BIOS 663: Diabetes Investigation Final Report

Group 9: Joyce Choe, Joyce Liu, Meng Pan, Anthony Wang

Introduction

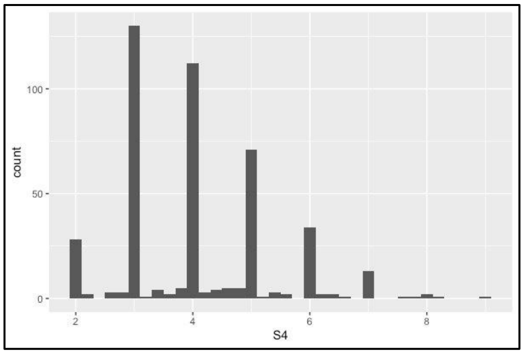
Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar) due to insufficient insulin, which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves (CDC, 2023). Early detection of diabetes progression is essential for secondary prevention. In the present study, we aim to determine the predictors that are most important for diabetes progression and associated with another.

Data

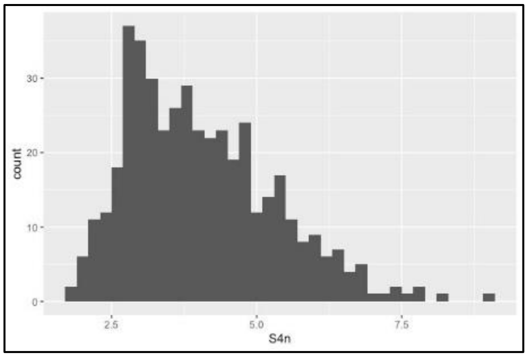
We obtained data from a diabetes study introduced in the 2004 article *Least Angle Regression* (Efron et al., 2004). The data set consists of 10 predictor variables at baseline and one response variable for 442 diabetes patients. There are no missing entries in the dataset. We provide the summary statistics of the variables in our dataset as follows:

Variable		Description	Mean (SD)	Range
AGE		Age in years	48.5 (13.11)	(19, 79)
SEX		Categorical, binary, de-identified, value = {1, 2}	1: 235 (53%) 2: 207 (47%)	
BMI		Body Mass Index	26.38 (4.42)	(18, 42.2)
BP		Average blood pressure (mmHg)	94.65 (13.83)	(62, 133)
Blood Serum Measurements S1-S6	**S1**. TC	Total serum cholesterol (mg/dL)	189.14 (34.61)	(97, 301)
	S2. LDL	Low-density lipoproteins (LDL) (mg/dL)	115.44 (30.41)	(41.6, 242.4)
	S3. HDL	High-density lipoproteins (HDL) (mg/dL)	49.79 (12.93)	(22, 99)
	S4. TCH	Total cholesterol/HDL ratio	4.05 (1.26)	(2, 9.1)
	S5. LTG	Log of serum triglycerides (log mg/dL)	4.64 (0.52)	(3.3, 6.1)
	S6. GLU	Blood glucose level (mg/dL)	91.26 (11.50)	(58, 124)
Y (response variable)		Quantitative measure of disease progression one year after baseline	152.13 (77.10)	(25, 346)

One caveat in our predictor variables is a potential measurement error associated with variable S4. In the dataset, most observations have integer values for S4, but the rest have fraction values for S4, as seen in the histogram of S4 on the left. We suspect this is likely a measurement error caused by the differing preference for rounding method for each health practitioners while recording the S4 value. To confirm our suspicion, we constructed a new S4 variable, namely S4n, based on the description of S4—ratio of total cholesterol over HDL—which corresponds to S1 divided S3 in our dataset. The resulting values of S4n are all fractions, which confirms our suspicion. As seen in the histogram of S4n, the distribution of S4n is smoother than S4 but roughly maintains the same shape. This is because the observations with integer values for S4 have spread out between integers for S4n. We decided to proceed with modeling using S4n in place of S4.



S4 values



S4n values

Additionally, we compared Pearson’s correlation coefficients between all potential predictors before fitting the model. In particular, S1 and S2 had a very high positive correlation (r = 0.897) and S3 and S4 had a high negative correlation (r = - 0.738). Therefore, we made sure to check for multicollinearity as we fit the models.

Methodology

Before fitting models, we mean-centered continuous predictor variables by group. This helped us interpret our model estimates for prediction. Also, instead of x=0, the intercept is a more realistic and meaningful value when at the predictor mean (Newsom, 2024). We chose our final model based on the following six steps:

1) Multiple linear regression: full model

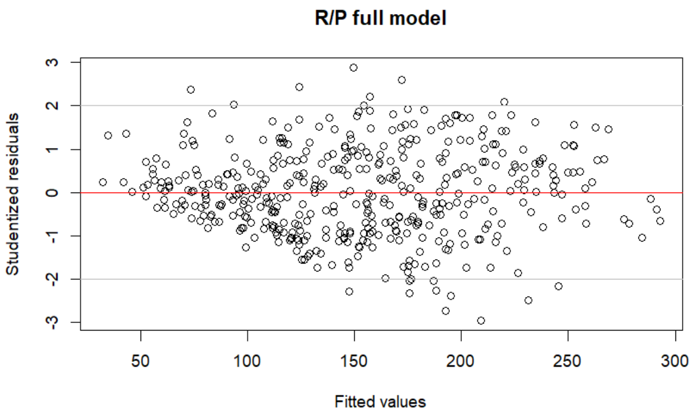
For the full model, we used quantitative measure of diabetes progression one year after baseline (Y) as the response variable and all predictor variables (age, BMI, BP, S1-S6, and sex) as independent variables and fitted a linear model (adj. $R^2 = 0.5065$, RMSE = 54.16).

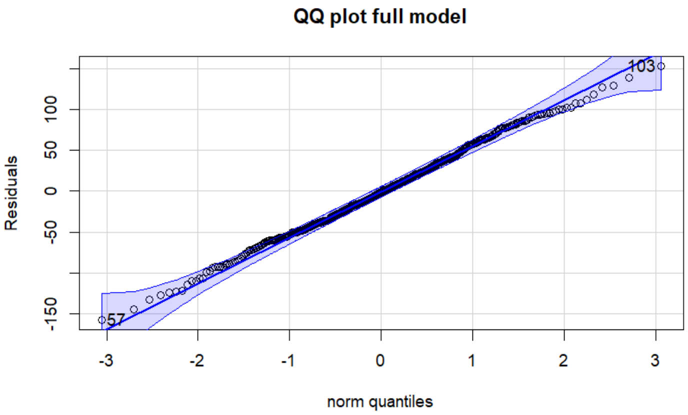
2) Diagnosis: multicollinearity and slight heterogeneity of error variance

The full model had the following variance inflation factors (VIFs) for the predictors:

```
      AGE       BMI        BP        S1        S2        S3        S4        S5 
 1.217854  1.509828  1.454147 63.717794 39.743676 19.427811 13.552361 10.120711 
       S6       SEX 
 1.485601  1.274206 
```

We noticed S1 and S2 had especially high VIFs (> 30) and S3, S4, and S5 exhibited VIFs greater than 10. This indicated a high level of multicollinearity. Hence, S1 and S2 variables were excluded from further modeling, and we would check VIFs for S3, S4, and S5 after fitting newer models.

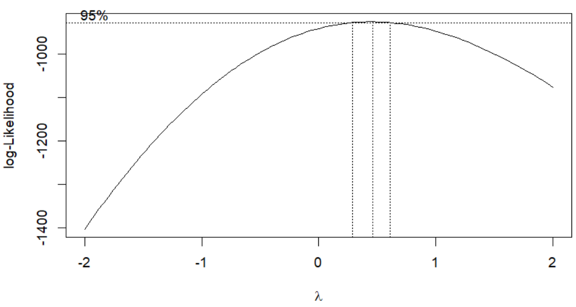


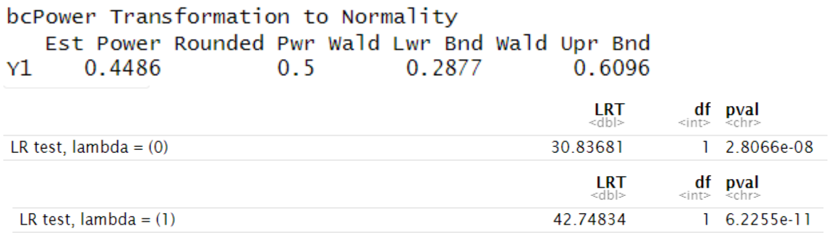


The full model residual plot shows minor violations of constant error variance because the errors are not evenly distributed about fitted values (50 to 150). However, the residual plot shows linearity and the QQ plot shows Gaussian errors.

3) Transformation

Since we wanted to improve error variance by minimizing sum of squared errors (SSE), we applied a Box-Cox transformation on Y and found the ideal transformation for Y was √Y (λ = 0.5). Additionally, the power transform test confirmed that we should reject the null hypotheses, λ = 0 or λ = 1, and transform our response variable. The sqrt transformation is known to stabilize the error variance as the variance is proportional to the mean. However, interpreting coefficients may be difficult after transformation.





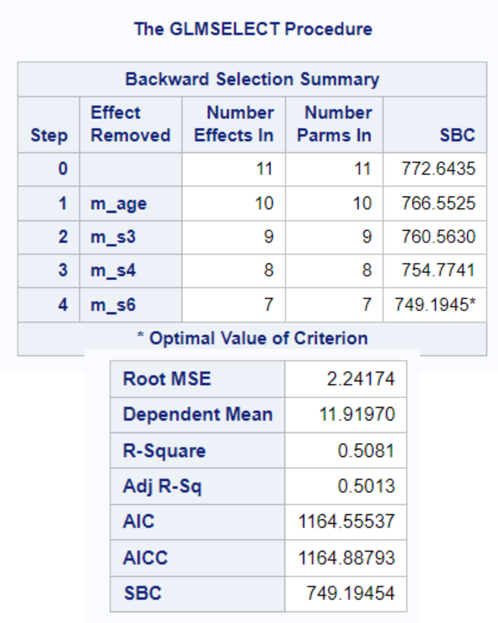
```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.4486          0.5       0.2877       0.6096

                                        LRT      df  pval
                                       <dbl>   <int> <chr>
LR test, lambda = (0)               30.83681       1 2.8066e-08

                                        LRT      df  pval
                                       <dbl>   <int> <chr>
 LR test, lambda = (1)              42.74834       1 6.2255e-11
```

4) Model selection

Our goal was to narrow down the most useful predictors for diabetes progression. As a result, we decided to select a model with few predictors. We used Schwarz Criterion (SBC) or Bayesian Information Criterion (BIC) as our model criteria measure because this statistic is determined by the number of predictors in a model and the model's SSE. SBC increases with larger models and increases as SSE increases, so we looked for the lowest SBC when choosing a model. We ran backwards selection, forward selection, and LASSO. The output for each model is shown below:



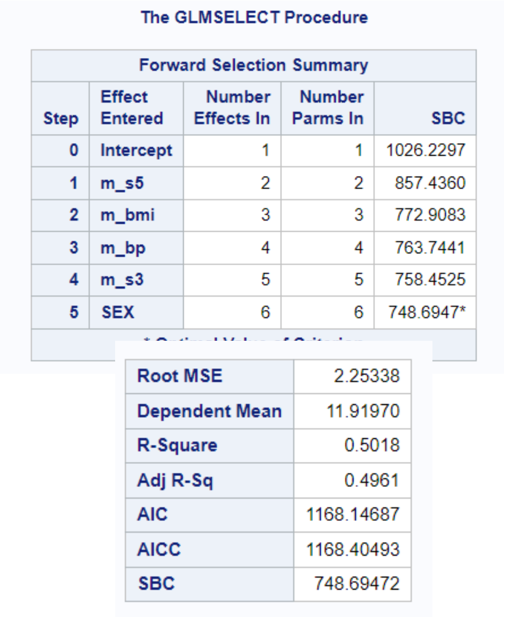
The GLMSELECT Procedure

Backward Selection Summary

Step	Effect Removed	Number Effects In	Number Parms In	SBC
0		11	11	772.6435
1	m_age	10	10	766.5525
2	m_s3	9	9	760.5630
3	m_s4	8	8	754.7741
4	m_s6	7	7	749.1945*

* Optimal Value of Criterion

Root MSE	2.24174
Dependent Mean	11.91970
R-Square	0.5081
Adj R-Sq	0.5013
AIC	1164.55537
AICC	1164.88793
SBC	749.19454



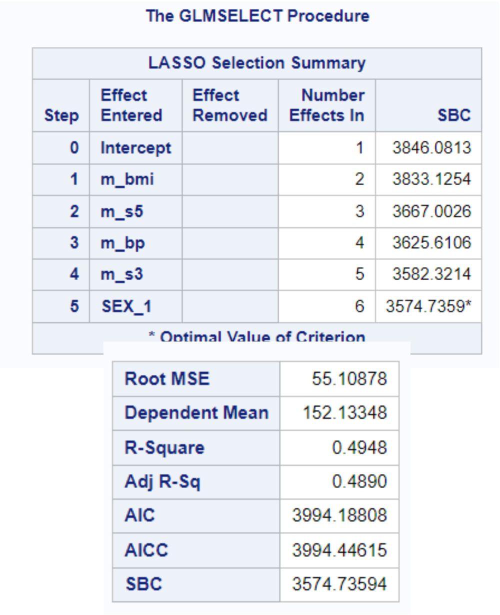
The GLMSELECT Procedure

Forward Selection Summary

Step	Effect Entered	Number Effects In	Number Parms In	SBC
0	Intercept	1	1	1026.2297
1	m_s5	2	2	857.4360
2	m_bmi	3	3	772.9083
3	m_bp	4	4	763.7441
4	m_s3	5	5	758.4525
5	SEX	6	6	748.6947*

* Optimal Value of Criterion

Root MSE	2.25338
Dependent Mean	11.91970
R-Square	0.5018
Adj R-Sq	0.4961
AIC	1168.14687
AICC	1168.40493
SBC	748.69472



The GLMSELECT Procedure

LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC
0	Intercept		1	3846.0813
1	m_bmi		2	3833.1254
2	m_s5		3	3667.0026
3	m_bp		4	3625.6106
4	m_s3		5	3582.3214
5	SEX_1		6	3574.7359*

* Optimal Value of Criterion

Root MSE	55.10878
Dependent Mean	152.13348
R-Square	0.4948
Adj R-Sq	0.4890
AIC	3994.18808
AICC	3994.44615
SBC	3574.73594

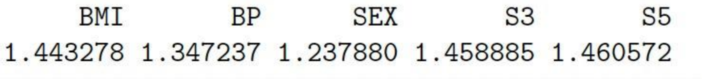
LASSO selection in SAS did not accept √Y since LASSO in SAS standardizes the values for us, so Y was added as the response variable. This explains a higher SBC in LASSO than in backward or forward selection, even though LASSO is recommended for multicollinear data because it attempts to minimize the MSE. Interestingly, forward and LASSO selected the same important predictors (BMI, sex, BP, S3, and S5). We chose forward selection for our final model because it had the lowest SBC. In addition, forward selection begins with the smallest possible model (intercept-only model) and builds on that using added-in-order tests to test a predictor's significance above and beyond preceding predictors already in the model. Using forward selection, we discovered a unique set of predictors that were important to contributing to changes in the response variable.

5) Interaction

We tested all potential interactions with five predictors in our model, but all of them had relatively low significance (Appendix Table 1). Since interaction effects were not as important as main effects, we excluded interaction terms to avoid data overfitting in the final model.

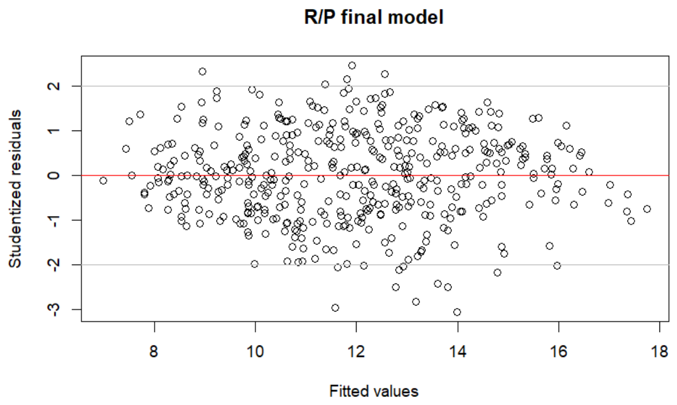
6) Diagnostic check of final model

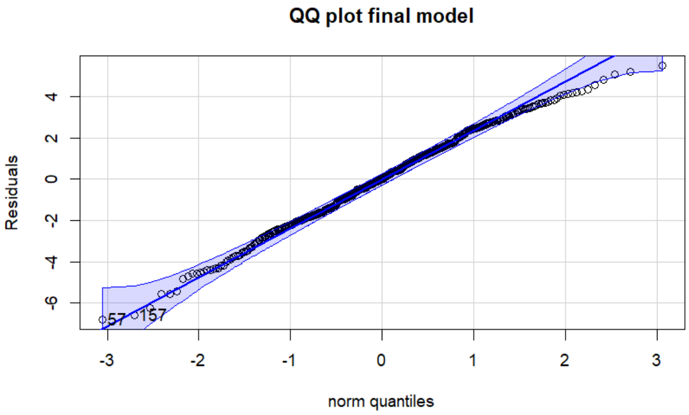
The VIFs for each predictor in the final model are shown below:

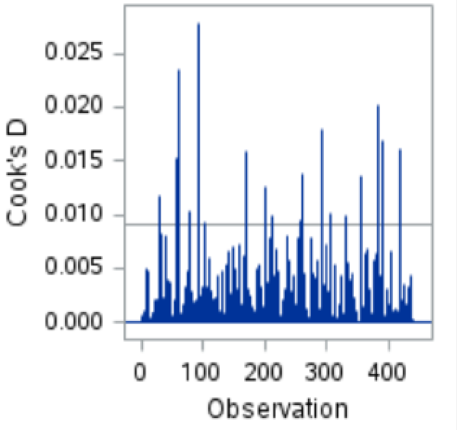


```
     BMI       BP      SEX       S3       S5 
1.443278 1.347237 1.237880 1.458885 1.460572 
```

Each predictor has a VIF much lower than 10, which indicates that multicollinearity is no longer an issue in the final model. The following residual plot for the final model also indicates homogeneity was met. In addition, the residual plot shows linearity and the QQ plot shows Gaussian errors. A side-by-side comparison of residual plots shows that the fanning pattern is less visible in the final model than in the full model. We also observed one less outlier in the final model than in the full model. (We define outliers in this report as residuals ± 2 standard deviations from the mean). For outliers in the final model, we tested their Cook's Distance to check their degree of influence on the model fit by measuring the standardized shift in $\hat{y}$ and $\hat{\beta}$ when that observation is deleted from the model. No outlier was excessively influential since none had a Cook's distance near or above 1.







Results

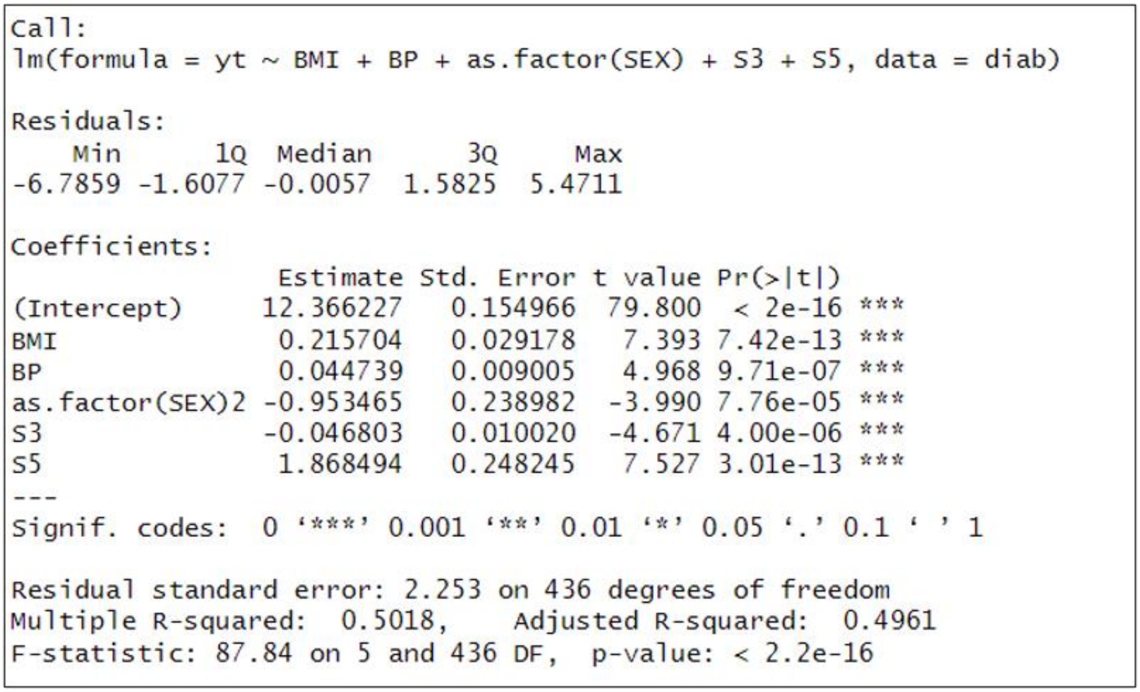
We determined that BMI, BP (blood pressure), sex, S3 (high-density lipoproteins), and S5 (triglyceride) were the most important predictors for diabetes progression. The final model met all regression assumptions, so the parameter estimates were defined as below:

$$\sqrt{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot BMI + \hat{\beta}_2 \cdot BP + \hat{\beta}_3 \cdot I[SEX = 2] + \hat{\beta}_4 \cdot S3 + \hat{\beta}_5 \cdot S5$$

$$\text{where } \hat{\beta}_0 = 12.3662; \hat{\beta}_1 = 0.2157; \hat{\beta}_2 = 0.0447; \hat{\beta}_3 = -0.9535; \hat{\beta}_4 = -0.0468; \hat{\beta}_5 = 1.8685$$

The model reduced to four continuous predictors and one categorical predictor. Fitting estimates gave an intercept that represents the mean diabetes progression value when all predictors are at their mean values if Y had not been transformed. However, we transformed Y to √Y in our model, and square root transformation does not allow linear or proportional relationships between variables. Instead, we interpreted estimates by positive or negative sign.

The R analysis output:



```
Call:
lm(formula = yt ~ BMI + BP + as.factor(SEX) + S3 + S5, data = diab)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7859 -1.6077 -0.0057  1.5825  5.4711 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     12.366227   0.154966  79.800  < 2e-16 ***
BMI              0.215704   0.029178   7.393 7.42e-13 ***
BP               0.044739   0.009005   4.968 9.71e-07 ***
as.factor(SEX)2 -0.953465   0.238982  -3.990 7.76e-05 ***
S3              -0.046803   0.010020  -4.671 4.00e-06 ***
S5               1.868494   0.248245   7.527 3.01e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.253 on 436 degrees of freedom
Multiple R-squared:  0.5018,	Adjusted R-squared:  0.4961 
F-statistic: 87.84 on 5 and 436 DF,  p-value: < 2.2e-16
```

Each parameter estimate resulted in a p-value of less than 0.001, which is significant at an alpha level of 0.005. This alpha level was calculated with the Bonferroni correction of 10 pairwise tests. Additionally, the adjusted R2 was 0.4961, so the model explained 49.61% of the variation in the data.

In general, we predict diabetes progression to increase on average for patients with a high BMI, high blood pressure, de-identified sex of 1, low HDL cholesterol, and high triglyceride levels one year after baseline. From this model, the most significant predictors of diabetes progression are BMI and triglyceride (t-value >7.3). Although there was high collinearity in S3 and S5 in the full model, we were surprised to see both variables without collinear issues in the final model. Reduced collinearity may be due to removal of confounder variables in the full model and variables with poor fit. Also, it makes sense that health measures like diet, weight, and blood pressure may be related to diabetes, but we did not expect sex to be related to diabetes as well.

Discussion and Conclusion

The results suggested that BMI, blood pressure, serum triglycerides are risk factors for diabetes progression, while female sex and high-density lipoproteins are protective variables of diabetes progression.

Our study has several limitations. Firstly, some of the predictor variables are not thoroughly documented in our data source. The sex variable was deidentified, so we assumed 1 represented male and 2 represented female in the analysis. Similarly, the response variable Y was only described as a measure of diabetes progression and we assumed higher values signified greater disease progression (i.e., worse health). If either of the assumption of variables was not met, our interpretation of findings would have to change accordingly. Secondly, we fitted the model based on a relatively small sample size, some of the predictors would be accidentally included or excluded using the forward selection method based on the p-values. Thirdly, there are potential unmeasured variables outside of our data that can be significant predictors for diabetes disease progression, such as race, smoking status, dietary factors, and family history of diabetes. Lastly, our study result may have limited external validity, which means the final model we proposed may not be fit for other populations of diabetes patients that display distinct characteristics from our study population.

To address these limitations, more data needs to be collected from diabetes patients with diverse range of demographics and biomarkers, so we can obtain a potentially more accurate statistical model and properly assess adaptability of the use of our model. Additionally, interviews with clinicians and medical practitioners can provide us with domain knowledge to inform our model selection process, which can be helpful improving the interpretability of our disease progression model in secondary diabetes prevention.

Reference

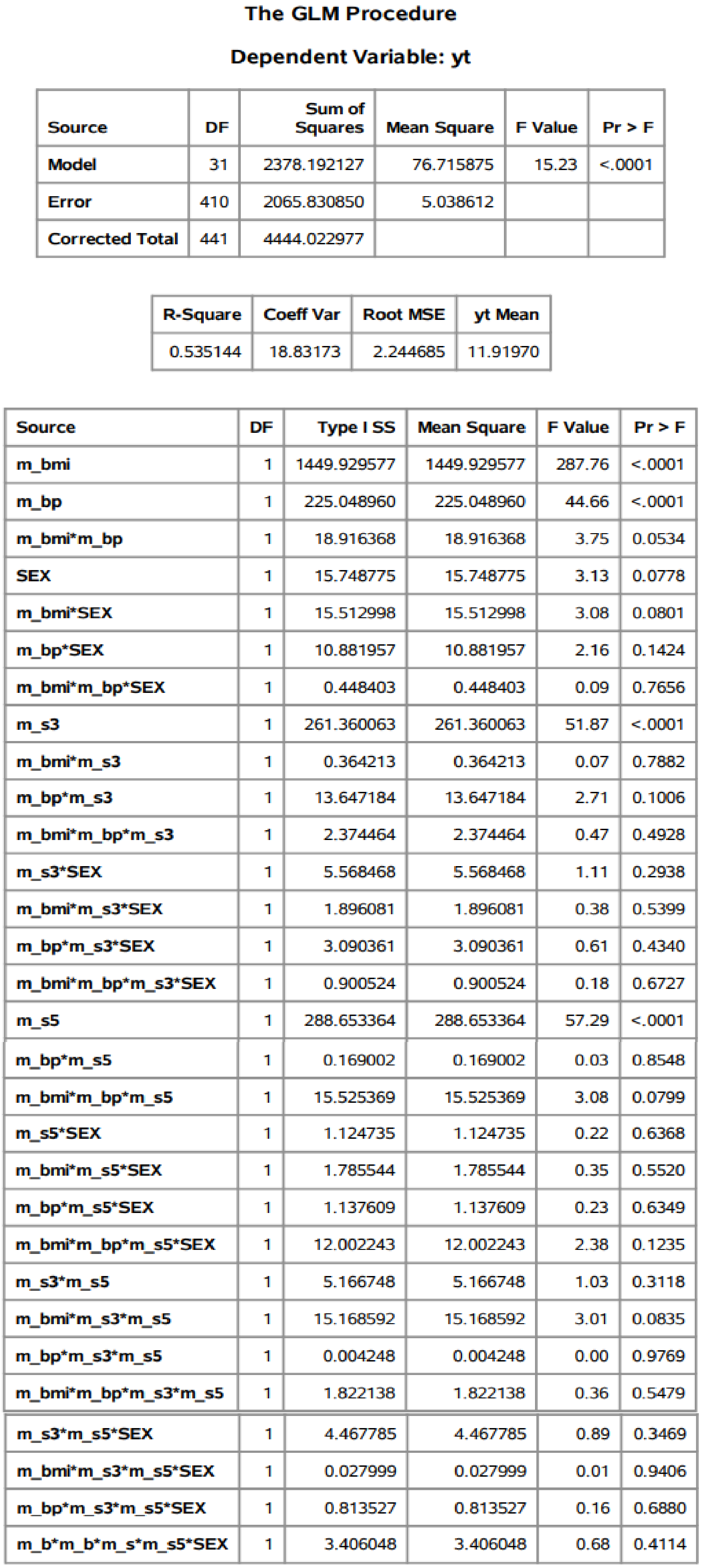
Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani "Least angle regression," The Annals of Statistics, Ann. Statist. 32(2), 407-499, (April 2004)

Centers for Disease Control and Prevention. (2023, September 5). What is diabetes?. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/diabetes.html

Newsom, J. T. (2024). Centering in multilevel regression.
https://web.pdx.edu/~newsomj/mlrclass/ho_centering.pdf

Appendix

Table 1: Results from Interaction tests



The GLM Procedure

Dependent Variable: yt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	31	2378.192127	76.715875	15.23	<.0001
Error	410	2065.830850	5.038612		
Corrected Total	441	4444.022977			

R-Square	Coeff Var	Root MSE	yt Mean
0.535144	18.83173	2.244685	11.91970

Source	DF	Type I SS	Mean Square	F Value	Pr > F
m_bmi	1	1449.929577	1449.929577	287.76	<.0001
m_bp	1	225.048960	225.048960	44.66	<.0001
m_bmi*m_bp	1	18.916368	18.916368	3.75	0.0534
SEX	1	15.748775	15.748775	3.13	0.0778
m_bmi*SEX	1	15.512998	15.512998	3.08	0.0801
m_bp*SEX	1	10.881957	10.881957	2.16	0.1424
m_bmi*m_bp*SEX	1	0.448403	0.448403	0.09	0.7656
m_s3	1	261.360063	261.360063	51.87	<.0001
m_bmi*m_s3	1	0.364213	0.364213	0.07	0.7882
m_bp*m_s3	1	13.647184	13.647184	2.71	0.1006
m_bmi*m_bp*m_s3	1	2.374464	2.374464	0.47	0.4928
m_s3*SEX	1	5.568468	5.568468	1.11	0.2938
m_bmi*m_s3*SEX	1	1.896081	1.896081	0.38	0.5399
m_bp*m_s3*SEX	1	3.090361	3.090361	0.61	0.4340
m_bmi*m_bp*m_s3*SEX	1	0.900524	0.900524	0.18	0.6727
m_s5	1	288.653364	288.653364	57.29	<.0001
m_bp*m_s5	1	0.169002	0.169002	0.03	0.8548
m_bmi*m_bp*m_s5	1	15.525369	15.525369	3.08	0.0799
m_s5*SEX	1	1.124735	1.124735	0.22	0.6368
m_bmi*m_s5*SEX	1	1.785544	1.785544	0.35	0.5520
m_bp*m_s5*SEX	1	1.137609	1.137609	0.23	0.6349
m_bmi*m_bp*m_s5*SEX	1	12.002243	12.002243	2.38	0.1235
m_s3*m_s5	1	5.166748	5.166748	1.03	0.3118
m_bmi*m_s3*m_s5	1	15.168592	15.168592	3.01	0.0835
m_bp*m_s3*m_s5	1	0.004248	0.004248	0.00	0.9769
m_bmi*m_bp*m_s3*m_s5	1	1.822138	1.822138	0.36	0.5479
m_s3*m_s5*SEX	1	4.467785	4.467785	0.89	0.3469
m_bmi*m_s3*m_s5*SEX	1	0.027999	0.027999	0.01	0.9406
m_bp*m_s3*m_s5*SEX	1	0.813527	0.813527	0.16	0.6880
m_b*m_b*m_s*m_s5*SEX	1	3.406048	3.406048	0.68	0.4114