

Project 1 (Due 3/10 at 10pm)

Implement simple linear regression methodology to evaluate the relationship between the number of homes sold and the total sales revenue generated by those sales in three housing markets in Texas. The project makes use of the `txhousing` data (available in the `tidyverse` package) in R. Submissions should be in the form a report, no longer than four pages. All your code used to perform different components of the analysis should be included in an appendix (not counted towards page limit).

The full dataset consists of observations collected monthly from January 2000 through July 2015 in 46 different housing markets across the state. For this project, you should select three markets to consider from the following list: Abilene, Austin, Fort Worth, Garland, Houston, Paris, San Antonio, Sherman-Denison, and Wichita Falls.

In each housing market considered:

- Fit a simple linear regression model using `log(sales)` (the logarithm of number of units sold) to predict `log(adjusted volume)`.
- Calculate and interpret a metric that assesses the quality of the regression model.
- Perform a residual analysis of the model, discussing all findings and highlighting any anomalies that you encounter.

After performing the three separate regression analyses, compare the slope parameters for the three housing markets considered to assess whether the observed relationship between `sales` and `adjusted volume` is comparable across markets. One way in which you can do this is by constructing confidence intervals for the slope parameters.

Finally, it is projected that all of the housing markets under consideration will have sales totalling 500 units in March 2023. For each of the three markets, construct a 95% prediction intervals (no adjustment for multiple intervals required) considering the `volume` (total sales) in this month assuming the model has remained accurate.

(**Note:** The coverage accuracy of prediction intervals is maintained when applying monotone transformations. So, for example, if you construct a prediction interval and the endpoints are both positive, then taking either the square root or squaring both endpoints – monotone transformations on the positive real line – results in appropriate prediction intervals on the transformed scale.)

Your report should include mathematical equations as needed to aid a general audience (think Calc I background) in understanding important concepts. Equations should never be included without explanation. Additionally, the use of appropriate data visualizations is highly encouraged. Also discuss any assumptions that you make in the model, both those related to the linear regression model and any other that seem relevant to the application.