STA 108 Project 2

Finding the Best Multiple Linear Regression Model:

Predicting the Number of Active Physicians in a County with Multiple Predictors

https://www.google.com/url?sa=i&url=https%3A%2F%2Fpatientengagementhit.com%2Ffeatures%2Fpreventing-physician-burnout-from-impacting-the-patient-experience&psig=AOvVaw0wDk98EzYGvv_-4JlP-tnB&ust=1685843664838000&source=images&cd=vfe&ved=0CAQQjB1qFwoTCMCw5Zj_pf8CFQAAAAAdAAAAABAJ

Jasper Dong
Allison Peng

Amy Kim
STA 108

I) Introduction:

The dataset, CDI2, consists of numerical data that consists of 7 variables, with a sample size of 440 observations. Population (X_1) is the estimated total population, Income (X_2) is the total personal income in dollars, Physician is the number of professionally active non-federal physicians, Bed (X_3) is the total number of beds, cribs, and bassinets, Area (X_4) is the land area in square miles, Senior (X_5) is the percent of population aged 65 years old or older, Crime (X_6) is the total number of serious crimes, and. Our goal is to predict the number of active physicians in a county (Y) using a multiple linear regression model. With 6 variables available to predict Y, we will determine which variables are the most significant to build the best model. By comparing each model, we will determine which multiple regression model is the best to predict the number of physicians.

II) Summary:

We first conduct exploratory data analysis to inspect the individual data types of each variable, as well as the initial relationship between each predictor variable and the response variable. We observed five number summaries, means, and standard deviation values. We observe high standard deviations for each variable except for Senior, indicating that values are generally more spread out away from the mean. We also observe extremely high maximum data points compared to the third quartile of each variable, indicating the presence of possible outliers in our dataset.

We also analyze the relationship between each variable with one another by use of a correlation plot and a correlation matrix. From both the correlation plot and the correlation matrix, we see that there is a strong positive linear correlation between response Physician and predictors Population, Bed, Crime, and Income, indicating that these predictors could be the best to fit a multiple linear regression model. However, we also observe strong correlation between the predictor variables, indicating that there may be high multicollinearity present, which could make interpretation of coefficients used in the regression model more difficult.

III) Variable Selection:

We will be using base multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \qquad i = 1 \ldots n$$

We can use the extra sum of squares to determine the coefficients of partial determination to measure the effect of an added predictor variable in addition to other variables (in our case, the base model). Looking at the table with all of the coefficients of partial determination, we see that adding the variable Bed is the best for the multiple linear regression model, as the highest was $Y^2_{3|1,2} = 0.554$. Furthermore, a General Linear F Test to test the hypotheses H_0: $\beta_3 = 0$ vs. H_A: $\beta_3 \neq 0$, with a significance level of $\alpha = 0.0002$. Thus, we reject the null hypothesis and we conclude that the full model, or the model with the predictor Bed is a better fit.

IV) Model Comparison and Fit:

We are given two proposed models:

Proposed Model 1) $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 X_5 + \epsilon \qquad i = 1 \ldots n$

Proposed Model 2) $Y_i = \beta_0 + \beta_1 (X_1/X_4) + \beta_2 X_2 + \beta_3 X_3 + \epsilon \qquad i = 1 \ldots n$

Looking at the R^2, or coefficient of determination values for each model, we can see that Model 1 has a higher R^2 value of 0.955. Since our number of regressors is the same in both models, we do not need to use the R^2_{adj} value to compare. Thus, we will continue and inspect Model 1 further by performing diagnostics.

V) Model Diagnostics:

Using the Proposed Model 1, with Population, Income, and Bed as predictors for our best model, we performed model diagnostics to see if the assumptions of the Normal linear regression hold, as well as detecting the presence of outliers and high leverage points in our dataset. The assumptions we test for are:

1) Error terms are independent

2) Error terms are normally distributed
3) Error terms have constant variance (homoscedasticity)

When assessing independence of error terms, we created a residual index plot, and observed a pattern present in the plotted values: as index increases, errors become more centered around 0. This indicates that the error terms may not be independent of each other.

When assessing normality of error terms, we created a Normal Q-Q plot, and observed that plotted values do not follow the straight line. Although error terms seem to be symmetrical about the center, plotted values show greater deviations at the ends. This, in addition to the small Shapiro-Wilks p-value, suggests a non-Normal distribution.

When assessing constant variance of error terms, we created a plot of residuals vs. fitted values, and observed that residuals seem to cluster around smaller fitted values, before becoming less frequent as fitted values become larger. We also observe the presence of possible outliers in this plot, as there seem to be a few points that deviate from the mean. In addition, the small Fligner-Killeen Test p-value indicates there is not constant variance.

To detect for possible outliers, we looked for any studentized residuals greater than 3 and high leverage points. We found 12 possible outliers in our dataset with the studentized residuals and 55 possible high leverage points. We will be considering the studentized residuals as our outliers as it removes a lesser proportion of the dataset, at 2.727%.

We can also assess if the model has multicollinearity by looking at the VIFs, or variance inflation factors. The variables Population and Income have VIFs that are higher than 10, thus the Population and Income variable are likely correlated with other predictor variables in the model. A solution would be to remove the variables from the model.

Based on our diagnostics, we conclude that the assumptions of the Normal linear regression do not hold for our chosen model. However, we will continue to use this model for interpretation and prediction.

VI) Interpretation:

When estimated total population increases by 1 unit, we expect the number of physicians to decrease by -0.002 on average, holding all other predictor variables constant.

When total personal income increases by 1 unit, we expect the number of physicians to increase by 0.138 on average, holding all other predictor variables constant.

When total number of beds, cribs, and bassinets increases by 1 unit, we expect the number of physicians to increase by 0.487 on average, holding all other predictor variables constant.

We do not interpret our intercept of -89.105, as in reality, it would be impossible to have a negative number of physicians.

The R^2 value of 0.955 indicates that 95.5% is the proportionate reduction of total variation in Y associated with the use of the set of X variables, Population, Income, and Bed. The partial coefficient of determination $Y^2_{3|1,2} = 0.554$ indicates the proportion of decrease in SSE when the X_3 variable is added to the model with X_1 and X_2.

We are 95% confident that the estimated coefficient for population is between (-0.002, -0.001), the estimated coefficient for income is between (0.121, 0.155), and the estimated coefficient for bed is between (0.429, 0.544). Since all coefficients do not include 0, the estimates are significant.

VII) Prediction

$$\hat{Y}_i = -89.105 - 0.002X_1 + 0.138X_2 + 0.487X_5 \text{ , } X_1 = 394000, X_2 = 8500, X_5 = 300$$

Using the estimated coefficients, we found the predicted value for physicians to be 509.559, or 509 physicians.

VIII) Conclusion:

Based on our findings, we found the multiple linear regression model between response variable Physician and predictor variables Population, Income, and Bed to be our statistically best model, with its high R^2 compared to other models.

However, limitations of our model include strong multicollinearity, as well as the presence of outliers and high leverage points, that could make interpretation of our model difficult.