BANA 7046-001

Professor Yan Yu

Homework #2

February 5th, 2019

Bhavya Bansal

Anupriya Kushwanshi

Kunal Indore

Macy McCollum

**Boston Housing Data**

*Executive Summary*

We have used the Boston dataset for our analysis, it is available in the R library Mass. It contains 506 observations of 14 variables, the table below contains the name, class, and definition of each variable.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Class** | **Definition** |
| crim | Numeric | per capita crime rate by town |
| zn | Numeric | proportion of residential land zoned for lots over 25,000 sq.ft |
| indus | Numeric | proportion of non-retail business acres per town |
| chas | Integer | Charles River dummy variable (1 if tract bounds river; 0 otherwise) |
| nox | Numeric | nitrogen oxides concentration (parts per 10 million) |
| rm | Numeric | average number of rooms per dwelling |
| age | Numeric | proportion of owner-occupied units built prior to 1940 |
| dis | Numeric | weighted mean of distances to five Boston employment centres |
| rad | Integer | index of accessibility to radial highways |
| tax | Numeric | full-value property-tax rate per \$10,000 |
| ptratio | Numeric | pupil-teacher ratio by town |
| black | Numeric | *1000(Bk - 0.63)^2* where *Bk* is the proportion of blacks by town. |
| lstat | Numeric | lower status of the population (percent) |
| medv | Numeric | median value of owner-occupied homes in \$1000s |

Our goal is to discover what factors affect the price of homes, hence ‘medv’ will be our variable of interest or response variable. We will then determine which of the remaining variables are the most influential in determining home value (medv), the outcome will determine our regression model. According to our analysis, best subset variable selection gives the best results hence, we will choose that model.

**A Simulation Study (Linear Regression)**

*Executive Summary*

The mean function E(y|x)=4+0.9\*x1+3\*x2 will be used to generate 200 random numbers, those numbers will then be used to create a linear line consequently estimating the coefficients of the given equation. An error term will also be included, initially the value is 1. To test the impact of sample size and error the model was regenerate at different values of n and sigma. The closest estimate had the largest sample size and smallest sigma value, 5000 and 0.1 respectively, the resulting coefficients were (4.03, 3.02, 0.90) Four major trends were observed during the study.

1. We observe that for almost every combination only two variables are selected i.e x2 and x1 with exception of sample size 25 and sigma 0.5 and sample size 500 and sigma 1.
2. As sample size increases, we observe the Mean square error get closer to square of sigma which means that residual error gets closer to its true value.
3. R square and adj. R-square decreases with increase in sigma values. This is because increase in error causes increase in variability and hence decreases the accuracy of the model.
4. We also observe that as sample size increases adj R-square gets closer to actual R-square because accuracy of the model increases with the sample size.

**Monte Carlo Simulation Study**

*Executive Summary*

The same expression as the previous study was used in the Monte Carlo Simulation Study, E(y|x)=4+0.9\*x1+3\*x2. The sample size of 200 random numbers and the numerical ranges of x1 and x2 remain the same. The primary difference between the two is that Monte Carlo performed the linear regression estimate 100 times, a new error term was generated each time. With each iteration a matrix was populated by the resulting estimate coefficients, at the end the mean and variance of the matrix was reported. The matrix contains estimates of each coefficient, when each column is plotted it will result in a bell curve. Each iteration may have over or underestimated the actual value so the mean of the results should be close to the actual value. This method was more accurate than the Simulation Study at any sample size or noise level.

The results are as follows:

|  |  |  |
| --- | --- | --- |
| **Coefficients** | **True** | **Mean Fitted Coefficients** |
| Intercept (beta0) | 4.0 | 4.0019983 |
| x1 (beta1) | 0.9 | 0.8999554 |
| x2 (beta2) | 3.0 | 3.0008277 |

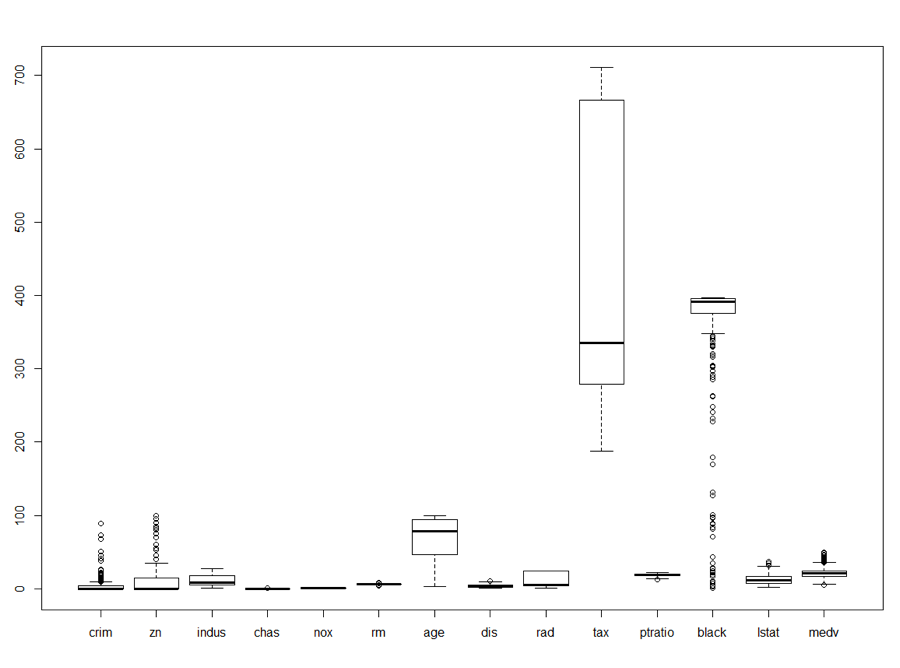
**Boston Housing**

*Exploratory Data Analysis*

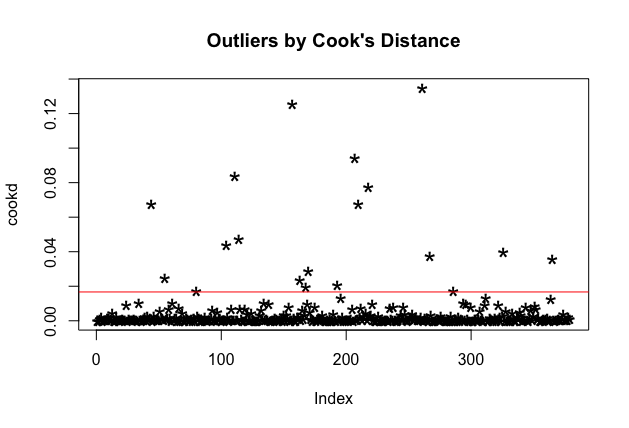
Our initial exploratory data analysis reveals that there are no empty values in our datasets. The following table shows the summary statistics of each variable in our dataset, categorical variables represented by integers were not included.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Min** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max** |
| crim | 0.0063 | 0.0820 | 0.2565 | 3.6135 | 3.6771 | 88.9762 |
| zn | 0.00 | 0.00 | 0.00 | 11.36 | 12.50 | 100.00 |
| indus | 0.46 | 5.19 | 9.69 | 11.14 | 18.10 | 27.74 |
| nox | 0.3850 | 0.4490 | 0.5380 | 0.5547 | 0.6240 | 0.8710 |
| rm | 3.561 | 5.886 | 6.208 | 6.285 | 6.623 | 8.780 |
| age | 2.90 | 45.02 | 77.50 | 68.57 | 94.08 | 100.00 |
| dis | 1.130 | 2.100 | 3.207 | 3.795 | 5.188 | 12.127 |
| tax | 187.0 | 279.0 | 330.0 | 408.2 | 666.0 | 711.0 |
| pratio | 12.60 | 17.40 | 19.05 | 18.46 | 20.20 | 22.00 |
| black | 0.32 | 375.38 | 391.44 | 356.67 | 396.23 | 396.90 |
| lstat | 1.73 | 6.95 | 11.36 | 12.65 | 16.95 | 37.97 |
| medv | 5.00 | 17.02 | 21.20 | 22.53 | 25.00 | 50.00 |

Outliers

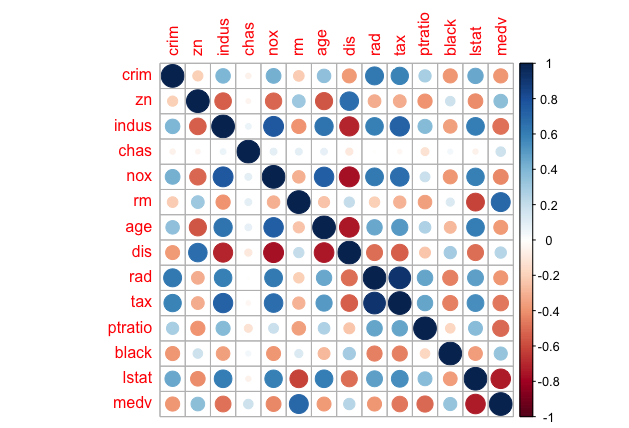


As we can see in the above box plot we can see quite a few outliers for variables like black, zn, crim. When a linear regression model using all variables to predict medv was performed, 19 of the 506 observations were determined to be outliers using Cook’s Distance.



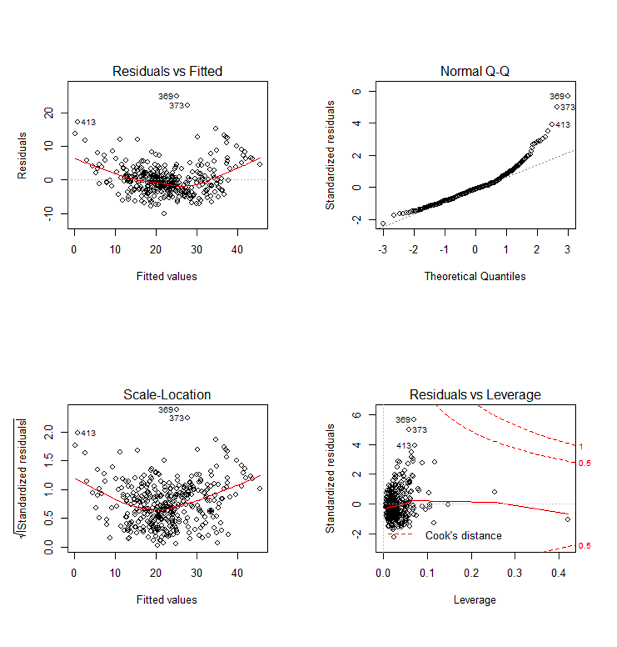
Pairwise Correlation

A pairwise correlation indicated a strong positive correlation between tax and rad, a strong negative correlations was identified between dis and age. We are most interested in correlations between medv and other variables. Most prominent is the positive relationship with rm and negative relationship with lstat.



*Best Model Selection*

Original Model with All the Covariates



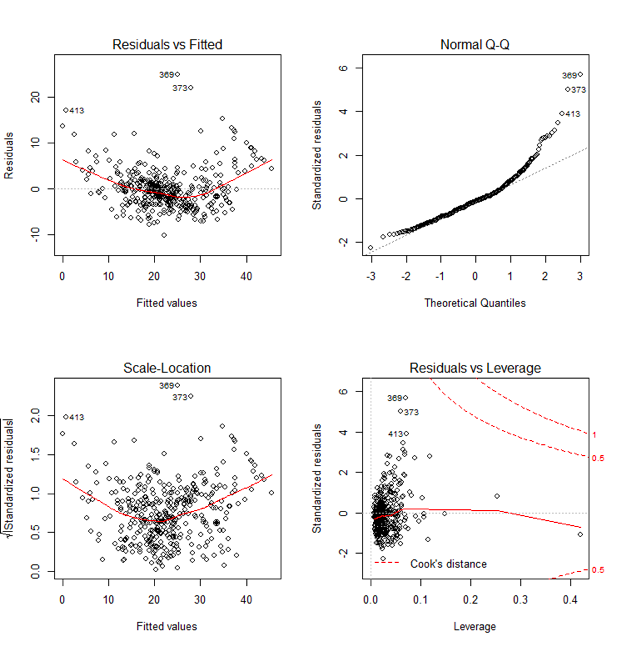
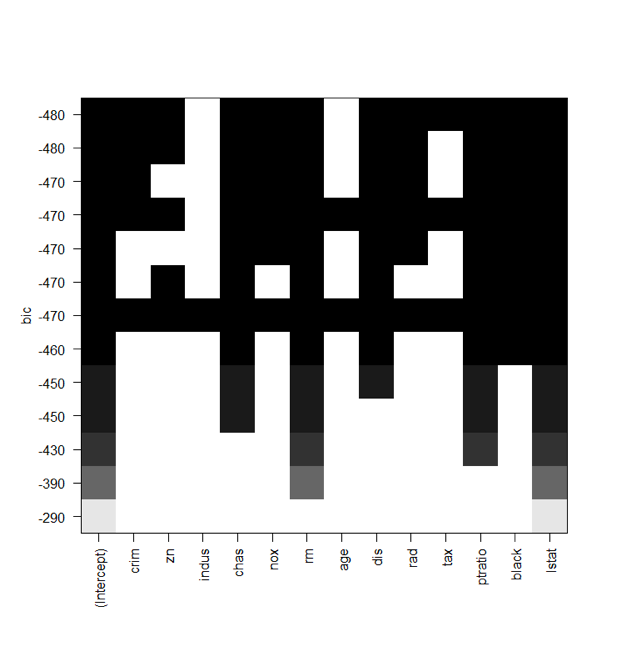
We can see outliers 369 and 373 very clearly. Normal Q-Q shows a slightly skewed distribution wherein almost all the data points appear to lie on straight line except its tail curves at the end. Residuals seem to follow some pattern yet randomly distributed.

Summary statistics:

|  |  |
| --- | --- |
| **R^2** | 76.51% |
| **Adjusted R^2** | 75.67% |
| **MSE** | 19.95 |
| **p-value** | <0.05 |
| **AIC** | 2240.077 |

Model Created After Best Subset Variable Selection

We use regsubsets function which is a part of leaps library to which uses exhaustive algorithms for variable selection. Output of best subset model shows that we have to include all the variables except Indus and age.



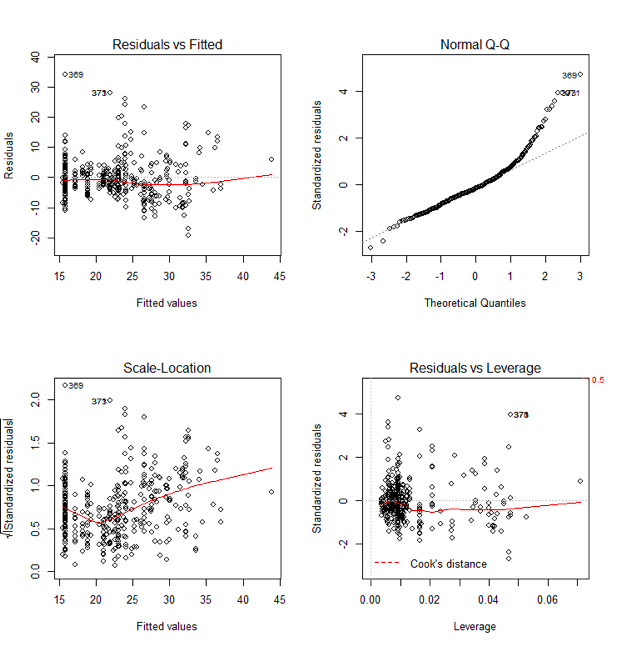
We can see that there is no significant difference in this model as compared to the previous one. However, Residual plot is giving better results this time.

Summary statistics:

|  |  |
| --- | --- |
| **R^2** | 76.40% |
| **Adjusted R^2** | 75.70%, |
| **MSE** | 20.04 |
| **p-value** | <0.05 |
| **AIC** | 2237.819 |

Stepwise Regression

We use bidirectional stepwise regression which considers AIC of each model to select the best possible model by employing both forward and backward stepwise and comparing AIC at each step.



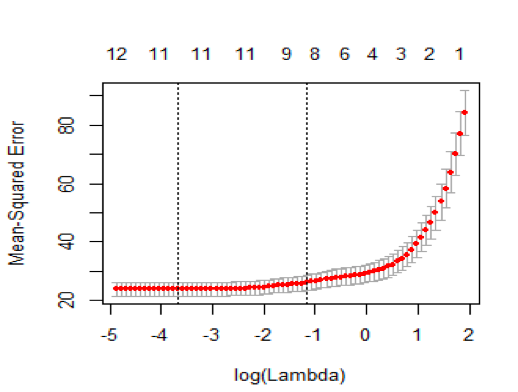
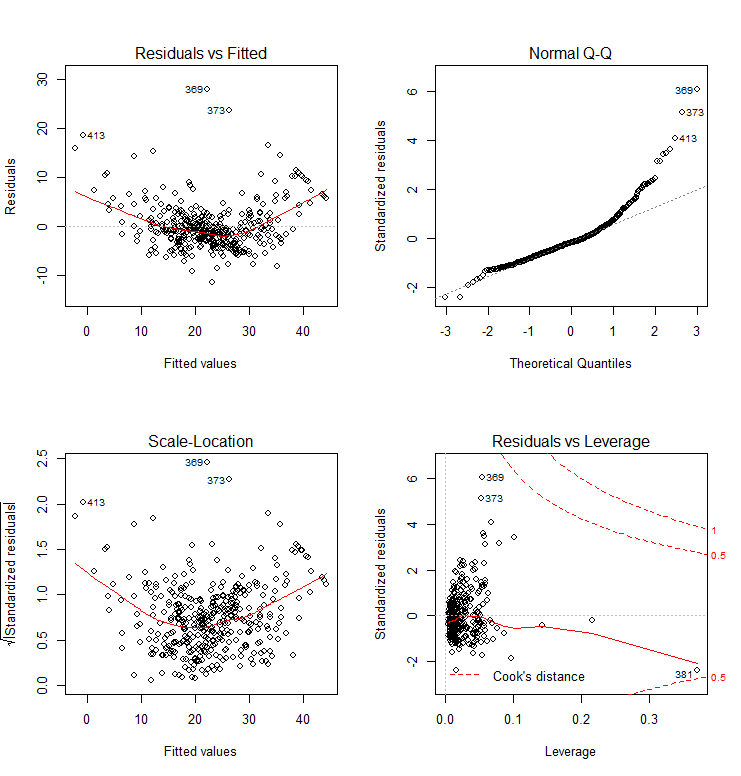
We can see that residual plot is quite disappointing as the plot is not very randomized.

Summary statistics:

|  |  |
| --- | --- |
| **R^2** | 38.45%, |
| **Adjusted R^2** | 37.80% |
| **MSE** | 52.28 |
| **p-value** | <0.05 |
| **AIC** | 2587.146 |

The performance of this model is the poorest among all as it has the highest MSE, AIC and lowest Adjusted R^2 values. We will not be selecting this as our final model.

LASSO

For LASSO regression we used glmnet library from R packages. LASSO and Ridge regression are used to cure overfitting by inducing a penalty in the model. The degree of the penalty is represented by lambda(λ).   
 In cv.glmnet function, if we take alpha = 0, it will become Ridge, alpha = 1 is LASSO and anything between 0–1 is Elastic net. Cv.Glmnet() takes x as matrix, hence we converted all the covariates in the matrix form using as.matrix() and ran the function on original Boston dataset.  
 The output, shows from left to right the number of nonzero coefficients (Df), the percent deviance explained (%dev) and the value of λ (Lambda). Although by default glmnet calls for 100 values of lambda the program stops early if `%dev does not change sufficiently from one lambda to the next.   
 The function glmnet() returns a sequence of models for us to choose from. We prefer the software to select one of them automatically. Cross-validation is perhaps the simplest and most widely used method for that task. cv.glmnet is the main function to do cross-validation here, along with various supporting methods such as plotting and prediction.   
The plot of the output of cv.glmnet:  
   
  
It includes the cross-validation curve (red dotted line), and upper and lower standard deviation curves along the λ sequence. Two selected λ’s are indicated by the vertical dotted lines. We can view the selected λ’s and the corresponding coefficients. lambda.min is the value of λ that gives minimum mean cross-validated error. The other λλ saved is lambda.1se, which gives the most regularized model such that error is within one standard error of the minimum. We choose the best 2 λ values i.e. ‘lambda.min’ where minimum error observed and another is ‘lambda.1se’ where error is within 1 standard error of minimum error. We chose ‘lasso$lambda.1se’ which in our case is 0.3789258. We ran the function glmnet() and found out that percentage deviation is 71.07%. Based on the LASSO regression result, we did not include indus, zn, age, rad, and tax.  


The residuals seem to be concentrated in the middle and seem to follow a pattern.

Summary statistics:

|  |  |
| --- | --- |
| **R^2** | 74.25% |
| **Adjusted R^2** | 73.70% |
| **MSE** | 21.28 |
| **p-value** | <0.05 |
| **AIC** | 2264.843 |

**A Simulation Study**

The MSE increases as noise increases and decreases as the sample size increase, with the exception of n=25. Sigma = 0.1 had the highest R-square and adjusted R-square values, >0.96 at every sample size. The remaining errors did not see a correlation higher than 0.57, these fields demonstrate the importance of minimizing error.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Noise** | **Sample Size** | **Selected Model** | **Model Coefficient Estimates**  **(int, x2, x1)** | **MSE** | **R square** | **adj. R square** |
| σ = 0.1 | 25 | y ~ x2 + x1 | b(est) = (4.10, 3.07,0.89) | 0.012 | 0.9684 | 0.9655 |
| 100 | y ~ x2 + x1 | b(est) = (4.18,3.20,0.92) | 0.010 | 0.9618 | 0.9610 |
| 200 | y ~ x2 + x1 | b(est) = (3.98,2.97,0.89) | 0.011 | 0.9673 | 0.9969 |
| 500 | y ~ x2 + x1 | b(est) = (3.93,2.94,0.90) | 0.011 | 0.9656 | 0.9654 |
| **5000** | **y ~ x2 + x1** | **b(est) = (4.03,3.02,0.90)** | **0.010** | **0.9686** | **0.9686** |
| σ = 0.5 | 25 | y ~ x3 + x1 | b(est) = (0.88,1.23,2.18) | 0.235 | 0.4906 | 0.4443 |
| 100 | y ~ x2 + x1 | b(est) = (4.89,4.01,0.98) | 0.279 | 0.5696 | 0.5607 |
| 200 | y ~ x2 + x1 | b(est) = (3.90,2.84,0.87) | 0.265 | 0.5266 | 0.5218 |
| 500 | y ~ x2 + x1 | b(est) = (3.67,2.71,0.92) | 0.267 | 0.5261 | 0.5242 |
| 5000 | y ~ x2 + x1 | b(est) = (4.17,3.12,0.88) | 0.242 | 0.5519 | 0.5517 |
| σ = 1 | 25 | y ~ x1 | b(est) = (0.85,0.92) | 0.969 | 0.1552 | 0.1185 |
| 100 | y ~ x2 + x1 | b(est) = (5.79,1.06,5.02) | 1.077 | 0.3110 | 0.2968 |
| 200 | y ~ x2 + x1 | b(est) = (1.13,4.21,2.07) | 1.059 | 0.2045 | 0.1964 |
| 500 | y ~ x3 + x1 | b(est) = (0.91,1.20,2.16) | 1.067 | 0.2182 | 0.2151 |
| 5000 | y ~ x2 + x1 | b(est) = (4.34,3.25,0.86) | 0.967 | 0.2352 | 0.2349 |

**Monte Carlo SImulation Study**

We sampled the values at random from the input response and predictor variables. The result is a probability distribution of possible outcomes. In this way, Monte Carlo simulation provides a much more comprehensive view of what may happen.  
We can see that there is no significant difference between true and fitted coefficients when we ran 100 simulations. This suggests that the risk of our data performing poor over time is reduced.

The average estimated model MSE is 0.9839276 which is close to 1.

Mean bias of the coefficients of the fitted model are:

intercept(beta0) beta1(x1) beta1(x2)

1.9906879 0.9456318 1.7188840

The difference between the estimator's expected value and the true value of the parameter being estimated is 0.94 for x1 and 1.71 for x2.

We did 100 random sampling and performed the regression analysis. The mean variability of x1 is 0.02 and x2 is 0.52.

intercept(beta0) beta1(x1) beta1(x2)

0.60252290 0.02021896 0.52209775

|  |  |  |
| --- | --- | --- |
| **Coefficients** | **True** | **Mean Fitted Coefficients** |
| Intercept (beta0) | 4.0 | 4.0019983 |
| x1 (beta1) | 0.9 | 0.8999554 |
| x2 (beta2) | 3.0 | 3.0008277 |