1 Introduction

The diabetes dataset contains 442 patient records on the progression of diabetes after one year. There are ten baseline variables, which are listed in Table 1, along with the target variable `Y`. The target variable is a quantitative measurement of the disease progression. The `SEX` variable is the only categorical variable. All other baseline variables are quantitative. The goal of this project is to develop a model that can predict the disease response variable using the baseline measurement variables.

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: Baseline variables for diabetes patients.

2 Executive Summary

A linear regression model predicting the disease progression `Y` was able to explain 53.2% of the variability in `Y` with the patient's age, sex, body-mass index (BMI), average blood pressure, and three blood serum measurements (`S1`, `S2`, and `S3`). The interaction of the patient's age and sex, and the interaction between the patient's BMI and average blood pressure were also used in the model. There was significant multicollinearity in the model, specifically with `S1` and `S2`. As a result, we cannot infer information about the relationships between individual variables and the disease response variable.

3 Predicted Model

The predicted model is given by Equation (1). To incorporate the categorical variable `SEX` into the model, the value of `SEX` is 1 for males and 0 for females.

$$\begin{aligned}\hat{y} = &-0.6952\texttt{AGE} - 6.6417\texttt{BMI} - 2.0761\texttt{BP} - 0.8956\texttt{S1} + 0.7469\texttt{S2} \\ &+ 72.2785\texttt{S5} - 98.2615\texttt{SEX} + 1.5151\texttt{AGE} \times \texttt{SEX} + 0.1248\texttt{BMI} \times \texttt{BP} \quad (1)\end{aligned}$$

A scatterplot of the hypothesized model, comparing the actual observed disease progression to the predicted progression is shown in Figure 1. The plot shows both the data used to train the model, and the data used in validation. Several observation from the training data with high-magnitude residuals are indicated.

4 Model Development

Before building the model the data was split into training and validation sets. Note that the last 50 records were withheld for testing purposes, leaving 392 for training and validation. The 392 records were then divided into training and validation data sets, with approximately 85% used for training. The size of each data set and percentages of the total are shown in Table 2.

Several methods were used to build multiple models to compare with the validation data set. The first method used was backward elimination with only first order terms. R_a^2, AIC and, BIC were all compared when building the initial models, but BIC was used as the primary metric to try and ensure a limited number of parameters. The p-value for each coefficient was also used to ensure all were statistically significant (with

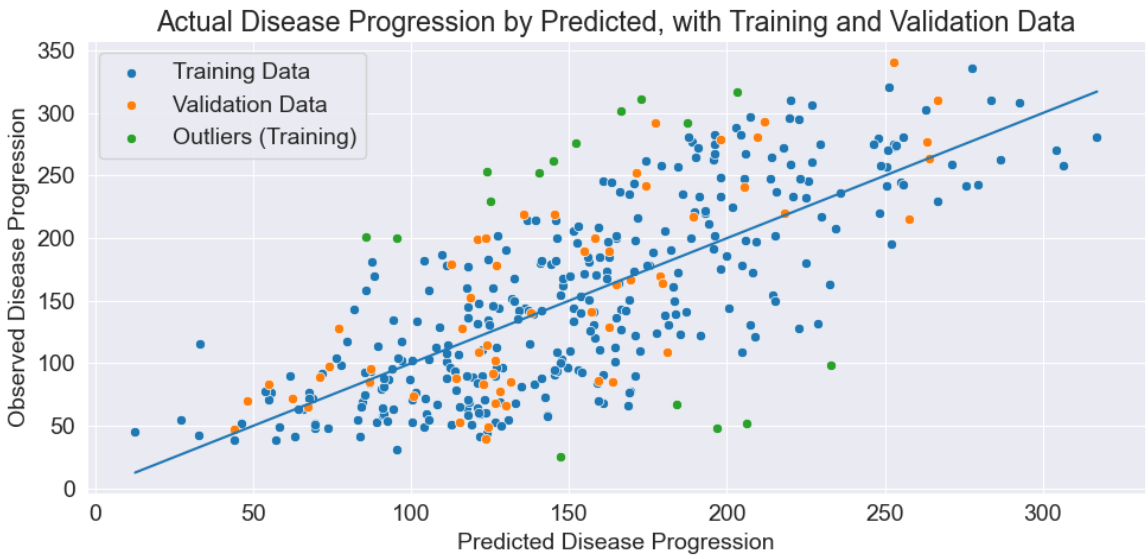

Figure 1

Data Set	Record Count	% of Total
Training	333	75.3%
Validation	59	13.3%
Test	50	11.3%

Table 2

$\alpha = 0.05$).

The first order model that minimized BIC with all remaining coefficients being significant included `BMI`, `BP`, `S1`, `S2`, `S5`, and `SEX`. `S1` and `S2` were highly correlated so `S2` was also removed. The resulting model had a higher BIC (3650 vs 3647), so both models were kept for later validation comparison. The simplest model (without `S2`) is referred to as Model 0 in Table 3, and the first model built is Model 1. The table shows the parameters used for each model and the BIC.

The next candidate model, Model 2, was built by using backward elimination again, but with quadratic terms for all numerical variables included with all the first order terms. The result of this method again included the variables `BMI`, `BP`, `S1`, `S2`, `S5`, and `SEX` as well as two new variables, `AGE` and $\texttt{AGE}^2$. The BIC (3655) was worse than the first models, but it was still kept for later comparison.

To add interaction terms, forward selection was used starting with all the parameters from Model 2 and looking at the two-way interactions. Two interaction terms, `AGE:SEX` and `BMI:BP`, were added to the model before the BIC stopped improving. $\texttt{AGE}^2$ was no longer significant, so it was removed and the BIC improved to 3644, the lowest yet. The `BMI` and `BP` coefficients were no longer significant ($p > 0.05$), but they were left in because of the interaction terms. This model is referred to as Model 3. The p-value for the intercept was 0.727. Since it was not significant another model was fit without the intercept. The BIC improved to 3639, so this was added as Model 4, and it would later be selected as the prediction model.

One more method was used to build models for comparison and that was a transform to the response variable `Y`. To improve the normality of `Y`, a y^λ transform was used with $\lambda = 0.394$. Figure 2 shows the distribution of `Y` before and after the transformation.

Rather than going through more step-wise regression, the transform was applied and the same model parameters for Models 0-4 were used to fit new models. The transformed models are indicated with a 'T' prefix before the model number in Table 3. One additional model was added because another intercept term was not significant after the transform.

Model	Model Parameters											BIC
	AGE	BMI	BP	S1	S2	S5	SEX	AGE^2	AGE:SEX	BMI:BP	Intercept	
0		×	×	×		×	×				×	3650
1		×	×	×	×	×	×				×	3647
2	×	×	×	×	×	×	×	×			×	3655
3	×	×	×	×	×	×	×		×	×	×	3644
4	×	×	×	×	×	×	×		×	×		3639
T0		×	×	×		×	×				×	1034
T1		×	×	×	×	×	×				×	1028
T2	×	×	×	×	×	×	×	×			×	1033
T3	×	×	×	×	×	×	×		×	×	×	1027
T4	×	×	×	×	×	×	×		×	×		1023
T5	×	×	×	×	×	×	×	×				1029

Table 3: Model parameter and BIC comparison. Models with a 'T' prefix used a y^λ transform.

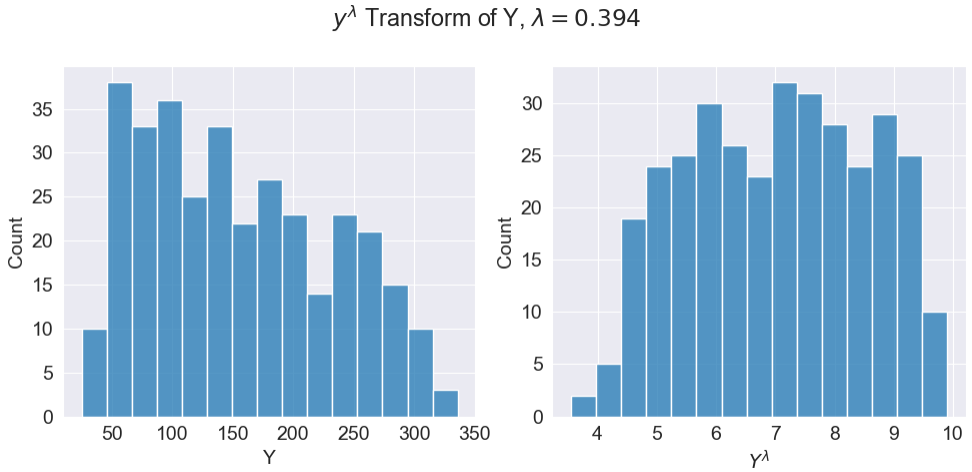

Figure 2: Comparison of the response variable before and after applying a y^λ transform for normalization.

Test for validity: With the exceptions already mentioned concerning the significance of individual parameters, all other individual p-values for each model were $p < 0.05$. So we reject the null hypotheses that the coefficient of each equals zero. The coefficient values, test statistics, p-values and 95% confidence intervales for the selected model (Model 4) are shown in Table 4. The null hypothesis of each candidate model is that at all of the coefficients for the model are equal to zero. For each one the p-value was $p < 0.0001$, so we reject the null hypothesis for each model. The selected model had a test statistic of $F = 348$, and $p < 0.0001$.

Validation of Assumptions
The assumptions for each candidate model were validated as they were built. Only the selected model will be covered here. Most of the candidate models had similar results. Nothing was discovered in any of the models that precluded it from further comparison.

Parameter	Coefficient	t	p-value	Lower CI	Upper CI
`AGE`	-0.6952	-2.16	0.0309	-1.33	-0.06
`BMI`	-6.6417	-5.02	< 0.0001	-9.24	-4.04
`BP`	-2.0761	-6.12	< 0.0001	-2.74	-1.41
`S1`	-0.8956	-3.52	0.0005	-1.40	-0.40
`S2`	0.7469	2.91	0.0038	0.24	1.25
`S5`	72.2785	8.80	< 0.0001	56.13	88.43
`SEX`	-98.2615	-4.28	< 0.0001	-143.36	-53.16
`AGE:SEX`	1.5151	3.35	0.0009	0.63	2.40
`BMI:BP`	0.1248	10.92	< 0.0001	0.10	0.15

Table 4: Coefficients for the selected prediction model (Model 4), with 95% confidence intervals.

The residuals are shown in Figure 3, and there is no significant pattern in the mean. There appears to be less variance at both the low and high ends of the predicted values. However, there is much less data to compare in those regions.

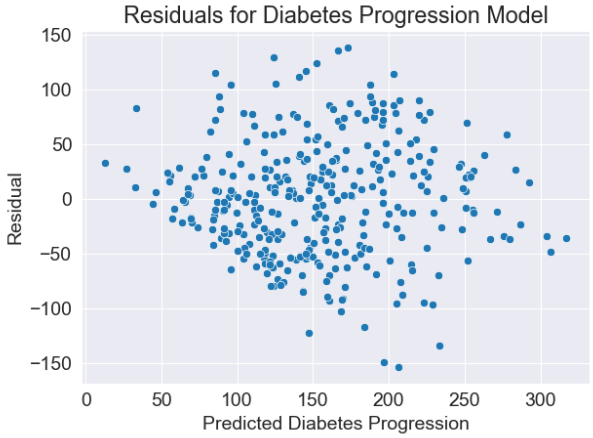

Figure 3

Figure 4 shows that the residuals are approximately normally distributed with both a histogram and a QQ-plot. The Omnibus test returned a p-value of $p = 0.73$, so we can not reject the null hypothesis that the residuals are normally distributed.

There is no way to test for the independence of errors because there is no data on the timeliness or geographic location of data collection.

Other Diagnostics The outliers identified earlier are show in the standardized and studentized residual plots in Figure 5. The outliers are shown as the high residuals on the plots and the high leverage points are also highlighted. These are also shown in the influence plot in Figure 6, where it is easy to distinguish the high residual and high leverage points.

The multicollinearity of each candidate model was also checked using the Variance Inflation Factors (VIF). For all but Model 0 there were concerns with multicollinearity. Table 5 shows the VIFs for the selected model (Model 4). This is expected for several models because of the interactions and second order terms. It was also known that there was correlation between `S1` and `S2`, but both were kept in most models because it improved the model fit, and could help model prediction. When multicollinearity is high, it is difficult to infer information about the parameters in the model. Since the purpose of this model is for

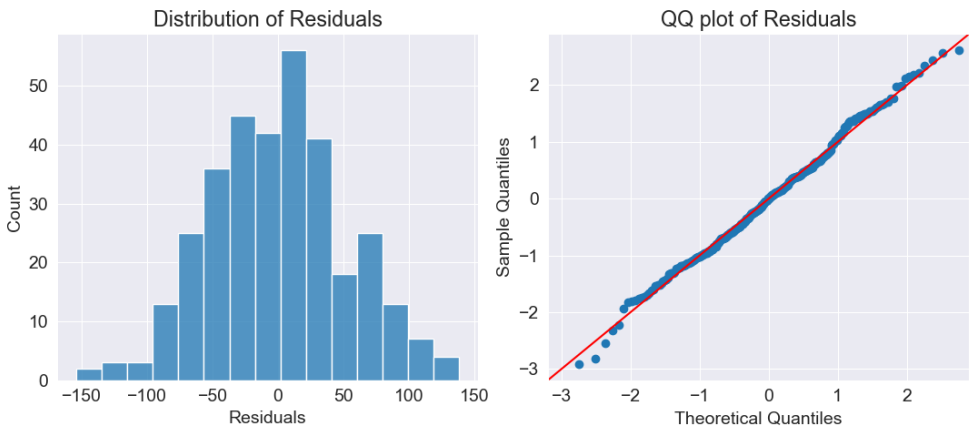

Figure 4

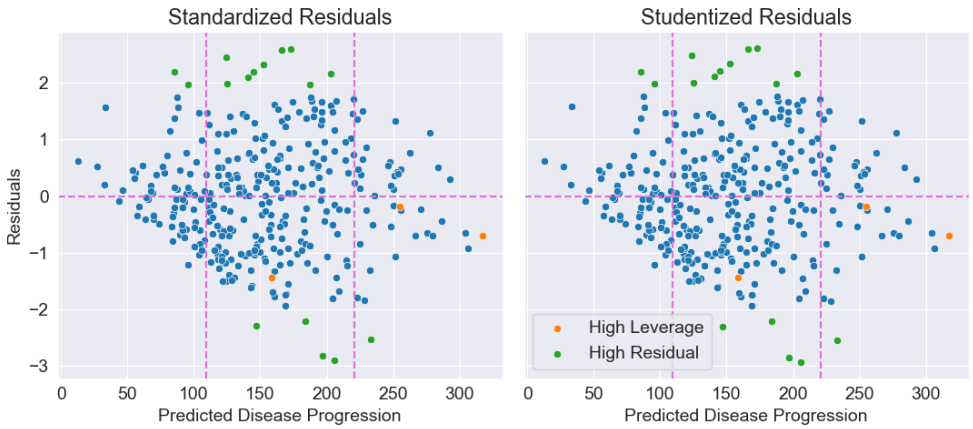

Figure 5

prediction, the multicollinearity issues were be accepted.

Variable	VIF
`AGE`	30.34
`BMI`	147.90
`BP`	123.55
`S1`	280.36
`S2`	110.83
`S5`	172.36
`SEX`	28.64
`AGE:SEX`	30.17
`BMI:BP`	105.93

Table 5

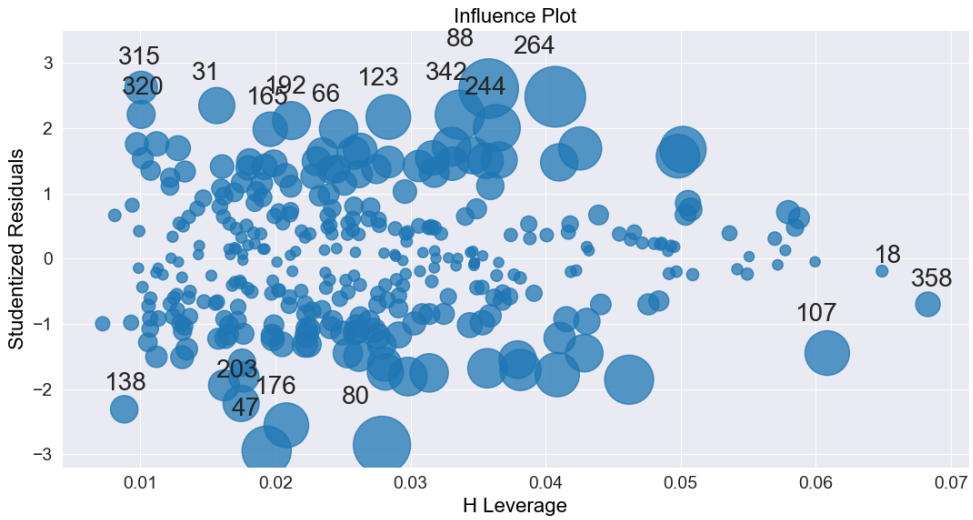

Figure 6

5 Model Selection

Model 4 has already been identified as the selected model, but it wasn't selected until after further comparison using the validation data set. After building all the candidate models, each model predicted the disease progression for each record in the validation data set. The root mean squared error (RMSE) for each model was calculated and compared to see how well the trained models could predict new data. The results are shown in Table 6. Model 4 had the lowest RMSE for both the training data (53.50) and for predicting the validation data (50.39). Both the training and validation RMSE are similar, indicating that the model is not over fit.

Model	$\text{RMSE}_{\text{Training}}$	$\text{RMSE}_{\text{Validation}}$
0	55.62	52.90
1	54.98	50.89
2	54.81	51.28
3	53.57	50.40
4	53.50	50.39
T0	56.29	54.24
T1	55.28	51.25
T2	55.36	51.96
T3	55.61	52.80
T4	54.18	51.11
T5	54.16	50.90

Table 6

6 Assessment of Final Model

The following relationships were significant enough to include in the final model. As mentioned it is difficult to infer information about parameters with high multicollinearity. Because of the high multicollinearity of this model, how the individual parameters vary together with the disease progression, may not be accurate. The discussion below is how the relationships would be interpreted, if multicollinearity were not an issue.

`AGE`: *Patient's age in years.* The predicted slope of the relationship between `AGE` and disease progression is -0.0695, indicating that for every year older a person is, the disease progression is expected to be 0.0695 units lower.

`BMI`: *Patient's body mass index.* The predicted slope of the relationship between `BMI` and disease progression is -6.64, indicating that for every unit increase in BMI, the disease progression is expected to be 6.64 units lower.

`BP`: *Patient's average blood pressure.* The predicted slope of the relationship between `BP` and disease progression is -2.08, indicating that for every unit increase in average blood pressure, the disease progression is expected to be 2.08 units lower.

`S1`: *Blood serum measurement 1.* The predicted slope of the relationship between `S1` and disease progression is -0.896, indicating that for every unit increase in `S1`, the disease progression is expected to be 0.896 units lower.

`S2`: *Blood serum measurement 2.* The predicted slope of the relationship between `S2` and disease progression is 0.747, indicating that for every unit increase in `S2`, the disease progression is expected to be 0.747 units higher.

`S5`: *Blood serum measurement 5.* The predicted slope of the relationship between `S5` and disease progression is 72.28, indicating that for every unit increase in `S5`, the disease progression is expected to be 72.28 units higher.

`SEX`: *Patient's sex, male or female (male used as true in the model).* The predicted slope of the relationship between `SEX` and disease progression is -98.26, indicating that if everything else is constant, the disease progression is expected to be 98.26 units lower for a male than a female.

`AGE:SEX`: *Interaction of the patient's age and sex.* The predicted slope of the relationship between `AGE:SEX` and disease progression is 1.52, indicating that for every year older a patient is, the disease progression is expected to be 1.52 units higher for males than for females.

`BMI:BP`: *Interaction of the patient's BMI and average blood pressure.* The predicted slope of the relationship between `BMI:BP` and disease progression is 0.125, indicating that for every unit increase in BMI, the disease progression is expected to be 0.125 units higher for each unit increase in average blood pressure than if there wasn't a change in BMI.