NC STATE UNIVERSITY

BUS 351 Introduction to Business Analytics

Regression Analysis of Stroke Death Rates in North Carolina

According to the Internet Stroke Center, stroke is the third leading cause of death and the leading cause of serious, long-term disability in the United States, resulting in annual deaths of more than 140,000 individuals. In this project, you will conduct a study to determine what specific factors have an effect on stroke death rates in North Carolina. You will use descriptive statistics and regression analysis to complete this task using R. You will use the NCStrokeDeaths.csv dataset on Moodle (in Project 3 Resources folder) to analyze the stroke death rate per 100,000 during the timeframe of 2014-2016 for individuals living in all counties of North Carolina (NC), aged 35 years and older. These data came from the Centers for Disease Control (CDC) website https://nccd.cdc.gov/DHDSPAtlas/Reports.aspx and include all races, ethnicities, and genders. These specific variables were selected:

- NC County identifier and name;
- Stroke death rate per 100,000 (*Stroke_Death_Rate_Per_100000*). (For example, Wake County had a stroke death rate of 73.5 deaths per 100,000 individuals from 2014-2016, while Edgecomb County had a stroke death rate of 180.7 deaths per 100,000 individuals.);
- Risk factors:
    - Diagnosed Diabetes Percentage (*dm_prev_adj*);
    - Obesity Percentage (*ob_prev_adj*);
    - Leisure Time-Physical Inactivity Percentage;
- Social and economic data:
    - Education (Less than High School %; Less than College %);
    - Female-Headed Households % (female_hd);
    - Food stamp/SNAP recipients %;
    - Median Home Value ($) (home_val);
    - Median Household Income ($) (income);
    - Income Inequality, Poverty %, and Unemployment Rate;
- Demographic data: American Indian/Alaska Native %, Asian/Native Hawaiian/Other Pacific Islander %, Black %, White %, Hispanic/Latino %, and Aged 65 or Older %;
- Urban-Rural status; and
- Heath Care Total Costs Per Capita ($).

(30 points) Descriptive Statistics:

Use descriptive statistics to determine if Diagnosed Diabetes Percentage (*dm_prev_adj*) had an effect on Stroke death rate per 100,000 individuals (*Stroke_Death_Rate_Per_100000*) in North Carolina.

1. If you haven't already done so, download the Project 3 Moodle dataset to your Desktop as **NCStrokeDeaths.csv**.
2. Create some descriptive statistics on these two variables. Include various descriptive measures (get the 5 number summary using the ***summary*** function). Use histograms (at least one) and analyze any outliers you find in the data.
3. Create a scatter plot. Does there appear to be a relationship between the two variables?
4. Create a correlation matrix to determine which variables (including Diagnosed Diabetes Percentage) are most correlated to stroke death rate per 100,000 individuals in NC (numeric variables only).
    - Determine which possible independent variables are most correlated to the dependent variable (stroke death rate).
    - Determine possible multicollinearity between independent variables.
    - Based on this assessment, determine possible models to test which do not include multicollinearity. Which variables could you eliminate as additional independent variables due to possible multicollinearity?

Regression Models:

(30 points) Model 1 (SLR): For the first regression model, run a simple linear regression (SLR) to determine if diagnosed diabetes % predicts stroke death rate in NC. Run the regression, use a 0.05 level of significance to perform statistical inference, analyze the results by recording the following:

- Adjusted R-Squared %.
- Hypothesis associated with the F-test, and what is your conclusion.
- P-value of the independent variable *Diagnosed Diabetes Percentage* (dm_prev_adj). Is it statistically significant?
- Type the actual equation and explain the equation in everyday language using the coefficients.

(30 points) Model 2 (MLR): Based on your correlation matrix assessment, run a MLR model. Be sure to record and explain each of the following:

- Find two variables that could together better predict stroke death rate than just diagnosed diabetes % alone (with collinearity to the dependent variable but without multicollinearity). Create a MLR with those two variables as independent variables.

- Perform statistical inference, analyze the results. Be sure to assess the same regression statistics as Model 1. Is this a better model? Did the R- squared % increase with the model being statistically significant? Use a 0.05 level of significance.

(10 points) Using your best regression equation (SLR or MLR), what is your stroke death rate prediction for a North Carolina County whose diagnosed diabetes % is 7.4, whose median home value is 85.5, Education-Less than high school % of 14.3, and whose female headed household % is 23.9? (Use only the variables present in your best model and disregard the rest)