Several State Health Statistics and Deaths from Diabetes and Heart Disease: Prediction Through Linear Regression and a Neural Net

Elizabeth Javor

1 List of Statistics and Overview

This project uses linear regression and a neural net implemented in Pytorch (with ReLU providing nonlinearization) to predict deaths per million from diabetes and heart disease for each state and year, for years 2011-2017. Data is preprocessed in SQLite.

The features used in the model draw from a set of statistics publicly available at data.gov; these statistics (numbered starting at 0 to align with the code) are:

0. Percent of adults aged 18 years and older who have an overweight classification
1. Percent of adults aged 18 years and older who have obesity
2. Percent of adults who engage in no leisure-time physical activity
3. Percent of adults who engage in muscle-strengthening activities on 2 or more days a week
4. Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic activity
5. Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic physical activity and engage in muscle-strengthening activities on 2 or more days a week
6. Percent of adults who achieve at least 300 minutes a week of moderate-intensity aerobic activity or 150 minutes a week of vigorous activity
7. Percent of adults who report consuming fruit less than one time daily
8. Percent of adults who report consuming vegetables less than one time daily

Figure 1: Health behavior questions used as features

We seek to use some of statistics as features to predict the number of deaths due to diabetes and heart disease in each state and in each year, and in our analysis of such also examine the effects highly correlated features have on a linear model.

The purpose of this document is to provide a short explanation as to the results of my personal project which can be found at https://github.com/lrj8881/healthstatpredictions. For details on running the code and how the program itself works, please see the README file on GitHub.

2 Preprocessing and Data Exploration

Data preprocessing was done through SQLite, in an attempt to demonstrate the author's newfound ability to query in SQL. The statistics from Figure 1 were pulled from [2], and data on deaths by heart disease and diabetes from [1]. Population data from all states was pulled from [3] (note: the line

"State,Year,Population" was added at the top of the .csv from that source, to match the formatting of the other 2 datasets. This line appears in the csv stored on github, but does not appear in the original source [3]).

We sought to predict the number of deaths per million people from each cause (the causes being, diabetes and heart disease) by using the statistics in Figure 1, with deaths *per million* being chosen simply for readability. As a preliminary measure, scatterplots were created to examine which features were most strongly correlated with deaths per million from each cause:

*Note that the horizontal axis corresponds to each question in 1; above a shorthand is used.

For "Overweight", "No fruit," and "No vegetable" (i.e. 0, 7 and 8 in Figure 1), there seems to be little if any correlation to deaths from either cause. Going forward, these statistics are omitted (though still testable, for curiosity's sake, in the codebase). We also note that features 0, 1, and 2 ("Overweight", "Obese", "No physical activity") were the only features that had data for all states

and all years, around 350 datapoints. When features 4-6 were used, even in conjunction with 1, 2 and 3, only those points where all features were present were included in the dataset (this still made for around 200 datapoints).

3 Model training and tuning of hyperparameters

Using grid search, we explored the possible learning rates 0.01, 0.001, and 0.0001, possible numbers of epochs of 500, 1000, and 2000, and possible training/validation data splits of 75/25, 80/20, and 90/10. Grid search was run in the case of using all viable features (1,2,3,4,5, and 6 in Figure 1). This produced the following optimal parameters:

Cause	Model Type	Train/Val Split	Epochs	Learning Rate	Avg. Loss
Diabetes	Linear Regression	75/25	1000	0.01	0.383
	Neural Net	80/20	2000	0.01	0.339
Heart Disease	Linear Regression	75/25	500	0.01	0.464
	Neural Net	75/25	2000	0.01	0.242

Table 1: Best model configurations and performance for predicting diabetes and heart disease deaths.

Note that the neural net improved performance of the model in both cases, significantly for heart disease. Below are the graphs of actual vs. predicted DPM for each model, with the red line representing a perfect (100% correct) model.

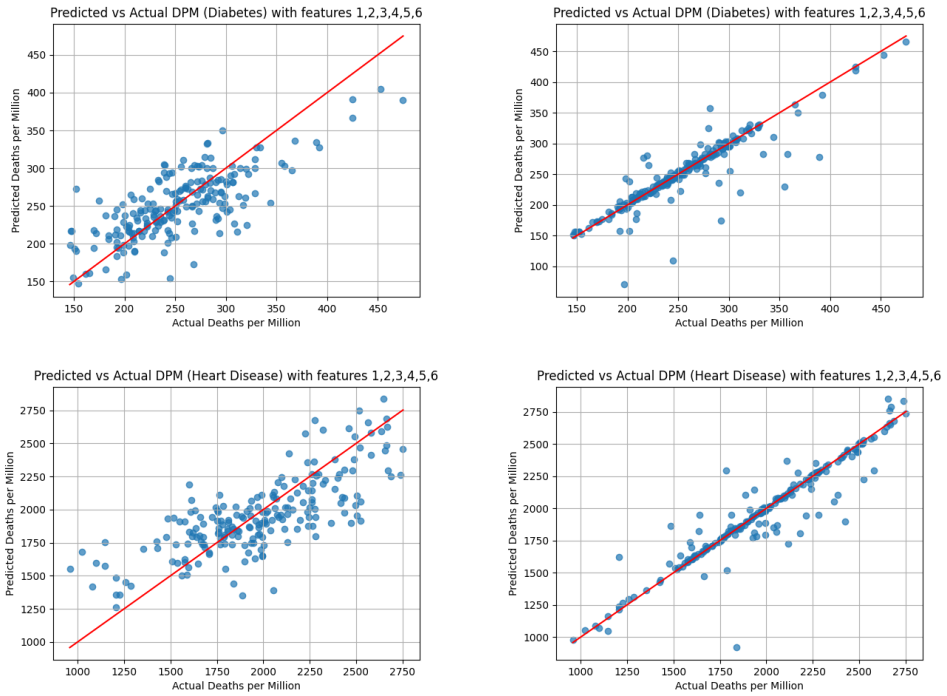
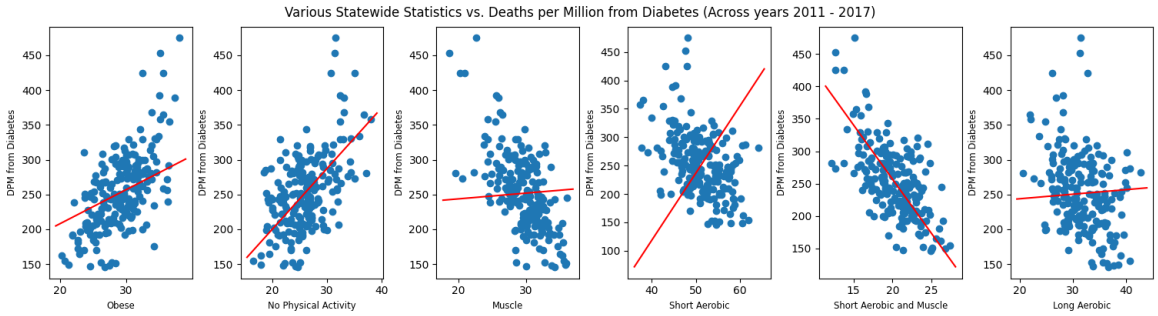


Figure 2: Left: Linear model. Right: Neural net. Note the improvement in accuracy when using the NN.

Grid search was not performed for each feature separately or every possible combination of features in the interest of time (as this would involve 252 grid searches). The pre-set "optimal" parameters, as of now, are the ones shown above, for any combination of features. However, running grid search on a new combination of features (i.e. any subset of the features 1,2,3,4,5, and 6, besides all of them) will automatically save a preset. (The code can be run without a grid search, and the parameters used are those saved in an SQLite database of optimal parameters).

4 Effects of Correlated Features

The optimal model from grid searching (diabetes, linear model) produced these graphs, which include the scatterplots from Section 2 and the prediction line from the model:



(This graph is the result of using the -g1 tag in code).

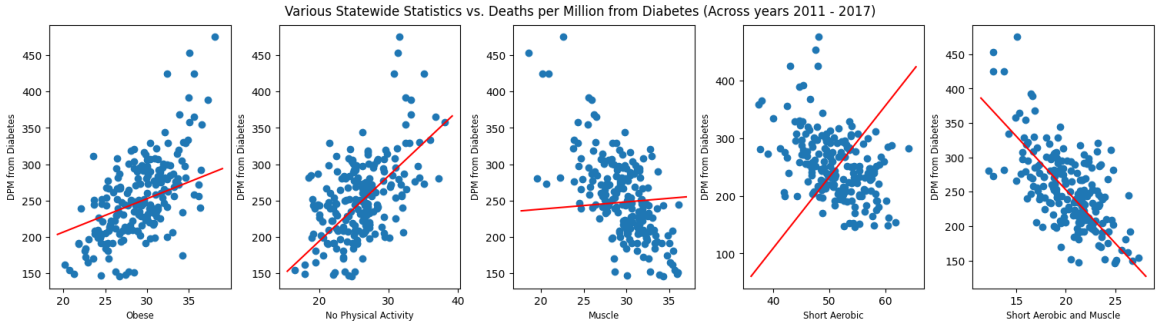
Note that uses the cause of death "Diabetes" and the model type "linear" is simply an example; this discussion pertains to heart disease and the neural net as well.

These graphs show the effects of varying a singular feature while holding the others at 0 (which, since the data is normalized, is the mean). Considering the high correlation between the features, this is not a sensical assumption: we cannot increase the percentage of adults who perform muscle-strengthening activities while holding the percentage of adults who perform no leisure-time physical activity constant.

The correlation between features allows the model to pay little heed to some variables while relying more heavily on others that encode similar information. For the model used to generate the above graphs, the weight tensor is (0.3377, 0.1594, -0.2544, 0.1210, -0.4246, 0.3966). Here, Feature 5 ("Short Aerobic and Muscle") is weighted strongly, while Features 3 and 4 receive much less emphasis. As feature 5 is simply a composite of Features 3 and 4 (it represents the percentage of individuals who meet both criteria), the model learned that a strong fit for Feature 5 can effectively substitute for separate weightings on 3 and 4 (see Figure 1 for clarity on the exact statistics these features represent).

A similar issue arises with Feature 6 ("Long Aerobic"), which has a poor individual fit. However, feature 6 is a subset of feature 4, and thus it overlaps with 5. Because of this redundancy, Feature 6 adds little new information and introduces multicollinearity. The model's attempt to reconcile these overlapping features leads to unstable or unintuitive weight assignments. This weight redistribution — driven by highly correlated inputs — contributes to confusion in model interpretation and performance.

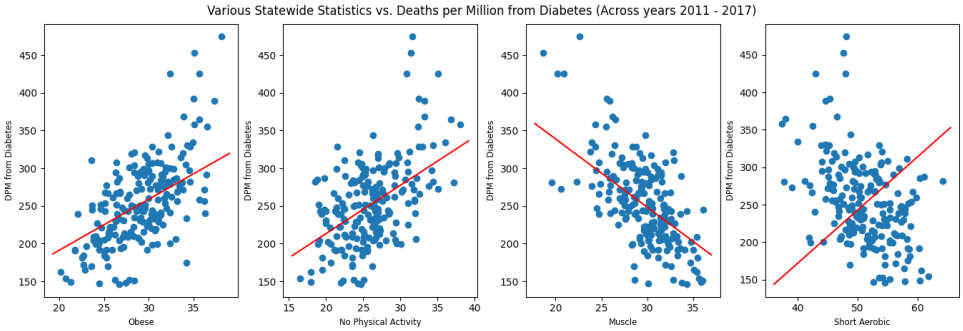
This pattern is further illuminated when feature 6 is removed:



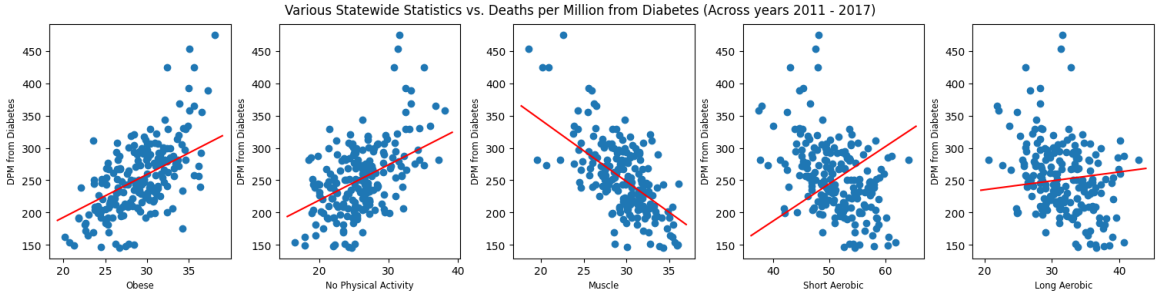
Here, the weight tensor is (0.4105, 0.0337, -0.6183, 0.2418, 0.0830), and the loss achieved is 0.393 (slightly worse than for 6 features).

Note how closely the prediction follows the scatterplot data despite the model not relying on the features 2 and 5, and how the red line deviates significantly from the scatterplot for features 3 and 4. The high correlation between 3,4 and 5 means the model can safely distribute weight across those two features, ignoring 5, and still get a good fit.

Removing feature 5 due to its redundancy (and grid searching models for optimal results) leads to a model with a loss of 0.392:



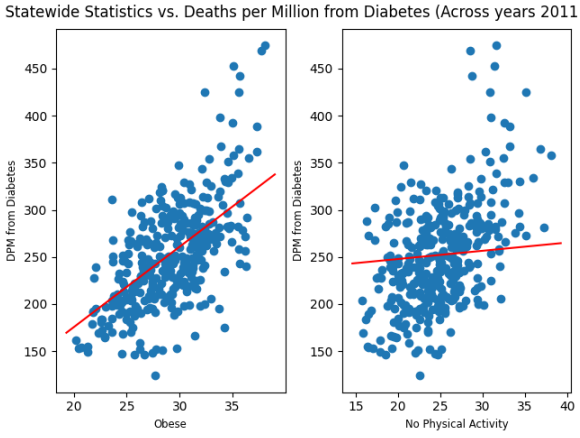
This loss is about the same as the 5-feature model, as expected from the redundancy of the 5th feature. However, performance here is still lacking when compared to the 6-feature model; some information was perhaps lost when we ignored the relationship between long aerobic exercise and diabetes deaths. Trying with features 1,2,3,4 and 6 (and running grid search on some possible models with these features) we get a loss of 0.384:



This is almost the same loss as for all (6) features, and further confirms the unimportance of feature 5.

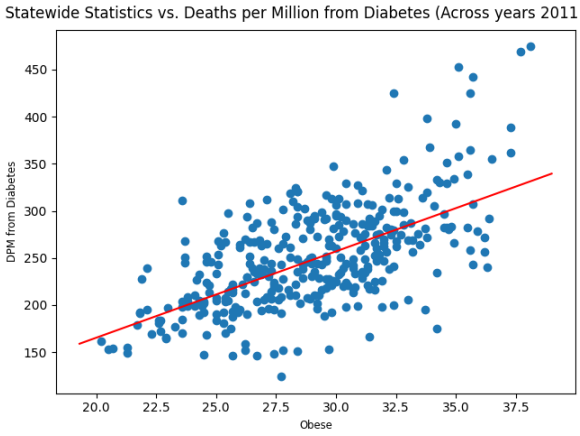
However, there is of course some correlation between these remaining 5 features, leading to the poor fit for features 5 and 6 above.

Simplifying the model further achieves a loss of only 0.379 (when models involving only features 1 and 2 ("Obesity" and "No physical activity")are grid searched for optimal parameters, found to be a 0.9 split, 2000 epochs, and 0.01 learning rate):



This is the lowest loss achieved by any linear model in this document thus far.

In this case, the first feature is weighted much more strongly, as the fit on the scatterplots above might suggest (the weight tensor is (0.5046, 0.1772)). Yet, the second feature still contributes to the success of the linear model; running grid search on feature 1 by itself produces an optimal average loss of 0.420, with 0.9 split, 2000 epochs, 0.01 learning rate, and the following plot:



The loss is of course higher than in the case of all features but the difference is perhaps surprisingly small, demonstrating that obesity itself, without information physical exercise levels, is a strong predictor of diabetes deaths, as one would expect, but any addition of physical activity data does improve prediction power.

5 Future Work

For using real-world health data, we believe the models worked relatively well, but there is definite room for improvement. we used a very simple 3 layer neural network (simply 3 linear layers with ReLU between). There are a myriad of options for layers, activation functions, other model improvements, etc., and given more time, we believe much more accurate predictions could be achieved.

The correlation between model features merits more investigation. Performing grid search on each feature separately and comparing performance, and investigating certain combinations of features, could prove fruitful in studying how such correlated features affect a model's performance and placement of weight.

Datasets:

[1] Centers for Disease Control and Prevention (CDC). *NCHS - Leading Causes of Death: United States.* Available at: https://catalog.data.gov/dataset/nchs-leading-causes-of-death-united-states

[2] Centers for Disease Control and Prevention (CDC). *Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System* Available at: https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system

[3] *Historical Population of the U.S. States*, compiled by JoshData with data from the Federal Reserve Bank of St. Louis, available at: https://github.com/JoshData/historical-state-population-csv