Understanding Diabetes Progression

1. INTRODUCTION

According to the World Health Organization's (WHO) September 2022's Key Facts on Diabetes, the number of people with Diabetes has increased 74% between 1980 and 2014. WHO also states that "Diabetes is a major cause of blindness, kidney failure, heart attacks, strokes and lower limb amputation." For health insurance companies and the medical professions, it is important to know the factors that impact such a devastating disease and how to expand the number of patients who successfully live with well-controlled diabetes. That is why the chosen dataset has us examining which factors are most important in their contribution to Diabetes *(from Efron, B., Hastie, T., Johnstone, J., and Tibshirani, R. (2004). Least Angle Regression. Annals of Statistics, 32, 407-499).*

This dataset contains ten baseline variables that are possible predictors to whether or not the disease is worse (greater than 200) or better (less than 200) after a one year baseline. The number 200 was chosen because it is the accepted "line in the sand" in which blood sugars higher than 200 on the Glucose Tolerance Test or the Random Blood Sugar Test indicate the patient is diabetic (*https://www.cdc.gov/diabetes/basics/getting-tested.html*). The variables included in the dataset include the patient's age, Gender, Body Mass Index (BMI), and Blood Pressure (BP). It also includes test results found on the standard Lipid Panel Blood Test: Total Cholesterol, Low-Density Lipoproteins (LDL), High-Density Lipoproteins (HDL), and the Total Cholesterol ratio (Total Cholesterol divided by HDL or simply, TCH). Then there are the Logarithm of Triglycerides (LTG) and a Fasting Glucose Score, both of which are two types of sugars located in the body.

Our chosen response variable is a new variable, Y Binary, which has been calculated using the Test Score from the Glucose Tolerance Test or the Random Blood Sugar Test which indicate the patient is diabetic. Through a formula which assigns "Higher" to a test score above 200 and "Lower" to a score that is lower o equal to 200. SAS's JMP software automatically turns this into a binary integer since the applied formula only has two categories.

By applying 5 different estimation models to our dataset, we aim to determine which of these variables most impact the one-year score so that it is lower than 201. The first estimation model is the Standard or Ordinary Logistic Regression (OLS). Then 4 Penalized Logistic Regression Estimation models will be chosen: Lasso, Adaptive Lasso (ALasso), Elastic Net, and Adaptive Elastic Net (AENet). The advantage of the Penalized Estimation Models is that they reduce unimportant variables down to zero, effectively removing them from the model. This is sometimes called "Applying Regularization" and it's goal is to overcome overfitting the model and minimizing outliers.

2. ANALYSIS

The following chart quickly summarizes the differences between the 5 chosen Estimation Models:

Type of Method	Advantages	Disadvantages
Ordinary / Standard Least Squares	It is well-known, easy to explain and easy to understand.	It is sensitive to outliers and does not deal well with singularities.
Lasso Regression	Avoids overfitting models and removes variables that do not affect the response variable	Shrinks uninformative coefficients all the way to zero. Arbitrarily picks variables that are correlated
Adaptive Lasso Regression	Consistent variable selection where adaptive weights are used for penalizing different coefficients in the L1 penalty	Like Lasso, it cannot do group selection with highly correlated variables.
Elastic Net Regression	Keeps correlated variables and adds weighted penalties to them.	Takes more time to compute since it combines Ridge and Lasso.
Adaptive Elastic Net Regression	Encourages a grouping effect: either selects the correlated group or omits them	Increase in computational time.

For the overall analysis of the data set, cross validation was used to determine the best parameters for the variable in the data set. So, a random Validation Column was created with a Seed of 123. The data was split into three subsets of data: Training Subset which was 60% of the data, Validation Subset which was 20% of the data set and Testing Subset which was also 20% of the data set. While the OLS Model keeps all of the ten of the data variables previously described, the Penalized Logistic Regression Methods concurrently train and estimate those parameters so that the end result includes only the variables which have a larger impact on the response variable, Y Binary in this case, are left.

Understanding Diabetes Progression

3. MODEL COMPARISON

Starting with the OLS estimation model, and going through the Lasso, Adaptive Lasso, Elastic Net and Adaptive Elastic Net, columns were inserted into the data set representing each method's estimation. Then JMP's Model Comparison tool was used to collectively compare the columns which designated the probability that the test score was "High" or over 200. Since we only wanted to look at the Test subset portion of the cross validation, we examined the Misclassification Rate and the Area Under the Curve (AUC) for the 89 observations in the Test subset. As seen below, the Adaptive Lasso and the Adaptive Elastic Net resulted in equal Misclassification and AUC values as shown below).

Measures of Fit for Y Binary

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Fit Nominal Logistic		0.4012	0.5458	0.3568	0.3387	0.2294	0.1774	265	0.8937
Training	Fit Generalized Lasso		0.3854	0.5289	0.3661	0.3446	0.2460	0.1811	265	0.8871
Training	Fit Generalized Adaptive Lasso		0.3577	0.4984	0.3827	0.3497	0.2635	0.2000	265	0.8809
Training	Fit Generalized Elastic Net		0.3847	0.5281	0.3666	0.3447	0.2470	0.1811	265	0.8880
Training	Fit Generalized Adaptive Elastic Net		0.3575	0.4982	0.3828	0.3498	0.2636	0.2038	265	0.8809
Validation	Fit Nominal Logistic		0.3027	0.4111	0.3423	0.3228	0.2034	0.1023	88	0.8351
Validation	Fit Generalized Lasso		0.3236	0.4352	0.332	0.3179	0.2127	0.1364	88	0.8434
Validation	Fit Generalized Adaptive Lasso		0.3416	0.4556	0.3232	0.3103	0.2240	0.1136	88	0.8799
Validation	Fit Generalized Elastic Net		0.3237	0.4354	0.3319	0.3178	0.2136	0.1364	88	0.8434
Validation	Fit Generalized Adaptive Elastic Net		0.3416	0.4556	0.3232	0.3102	0.2241	0.1136	88	0.8799
Test	Fit Nominal Logistic		0.3307	0.4760	0.4224	0.3749	0.2355	0.2135	89	0.8856
Test	Fit Generalized Lasso		0.3237	0.4679	0.4269	0.3740	0.2512	0.2022	89	0.8747
Test	Fit Generalized Adaptive Lasso		0.3737	0.5245	0.3953	0.3588	0.2601	0.1910	89	0.8960
Test	Fit Generalized Elastic Net		0.3253	0.4697	0.4258	0.3737	0.2522	0.2022	89	0.8753
Test	Fit Generalized Adaptive Elastic Net		0.3737	0.5246	0.3953	0.3588	0.2602	0.1910	89	0.8960

ROC Curve

Model Comparison

Target Y Binary missing a predictor for category Low
Target Y Binary missing a predictor for category Low
Target Y Binary missing a predictor for category Low

Predictors

Measures of Fit for Y Binary

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Fit Nominal Logistic		0.4012	0.5458	0.3568	0.3387	0.2294	0.1774	265	0.8937
Training	Fit Generalized Adaptive Lasso		0.3577	0.4984	0.3827	0.3497	0.2635	0.2000	265	0.8809
Training	Fit Generalized Adaptive Elastic Net		0.3575	0.4982	0.3828	0.3498	0.2636	0.2038	265	0.8809
Validation	Fit Nominal Logistic		0.3027	0.4111	0.3423	0.3228	0.2034	0.1023	88	0.8351
Validation	Fit Generalized Adaptive Lasso		0.3416	0.4556	0.3232	0.3103	0.2240	0.1136	88	0.8799
Validation	Fit Generalized Adaptive Elastic Net		0.3416	0.4556	0.3232	0.3102	0.2241	0.1136	88	0.8799
Test	Fit Nominal Logistic		0.3307	0.4760	0.4224	0.3749	0.2355	0.2135	89	0.8856
Test	Fit Generalized Adaptive Lasso		0.3737	0.5245	0.3953	0.3588	0.2601	0.1910	89	0.8960
Test	Fit Generalized Adaptive Elastic Net		0.3737	0.5246	0.3953	0.3588	0.2602	0.1910	89	0.8960

ROC Curve

So, the top three models with the highest AUC (OLS, Adaptive Lasso and Adaptive Elastic Net) were put into a comparison on their own. The results showed that the Adaptive Lasso and the Adaptive Elastic Net were extremely similar. After adding in the AUC Comparison, again only the Testing subset was examined. When this tool compared the OLS Model to the Adaptive Model, and to the Adaptive Elastic Net Model, the values were identical (see the top picture on the next page).

When it compared the Adaptive Lasso to the Adaptive Elastic Net, it literally found no difference between the two methods of Estimation. Ultimately, the Adaptive Lasso was chosen because the Adaptive Elastic Net Method is simple a combination of the Lasso and the Ridge Model with a weighted L1 and L2 errors. The results basically said that the Elastic Net model eliminated the Ridge Method in its calculations.

AUC Comparison

AUC Comparison for Y Binary=High for Validation=Training

AUC Comparison for Y Binary=High for Validation=Validation

AUC Comparison for Y Binary=High for Validation=Test

Predictor	AUC	Std Error	Lower 95%	Upper 95%
OLS Prob[High]	0.8856	0.0337	0.8015	0.9369
ALasso Probability (Y Binary=High)	0.8960	0.0324	0.8133	0.9445
AENet Probability (Y Binary=High)	0.8960	0.0324	0.8133	0.9445

Predictor	vs. Predictor	AUC Difference	Std Error	Lower 95%	Upper 95%	ChiSquare	Prob>ChiSq
OLS Prob[High]	ALasso Probability (Y Binary=High)	-0.010	0.0160	-0.042	0.0209	0.4203	0.5168
OLS Prob[High]	AENet Probability (Y Binary=High)	-0.010	0.0160	-0.042	0.0209	0.4203	0.5168
ALasso Probability (Y Binary=High)	AENet Probability (Y Binary=High)	0.0000	0.0000	0.0000	0.0000	.	.

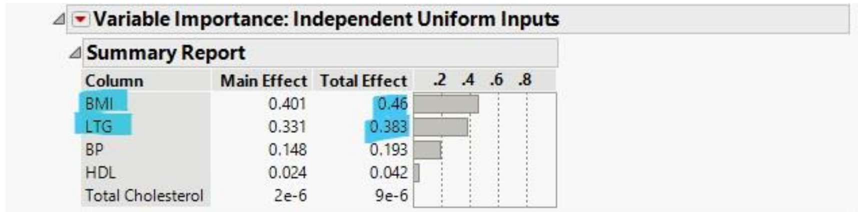
4. INTERPRETATION

After the Adaptive Lasso Model was chosen, it was run again. Looking at the Parameter Estimates in the chart to the right, there were five unimportant variables eliminated as designated by zeroes in the chart on the right: Age, Gender, Low-Density Lipoproteins (LDL), Total Cholesterol Ratio (TCH), and Fasting Glucose. There were five variables left in the model: Body Mass Index (BMI), Blood Pressure (BP), Total Cholesterol (TCH), and Logarithm of Triglycerides (LTG). It could be construed that the Total Cholesterol Ratio (TCH) was nearly zero with a negative parameter estimate of -7.231e-5, but it was kept because Adaptive Lasso felt it was slightly important to predict our response variable. As it turned out, the Total Cholesterol Ratio remained the 5th most important indicator that patients would remain diabetic one year after their baseline.

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-13.9585	2.4468621	32.543056	<.0001*	-18.75426	-9.162735
Age	0	0	0	1.0000	0	0
Gender[1-2]	0	0	0	1.0000	0	0
BMI	0.1479024	0.0372935	15.7284	<.0001*	0.0748085	0.2209963
BP	0.0375169	0.0126637	8.7767838	0.0031*	0.0126966	0.0623372
Total Cholesterol	-7.231e-5	0.0059682	0.0001468	0.9903	-0.01177	0.0116252
LDL	0	0	0	1.0000	0	0
HDL	-0.019621	0.0143958	1.8576481	0.1729	-0.047836	0.0085944
TCH	0	0	0	1.0000	0	0
LTG	1.2948845	0.4490309	8.3159062	0.0039*	0.4148001	2.1749689
Glucose	0	0	0	1.0000	0	0

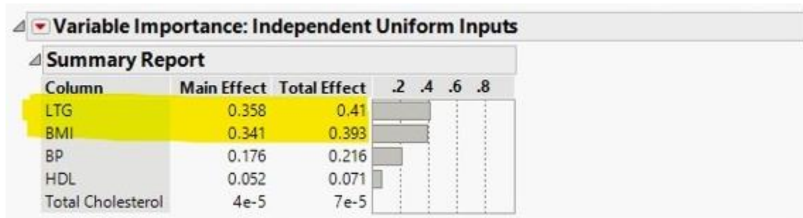
While utilizing the Variable Importance Tool under JMP's Profiler, it was noted that LTG and BMI made up around 80% of the Total Effect on the Response Variable. The Adaptive Lasso Model was run a few times and the Most Important Variable Model flip flopped between these two import variables. Blood Pressure remained the 3rd most important contributor to the Response on diabetes with High-Density Lipoproteins slightly more important than Total Cholesterol.



Variable Importance: Independent Uniform Inputs

Summary Report

Column	Main Effect	Total Effect
BMI	0.401	0.46
LTG	0.331	0.383
BP	0.148	0.193
HDL	0.024	0.042
Total Cholesterol	2e-6	9e-6

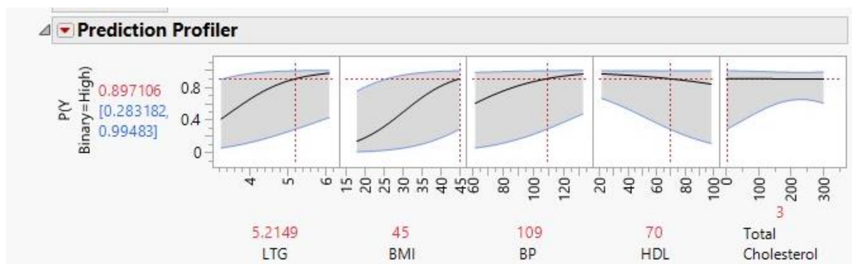


Variable Importance: Independent Uniform Inputs

Summary Report

Column	Main Effect	Total Effect
LTG	0.358	0.41
BMI	0.341	0.393
BP	0.176	0.216
HDL	0.052	0.071
Total Cholesterol	4e-5	7e-5

5. NEW CASE



There are a few options in JMP to test the chosen model by inserting new observations. A single patient was chosen with the following data: Age: 47, Gender=1, BMI=45, BP=109, Total Cholesterol=237, LDL=100.2, HDL=70, TCH=3, LTG=5.2149, Glucose=107. One option to test these values would be to manipulate the Prediction Filer's graphs to align with data presented above. As the picture directly above demonstrates, the Adaptive Lasso Model would only use the five most important variables. It gives an estimate of approximately a 90% chance that the patient would test as diabetic one year after the baseline score. It is worth noting that this estimate varies with subsequent Model runs.

An alternative was of testing values would be to insert a new row into the data set itself, using the above values as corresponding column entries. An additional data entry of Y was necessary for JMP to calculate the estimate directly into the table and the chosen value for this was 300. The results were similar to when the model itself was run with an estimate of approximately 90%. However, this method would allow instance comparison in all the models in the table.

	Y	Y Binary	Age	Gender	BMI	BP	Total Choleste...	LDL	HDL	TCH	LTG	Glucose	Validation	OLS Lin[High]	OLS Prob[High]	OLS Prob[Low]	OLS Most Likely Y Binary	ALasso Probability (Y Binary=High)
440	132	Low	60	2	24.9	99.67	162	106.6	43	3.77	4.1271	95	Training	-2.764229...	0.05928805...	0.94071194...	Low	0.1142842402
441	220	High	36	1	30.0	95	201	125.2	42	4.79	5.1299	85	Validation	0.2975018...	0.57383170...	0.42616829...	High	0.4617503657
442	57	Low	36	1	19.6	71	250	133.2	97	3	4.5951	92	Validation	-6.936688...	0.00097053...	0.99902946...	Low	0.0125292391
443	300	High	47	1	45.0	109	237	100.2	70	3	5.2149	107	•	0.5602968...	0.63652122...	0.36347877...	High	0.8955335851