Model Card: Cancer Mortality Rates by United States Counties

Model Details

This model was developed by students at University of Montreal Mila Lab to further explore the validity of linear models using Ordinary Least Squares also known as linear regression. The goal being to develop our intuition around linear models and explore its limitations with real world data.

Model Date

March 2021

Model Type

Looking at the distribution of our interest variable (TARGET_deathRate), it was imperative to find out if our model could adequately predict unseen data. Furthermore, RFE (Recursive Feature Elimination) was used to select the features based on their p-value. Based on our initial assumptions, we began by stating the data had a predictable linearity which allowed us to use OLS. It is important to realize that such a model will not perform well if the data is non-linear.

Model Version

So far we have produced a single version but we are keeping in mind that further additions could be added in order to increase the validity of our model. For future reference if the dataset increases in size, we could decide to keep more attributes. We could also apply more complex transformations to make our data more predictable.

Model Use

Intended Use

Our model was initially created to help us develop a better intuition around linear models. Using real data, we were able to see first-hand how to deal with missing values, discard irrelevant variables and improve our understanding regarding the validity of our model using k-fold cross validation. Although simple, this technique has proven to be sufficiently robust to obtain a better picture of the variance of our errors. Do we have the right model or are we leaving it to chance? Among others, these were questions that piqued our curiosity and we were able to answer.

Primary intended uses

Our objective as well as our hope was to write and explain our model in a way that makes it accessible to a large crowd of individuals who seek to start understanding OLS. We do not suggest our model should be use for anything else than academic exploration. Unfortunately, the small sample size of our data makes our model lose credibility and therefore could contain a large variance.

Out-of-Scope Use Cases

It would be considered out-of-scope if the model was used to predict real data in a professional environment. Furthermore, the data that we used to train and test our model was only for American counties. Therefore, using this model to try to forecast the cancer death rate of other countries would be irresponsible. Factors such as poor diet, exercise and preexisting gene mutations could drastically skew the data and forecast poor results in the new dataset.

Data

The model was trained on publicly available data that was compiled and extracted from different government websites. You can find the dataset and different databases used in the Sources section.

Data Mission Statement

Our expectation with this dataset was to explore real world data and see what transformations could be beneficial for increasing the validity of our model. By properly evaluating the linearity of our data, we are better able to generalize and diminish our variance on unseen data. The data itself was collected and placed together by the website data.world. Therefore, all credit pertaining to the assembling of the database should go to them. A future model could be deployed by scrapping multiple websites and combining data from past decades in order to obtain a better idea of cancer mortality rates.

Performance and Limitations

Performance

Our model performance was obtained using the R^2 measure. This statistic is especially used in linear models as it adequately indicates how much the chosen model is able to explain the variation in our variable of interest.

Limitations

This model contains several limitations one of which is the small size of the dataset that was used during the training. As previously stated, we do not support the use of this model under a commercial/professional context. The model would most certainly fail to generalize on data that is from another country than the US. Not to mention that certain counties contain much more data than others, ultimately skewing the data one way or another. We doubt it will be able to perform well with unseen values. Additionally, there are no timestamps embedded in our attributes which could make our model lose value over time.

Bias and Fairness

Finally, as with all linear models, outliers will drastically modify our weights for each variable. Maybe we should have been more careful in selecting which data to keep. It is not an unreasonable consideration to make. The very first model we had built included as one of the variables the geographic location of the observation. We noticed the OLS method ended up keeping only a handful of states. This made us realized it was best to remove the entire variable as this was creating noise in our predicted values. By doing so, we were able to better generalize with our validation sets. At some point we also noticed that certain variables had a significant number of missing values. We then decided to look at the distribution of those observations and thought it was fair to replace them with the mean value as there were not many outliers present. The final R^2 measure that we obtained was just under 80% which makes us wonder how well it will continue to perform with more data. As all machine learning engineers end up asking themselves, we fear that our model might be too close to the noise present in the data and will ultimately be overfitting. Ultimately failing to generalize adequately if the new data contains many outliers.

Sources

- https://data.world/exercises/linear-regression-exercise-1/workspace/file?filename=cancer_reg.csv
- census.gov
- clinicaltrials.gov
- cancer.gov