Christopher Dillard 9/17/2023

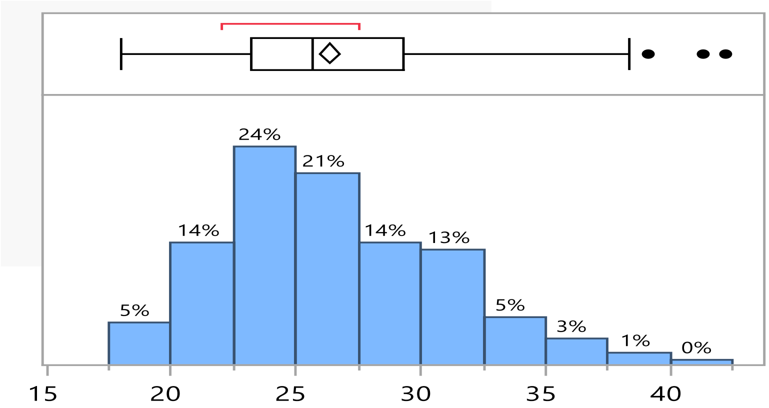
BAN 525

Module 3: Assignment 1 – Understanding Diabetes Progression

In this assignment, I analyzed how levels of ten baseline variables could contribute to the progression of diabetes, which is stored in the response variable's column, "Y Binary" with a binomial classification of LOW, indicating better patient outlook one year after baseline levels are recorded, and HIGH, representing higher disease progression. There are a total of 442 patients in the dataset, aged 19-79 (recorded within the "Age" column). Gender is encoded by 1 or 2, but information about which number corresponds to which gender is not disclosed. The gender split is about a 0.53 and 0.47 for genders "1" and "2", respectively; so, this is a fairly even split. Nondisclosure of such a variable could be helpful to discourage confirmation bias when running predictive models. That being said, the absence of clear gender information may hinder deeper interpretability of the models in case there are any medical implications as for whether gender is an important predictor variable, which is determined by how the models select variables of importance in JMP.

Another variable that is explored in this assignment is BMI (Body Mass Index). Any BMI between 25.0 to 29.9 is considered overweight, while anything above that range is considered obese (cdc.gov, 2022). In this dataset, the mean BMI is 26.4, meaning that the average BMI is well within the overweight range.

Body Mass Index (BMI) Distribution

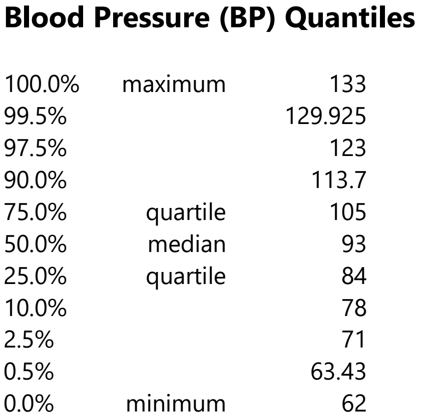


Quantiles

100.0%	maximum	42.2
99.5%		40.827
97.5%		36.1
90.0%		32.37
75.0%	quartile	29.325
50.0%	median	25.7
25.0%	quartile	23.175
10.0%		21
2.5%		19.3075
0.5%		18.186
0.0%	minimum	18

The above histogram reveals that the distribution of BMIs is skewed to the right, meaning that the spread of these data is further to the right side of the median (i.e., the higher values drive the mean to be higher than the median). The data within the “Quantile” table tells us that 50% of the patients have BMIs above 25.7, and almost 25% are considered obese. With so many of the patients being overweight, in addition to the large spread of values above the mean, the model’s selection of BMI and its estimation formula could give better insight into just how significant BMI may be as a predictor variable of diabetic disease progression.

One variable that unfortunately does not appear to make any medical sense in how it was recorded in the dataset, and therefore possibly statistically unreliable by default, is BP (Blood Pressure). Blood pressure is recorded as a fraction with the systolic blood pressure, the numerator in the fraction, and diastolic blood pressure, the denominator (Heart.org, 2023). Both numbers need to appear together since the systolic reading is the blood pressure when the heart is contracting and diastolic is the “resting” blood pressure. Furthermore, any division of systolic readings by their diastolic counterparts would not realistically produce the values seen in the BP column. Upon further examination of the data’s source (cited as “Efron, Bradley, et. al”., 2004), I could not find any clarification regarding the BP variable; so, I would presume that the values correspond to the diastolic blood pressure, again the denominator of a typical BP reading, since a diastolic reading of less than 80 is normal and over 120 is considered a hypertensive crisis, otherwise constituting a medical emergency. In the table below, we can see that only 2.5% of patients in the sample had values above 123 while the rest fell within a reasonably realistic range for diastolic BP readings. This interpretation is purely my speculation, so, if BP is selected by the best model, I would limit my discussion of any finding of significance in terms of BP’s relevance as a predictor variable, rather than what the individual values mean.



Blood Pressure (BP) Quantiles

100.0%	maximum	133
99.5%		129.925
97.5%		123
90.0%		113.7
75.0%	quartile	105
50.0%	median	93
25.0%	quartile	84
10.0%		78
2.5%		71
0.5%		63.43
0.0%	minimum	62

The other 6 “baseline” variables include Total Cholesterol, LDL (Low Density Lipoproteins), HDL (High-Density Lipoproteins), TCH (Total Cholesterol divided by HDL), LTG (Logarithm of Triglyceride level), and Glucose. Should these variables be selected by the model, I will interpret their significance in the closing paragraphs of this paper.

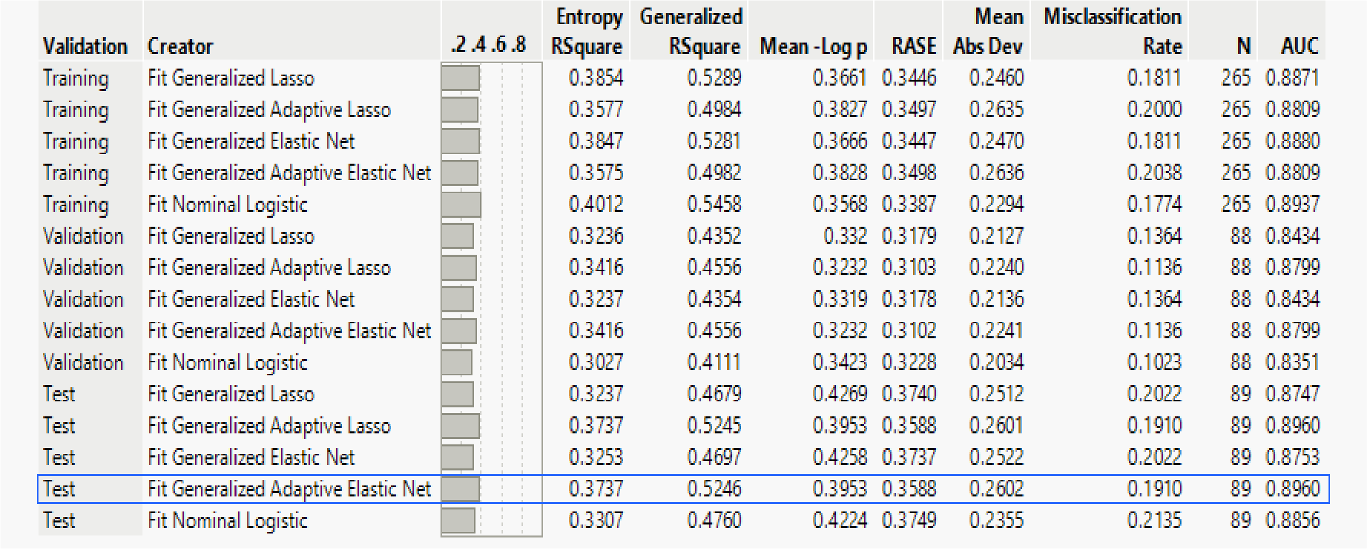
Classification problems look for the probability of different outcomes that are not continuous (not time-series data) and results are interpreted as the probability of certain outcomes, which in this case would only be either LOW or HIGH. Given that my interest is in seeing how the various variables could predict whether a patient’s diabetic condition progresses (gets worse), I will be focusing on the prediction formulas that correspond to HIGH predictions after a year for all patients and disregard the predicted LOWs.

I will employ 5 different methods in this assignment, one being Ordinary Logistic Regression and the rest belong to the Penalized Regression class of models: Lasso, Adaptive Lasso, Elastic Net, and Adaptive Elastic Net. One factor that must be addressed when selecting the best model, is how the ratio of rows to variables, 442 to 10, leaves me with the concern that there is a potential for overfitting. So, variable selection is of utmost importance when there is a chance of this phenomenon. Previous assignments have proven that Lasso and Elastic Net models and their adaptive versions do very well with variable selection by eliminating less important variables, something that the Ordinary Logistic Regression model would not be as efficient at doing.

Given that cross-sectional nature of the data (not a time-series sample) I will be implementing a random seed split of 123 on patients' data, using a 60-20-20 split for the training, validation, and testing sets. Much like in the other assignment, a new column called "Validation" will be used to identify each group.

The Area Under the Curve (AUC) will be my primary focus when analyzing the output for the Modeling Comparison tool in JMP. AUCs explain how efficient classification models are at predicting correct outcomes on a range between 0 and 1, where 1 would indicate that the model is a perfect fit. Any model above 0.5 can be interpreted as being more effective at predicting outcomes than a coin flip (i.e., random chance), and anything 0.5 and below is automatically useless for the sake of this assignment.

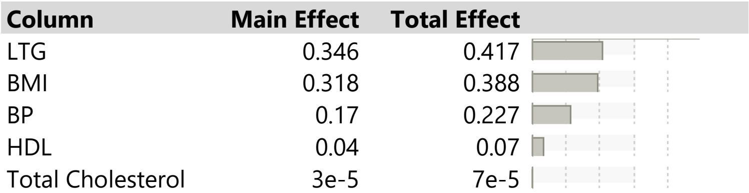
The image below captures the model comparison output, where my focus is on the AUCs for each model on the test set.



Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N	AUC
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
Training	Fit Nominal Logistic		0.4012	0.5458	0.3568	0.3387	0.2294	0.1774	265	0.8937
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
Validation	Fit Nominal Logistic		0.3027	0.4111	0.3423	0.3228	0.2034	0.1023	88	0.8351
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
Test	Fit Nominal Logistic		0.3307	0.4760	0.4224	0.3749	0.2355	0.2135	89	0.8856

As seen in the table, the AUCs are identical for both the Adaptive Lasso model and the Adaptive Elastic Model on the Testing set, which is of course the “new” and unbiased 20% of the data that the model was not trained on. This constitutes the need for further analysis of both models to see which one would be the best fit for the data. The only value that would favor one of these models over the other in this case has been shown by the blue box in the table: the generalized R Square for the Adaptive Elastic Net is 0.0001 higher than the Adaptive Lasso’s value. So, I opted to choose the Adaptive Elastic Net as the best model for this case. It is also worth noting that the performance metrics for all the models do not appear to have a large degree of variation across the board; but, the Ordinary Logistic Regression (Nominal Logistic) did perform the worst out of the 5 models in its predictive capabilities.

After making this decision, I then explored each variable’s importance in the model to see where analysts should primarily direct their attention to when discussing patient outlook and diabetic disease progression.

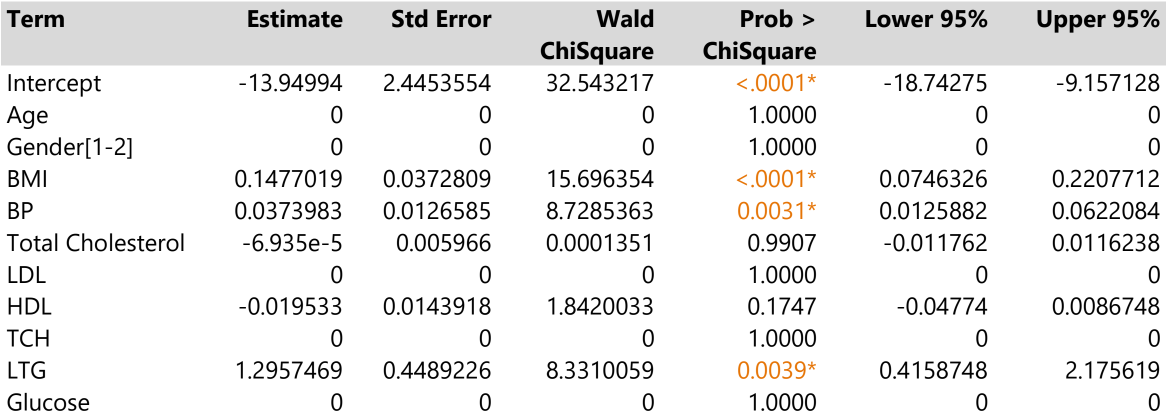


Column	Main Effect	Total Effect
LTG	0.346	0.417
BMI	0.318	0.388
BP	0.17	0.227
HDL	0.04	0.07
Total Cholesterol	3e-5	7e-5

The Adaptive Elastic Net model identified, in descending order, LTG, BMI, BP, HDL, and, to a very small degree, Total Cholesterol as the most important variables out of the ten baseline variables in the dataset, essentially deeming the rest to be statistically irrelevant in predicting disease progression compared to the 5 variables in the above table.

As shown by the parameter estimates, Age, Gender, LDL, TCH, and Glucose were dropped from the model. With gender being dropped by the model, this makes my previous concern about interpretability without the knowledge of which number corresponds to which gender irrelevant.

Parameter Estimates for Original Predictors: Adaptive Elastic Net Model



Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-13.94994	2.4453554	32.543217	<.0001*	-18.74275	-9.157128
Age	0	0	0	1.0000	0	0
Gender[1-2]	0	0	0	1.0000	0	0
BMI	0.1477019	0.0372809	15.696354	<.0001*	0.0746326	0.2207712
BP	0.0373983	0.0126585	8.7285363	0.0031*	0.0125882	0.0622084
Total Cholesterol	-6.935e-5	0.005966	0.0001351	0.9907	-0.011762	0.0116238
LDL	0	0	0	1.0000	0	0
HDL	-0.019533	0.0143918	1.8420033	0.1747	-0.04774	0.0086748
TCH	0	0	0	1.0000	0	0
LTG	1.2957469	0.4489226	8.3310059	0.0039*	0.4158748	2.175619
Glucose	0	0	0	1.0000	0	0

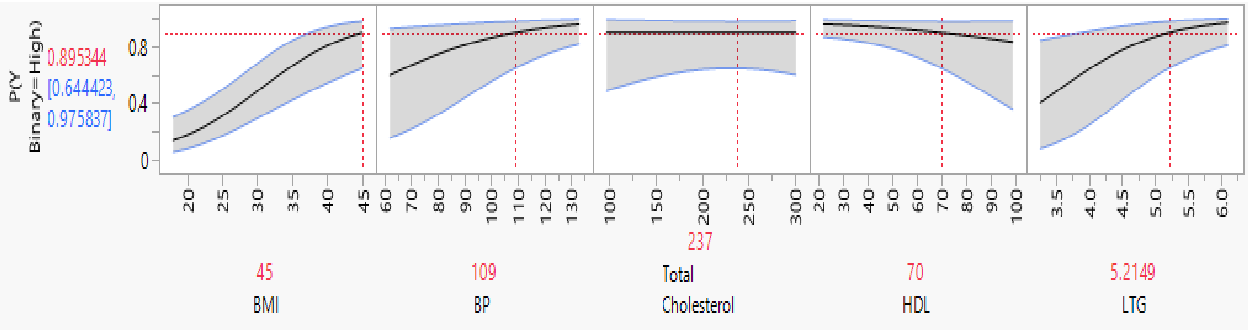
As for the most important predictor variable, LTG, or the logarithm of triglyceride level, high levels of triglycerides in the blood are often indicative of metabolic disorders (National Heart Lung and Blood Institute, 2023). This implication is significant since they are a type of lipid (fat) that the body uses to store calories that it doesn’t immediately need after eating. Conditions like Diabetes directly impact how the body processes and stores energy and in patients with this disease, particularly Type 2 Diabetes, insulin resistance is a factor that leads the body to resort to other sources of energy when there is some inability to correctly use insulin to bring glucose, the general primary source of energy, to the cells. In other words, the body releases triglycerides into the bloodstream as an alternative source of energy to glucose and higher levels of triglycerides can be indicative of higher resistance to insulin.

I was initially surprised to see that Glucose was not selected as a predictor variable since taking glucose levels is usually routinely done by diabetic patients to manage their disease. One potential reason for the exclusion of this variable in the model could have to do with how glucose levels are taken in the laboratory setting, along with how glucose levels fluctuate throughout the day, depending on what a patient is eating and when they're last meal was relative to when their bloodwork was taken (Nekrani et. al., 2023). Since no mention of timing is available in the dataset my interpretation would be purely speculatory but, if lab work was done in the morning, patients with worse disease progression could have elevated numbers if they're last meal was closer to when they fell asleep, or whether they ate anything for breakfast prior to getting bloodwork done. So, a diabetic patient could reasonably be aware of this and already have adjusted their eating schedule accordingly which could give a lower glucose level. Furthermore, we could assume that some patients did not eat in the hours prior to their appointment, but otherwise are less conscientious about their eating habits, thus facing higher disease progression despite having a low blood glucose reading in a clinical setting. So, the exclusion of glucose from the model could be reflective of how blood glucose readings significantly fluctuate in any given day, which could render this variable to be too unreliable for a predictive model.

Now, I can use JMP to develop a hypothetical prognosis on a patient that has the following profile:

Age: 47, Gender: 1, BMI: 45, BP: 109, Total Cholesterol: 237, LDL: 100.2, HDL: 70, TCH: 3, LTG: 5.2149, Glucose: 107.

JMP streamlines this process by allowing analysts to interpret the parameter estimates by viewing the prediction profilers of their model. Then, hypothetical values can be inputted to automatically view the probability of a certain outcome, being the HIGH predictions in this case:



This result tells us that the patient in question would be at an 89.5% risk (probability) of advancing to higher disease progression in one year based on their baseline values. As I discussed previously from the parameter estimates table, some of the information on the patient was deemed to be irrelevant in the model and this is apparent where their values are substituted by a zero across the rows corresponding to the patient's age, gender, LDL, TCH, and Glucose levels.

Work Cited

"Assessing Your Weight." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 3 June 2022, www.cdc.gov/healthyweight/assessing/index.html#:~:text=If%20your%20BMI%20is%20less,falls%20within%20the%20obese%20range.

Efron, Bradley, et al. "Least angle regression." *The Annals of Statistics*, vol. 32, no. 2, 2004, https://doi.org/10.1214/009053604000000067.

"High Blood Triglycerides." *National Heart Lung and Blood Institute*, U.S. Department of Health and Human Services, www.nhlbi.nih.gov/health/high-blood-triglycerides. Accessed 14 Sept. 2023.

Nakrani, Mihir N., et al. "Physiology, Glucose Metabolism - Statpearls - NCBI Bookshelf." *National Library of Medicine*, National Institutes of Health, 17 July 2023, www.ncbi.nlm.nih.gov/books/NBK560599/.

"Understanding Blood Pressure Readings." *Www.Heart.Org*, 30 May 2023, www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.