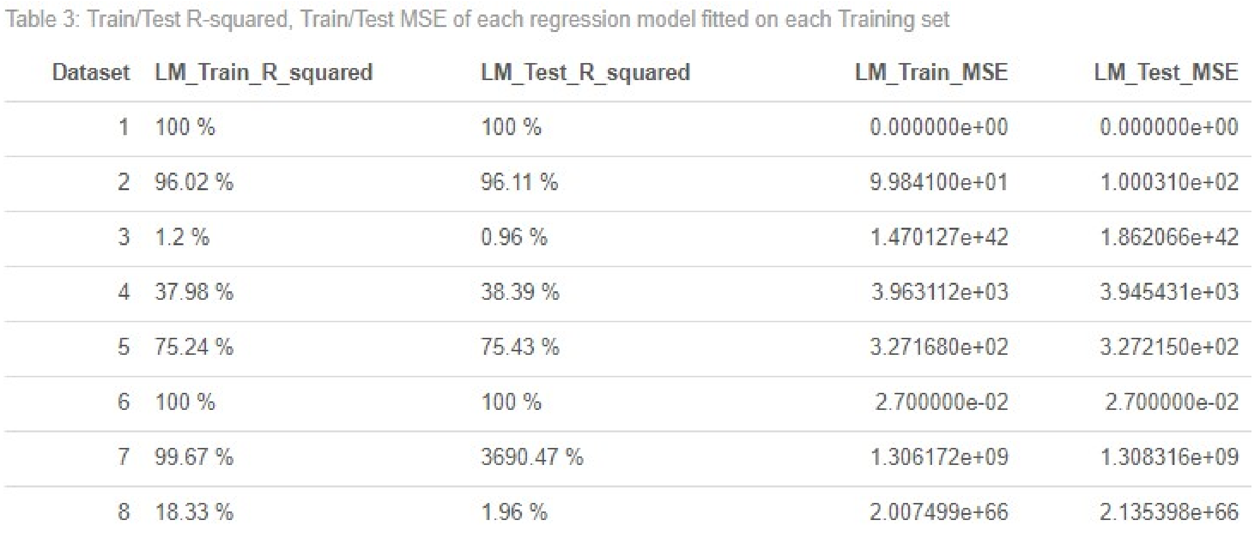


Table 3: Train/Test R-squared, Train/Test MSE of each regression model fitted on each Training set

Dataset	LM_Train_R_squared	LM_Test_R_squared	LM_Train_MSE	LM_Test_MSE
1	100 %	100 %	0.000000e+00	0.000000e+00
2	96.02 %	96.11 %	9.984100e+01	1.000310e+02
3	1.2 %	0.96 %	1.470127e+42	1.862066e+42
4	37.98 %	38.39 %	3.963112e+03	3.945431e+03
5	75.24 %	75.43 %	3.271680e+02	3.272150e+02
6	100 %	100 %	2.700000e-02	2.700000e-02
7	99.67 %	3690.47 %	1.306172e+09	1.308316e+09
8	18.33 %	1.96 %	2.007499e+66	2.135398e+66

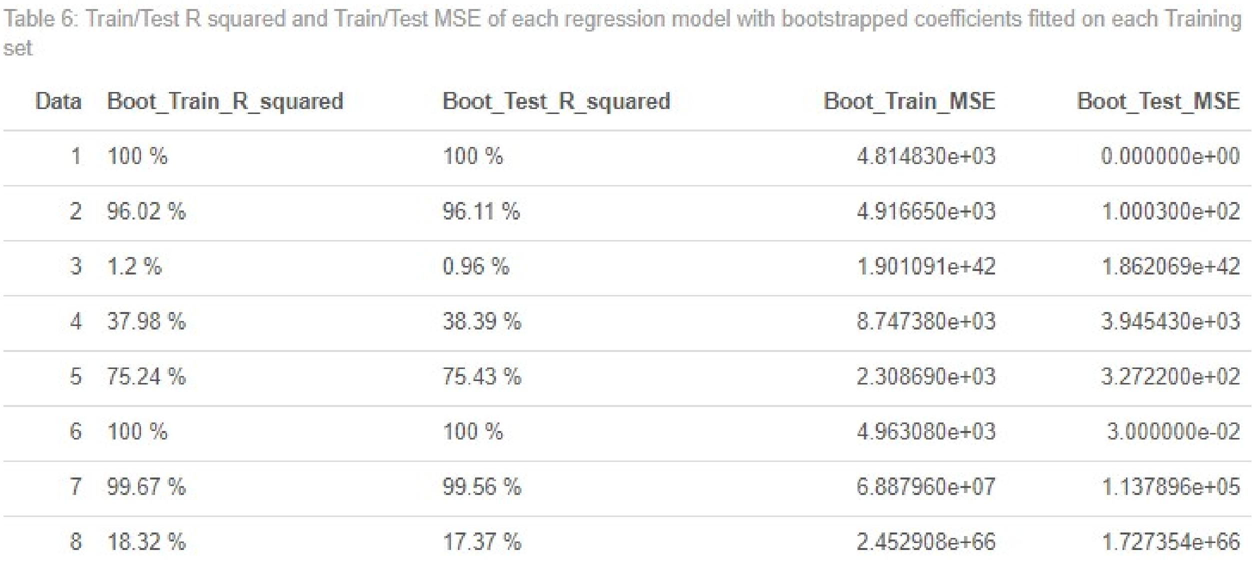


Table 6: Train/Test R squared and Train/Test MSE of each regression model with bootstrapped coefficients fitted on each Training set

Data	Boot_Train_R_squared	Boot_Test_R_squared	Boot_Train_MSE	Boot_Test_MSE
1	100 %	100 %	4.814830e+03	0.000000e+00
2	96.02 %	96.11 %	4.916650e+03	1.000300e+02
3	1.2 %	0.96 %	1.901091e+42	1.862069e+42
4	37.98 %	38.39 %	8.747380e+03	3.945430e+03
5	75.24 %	75.43 %	2.308690e+03	3.272200e+02
6	100 %	100 %	4.963080e+03	3.000000e-02
7	99.67 %	99.56 %	6.887960e+07	1.137896e+05
8	18.32 %	17.37 %	2.452908e+66	1.727354e+66

In can be seen that both methods produce perfect and consistent R^2 and MSE when the assumptions are not violated. Although the variables x4,x5,x6,error are highly correlated(0.98) in Data 6, the violation does not affect the predictive accuracy of the linear model in both methods. However we can notice that the Train MSE and Test MSE for bootstrapped linear model are not consistent for the Data 6 which signifies that these estimates are not good estimates. Based on the results, it might be argued that bootstrapped coefficient estimates produce more consistent results when the assumptions of normality of residuals(Data 7) and multiple assumptions(Data 8) are violated. However our results does not indicate that it improves the predictive accuracy in such cases.

6. Conclusion

As per the expectations, when the data conforms to the assumptions of linear regression, the coefficient estimates of linear model are as good as the coefficient estimates of the bootstrapped linear model.

When the assumption of linear relationship is violated, both methods perform extremely poor.
When the assumption of independence is violated, both methods have around 38% R^2 and when the assumption of homoscedasticity is violated, both methods have around 75% R^2. The Train and Test metrics of R^2 and MSE are consistent for both methods when these three assumptions of linear relationship, independence, homoscedasticity are violated.

As seen in results above when multicollinearity exists in the data, the R^2 might give deceiving results and so the Train and Test MSE can be compared to see the consistency of the results. We can notice that the Test and Train MSE of our bootstrapped model is not consistent for Data 6.

The bootstrapped coefficient estimates give more consistent results when the assumptions of normality and multiple assumptions are violated.

It can be concluded that in cases where assumption of normality or where many assumptions are violated, bootstrapped results does not improve the predictive power of the model but can lead to slightly improved coefficients which give a consistent result across different datasets.

7. Appendix

Check the html report for the Appendix