ADSP 31014 Statistical Models for Data Science

Autumn Quarter 2024 Assignment 2

Question 1 (50 points)

We continue to analyze the **NorthChicagoTownshipHomeSale.csv**. This data is about 403 single-family homes sold in the North Chicago township of Cook County between 2018 and 2020. The North Chicago township lies entirely within the City of Chicago. The township is bounded on the east by Lake Michigan, the north by Fullerton Avenue, and surrounded by the Chicago River. It contains many affluent neighborhoods, e.g., Gold Coast, Magnificent Miles, Streeterville, and Lincoln Park, in Chicago.

In Assignment 1, we determined to apply the Box-Cox transformation with a power lambda value of -0.5 to the *Sale Price* which is the real estate transaction amount in thousands of dollars. We will continue to use this transformed *Sale Price* as our response variable for our analysis.

We only use complete observations with no missing values for training. Therefore, we will exclude a record if missing values (i.e., NaN) exist in any predictors or the response variable.

Besides the Intercept term, the model may include the following predictors:

Categorical Predictors:

1. *Wall Material*: "Stucco", "Frame", "Frame + Masonry", and "Masonry"
2. *Roof Material*: "Other", "Shingle + Asphalt", and "Tar + Gravel"
3. *Basement*: "Partial", "Crawl", "Slab", and "Full"
4. *Central Air Conditioning*: "No Central A/C" and "Central A/C"

We will reorder the categories of the categorical predictors in ascending number of observations.

Continuous Predictors:

1. *Age*: Number of years since the home was last built.
2. *Bedrooms*: Number of bedrooms.
3. *Building Square Feet*: Footage of the living space in thousands of square feet.
4. *Full Baths*: Number of full baths. A full bathroom contains exactly one sink, one bathtub, one shower, and one toilet.
5. *Garage Size*: Number of vehicle-equivalent spaces in the garage.
6. *Half Baths*: Number of half baths. A half bathroom contains only two items (usually a sink and a toilet) in a full bathroom.
7. *Land Acre*: Size of the property lot in acres.
8. *Tract Median Income*: In thousands of dollars, the median income is from the U.S. Census tract where the property belongs.

Unless otherwise stated, please round your numeric answers to the fourth decimal place.

(a) (20 points) Use the Backward Selection method to train a linear regression model. The removal threshold is 0.05. What predictors does the method retain? Please provide us with the selection summary table.

(b) (10 points) Show a histogram and a horizontal box-plot of the leverage values. The common bin-width is 0.05 for the histogram.

(c) (10 points) Generate the vertical box plots of the simple, standardized, deleted, and studentized residuals. We preferred to have the four graphs in a single chart frame.

(d) (10 points) What influential observations do you identify? What is your threshold for high-leverage observations? What are your thresholds for outliers? In your opinion, why are these observations influential?

Question 2 (50 points)

In insurance ratemaking, the term Frequency is defined as the number of claims divided by the exposure of a policy. Exposure is the duration of the policy in a calendar year. Using the **claim_history.csv**, we will train a Poisson regression model to study how policy attributes affect Frequency. Since Frequency values are fractional, it is challenging to find an appropriate distribution. Therefore, we will take the number of claims as our response variable and assume the exposure values are fixed and non-stochastic. The model has the following specifications.

- Response Variable: CLM_COUNT
- Distribution: Poisson
- Link Function: Natural logarithm
- Offset Variable: Natural logarithm of EXPOSURE
- Categorical Predictors: CAR_TYPE, CAR_USE, EDUCATION, GENDER, MSTATUS, PARENT1, RED_CAR, REVOKED, and URBANICITY. **Reorder the categories of each predictor in ascending order of the number of observations**.
- Interval Predictors: AGE, BLUEBOOK, CAR_AGE, HOME_VAL, HOMEKIDS, INCOME, YOJ, KIDSDRIV, MVR_PTS, TIF, and TRAVTIME. **Please divide BLUEBOOK, HOME_VAL, and INCOME by 1000 before training the model**.
- The model always includes the Intercept term.

We only use complete observations with non-missing positive exposure for training. Therefore, we will exclude a record if missing values (i.e., NaN) exist in any predictors or the response variable.

Unless otherwise stated, please round your numeric answers to the fourth decimal place.

(a) (20 points) Use the Forward Selection method to train a Poisson regression model. The entry threshold is 0.05.? What predictors does the method select? Please provide me with the selection summary table.

(b) (10 points). Show a table of all parameters (including the aliased ones) of your final model. Besides the parameter estimates, we also include the standard errors, the 95% asymptotic confidence intervals, and the exponentiated parameter estimates. Conventionally, aliased parameters have zero standard errors and confidence intervals.

(c) (10 points). Calculate the Root Mean Squared Error, the Relative Error, the Pearson correlation between the observed and the predicted Number of Claims, and the Distance correlation between the observed and predicted Number of Claims for your final model in part (b). What are these values?

(d) (10 points). Plot separately the Pearson and Deviance residuals versus the observed number of claims. Please comment on how well your model fits the observations.