Project 4: Regression Analysis of Big Data

INTRODUCTION

This project aims to build upon such insights, incorporating the divide-and-conquer algorithm and advanced regression techniques to enhance the accuracy of housing price predictions. The application of divided regression analysis to predict housing prices, considering factors such as lot size, living area, and selling prices within datasets is also performed by the researchers [1]. Additionally, studies on California housing, such as the work by Yann and Yeh (2022) in sparse spatial autoregressions [2], provide a foundation for understanding the dynamics of house prices in diverse neighborhoods.

Task:

1. Properly select one independent variable from the data and carry out a polynomial regression analysis with Y as the response by using the "divide-and-conquer" algorithm.

RESULTS

Summary and Coefficient of a final model:

Residuals:  
## Min 1Q Median 3Q Max   
## -450948 -55535 -16549 37400 453242   
   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## [1,] 1562330.7 580.3 2692.11 <2e-16 \*\*\*  
## [2,] 1624949.1 83375.0 19.49 <2e-16 \*\*\*  
## [3,] 1671907.0 83375.0 20.05 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
## Residual standard error: 83370 on 20637 degrees of freedom  
## Multiple R-squared: 0.478, Adjusted R-squared: 0.478

## F-statistic: 9450 on 2 and 20637 DF, p-value: < 2.2e-16

FIGURES

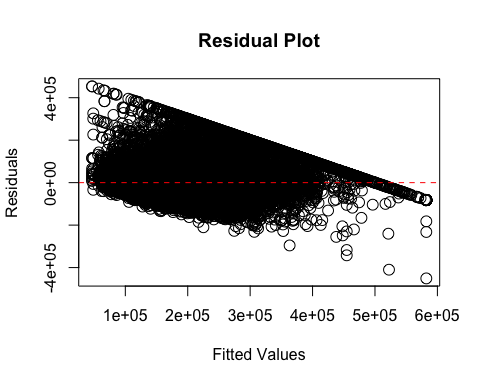
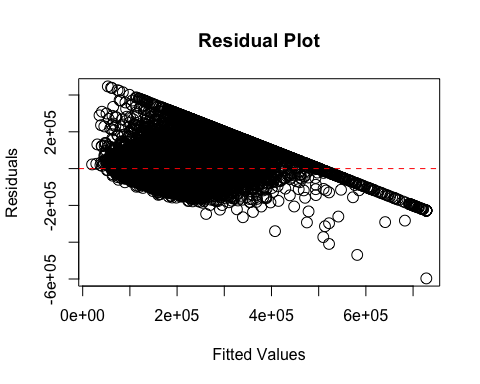


Figure1: (a) Scatter plot shows a residual values of a fitted polynomial regression model, where Y is a median\_house\_value (dependent variable) and median\_income as an independent variable. (b) Scatter plot shows residual values for a multiple linear regression model where two independent variables, median\_income and housing\_median\_value was taken.

(a)

(b)

DISCUSSION

Interpretation of polynomial regression analysis

The provided output from a polynomial regression model predicts the dependent variable (median house value) based on the independent variable (median income), using a polynomial of degree 2. The coefficients represent the estimated parameters for the polynomial terms. Here, the coefficients are highly significant (p-value < 0.001). The residual standard error is a measure of the variability of the residuals. Multiple R-squared measures how well the model explains the variability in the response variable. In this case, about 47.8% of the variability in the dependent variable is explained by the model. The p-value associated with the F-statistic is very close to zero, indicating that the model is statistically significant.

1. Properly select more than one independent variable from the data and conduct a multiple linear regression analysis with Y as the response using the "divide-and-conquer" algorithm.

RESULTS

Summary and Coefficient of a Multiple Linear Regression (mlr) model:

Residuals:  
## Min 1Q Median 3Q Max   
## -596748 -53834 -15000 36719 446725   
   
## Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -10189.03 1915.41 -5.32 1.05e-07 \*\*\*  
## median\_income 43169.19 298.36 144.69 < 2e-16 \*\*\*  
## housing\_median\_age 1744.13 45.04 38.73 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   
## Residual standard error: 80850 on 20637 degrees of freedom  
## Multiple R-squared: 0.5091, Adjusted R-squared: 0.5091   
## F-statistic: 1.07e+04 on 2 and 20637 DF, p-value: < 2.2e-16

DISCUSSION

Interpretation of Multiple linear regression model

The model predicts the dependent variable (presumably median house value) based on two independent variables, median\_income and housing\_median\_age. Here, the residuals are the differences between the observed values and the predicted values. They represent the errors of the model. The coefficients represent the estimated parameters for the model. The intercept represents the estimated median house value when both median\_income and housing\_median\_age are zero. For a one-unit increase in median\_income, the estimated change in the median house value is $43,169.19. For a one-unit increase in housing\_median\_age, the estimated change in the median house value is $1,744.13. All coefficients are statistically significant (p-value < 0.001). The residual standard error (RSE) is approximately 80,850. RSE measures the average amount by which the actual values deviate from the predicted values. A lower RSE indicates a better fit of the model to the data. Multiple R-squared measures how well the model explains the variability in the response variable. Adjusted R-squared considers the number of predictors in the model. In this case, about 50.91% of the variability in the dependent variable is explained by the model. The F-statistic tests the overall significance of the model. The p-value associated with the F-statistic is very close to zero, indicating that the model is statistically significant. See Figure1(b).

1. Show the response variable Y (median house value) is highly skewed to the right, and then apply a logarithm transformation of Y as the response and repeat (1) and (2).

RESULTS

## Skewness of median\_house\_value: 0.9776922

# Apply logarithm transformation to median\_house\_value  
housing$log\_median\_house\_value <- log(housing$median\_house\_value)

## Skewness after log transformation: -0.1731628 # after log transformation

coefficients and summary of the final model after log transformation of Y-variable

Residuals:  
## Min 1Q Median 3Q Max   
## -2.59927 -0.26899 -0.00986 0.24990 1.97192   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## [1,] 9.482476 0.002883 3289.46 <2e-16 \*\*\*  
## [2,] 9.797478 0.414145 23.66 <2e-16 \*\*\*  
## [3,] 10.053751 0.414145 24.28 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4141 on 20637 degrees of freedom  
## Multiple R-squared: 0.4705, Adjusted R-squared: 0.4705   
## F-statistic: 9170 on 2 and 20637 DF, p-value: < 2.2e-16

coefficients and summary of the mlr model after log transformation of Y-variable

Residuals:  
## Min 1Q Median 3Q Max   
## -2.72248 -0.25516 0.00484 0.25998 1.74835   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.110e+01 9.929e-03 1117.60 <2e-16 \*\*\*  
## median\_income 2.028e-01 1.547e-03 131.14 <2e-16 \*\*\*  
## housing\_median\_age 7.082e-03 2.335e-04 30.33 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4191 on 20637 degrees of freedom  
## Multiple R-squared: 0.4577, Adjusted R-squared: 0.4577   
## F-statistic: 8709 on 2 and 20637 DF, p-value: < 2.2e-1

FIGURES

(b)

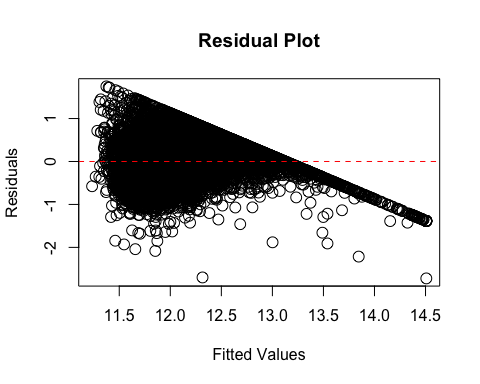
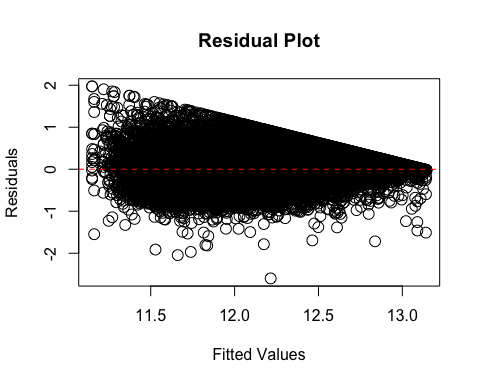
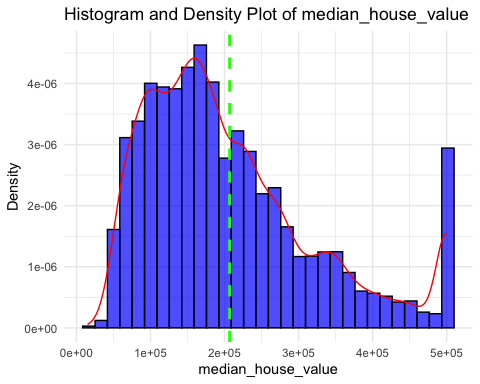
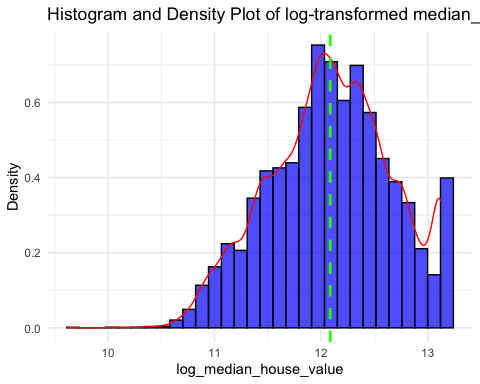


Figure2: (a) Histogram shows that Y- response variable (median\_house\_value) is highly skewed towards right. (b)Histogram shows the distribution of Y- variable towards mean = 0 after log transformation. (c) The scatter plot shows residual and fitted values for polynomial regression model after using log transformed Y- variable. (d) The scatter plot shows residual and fitted values for multiple linear regression model after using log transformed Y- variable

(a)

(c)

(d)

DISCUSSION

The skewness value of 0.9776922 for the median\_house\_value variable indicates that the distribution of house values is positively skewed. In the context of median\_house\_value, this implies that there is a tail on the right side of the distribution, meaning that there might be relatively few houses with very high values, creating a rightward skew. To correct the skewness of the Y-variable i.e. median\_house\_value, log transformation was applied and further the data was visualized for the skewness. (see in the appendix for before and after log transformation). Further, polynomial regression and multiple linear regression models were applied using log-transformed Y-variable. Results of Polynomial regression analysis to the log-transformed Y- variable and the median\_income as an independent variable give insight into the fitted model. From the significance of coefficient values, we can observe that all coefficients have extremely low p-values (< 2e-16), indicating that they are statistically significant. Residual standard error is also low. Furthermore, Multiple R-squared, and F-statistic test results also state the overall significance of the model. A low p-value (< 2.2e-16) indicates that the model is statistically significant. Overall, the model seems to fit the data well, with statistically significant coefficients and a good fit based on R-squared values.

When multiple linear regression analysis performed to the log-transformed Y- variable and independent variables (here, median\_income and housing\_median\_age) showed the coefficient for median\_income as 0.2028. This indicates the estimated change in log median house value for a one-unit increase in median income, holding other variables constant. The coefficient for housing\_median\_age as 0.0071 indicates the estimated change in log median house value for a one-unit increase in housing median age, holding other variables constant. The model is statistically significant, as indicated by the low p-values for coefficients and the F-statistic. The coefficients provide information about the direction and magnitude of the relationship between independent and dependent variables. The RSE is relatively low, suggesting that the model's predictions have a reasonable amount of precision.

1. Comment on the analyses above and the "divide-and-conquer" algorithm.

CONCLUSION

After the log transformation, both median\_income and housing\_median\_age have significant coefficients, suggesting they are important predictors of the log-transformed median house value. The model explains approximately 45.77% of the variability in log-transformed median house value, as indicated by the multiple R-squared. The p-values associated with the coefficients are very small, suggesting that both independent variables are statistically significant. The residuals have spread around zero, indicating that the model is a reasonable fit for the data. Overall, the model seems to be a good fit, and the included variables are meaningful predictors of log-transformed median house value.

Divide and conquer is a fundamental algorithmic paradigm that enhances problem-solving efficiency by breaking down complex problems into smaller, more manageable subproblems. The process involves three main steps: divide, conquer, and combine. Firstly, the problem is divided into smaller sub-problems, simplifying the overall complexity. Subsequently, each subproblem is solved independently, often through recursive calls, until reaching base cases that can be easily resolved. Finally, the solutions to the subproblems are combined to derive the solution for the original, more complex problem. Though this algorithm is used for linear regression, it fits best for logistic regression analysis. Here, instead of dividing the problem into subproblems, the algorithm divides the sample into subsets and then applies the linear regression algorithm.

REFERENCES

1. Chen, Yucong. (2023). Analysis and Forecasting of California Housing. Highlights in Business, Economics and Management. 3. 128-135. 10.54097/hbem.v3i.4704.
2. Goh, Yann & Goh, Yeh & Chun Chieh, Yip & Ng, Kooi. (2022). Housing Price Prediction by Divided Regression Analysis. Chiang Mai Journal of Science. 49. 1669-1682. 10.12982/CMJS.2022.102.
3. Douglas R. Smith, The design of divide and conquer algorithms, Science of Computer Programming, Volume 5, 1985, Pages 37-58, ISSN 0167-6423, <https://doi.org/10.1016/0167-6423(85)90003-6>.
4. Kim, Y., & Oh, H. (2021). Comparison between Multiple Regression Analysis, Polynomial Regression Analysis, and an Artificial Neural Network for Tensile Strength Prediction of BFRP and GFRP. *Materials (Basel, Switzerland)*, *14*(17), 4861. <https://doi.org/10.3390/ma14174861>
5. Eva Ostertagová, Modelling using Polynomial Regression, Procedia Engineering, Volume 48, 2012, Pages 500-506, ISSN 1877-7058, <https://doi.org/10.1016/j.proeng.2012.09.545>.
6. M. Ajona, P. Vasanthi, D.S. Vijayan, Application of multiple linear and polynomial regression in the sustainable biodegradation process of crude oil, Sustainable Energy Technologies and Assessments, Volume 54, 2022, 102797, ISSN 2213-1388, <https://doi.org/10.1016/j.seta.2022.102797>.
7. Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, *26*(2), 105–109. <https://doi.org/10.3969/j.issn.1002-0829.2014.02.009>
8. Lee D. K. (2020). Data transformation: a focus on the interpretation. *Korean journal of anesthesiology*, *73*(6), 503–508. <https://doi.org/10.4097/kja.20137>
9. <http://library.virginia.edu/data/articles/interpreting-log-transformations-in-a-linear-model>