Introduction

Diabetes mellitus is a chronic health condition characterized by abnormalities in processing glucose. Glucose processing involves insulin which is produced in the pancreas. Insulin is a hormone that regulates blood glucose levels by moving glucose from the food we eat into our cells. There are two types of diabetes: Type I and Type II. Individuals with Type I diabetes have immune systems that attack their own insulin-producing pancreatic cells. While, individuals with Type II diabetes either do not produce enough insulin or cannot effectively use it. Over time, diabetes causes health complications that affect various major organ systems and increases the risk for other health conditions, such as heart and kidney disease (CDC, 2023).

Efficient monitoring of the progression of diabetes is important because timely treatment and lifestyle changes can prevent or delay serious health complications. Diabetes is a condition that can change over time, and by staying on top of these changes, people with diabetes can take proactive steps to manage their health.

In this study, two linear statistical models are proposed to predict a quantitative measure of an individual's disease progression one year after baseline. Using these models, medical professionals will be able to input certain variables from an individual's medical history and identify those at risk for diabetes.

Theory of Regression

Linear Regression is used for explaining or modeling the relationship between a dependent variable and one or more independent variables (Faraway, 2014).

Simple Linear Regression:

A population regression model can be estimated using a specified data set.

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_0 is the intercept parameter and β_1 is the slope parameter. x is a variable that is hypothesized to have an impact on y. ϵ represents all random peturbations of our data and any unobserved/unexplained phoenomenon. The simplest regression model is a straight line and is denoted below

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$$

One point of confusion we aim to address is that a linear model requires the β coefficients to enter linearly, but not our variable(s) of interest.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^2 + \epsilon$$

is a valid model to estimate using linear regression. In regression, the β coefficients are found using a quadratic loss function.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \text{ where } \hat{y}_i \text{ is the fitted value}$$

The goal is to find the curve that lies the closest to all data points. The derivation is below.

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$
$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

The normal equations follow

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$
$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Solving this system of equations we get that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{y} y_i (x_i - \bar{x})} = \frac{S_{xy}}{S_{xx}}$$

Medical studies commonly seek the relationship between a dependent variable and multiple independent variable. Let's extend the intution from simple linear regression to multiple linear regression. In multiple linear regression we have

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

We can write this in matrix form.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{22} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$
$$y = X\beta + \epsilon$$

where y is an $n \times 1$ vector of observations, X is an $n \times p$ matrix of observed values for the regressor variables, β is a $p \times 1$ vector of regression coefficients, and ϵ is an $n \times 1$ vectors of random errors. The derivation of the β coefficients follows in a similar manner and it requires the use of matrix algebra. The results of the derivation are below.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$$

For a geometric intution of multiple linear regression. Consider the case where we are regressing against two variables of interest. If simple linear regression seeks to find the line that minimizes the distance across all data points, multiple linear regression finds the plane. Now that we have an understanding of what regression is and how it is implemented. It's crucial that we understand the assumptions that linear regression makes. There are five major assumptions:

1. There is a linear relationship between y and the regressor variables.
2. ϵ is uncorrelated (oftentimes we make a stronger assumption of independence)
3. ϵ is normally distributed
4. ϵ has constant variance (homoscedasticity)
5. ϵ has a mean of zero

In order for a linear regression model to be stable and useful, it must satisfy these assumptions. A stable model is one that doesn't change too much when trained on a different data set.

For any constant $\hat{\beta}_i$, it can be interpreted as the amount that the expected value of the response variable changes given an unit increase of x_i when all other variables are held constant. Hypothesis testing and the distributional results of multiple linear regression are presented below.

The $100(1-\alpha)\%$ confidence interval for β_j, $100(1-\alpha)\%$ confidence interval on the mean response at point x_0, and the $100(1-\alpha)\%$ prediction interval for a future observation can be found in the appendix. A deeper understanding of these results can be found in Linear Models with R by Faraway.

Given that our dataset is adequate, the next step after fitting our model is to validate our assumptions. The first assumption we should validate is the absence of multicollinearity since this causes computational errors. We will then validate the distributional assumptions. Constant variance and linearity can be validated using a diagnostic tool known as a residual plot. Residuals are defined as the difference between our fitted values and the actual values. It is convenient to think of residuals as the observed values for our ϵ term. Therefore, any violations in the residual plot suggests violations of our model assumptions (Faraway, 2014). Normality can be validated using a QQ-Plot. If the ϵ is normally distributed the points will lie on a straight diagonal line. We will find that multicollinearity is the major issue that we will come across in our diabetes dataset.

Multicollinearity occurs when one of the regressor variables is approximately a linear combination of the other columns in our dataset.

$$\sum_{i=1}^{P} t_i X_i = 0$$

where t_i is not all zero. This means that there exists a column in our dataset that is an exact linear combination of one or more of the columns. When this occurs $(X^T X^{-1})$ does not exist. When this is only approximately true we say that multicollinearity is present. This can cause computational errors and our β constants to be inflated. Multi-collinearity can be diagnosed using Variance Inflation Factors (VIFs). For intuition, VIFs, measures how much a regressor coefficient's variance is inflated due to multi-collinearity. VIFs above 5 signal that the variable is strongly correlated with at least one other variable. One solution to this issue is to amputate the variables that are strongly correlated with each other (Faraway 2014).

Another solution is to use a different type of linear statistical model that is insensitive to multi-collinearity. We will utilize ridge regression in our analysis of the dataset.

The ridge estimator $\hat{\beta}_R$ is the solution to

$$(X^T X + kI)\hat{\beta}_R = X^T y \implies \hat{\beta}_R = (X^T X + kI)^{-1} X^T y$$

$k \geqslant 0$ is known as the biasing or regularization parameter and is chosen by the analyst. Notice that ridge regression is a linear transformation of the ordinary least squares estimator $\hat{\beta}$

$$\hat{\beta}_R = (X^T X + kI)^{-1} X^T y = (X^T X + kI)^{-1}(X^T X)\hat{\beta} = Z_k \hat{\beta}$$

By choosing an appropriate value of k we can lower the variance of our estimates. For intuition, that as $\hat{\beta}_R \xrightarrow{k \to 0} \hat{\beta}$

Methods

The dataset used in this study comes from Least Angle Regression by Efron B., Hastie T., Johnstone I., Tibshirani R. All ten baseline predictor measurements were recorded for each patient (Efron & Hastie & Johnstone & Tibshirani, 2004). The ten baseline predictor variables considered are 1) age, 2) sex, 3) body mass index (BMI), 4) average blood pressure (MAP), 5) total serum cholesterol level (TC), 6) serum low-density lipoprotein level (LDL), 7) serum high-density lipoprotein level (HDL), 8) ratio of total serum cholesterol level/serum HDL level (TCH), 9) log of serum triglyceride level (LTG), and 10) serum glucose level (GLU). The sex variable is an indicator varaible that denotes 1 for males and 0 for females. There are 442 diabetes patients included in this dataset (n=442).

Procedures

y (quantitative measure of diabetes progression 1 year after baseline) was first regressed against all variables in the data set. Then, the major assumptions of Linear Regression were validated. First, we will validate the absence of multi-collinearity since this can cause computational issues. Based on our analysis of the Variance Inflation Factors (VIFs), there are several variables whose VIFs are above 5. While, it is not made clear in the paper this data set was taken from whether some of the variables were generated through calculations, it is possible since the definition of $TC = LDL + HDL + 0.2LTG$. From our analysis, it seems that the blood serum measurements are correlated. In this first linear regression model, we will solve this issue of multicollinearity through variable amputation. All blood serum variables except for glucose were amputated , and the resulting matrix has no concerning levels of multicollinearity. The adjusted model regresses against AGE, SEX, BMI, MAP and GLU. Next, lets validate the assumptions of a linear relationship and constant variance.

From the QQ-plot there is a random scatter of points. Therefore, the errors are normally distributed. The residual plot shows no evidence of non-linearity and the assumption of constant variance is reasonable as well.

Next, The data was split into a test and training set. 10-fold cross validation was utilized to see how well the model performs on unseen observations. Abu-Mostafa, Magdon-Ismail, and Lin outline the procedure for 10-fold cross validation as follows.

1. Randomly split the data set into 10 folds

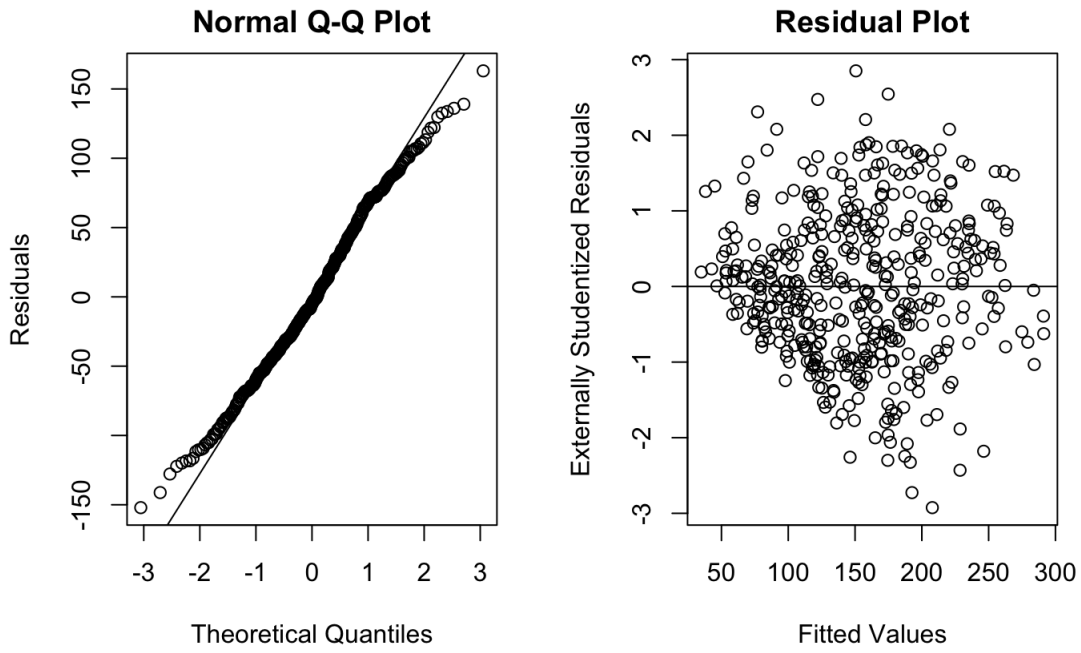

Figure 1: (a) Normal QQ-Plot, (b) Residual Plot

2. Choose one fold for validation
3. The remaining 9 folds will be used as the training set
4. Train the model on the training set
5. Compute the error of the model on the validation set
6. Repeat until all folds have been used as the validation set.
7. The final error will be the average of all the errors on the each validation set.

In cross validation the Root Mean Squared Error (RMSE) was computed, which is a measure of the average deviation of our fitted model from an unseen data point.

Next, the final model uses ridge regression because there exist many correlated predictors among the dataset especially among the blood serum measurements. Ridge regression can address these issues without having to amputate the correlated variables.

Results

The multiple linear regression model with variable amputation is as follows.

$$\mathsf{y} = 152.133 + 1.349 \cdot \mathsf{AGE} - 129.754 \cdot \mathsf{SEX} + 728.942 \cdot \mathsf{BMI} + 372.594 \cdot \mathsf{MAP} + 217.025 \cdot \mathsf{GLU}$$

This first model produced a RMSE of 59.232 and a R^2 of 0.4157. Based on these values this model would not function in practice. On average the prediction of an individual's quantitative progression one year after baseline would deviate by 59.232 and the model only explains 41.57% of the variation in our data.

The ridge regression model is as follows.

$$\begin{aligned} \mathsf{y} = 149.99 - & 11.331 \cdot \mathsf{AGE} - 156.91 \cdot \mathsf{SEX} + 374.450 \cdot \mathsf{BMI} + 264.900 \cdot \mathsf{MAP} - 31.969 \cdot \mathsf{TC} \\ & - 66.897 \cdot \mathsf{LDL} - 174.012 \cdot \mathsf{HDL} + 123.972 \cdot \mathsf{TCH} + 307.686 \cdot \mathsf{LTG} + 134.48 \cdot \mathsf{GLU} \end{aligned}$$

This model produced a RMSE of 54.594 and a R^2 of 0.5050. This model would also not be fit to function in practice. Based on the analysis of this dataset, it's difficult to produce a functional model. The ridge regression model contradicts previous medical experience in that it suggests that an older individual would have a slower progression of diabetes one year after baseline. From this data set it is unclear if confonding variables were controlled during the one year between base line measurements and the quantative measure of diabetes progression was taken.

Conclusions

The purpose of this research was to propose two statistical learning models that predict a measure of diabetes progression using regression techniques. The hope was that a statistical model that physicians could utilize to provide earlier intervention to patients could be developed. However the models generated from this dataset don't generalize well to unseen observations. This is most likely due to the quality of the data set. The data set is on the smaller end with only 442 samples. Based on other studies of diabetes predictive models, there have been models built with high accuracy using other statistical learning approaches. I propose that further research could be done on a different dataset.

Works Cited

Faraway, J. L. (2014) *Linear Models with R.* CRC Press

Montgomery D. C., Peck A. E., Vining G. G. (2012). *Introduction to Linear Regression Analysis.* Wiley

CDC. (April 24, 2023). *What is diabetes?.* cdc.gov. https://www.cdc.gov/diabetes/basics/diabetes.html.

Efron B., Hastie T., Johnstone I., Tibshirani R. (2004). Least Angle Regression. *The Annals of Statistics*, *2*(32), 407-499. https://tibshirani.su.domains/ftp/lars.pdf

Abu-Mostafa Y. S., Magdon-Ismail M., Lin H. (2012). *Learning From Data.*

Appendix

A $100(1-\alpha)\%$ confidence interval for β_j is given by

$$\hat{\beta}_j - t_{\alpha/2,n-p} se(\hat{\beta}_j) \leqslant \beta_j \leqslant \hat{\beta}_j + t_{\alpha/2,n-p} se(\hat{\beta}_j) \text{ where } se(\hat{\beta}_j) =$$

A $100(1-\alpha)\%$ confidence interval on the mean response at point x_0 is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0} \leqslant \mathbb{E}(y \mid x_0) \leqslant \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0}$$

A $100(1-\alpha)\%$ prediction interval for a future observation is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 (1 + x_0^T (X^T X)^{-1} x_0)} \leqslant y_0 \leqslant \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 (1 + x_0^T (X^T X)^{-1} x_0)}$$