Diabetes Screening Model

Executive Summary

This paper presents the development of a predictive model of diabetes disease progression. The final predicted model included inputs of body-mass index, blood pressure, age, and two blood serum measurements,all of which can be said to have a direct relationship with disease progression with 99% confidence. Increased BMI, BP, and S5 were associated with higher diabetes progression, and decreased S3 was associated with higher diabetes progression. Of the four, BMI had the strongest relationship with diabetes progression, and S3 had the weakest relationship.

This model performed effectively when applied against data that had not been used in training the model. Within the training data, it was able to explain 50.3% of the total variation in disease progression based on those factors, indicating that more research still needs to be conducted to identify predictive variables to better screen for diabetes risk. Within the held-out validation data, it was able to explain 47.3% of the total variation.

Introduction

Diabetes is a deadly problem. According the the CDC, it is the 8th leading cause of death in the United States[1]. Effective screening and prediction of diabetes risk can save lives, and the model presented herein provides some hope of doing just that based on data from 392 diabetes patients. All variables listed in table 1 were considered to develop a model that predicted a measure of disease progression after one year.

Variable	Description
AGE	Patient's age (years)
SEX	Patient's sex
BMI	Body-Mass Index
BP	Average Blood Pressure
S1 – S6	Blood Serum Measurements
Y	Response Variable (disease progression after one year)

Table 1: Variables used in this analysis

We developed three candidate ordinary least squares models through combinations of lasso regression, stepwise regression using backward elimination, and cross-validation between training and validation datasets. The final model passed all model adequacy checks without major issues, suggesting that it may be effective for screening new patients on their diabetes risk, although only 50.3% of disease progression is explained by this model. Future research should explore further measurements that could increase the accuracy of this model.

Methodology

Data Preparation

The full dataset consisted of data from 442 patients, but 50 patients' data was randomly held aside by the customer for a test set. Of the remaining data, we used a random selection of 262 patients' data to train the model and we held the remaining 130 patient's data for model validation purposes.

[1] https://www.cdc.gov/nchs/fastats/diabetes.htm

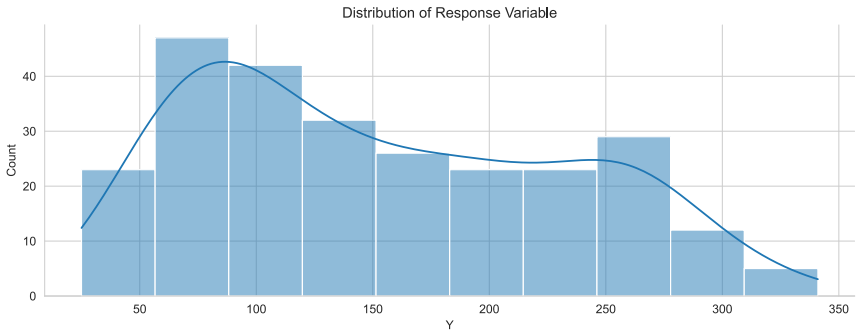

Figure 1: Distribution of response variable without transformation

As shown in Figure 1, the response variable has a right-skewed distribution suggesting that a transformation may be required to properly model it. A Box-Cox transform seemed to over-correct, yielding a slightly left-skewed distribution, so we landed on a square root transformation as shown in Figure 2.

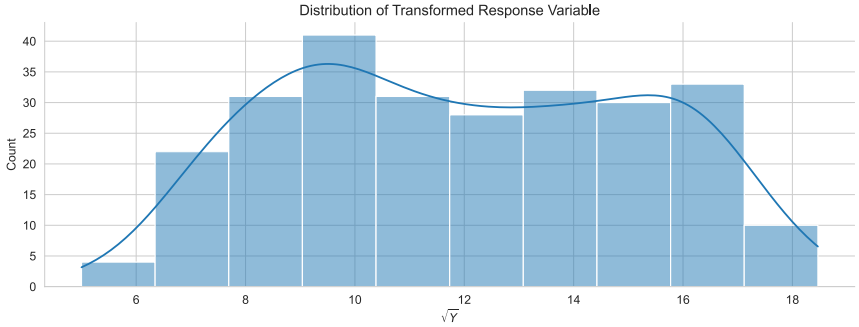

Figure 2: Distribution of response variable after transformation

Feature Selection

For this modeling effort, all main effects, 2-way interactions, and quadratic effects were considered. Models were required to maintain hierarchy and all included an intercept term.

Candidate Model 1 resulted from using lasso regression as a feature selection technique while optimizing for Akaike Information Criterion (AIC). Searching over the range of $\alpha \in [0.01, 2.00]$, the AIC-optimal model came from setting $\alpha = 0.16$. We refined this model by performing stepwise backward elimination of terms until AIC was no longer decreasing. Candidate Model 1, shown in Equation 1 included 11 terms including all variables except blood serum measurements S1, S2, and S4 along with the quadratic terms AGE^2 and S6^2, and interactions BMI×BP and S3×S5.

$$\begin{aligned}\sqrt{\hat{Y}} = 33.76 - 0.9613(\text{SEX} = \text{MALE}) - 0.1718(\text{BMI}) - 0.0555(\text{BP}) - 0.1955(\text{S3}) \\ +0.0939(\text{S5}) - 0.1427(\text{AGE}) + 0.0015(\text{AGE}^2) - 0.3753(\text{S6}) + 0.0020(\text{S6}^2) \\ +0.0041(\text{BMI} \times \text{BP}) + 0.0329(\text{S3} \times \text{S5})\end{aligned} \tag{1}$$

Candidate Model 1 achieved $R^2_{adj} = 0.515$, AIC = 1181, and BIC = 1223. The model F test, testing the alternative hypothesis that the actual model $R^2 = 0$ against the alternative hypothesis that $R^2 \neq 0$ yielded

a p-value of < 0.0001, so we reject the null hypothesis and conclude that this model has some predictive power. Applying Candidate Model 1 to the validation dataset yielded a root mean square error (RMSE) of 55.22.

Candidate Model 2 resulted from using lasso regression as a feature selection technique while optimizing for Bayesian Information Criterion (BIC). Searching over the range of $\alpha \in [0.01, 2.00]$, the BIC-optimal model came from setting $\alpha = 0.69$. Candidate Model 2, shown in Equation 2 is very parsimonious, using four first-order terms: BMI, BP, S3, and S5.

$$\sqrt{\hat{Y}} = -4.9445 + 0.2225(\text{BMI}) + 0.0469(\text{BP}) - 0.0337(\text{S3}) + 1.9111(\text{S5}) \tag{2}$$

Candidate Model 2 achieved $R^2_{adj} = 0.473$, AIC = 1196, and BIC = 1214. The model F test, testing the alternative hypothesis that the actual model $R^2 = 0$ against the alternative hypothesis that $R^2 \neq 0$ yielded a p-value of < 0.0001, so we reject the null hypothesis and conclude that this model has some predictive power. Applying this model to the validation dataset yielded a root mean square error (RMSE) of 55.73, performing similarly to model 1 despite massively trimming the model.

Candidate Model 3 resulted from using lasso regression as a feature selection technique while optimizing the RMSE achieved when applied to the validation data. Searching over the range of $\alpha \in [0.01, 2.00]$, the validation RMSE-optimal model came from setting $\alpha = 0.03$, yielding RMSE of 55.30. However, this model displayed a massive gap between performance in training and validation sets, indicating that it reflected an overfit model. Candidate Model 3, shown in Equation 3 is an enormous model including 38 terms.

$$\begin{aligned}
\sqrt{\hat{Y}} = 86.80 &- 8.99(\text{SEX} = \text{MALE}) - 0.3695(\text{AGE}) + 0.0871(\text{BMI}) + 0.1260(\text{BP}) \\
&- 0.1880(\text{S1}) + 0.2340(\text{S2}) - 0.1212(\text{S3}) + 0.6265(\text{S4}) - 26.09(\text{S5}) - 0.5284(\text{S6}) \\
&+ 0.0015(\text{AGE}^2) + 0.0002(\text{S3}^2) + 0.0877(\text{S4}^2) + 3.340(\text{S5}^2) + 0.0020(\text{S6}^2) \\
&+ 0.0347(\text{SEX} = \text{MALE}) \times (\text{AGE}) + 0.1204(\text{SEX} = \text{MALE}) \times (\text{BMI}) \\
&+ 0.0154(\text{SEX} = \text{MALE}) \times (\text{BP}) + 0.0074(\text{SEX} = \text{MALE}) \times (\text{S3}) \\
&- 0.3566(\text{SEX} = \text{MALE}) \times (\text{S4}) + 0.0282(\text{SEX} = \text{MALE}) \times (\text{S6}) - 0.0016(\text{AGE} \times \text{BMI}) \\
&+ 0.0003(\text{AGE} \times \text{BP}) - 0.0004(\text{AGE} \times \text{S2}) + 0.0009(\text{AGE} \times \text{S3}) + 0.0483(\text{AGE} \times \text{S5}) \\
&+ 0.0073(\text{BMI} \times \text{BP}) - 0.0004(\text{BMI} \times \text{S1}) - 0.1016(\text{BP} \times \text{S2}) + 0.0007(\text{BP} \times \text{S5}) \\
&- 0.0026(\text{BP} \times \text{S6}) - 0.0057(\text{S1} \times \text{S4}) - 0.0090(\text{S1} \times \text{S5}) + 0.0002(\text{S1} \times \text{S6}) \\
&- 0.0008(\text{S2} \times \text{S3}) + 0.0389(\text{S3} \times \text{S5}) + 0.0019(\text{S3} \times \text{S6}) + 0.0596(\text{S5} \times \text{S6})
\end{aligned} \tag{3}$$

Candidate Model 2 achieved $R^2_{adj} = 0.522$, AIC = 1202, and BIC = 1345. The model F test, testing the alternative hypothesis that the actual model $R^2 = 0$ against the alternative hypothesis that $R^2 \neq 0$ yielded a p-value of < 0.0001, so we reject the null hypothesis and conclude that this model has some predictive power. Naturally this model achieved the best validation RMSE at 55.30.

Final Model Selection

A comparison of results is given in Table 2. Candidate Model (CM) 1 has the best AIC and validation RMSE, CM2 has the best BIC, and CM3 has the best R^2_{adj}. CM 1 achieved better RMSE than CM3 by using backward elimination from the base lasso model, eliminating CM 3 from consideration. Figure 3 shows how RMSE on training and validation sets varies with the lasso parameter α. This suggests that the low-α CM 1 represents over-training compared to CM2 as the performance varies widely between data sets, so we ultimately chose CM 2 as the final model. CM 2 performs almost identically in the training and validation data sets.

Model	# Terms	R^2	R^2_{adj}	AIC	BIC	Validation RMSE
1	10	0.536	0.515	1181	1223	55.22
2	4	0.481	0.473	1196	1214	55.73
3	38	0.594	0.522	1202	1345	55.30

Table 2: Summary of performance metrics for candidate models

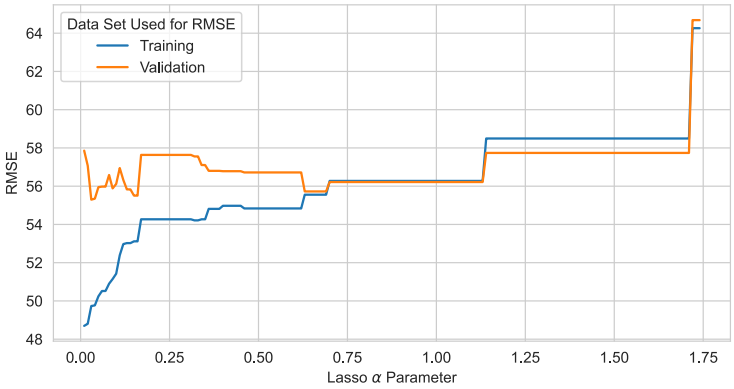

Figure 3: Training and Validation RMSE for Lasso models varying α

Adequacy

We assessed the final model in terms of the assumptions of linear regression. As seen in Figure 4, the mean of residuals was approximately constant and zero across the model space. As seen in Figure 5, the variance of residuals was slightly lower for the lowest and highest predicted values, indicating some issues with heteroscedasticity. Neither the Breusch-Pagan nor Goldfeld-Quandt tests indicated statistically significant heteroscedasticity with $\alpha = 0.1$. We decided to accept this moderate violation of assumptions because the issue was not severe and because this represents above-expected accuracy for identifying low- and high-progression patients.

Furthermore, as seen in Figure 6, the residuals appear approximately normally distributed, although the D'Agostino-Pearson Omnibus test yields a p-value of 0.099, indicating with 90% confidence that the underlying distribution of residuals is not quite normal. Again, we decided to accept this minor deviation from normality.

We also searched the residuals for significant influence points that may represent unusual observations that had undue effect on the model. No residual had an externally studentized residual above 3 or below -3, indicating no major issues with influence points.

We also assessed the model for issues with multicollinearity. After applying the model based on standardized inputs, the variance inflation factors (VIF) for every term was between 1 and 2, indicating no issues with multicollinearity that might impact our ability to make inference about the terms in the model.

Results

We can say with 99% confidence that each individual term in the model has an effect on Y. The results of each t-test and 99% confidence intervals on each term are given in Table 3.

As shown in Figure 7, the model performed similarly when applied to the training and to the validation datasets, indicating good capacity for out-of-sample prediction.

Figure 4: Fitted value and residual plot with LOWESS fit demonstrating constant mean of residuals

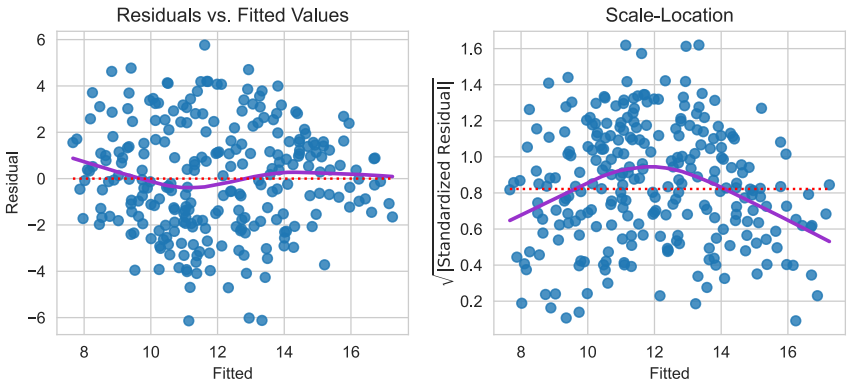

Figure 5: Fitted value and scale-location plot with LOWESS fit demonstrating slight residual heteroscedasticity

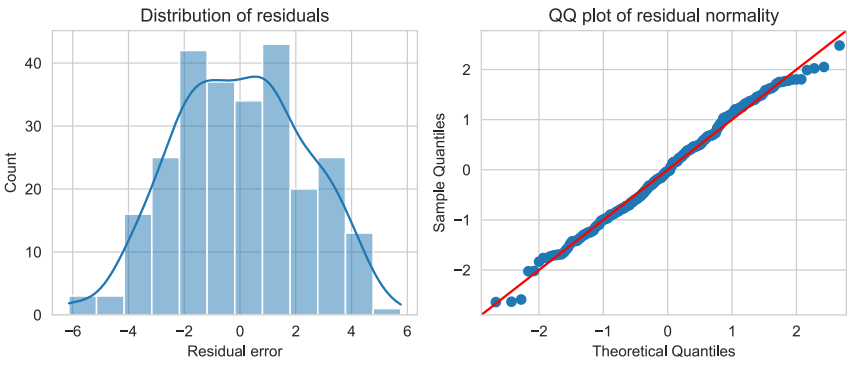

Figure 6: Histogram and QQ plot demonstrating approximate normality of residuals

	Coef.	Std.Err.	t	p-value	99% Confidence Interval Lower	Upper
Intercept	-4.945	1.885	-2.623	0.009	-9.837	-0.053
BMI	0.223	0.038	5.911	<0.001	0.125	0.320
BP	0.047	0.012	4.047	<0.001	0.017	0.077
S3	-0.034	0.012	-2.707	0.007	-0.066	-0.001
S5	1.761	0.337	5.227	<0.001	0.887	2.635

Table 3: Summary of hypothesis tests for each term in the final model

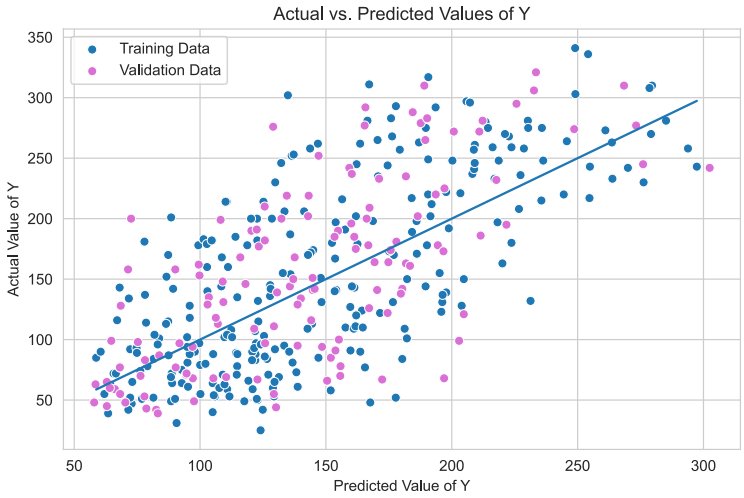

Figure 7: Actual data vs. predicted for both training and validation sets

90% prediction intervals for the customer-withheld test data points were generated and provided separately for assessment. By the magic of being the instructor, I also have results for that data, though I promise I did not run the results until after all the other modeling had been conducted. The RMSE when this model was applied to the test data was 49.89 — better than the performance for the other candidate models (52.07 for CM 1 and 52.73 for CM 3).This is actually slightly better than the training and validation RMSE. Furthermore, the generated intervals were slightly underconfident; 98% of the test data was covered by the generated 90% prediction intervals. Those intervals had an average width of 179, which probably means this wouldn't be diagnostically useful in practical settings as this represents a huge range of diabetes outcomes.

Conclusion

In this paper we have detailed the generation of a model predicting progression of diabetes based on supplied data. The final model included effects using BMI, BP, S3, and S5 to predict the square root of Y. All terms were significant with 99% confidence. BMI had the strongest association of the four, and S3 has the weakest association.

Three candidate models were generated and compared prior to selecting the final model, which was the most parsimonious of the three models. The winning model used lasso regression as a feature selection technique to select the model yielding minimum BIC.

The final model only explains 50.3% of the variance in disease progression in the training data and only 47.3% in the validation data, indicating that more research needs to be conducted to identify other predictive variables.