## TMATH 410: Regression Modeling With Applications

**Steve Mwangi , Tenki Ko**

Introduction:

In our study, we will be using similar variables as predictors. Our main objectives are going to be:

1. Assess the relationship between mortality rate and the type of cancer.
2. Determine whether this relationship is affected by a person’s age and gender.
3. Calculate whether there is any correlation between the household income in a county and the mortality rate of patients suffering from different types of cancer.
4. Gauge how much the total number of in situ tumors and malignant tumors affect the mortality of cancer patients.

In covering these objectives and conducting this study we hope to determine how significant the relationship is between the predictors and the explanatory variable (mortality rate), when it comes to cancer type, age, among other factors. Our data source is going to be the Surveillance, Epidemiology, and End Results Program (SEERS).

Methods:

1. The steps you intend to take to estimate your regression model a. The next steps would be to estimate the coefficients for the regression model we are going to use b. Conduct hypotheses tests for the model and testing if the regression coefficients are equal to 0

2. Identification of the response and explanatory variable(s)

a. Our response variables is: I. Mortality Rate

b. Our Explanatory variables are:

I. Cancer Type

II. AgeDiagnosed

III. MedHouInc

IV. Gender

V. INSvsMALI

3. Steps you intend to take to assess your regression model (e.g., residual analysis, etc.) a. Regression diagnostics using R to assess whether we are violating some model assumptions. We might have to standardize not just the residuals but also the mortality rate to be a percentage value. 4. Any transformations you anticipate might need to be performed. From the plots we have so far, we doubt that there will be any need for transformations.

Results:

Scatterplots with the response variable on the y-axis and each explanatory variable on the xaxis.

A close up of a map

Description automatically generated

The obvious patterns that might complicate our data is the X binary (1 or 2) values in the scatterplot of the Mortality rate vs. the In-situ or Malignant Tumors. Due to the absence of values other than 1 and 2, the predictor insitu or malignant could be undermined or overestimated in terms of its influence on our data. This could mean that this variable is deficient in some regards when it comes to using it in the model. Even though the three scatterplots point to a case where the linearity assumption might not necessarily be violated, there could still be underlying issues that cannot be determined from the scatterplots.

Scatterplots with the explanatory variables plotted against each other:

A screenshot of a computer

Description automatically generated

The deficient nature of the predictor variable insitu\_malignant is still echoed in these plots. It can be seen that some collinearity exists as seen in the AgeDiagnosed vs MedHouInc plot, though it could be said to be marginal, since the fitted line is not too steep.

A boxplot comparing the distribution of your data among the levels of the explanatory variable. Describe the distributions and compare them across levels of your explanatory variable:

A screenshot of a cell phone

Description automatically generated

From the boxplots above, we see that prostate cancer varies greatly compare to the other types of cancer, and stomach cancer seems to be the least varied, due to how short its boxplot height is compared to those of the other 3 cancer types. Additionally, we see that for Breast, Pancreatic and Prostate Cancer, the upper part of the boxplots is longer, pointing to cases where half of the proportion of people with this these cancers, end up having mortality rate that is higher from that of the other half(right skewed). For the gender boxplots, we see that there is greater variability in the mortality rate of female cancer patients compared to male cancer patients, even though there are two outliers in the upper portion of the male boxplot. Additionally, the female boxplot corresponds to the boxplot of patients who suffer from Breast Cancer, which could be attributable to the fact that most people who suffer from Breast Cancer are female. When it comes to the male boxplot, it corresponds to the Pancreatic Cancer Boxplot, which would seem counterintuitive, considering that prostate cancer mainly afflicts males. Upon investigating further, it has been noted that Pancreatic cancer could mainly affect males due to higher tobacco use in males than in females, which apparently exacerbates its effects (American Cancer Society, 2020).

Numerical summaries of each of the variables:

A screenshot of a cell phone

Description automatically generated

Discussion:

One of the peculiarities was in getting the mortality rate values. When we collected them, we had to use the Mortality Rate associated with each type of Cancer which was determined by two values (Gender and Age). Instead of formatting the values by dividing by 100000, we chose to keep the values as they are. Additionally, due to the binary nature of insitu or malignant—being 1 or 2, R summarized the entries in this column of the data frame as quantitative, hence having a mean of 1.088 even though the other values for median, minimum or maximum are either 1 or 2. The value was meant as a code to signify number of tumors, but in our model the values limited since they are either 1 or 2.

“Malignant and in situ neoplasms, including malignant CNS neoplasms, are assigned codes 00 through 35. If a patient has only one primary malignant or in situ neoplasm, the sequence number assigned is 00. If a patient has multiple primary neoplasms during a lifetime, the sequence number for the first tumor is 01, the sequence number for the second primary tumor is 02, and so forth” (National Cancer Institute, 2020).

References

Monica, K. (2018, July 30). *EHR Data Helps to Predict Short-Term Mortality for Cancer Patients*. Retrieved from ehrintelligence.com: https://ehrintelligence.com/news/ehr-data-helps-to-predict-short-term-mortality-for-cancer-patients

Ni, D., Xiao, Z., Zhong, B., & Feng, X. (2018, November 9). Multiple Human-Behaviour Indicators for Predicting Lung Cancer Mortality with Support Vector Machine. *Scientific Reports*. doi:https://doi.org/10.1038/s41598-018-34945-z

Parikh, R. B., Manz, C., Chivers, C., Regli, S. H., Braun, J., Draugelis, M. E., . . . Patel, M. S. (2019, October 2). Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. *JAMA Netw Open*.