An introduction to Measurement Error Theory and Bootstrapping in Regression

Rupa Kumari
Roll no: MA20MSCST11014

A Project report submitted to

Indian Institute of Technology Hyderabad

In Partial Fulfillment of the Requirement for

The Degree of Master of Science in Mathematics and Computing

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Mathematics

Indian Institute of Technology Hyderabad

Kandi, Sangareddy, Telangana, India – 502285

Advisor

Dr. Arunabha Majumdar

June 2022

Chapter 1

An introduction to Measurement Error Theory

Motivation

Our motivation behind this project is to study the consequences of measurement error under linear regression model and how can we correct them back. In statistical analysis , measuring the observations with errors may lead to incorrect statistical inferences and results. We will see the effects of measurement errors under the simple linear regression setup.

Linear Regression

Regression model allows us to describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line to estimate how a dependent variable changes as the independent variables change.

Simple Linear Regression. Simple linear regression is used to estimate the relationship between two quantitative variables. Simple linear regression model with one independent variable is given as :

$$Y = \beta_0 + \beta_1 x + \epsilon$$

Where

- Y is the dependent variable.
- β_0 is the intercept, the predicted value of Y when the x is 0.
- β_1 is the regression coefficient – how much we expect Y to change as x increases.
- x is the independent variable (the variable we expect is influencing y).
- ϵ is the error of the estimate, or how much variation there is in our estimate of the regression coefficient.

The least squares estimates are :

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

Measurement Error Model

In real world, sometimes one is unable to observe the true explanatory variable x directly , instead of observing x, one observes the sum X.
Where

$$X = x + u \tag{1}$$

- u is called as **measurement error or errors-in-variables**.
- X is observed predictor value.
- x is true latent variable but it is unobserved.
- Example: Relationship between yield of corn and available nitrogen in the field.

Effects of Measurement Error on estimation of parameters

Let's investigate effects of measurement error on the Least Squares estimation.
We assume ,

$$x \sim N(\mu_{xx}, \sigma_{xx})$$
$$\epsilon \sim N(0, \sigma_{\epsilon\epsilon})$$
$$u \sim N(0, \sigma_{uu})$$

Where

- μ_{xx} is the mean of x.
- σ_{xx} is the variance of x. And
  $\sigma_x = \sqrt{\sigma_{xx}} =$ standard deviation of x.
- x is a random variable with $\sigma_{xx} > 0$

Then we formulate the **True model** as :

$$Y = \beta_0 + \beta_1 x + \epsilon$$

And the **Model with observed X variable** as :

$$Y = \gamma_0 + \gamma_1 X + e$$

Where X is the observed predictor value .
Now the regression coefficient computed from the observed variables is :

$$\hat{\gamma_1} = \frac{\sum_{i=0}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=0}^{n}(X_i - \bar{X})^2}$$

On taking expectation of $\hat{\gamma_1}$, we get :

$$E[\hat{\gamma_1}] = \frac{\sigma_{XY}}{\sigma_{XX}} = \frac{\beta_1 \sigma_{xx}}{\sigma_{xx} + \sigma_{uu}}$$

$$E[\hat{\gamma_1}] = \frac{\beta_1.\sigma_{xx}}{\sigma_{XX}}$$

$$E[\hat{\gamma_1}] = \beta_1.K_{xx}$$

Where , $K_{xx} = \frac{\sigma_{xx}}{\sigma_{XX}}$

- K_{xx} is known as the **Reliability Ratio.**

- $0 < K_{xx} < 1$.

Since , $E[\hat{\gamma_1}] \neq \beta_1$. So $\hat{\gamma_1}$ is not an unbiased estimator of β_1 and because of K_{xx}, regression coefficient has been attenuated by the measurement error.

Simulation Study

Data Generative Process: We have 100 observations from the normal distributions.

- x $\sim N(100, 0, 1)$, $\epsilon \sim N(100, 0, 0.1)$, $u \sim N(100, 0, 0.1)$.

- We set true parameters $\beta_0 = 0$ and β_1 in the range of [0 1] such as β_1=0 , β_1=0.2 , β_1=0.4 etc .

- We find Y from the equation of linear regression $Y = \beta_0 + \beta_1 x + \epsilon$. Then we have (Y,x) and (Y,X) pairs .

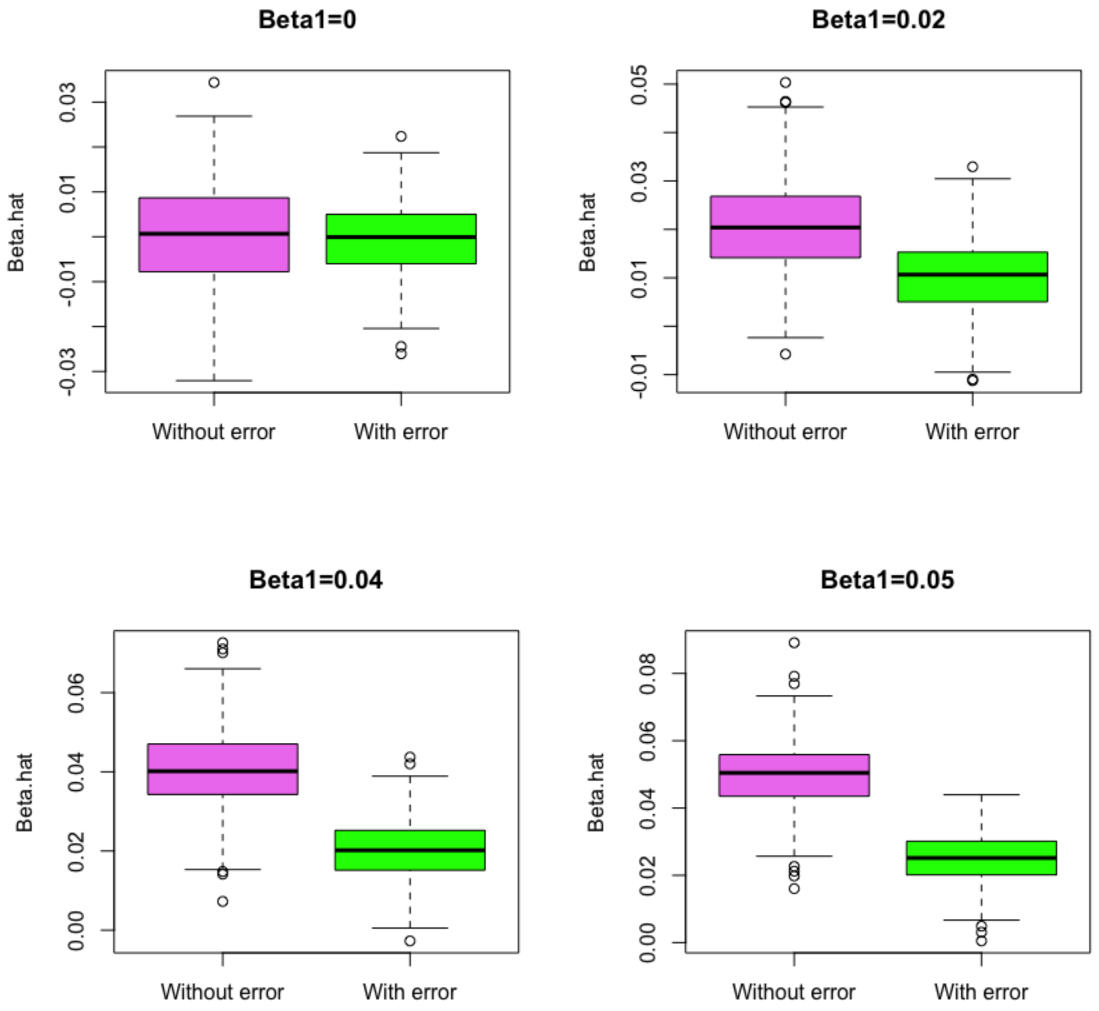
- We forget the true parameters and we apply the linear regression on both true model and model with observed varaible X .

- Find the least square estimators on both cases .

- Then calculate the power of the hypothesis testing .

- We repeat this process for 1000 times and we see the summarised pictures and tables .

Effects of measurement error on Estimation



Figures : Box plots for estimated parameter $\hat{\beta}_1$ with & without measurement error.

From above graphs it is clear that esitimated parameter $\hat{\beta}_1$ has been reduced in the presence of measurement error u in the predictor variable x .

Power Curve

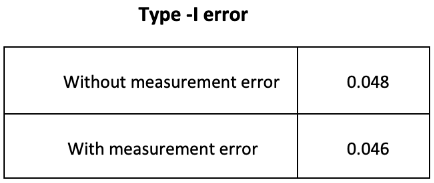
Our hypothesis testing is :

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

We define type I error , type II error and power as :

- **Type-I error** = Probability of rejecting null hypothesis when it is true .
- **Type-II error** = Probability of do not rejecting H_0 when it is false .
- **Power** = The probability of correctly rejecting H_0 when it is false is known as the power of the test. The larger it is, the better .
- We set the significance level $(\alpha) = 0.05$.
- Then we plot the power curve as Power Vs β_1 .

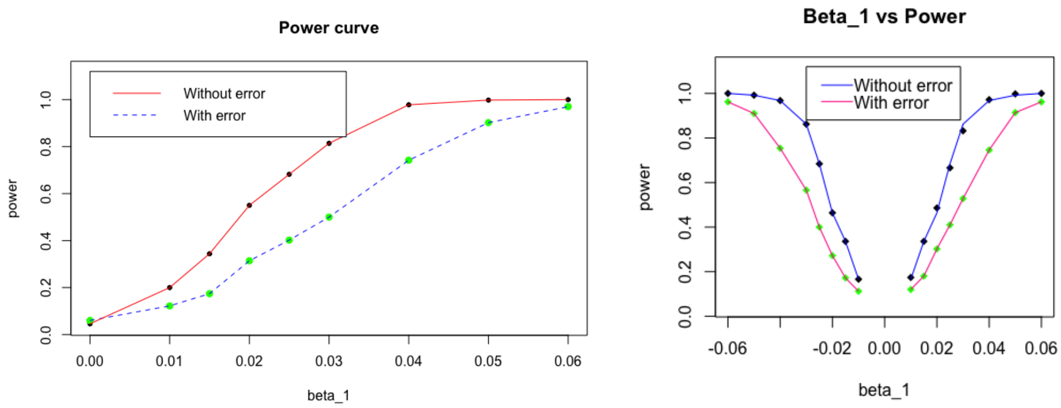
Simulation Results



Type -I error

Without measurement error	0.048
With measurement error	0.046

Effects of measurement error on the Power of the hypothesis test

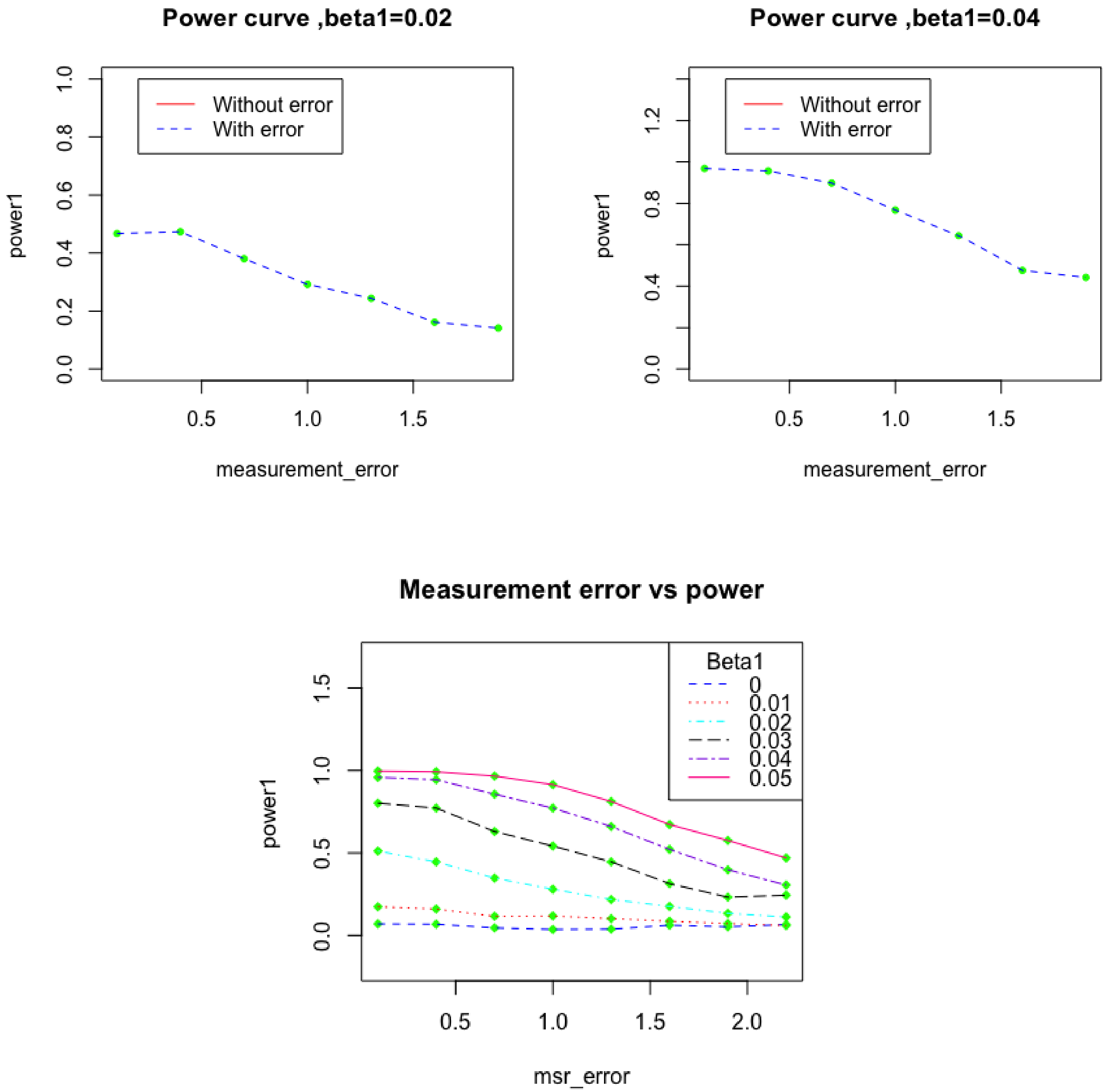


Figures : Power curve with & without measurement error.

From the graph we can see that the power of the hypothesis testing in the case of measurement error is less than the power of the test in the case of without measurement error.

Power Vs Measurement error u

Now we'll see the effect of magnitude of measurement error introduced in x ,on the power of the test.



Figures : Graphs of power Vs variance of the measurement error u.

We can see from the above graphs that as we increase the variance of the measurement error, the power of the test decreases. And we see this result for different values of true parameter β_1.

Correction of the Measurement Error Model

We consider the case when the reliability ratio K_{xx} is known . When K_{xx} is known , then it is possible to construct an unbiased estimator of β_1 .
Since

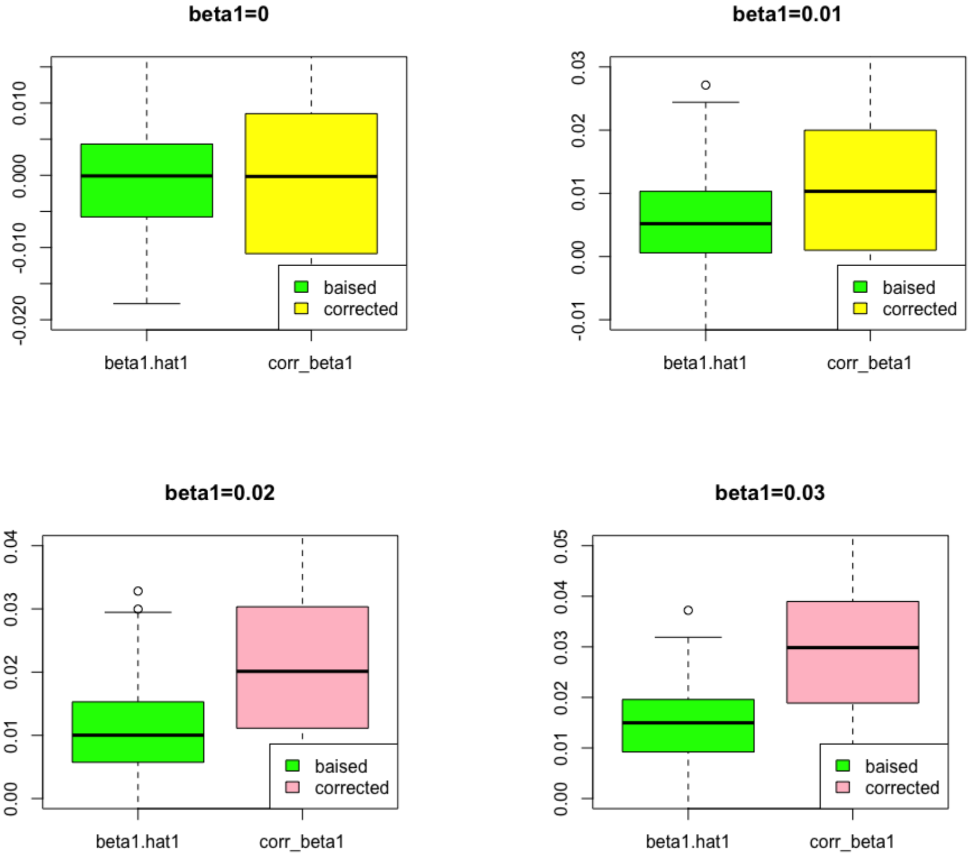
$$E[\hat{\gamma_1}] = \beta_1 .k_{xx}$$

where , $k_{xx} = \sigma_{xx}/\sigma_{XX}$.

Then we construct an unbaised estimator for β_1 as :

$$\boxed{\hat{\beta}_{1corr} = \frac{\hat{\gamma_1}}{K_{xx}}}$$

Simulation results for corrected parameters



Figures : Box plots of biased and corrected parameter $\hat{\beta_1}$.
From above graphs we can see that the corrected parameters are approximately close to the true parameters.

Conclusion

- The Least square estimation gets affected by the presence of measurement error in the independent variable x in linear regression .

- There is reduction in the power of the hypothesis testing.

- Type-I error gets controlled under linear regression model.

- We can correct back the biased parameter , when K_{xx} is known.

Chapter 2

Bootstrapping in Regression

Motivation

Our motivation behind this project is to study the bootstrap approach to approximate the distribution of an estimator, when driving the distribution of an estimator becomes mathematically very challenging. We will also see the bootstrap approach in the ridge regression and the lasso regression setup.

Bootstrapping

Bootstrapping is a resampling method which was introduced by Bradley Efron in 1979. It is used to approximate the sampling distribution of some statistic using the replacement technique. This approach relies on large amounts of computation rather than mathematical analysis and distributional assumptions of traditional parametric inference. Bootstrapping may make more accurate inferences than the traditional approach. We will see how does it work :

- In this procedure we resample a single data set to create many simulated samples.
- The bootstrap method has an equal probability of randomly drawing each original data point for inclusion in the resampled datasets.
- The procedure allows "with replacement" aspect of the process.
- The procedure creates resampled datasets that are the same size as the original data set.

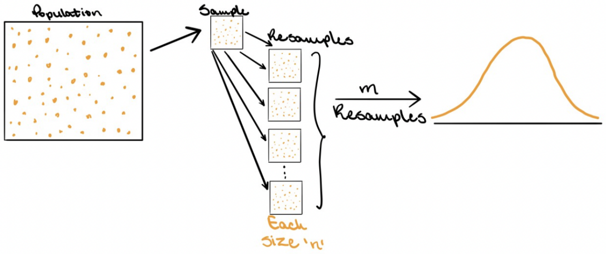


Figure : Pictorial representation of bootstrapping.

Types of bootstrapping

There are two types of bootstrapping namely bootstrapping cases and residual bootstrapping. But we will be using residual bootstrapping in our simulation study.

1. **Bootstrapping cases**

- It doesn't make any assumptions about the distribution.
- We resample the data with replacement, keeping size of the resample equal to the size of the original data set . Example :In linear regression, we will resample the (x,y) pairs.
- Then the statistic of interest is computed from the resample .
- Repeat this process for many times to get a more precise estimate of the distribution of the statistic.

1. **Residual Bootstrapping**

- Residual resampling assumes that the model is correctly specified and that the error terms in the model are identically distributed and independent.
- However, the errors do not need to be normally distributed.

Steps for Residual bootstrapping

- Fit a regression model.
- Save the predicted values as $y_{Pred} = \{y_{(1,Pred)}, y_{(2,Pred)}, ..., y_{(n,Pred)}\}$ and the residual values as $\hat{\epsilon} = \{\hat{\epsilon_1}, \hat{\epsilon_2}, ..., \hat{\epsilon_n}\}$.
- R is resampled (with replacement) from the residuals $\hat{\epsilon}$.
  Let R=$\{R_1, R_2, ..., R_n\}$.
- Creat a bootstrap sample by forming a new response vector as $y_{i,Boot} = y_{i,Pred} + R_i$, $1 \leq i \leq n$
- Fit a regression model that regresses y_{Boot} onto the original regressors x, this gives the bootstrapped estimator β^*.
- Repeat the procedure B-times to create to creat B samples $(\beta_1^*, \beta_2^*, ..., \beta_B^*)$ of $\hat{\beta}$ from each bootstrap sample.
- Analyze the bootstrap distribution to estimate standard errors.

Simple Linear Regression

The formula for the simple linear regression is :

$$y = \beta_0 + \beta_1 x + \epsilon$$

The least squares estimates are :

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

The Standard Error of the estimator $\hat{\beta_1}$ using usual method is given as :

$$SE(\hat{\beta_1}) = \frac{s}{\sqrt{S_{xx}}}$$

Where, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and σ is the (unknown) true standard deviation.

And s is an estimate of σ given as :

$$s = \sqrt{s^2}$$

Where

$$s^2 = \hat{\sigma}^2 = \frac{SSE}{(n-2)} = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{(n-2)}$$

Bootstrap SE : Standard error of the estimator $\hat{\beta_1}$ using bootstrap method is written as :

$$SE_B(\hat{\theta}) = \sqrt{(1/B)\sum_{i=1}^{B}(\theta_i^* - \bar{\theta^*})^2}$$

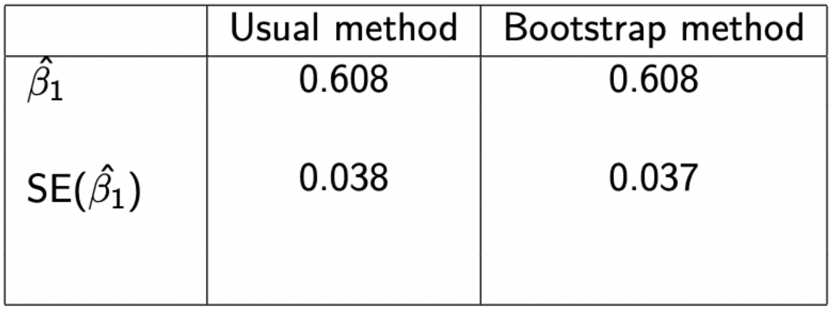
Where

- $\bar{\theta^*}$ is the mean of the bootstrap estimates $(\theta_1^*, \theta_2^*, ..., \theta_B^*)$.
- $(\theta_1^*, \theta_2^*, ..., \theta_B^*)$ are computed using the same computing formula as the one used for $\hat{\theta}$,where $\hat{\theta}$ is a full sample estimator of θ.

Simulation Setup

- Number of observations (n) = 200
- $\beta_1 = 0.6$
- Number of bootstrap samples (B) = 400

Simulation Results



	Usual method	Bootstrap method
$\hat{\beta}_1$	0.608	0.608
SE($\hat{\beta}_1$)	0.038	0.037

Table : $\hat{\beta_1}$ & $SE(\hat{\beta_1})$ by usual & bootstrap method.

From above table, it is clear that the bootstrap method gives close approximation to the usual SE.

Multiple Linear Regression

The multiple linear regression model can be written as :

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

(n $\times$1) $(n{\times}p^{'})$ ($p^{'}$ $\times$ 1) (n $\times$1)

The estimates of the regression coefficients are :

$$\hat{\beta} = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{Y}$$

$$\hat{\beta} = [(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}]\mathbf{Y}$$

Variance of $\hat{\beta}_1$ from usual method is given as :

$$\mathbf{Var}(\hat{\beta}) = (\mathbf{X}^{'}\mathbf{X})^{-1}\sigma^2$$

SD($\hat{\beta}$) is the square roots of the diagonal elements of **Var**($\hat{\beta}$) . And to get **SE**($\hat{\beta}$) , we plug in the estimates of σ by s .

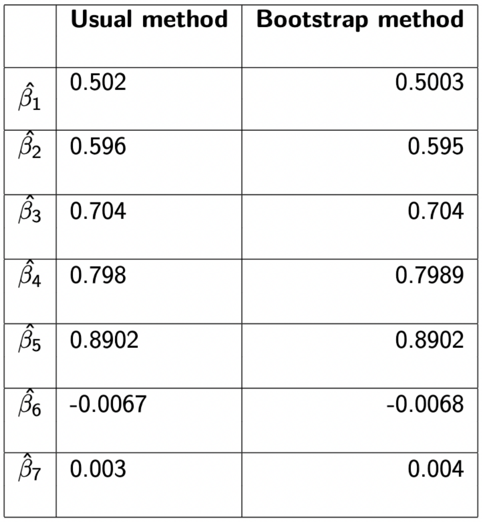
Where s given as :

$$s = \sqrt{\frac{SSE}{(n-(p+1))}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-(p+1))}}$$

Simulation Setup

- Number of observations (n) = 200
- Number of variables (p) = 100
- $\beta_1 = 0.5, \quad \beta_2 = 0.6, \quad \beta_3 = 0.7, \quad \beta_4 = 0.8, \quad \beta_5 = 0.9$
  $\beta_i = 0, \ \forall\, i > 5$
- Number of bootstrap samples = 400

Simulation Results for estimators

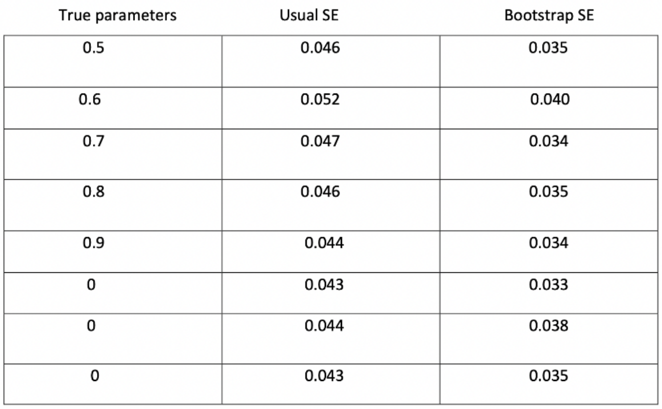


	Usual method	**Bootstrap method**
$\hat{\beta}_1$	0.502	0.5003
$\hat{\beta}_2$	0.596	0.595
$\hat{\beta}_3$	0.704	0.704
$\hat{\beta}_4$	0.798	0.7989
$\hat{\beta}_5$	0.8902	0.8902
$\hat{\beta}_6$	-0.0067	-0.0068
$\hat{\beta}_7$	0.003	0.004

Table : Estimated $\hat{\beta}_1$ by usual method and bootstrap method.

From above table we can infer that bootstrap estimators are approximately same to the usual estimators.

Simulation Results for SE



True parameters	Usual SE	Bootstrap SE
0.5	0.046	0.035
0.6	0.052	0.040
0.7	0.047	0.034
0.8	0.046	0.035
0.9	0.044	0.034
0	0.043	0.033
0	0.044	0.038
0	0.043	0.035

Table : Estimated SE by usual method and bootstrap method .

Above table shows that the bootstrap SE are approximately close to the usual SE.

Some scenario

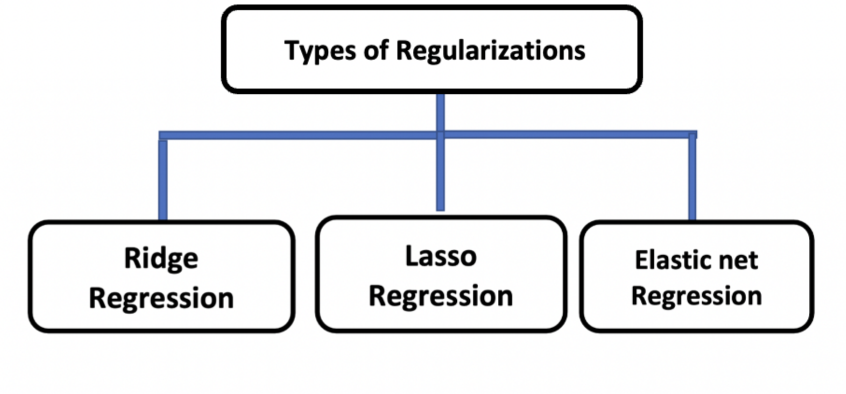
There arises some scenarios when the ordinary least squares estimation does not provide reliable solutions. We discuss some of the such cases .

- $\mathbf{p} > \mathbf{n}$ (Number of variables greater than number of observations).
  - In this case the OLS method becomes unstable.
  - It leads to overfitting . We will see what do we mean by overfitting .
  - The OLS estimates are not unique .
- Highly correlated predictors .
  - In this case the also, the OLS method becomes unstable as the columns of X matrix being not linearly independent .

Both of above situations are found in many applications in fields such as bioinformatics, chemical and drug analysis.

Overfitting : If we have too many features , the learned hypothesis may fit the training set very well, but fail to generalize to new test data.This case is known as overfitting. While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.
There are different types of regularization techniques in regression, but we will mainly see two types of

regularization that is Ridge and Lasso regularization.



Ridge Regression

- It is like least squares but shrinks the estimated coefficients towards zero. .
- The ridge regression coefficients are defined as :

$$\hat{\beta}_{Ridge} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \; + \; \lambda \sum_{j=1}^{p} \beta_j^2$$

  - y is response vector in $\mathbb{R}^n$.
  - $\mathbf{X}$ is predictor matrix in $\mathbb{R}^{n \times p}$.
  - $\lambda \geq 0$ is a tuning parameter, which controls the strength of the penalty term.

- Ridge Regression estimators have closed form solutions :

$$\hat{\beta}_{Ridge} = (\mathbf{X^T X} + \lambda \mathbf{I})^{-1} \mathbf{X^T y}$$

$$\mathbf{Var}(\hat{\beta}_{Ridge}) = \sigma^{\mathbf{2}} (\mathbf{X^T X} + \lambda \mathbf{I})^{-1} \mathbf{X^T X} (\mathbf{X^T X} + \lambda \mathbf{I})^{-1}$$

  $\mathbf{SD}(\hat{\beta}_{Ridge})$ is the square roots of the diagonal elements of $\mathbf{Var}(\hat{\beta})$.

  And to get the SE of estimators we plug in the estimates of $\sigma^{\mathbf{2}}$ and the optimal value of λ.

LASSO Regression

- It stands for the Least absolute Shrinkage and Selection Operator .
- It is like least squares but shrinks some of the βs to exactly zero, resulting in a regression model that's easier to interpret. And it also performs variable selection.

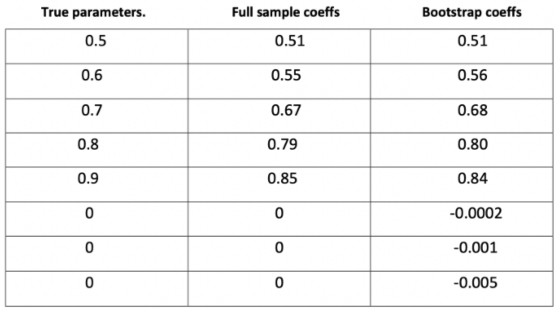
- The lasso regression coefficients are defined as :

$$\hat{\beta}_{lasso} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \ + \ \lambda \sum_{j=1}^{p} |\beta_j|$$

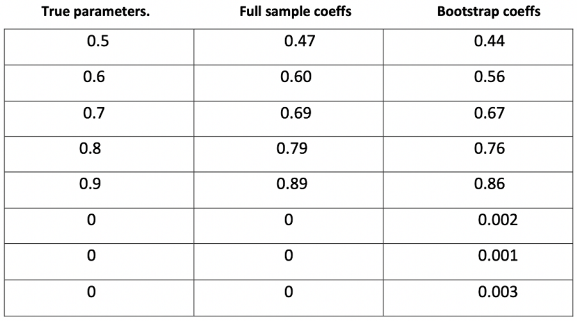
  - y is response vector in $\mathbb{R}^n$.
  - X is predictor matrix in $\mathbb{R}^{n \times p}$.
  - $\lambda \geq 0$ is a tuning parameter, which controls the strength of the penalty term.

- There is no closed form analytical solution , but we rely on numerical solution schemes like gradient descent .

Simulation results for estimators



True parameters.	Full sample coeffs	Bootstrap coeffs
0.5	0.51	0.51
0.6	0.55	0.56
0.7	0.67	0.68
0.8	0.79	0.80
0.9	0.85	0.84
0	0	-0.0002
0	0	-0.001
0	0	-0.005

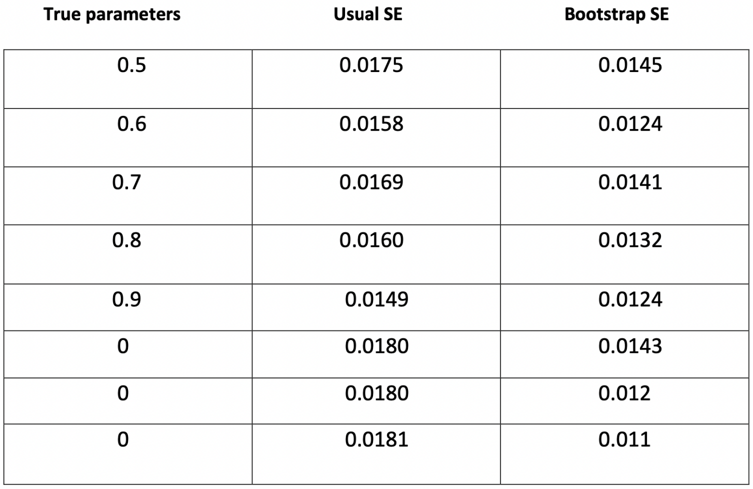


True parameters.	Full sample coeffs	Bootstrap coeffs
0.5	0.47	0.44
0.6	0.60	0.56
0.7	0.69	0.67
0.8	0.79	0.76
0.9	0.89	0.86
0	0	0.002
0	0	0.001
0	0	0.003

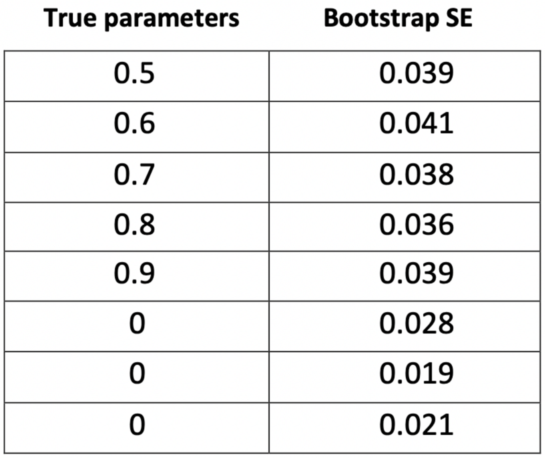
Tables : Estimated $\hat{\beta}_1$ from usual & bootstrap method. Left table is for the ridge regression and right table is for the lasso regression .

From above tables it is clear that the results are not very accurate but the bootstrap estimators provide decent approximation to the usual estimators .

Simulation results for SE



True parameters	Usual SE	Bootstrap SE
0.5	0.0175	0.0145
0.6	0.0158	0.0124
0.7	0.0169	0.0141
0.8	0.0160	0.0132
0.9	0.0149	0.0124
0	0.0180	0.0143
0	0.0180	0.012
0	0.0181	0.011



True parameters	Bootstrap SE
0.5	0.039
0.6	0.041
0.7	0.038
0.8	0.036
0.9	0.039
0	0.028
0	0.019
0	0.021

Tables: SE from usual & bootstrap method. Left table is for the ridge regression and right table is for the lasso regression .

From above tables it is clear that the results are not very accurate but the bootstrap SE provides decent approximation to the usual SE .

Conclusion

Bootstrapping is a powerful computer-based tool that creates statistical inferences without relaying on many assumptions. It is also an appropriate way to control and check the stability of the results. We can use bootstrap estimators and bootstrap SE in the case where deriving the distribution for the estimator gets very challenging like lasso regression . Bootstrapping is often applied in statistical inferences such as regression models, confidence intervals, and other machine learning topics. Using this method we can avoid the cost of repeating the experiment to get other groups of sample data in the case of small sized data.

References

[1] Measurement Error Models: Wayne A. Fuller New York: Wiley, 1987.

[2] Econometric Theory-Measurement Error Models - Salabh .

[3] An Introduction to the Bootstrap Bradley Efron Department of Statistics Stanford University and Robert J. Tibshirani Department of Preventative Medicine and Biostatistics and Department of Statistics, University of Toronto .

[4] Bootstrap: A Statistical Method Kesar Singh and Minge Xie Rutgers University .

[5] Statistical learning with Sparsity The Lasso and Generalization: Trevor Hastie Robert Tibshirani Martin Wainwright .