STAT 512: Applied Regression Analysis
Final Project Report

Due on Friday, Dec 8, 2017

Clint Alfaro, Jingjing Guo and Erin Tooley

Contents

Part I

0. Overview

Liver steatosis (LS) is a fatty liver disease that causes substantial liver inflammation and can eventually lead to tissue death and liver failure. The condition is reversible in its early stages. Therefore, the early and minimally invasive detection of LS would significantly improve ability to treat the disease before it progresses to more significant stages. In this dataset, the investigators in Wu et al. collected a range of variables, including metabolic panels and blood lipid composition, clinical measurements such as weight and height, and liver ultrasound imaging, to assess whether a collection of these variables is useful in predicting whether a patient is suffering from LS [1].

The current gold standard method for diagnosis of LS is liver biopsy followed by histopathology. Examination of thin stained tissue sections reveals the characteristic tissue morphology changes caused by the fatty liver disease. Lipid accumulation, primarily in the form of lipid droplets within the cellular cytoplasm, is ready visualized with light microscopy.(2) The response variable in the data set is the outcome of the liver biopsy test that was performed on the human subjects enrolled in the Wu et al. study. This variable, LS_Bi, takes values of 0 (negative for LS) , 0.5 (inconclusive test), and 1 (positive for LS). The variables nonalcoholic fatty liver disease activity score (NAS) as well as Fibrosis are also determined using liver biopsies.(3) All of the variables are listed in the table below. As far as building a useful model in the context of the clinical problem of diagnosing fatty liver diseases, the best model should be able to predict the outcome of the liver biopsy test using only variables that can be obtained without a liver biopsy. Biopsies are routine procedures but come with risk of complications such as infection, adverse reactions to anesthesia, and liver damage. Therefore, Part II of the project will focus on predicting the outcome of LS_Bi using the clinical, blood test, and imaging data.

Table 1: List of Variables from the Liver Steatosis Dataset

Variable name	How	Type	Codes, Units, Values	Abbreviation
Age	Clinical	Continuous	Years	Age
Gender	Clinical	Categorical	1 =male, 2=female	Sex
Height	Clinical	Continuous	cm	Height
Weight	Clinical	Continuous	kg	Weight
Body mass index	Clinical	Continuous	kg/mg2	BMI
Duration of obesity	Clinical	Continuous	years	Obes
Diabetes	Clinical	Categorical	0 = absent, 2 = present	DM
Metabolic syndrome	Clinical	Categorical	0 = absent, 2 = present	Met
Hypertension	Clinical	Categorical	0 = absent, 2 = present	HTN
Hyperlipidemia	Blood test	Categorical	0 = absent, 2 = present	HPL
Plasma triglycerides	Blood test	Continuous	%	TG
Cholesterol	Blood test	Continuous	mg/dL	CHOL
High-density lipoprotein cholesterol	Blood test	Continuous	mg/dL	HDL
Low-density lipoprotein cholesterol	Blood test	Continuous	mg/dL	LDL
Very-low-density lipoprotein cholesterol	Blood test	Continuous	mg/dL	VDL
Aspartate aminotransferase	Blood test	Continuous	U/L	AST
Alanine aminotransferase	Blood test	Continuous	U/L	ALT
Nonalcoholic fatty liver disease activity score	Liver biopsy	Categorical	Range from 0 to 8	NAS
Fibrosis	Liver biopsy	Categorical	0 = none	Fibrosis Fibrosis
			1 = perisinusoidal or periportal	
			2 = perisinusoudal and portal/periportal	
			3 = briding fibrosis	
			4 = cirrhosis	
Positive liver steatosis by ultrasound	Imaging	Categorical	0 = negative	LS_US
			0.5 = inconclusive	
			1 = positive	
Positive liver steatosis by biopsy	Liver Biopsy	Categorical	0 = negative	LS_Bi
			0.5 = inconclusive	
			1 = positive	

1. Piecewise simple linear regression of LS_Bi vs VDL

The plot for LS_Bi vs. VDL is shown with a smooth curve (Figure 1). It appears that there are two regions in the curve that could be modeled using piecewise regression. The inflection point occurs at around VDL = 25.

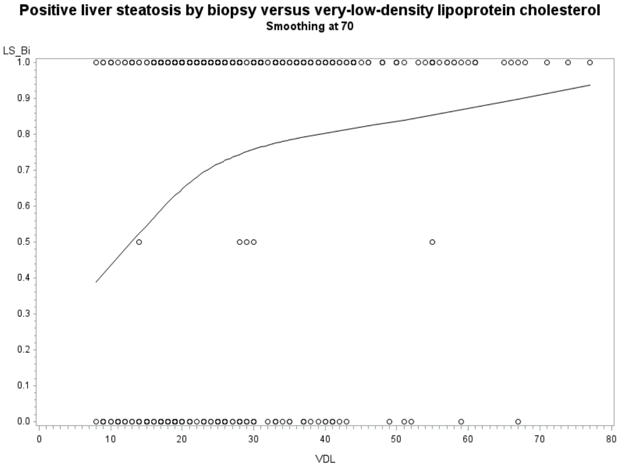


Figure 1: Plot of LS_Bi vs. VDL with a smooth curve. A curve is noted in the graph.

The piecewise regression plot is provided below with two different lines used in the model (Figure 2). There is one slope from VDL =5 to VDL = 25, and a different slope for VDL ¿ 25.

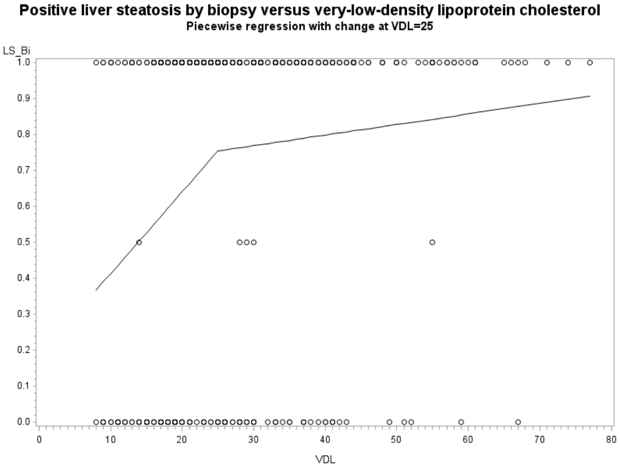


Figure 2: Piecewise regression plot for modeling relationship between LS_Bi and VDL

The model used is:

$$\begin{aligned}
LS\_Bi &= b_0 + b_1 \cdot (VDL) + b_2 X_3 (VDL - 25) \\
LS\_Bi &= b_0 + b_1 \cdot (VDL) + b_2 X_2 (VDL - 25) \\
&= b_0 - 25 b_2 X_2 + b_1 (VDL) + b_2 X_2 \cdot VDL \\
&= b_0 + b_1 X_1 X_2 = 0 (X_1 \leq 25) \\
&= (b_0 - 25 \cdot b_2 X_2) + (b_2 + b_1) \cdot (VDL) X_2 = 0 (X_1 > 25)
\end{aligned}$$

The test shows that the lines are different. The hypotheses are:

$$H_0 : b_1 = b_3 = 0$$
$$H_a : b_1 \text{ or } b_3 \text{ are not zero}$$

Table 2: Test sameline Results for Dependent Variable LS_Bi

Source	DF	Mean square	F Value	$Pr > F$
Numerator	2	2.77694	14.41	$< .0001$
Denominator	390	0.19272		

The F-statistic is 14.41. The numerator $df = 2$, denominator $df = 390$. The critical F-value is 3.087 with $\alpha = 0.05$. The corresponding p-value is very small (< 0.0001). There is sufficient evidence to reject the null hypothesis and conclude that the two lines are not the same for the piecewise regression. This is also consistent with the visual observation made from the piecewise regression plot shown above.

2. Extra Sums of Squares

To better investigate implications of the extra sums of squares and the effects of changing model variables, a variable named SUM was created. This variable was the summation of the predictors ALT (alanine aminotransferase) and AST (aspartate aminotransferase). One reason behind selecting these two specific variables to sum is that they display a degree of correlation.This makes sense as they are both measurements of enzyme activity, and are both aminotransferase enzymes at that. The units for the two are also the same.

In order to test the effects of the new variable SUM, linear regression was run on the model of all explanatory variables including SUM and the model of all explanatory variables except for SUM. Note that neither test included SUM's component variables ALT and AST as their addition would be redundant. To better illustrate the effects, the model sum of squares was calculated based on the ANOVA results of the two regression models. This value was then used to calculate the F-statistic to test the null hypothesis that the coefficient for SUM equaled zero. The alternative hypothesis was that it did not equal zero.
Let $X_1, ..., X_n$ represent all explanatory variables excluding AST, ALT, and SUM.

$$SSM(SUM|X_1, ..., X_n) = SSE(R) - SSE(F)$$
$$SSM(SUM|X_1, ..., X_n) = SSE(X_1, ..., X_n) - SSE(SUM, X_1, ..., X_n)$$
$$SSM(SUM|X_1, ..., X_n) = 31.85979 - 31.68680 = \boxed{0.17299}$$

$$F = \frac{SSM/(df_E(R) - df_E(F))}{SSE(F)/df_E(F)} = \frac{0.17299/(334 - 333)}{31.6868/333} = \boxed{1.81797}$$

The "test" statement was added to the regression statement in SAS to confirm the result. The conclusion drawn from these test statistics was to accept the null hypothesis. Running these models indicated that the inclusion of the variable SUM as a predictor did not improve the model.

3. Type I and II Sums of Squares

The regression process was again run in SAS, this time with the SS1 and SS2 options included in order to obtain the type I and type II sums of squares, respectively. The model included all explanatory variables except for SUM. The order of the variables was the same as they were in the original dataset.

For all predictors but LS_US, the sums of squares for type I and type II did not equal each other. For LS_US the reason for the two sums being equal was because they were the last calculated, therefore not giving any meaningful interpretation. The sum of the type I sum of squares was 33.47034. The sum of the type II sum of squares was 20.99982. The model sum of squares equaled that of the total summation of the type I sum of squares. This is because the type I sum of squares is calculated in a hierarchal method, each calculation an adjustment of error based on each additional predictor. Because the regression model used did not deviate from the original dataset in its variable order, the calculated sum of squares also followed the same order as type I was calculated, due to the model sum of squares being equal to the reduced model SSE minus the full model SSE.

4. Multiple linear regression models with different combinations of explanatory variables

A set of 16 models were run in SAS. The model DF (number of coefficents not including the intercept), the model statement, R^2, and adjusted R^2 are reported. The models are sorted in ascending order of R^2 adjusted. The R^2 adjusted ranges from 0.0689 to 0.5163, which shows that the the selection of explanatory variables significantly changes the amount of variation in the response variable, LS_Bi, that can be explained by the linear regression model.

Table 3: List of Variables from the Liver Steatosis Dataset

DF_M	Model	R^2	R^2_{Adj}
6	LS_Bi=Age Sex Height Weight BMI SUM;	0.0848	0.0689
5	LS_Bi=Age Sex Height Weight SUM;	0.0829	0.0697
16	LS_Bi=Age Sex Height Weight BMI Obes DM Met HTN HPL TG CHOL HDL LDL VDL SUM;	0.1519	0.1115
17	LS_Bi=Age Sex Height Weight BMI Obes DM Met HTN HPL TG CHOL HDL LDL VDL AST ALT SUM;	0.1654	0.123
6	LS_Bi=Age Sex Height Weight SUM LS_US;	0.3121	0.302
18	LS_Bi=Age Sex Height Weight BMI Obes DM Met HTN HPL TG CHOL HDL LDL VDL NAS LS_US SUM;	0.5158	0.4897
14	LS_Bi=Age Sex Height Weight BMI Obes DM Met HTN HPL NAS Fibrosis LS_US SUM;	0.5245	0.5048
14	LS_Bi=Age Sex Height Weight BMI Obes DM Met HTN HPL NAS Fibrosis LS_US SUM;	0.5245	0.5048
16	LS_Bi=BMI Obes DM Met HTN HPL TG CHOL HDL LDL VDL AST ALT NAS Fibrosis LS_US SUM;	0.5308	0.5085
15	LS_Bi=BMI Obes DM Met HTN HPL TG CHOL HDL LDL VDL NAS Fibrosis LS_US SUM;	0.5296	0.5087
15	LS_Bi=BMI Obes DM Met HTN HPL TG CHOL HDL LDL VDL NAS Fibrosis LS_US SUM;	0.5296	0.5087
16	LS_Bi=Age Sex Height Weight BMI Obes DM Met HTN HPL TG CHOL NAS Fibrosis LS_US SUM;	0.5349	0.5128
18	LS_Bi=Age Sex Height Weight BMI Obes DM Met HTN HPL TG HDL LDL VDL NAS Fibrosis LS_US SUM;	0.539	0.5142
19	LS_Bi=Age Sex Height Weight BMI Obes DM Met HTN HPL TG CHOL HDL LDL VDL NAS Fibrosis LS_US SUM;	0.5406	0.5143
17	LS_Bi=Age Sex Height Obes DM Met HTN HPL TG CHOL HDL LDL VDL NAS Fibrosis LS_US SUM;	0.5383	0.5149
16	LS_Bi=Age Sex Height DM Met HTN HPL TG CHOL HDL LDL VDL NAS Fibrosis LS_US SUM;	0.5383	0.5163

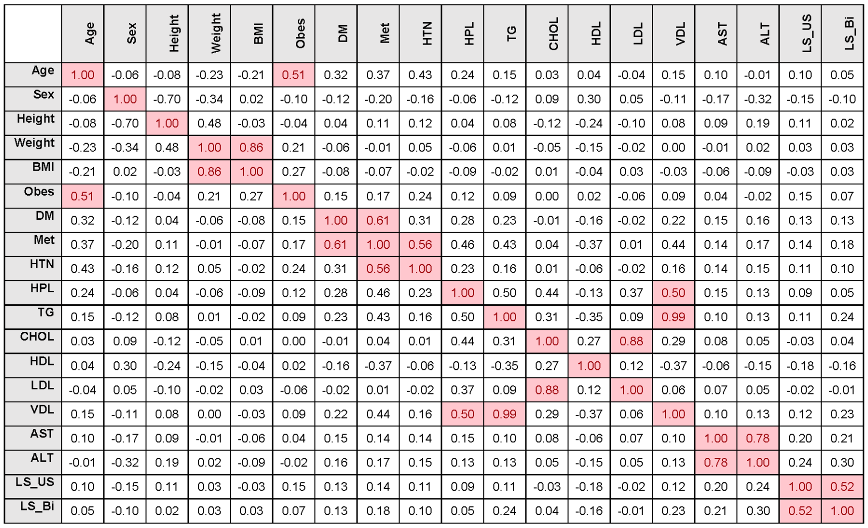
Part II

0. Overview

In Part II, we use techniques taught in class to diagnose any issues with the model, remedy these issues and select the best model. The proposed objective is to develop a model that uses non-invasive predictors to predict liver steatosis. NAS and Fibrosis are obtained from liver biopsy, so they were dropped from further analysis. The following model selection and analysis are based on the remaining 19 variables.

1. Scatterplot and correlation matrix

Figure 3 shows the correlation matrix of the 19 variables in consideration. The response variable LS_Bi is the last row and column on the correlation matrix. We also highlighted Pearson correlation equal or higher than 0.5 in red. Similarly the scatterplot matrix is 19 × 19 and it is shown in Figure 4.



	Age	Sex	Height	Weight	BMI	Obes	DM	Met	HTN	HPL	TG	CHOL	HDL	LDL	VDL	AST	ALT	LS_US	LS_Bi
Age	1.00	-0.06	-0.08	-0.23	-0.21	0.51	0.32	0.37	0.43	0.24	0.15	0.03	0.04	-0.04	0.15	0.10	-0.01	0.10	0.05
Sex	-0.06	1.00	-0.70	-0.34	0.02	-0.10	-0.12	-0.20	-0.16	-0.06	-0.12	0.09	0.30	0.05	-0.11	-0.17	-0.32	-0.15	-0.10
Height	-0.08	-0.70	1.00	0.48	-0.03	-0.04	0.04	0.11	0.12	0.04	0.08	-0.12	-0.24	-0.10	0.08	0.09	0.19	0.11	0.02
Weight	-0.23	-0.34	0.48	1.00	0.86	0.21	-0.06	-0.01	0.05	-0.06	0.01	-0.05	-0.15	-0.02	0.00	-0.01	0.02	0.03	0.03
BMI	-0.21	0.02	-0.03	0.86	1.00	0.27	-0.08	-0.07	-0.02	-0.09	-0.02	0.01	-0.04	0.03	-0.03	-0.06	-0.09	-0.03	0.03
Obes	0.51	-0.10	-0.04	0.21	0.27	1.00	0.15	0.17	0.24	0.12	0.09	0.00	0.02	-0.06	0.09	0.04	-0.02	0.15	0.07
DM	0.32	-0.12	0.04	-0.06	-0.08	0.15	1.00	0.61	0.31	0.28	0.23	-0.01	-0.16	-0.02	0.22	0.15	0.16	0.13	0.13
Met	0.37	-0.20	0.11	-0.01	-0.07	0.17	0.61	1.00	0.56	0.46	0.43	0.04	-0.37	0.01	0.44	0.14	0.17	0.14	0.18
HTN	0.43	-0.16	0.12	0.05	-0.02	0.24	0.31	0.56	1.00	0.23	0.16	0.01	-0.06	-0.02	0.16	0.14	0.15	0.11	0.10
HPL	0.24	-0.06	0.04	-0.06	-0.09	0.12	0.28	0.46	0.23	1.00	0.50	0.44	-0.13	0.37	0.50	0.15	0.13	0.09	0.05
TG	0.15	-0.12	0.08	0.01	-0.02	0.09	0.23	0.43	0.16	0.50	1.00	0.31	-0.35	0.09	0.99	0.10	0.13	0.11	0.24
CHOL	0.03	0.09	-0.12	-0.05	0.01	0.00	-0.01	0.04	0.01	0.44	0.31	1.00	0.27	0.88	0.29	0.08	0.05	-0.03	0.04
HDL	0.04	0.30	-0.24	-0.15	-0.04	0.02	-0.16	-0.37	-0.06	-0.13	-0.35	0.27	1.00	0.12	-0.37	-0.06	-0.15	-0.18	-0.16
LDL	-0.04	0.05	-0.10	-0.02	0.03	-0.06	-0.02	0.01	-0.02	0.37	0.09	0.88	0.12	1.00	0.06	0.07	0.05	-0.02	-0.01
VDL	0.15	-0.11	0.08	0.00	-0.03	0.09	0.22	0.44	0.16	0.50	0.99	0.29	-0.37	0.06	1.00	0.10	0.13	0.12	0.23
AST	0.10	-0.17	0.09	-0.01	-0.06	0.04	0.15	0.14	0.14	0.15	0.10	0.08	-0.06	0.07	0.10	1.00	0.78	0.20	0.21
ALT	-0.01	-0.32	0.19	0.02	-0.09	-0.02	0.16	0.17	0.15	0.13	0.13	0.05	-0.15	0.05	0.13	0.78	1.00	0.24	0.30
LS_US	0.10	-0.15	0.11	0.03	-0.03	0.15	0.13	0.14	0.11	0.09	0.11	-0.03	-0.18	-0.02	0.12	0.20	0.24	1.00	0.52
LS_Bi	0.05	-0.10	0.02	0.03	0.03	0.07	0.13	0.18	0.10	0.05	0.24	0.04	-0.16	-0.01	0.23	0.21	0.30	0.52	1.00

Figure 3: Correlation Matrix of Full Model Variables

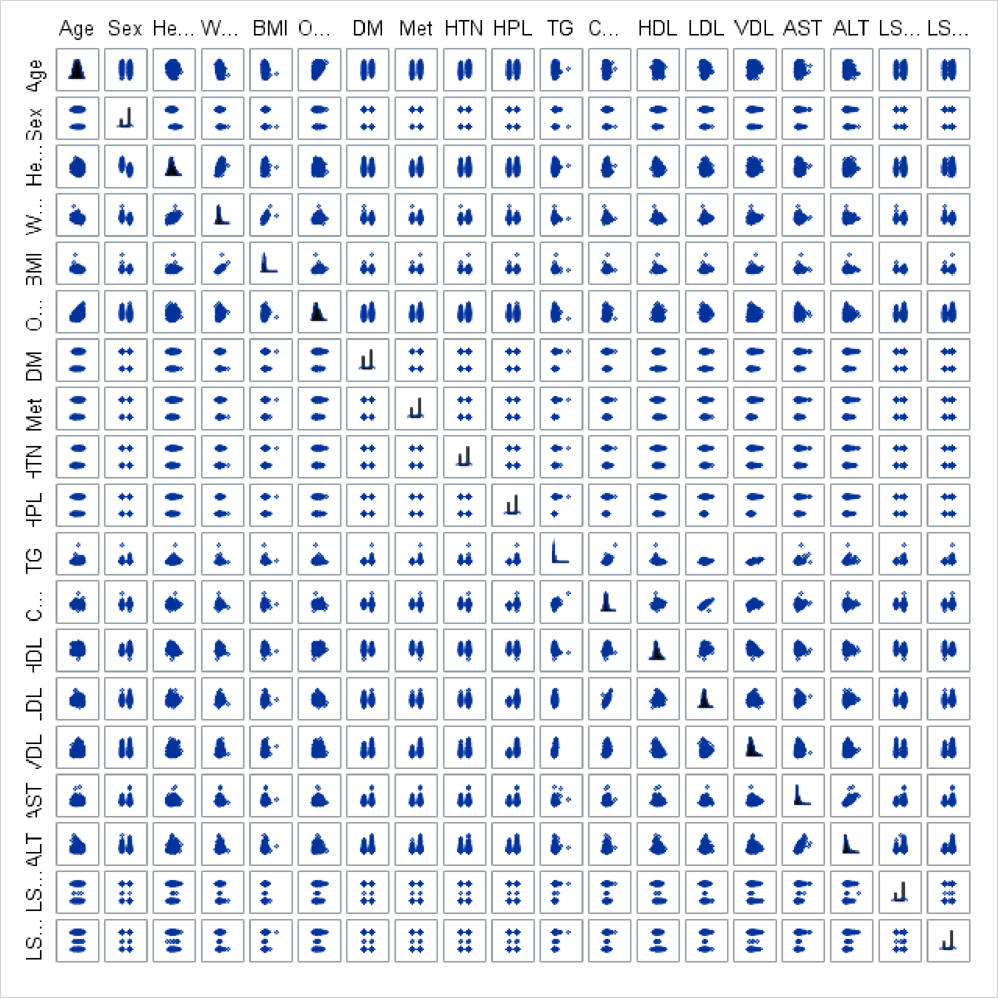


Figure 4: Scatter Plots of All Variables

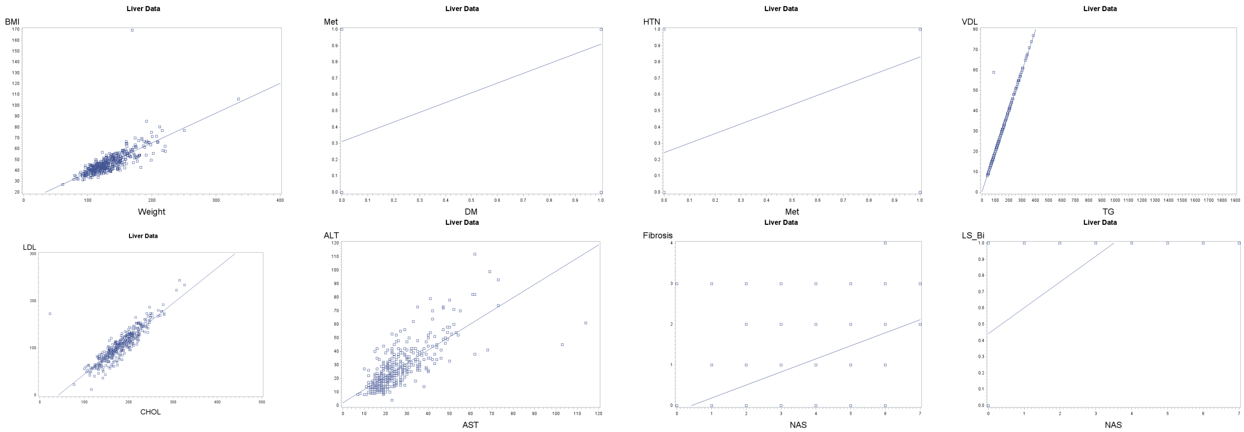


Figure 5: Scatter Plots of Correlated Variables

As is shown in the preceding plots, most variables are weakly correlated except for the follow pairs:

- BMI is highly correlated to Weight with Pearson correlation at 0.86.
- Met and DM are correlated, with Pearson correlation at 0.61.
- HTN and Met are moderately correlated with Pearson correlation at 0.56.
- VDL and TG are almost perfectly correlated with Pearson correlation at 0.99.
- LDL and CHOL are strongly correlated with Pearson correlation at 0.88.
- ALT and AST are correlated with Pearson correlation at 0.78.
- Fibrosis and NAS are correlated with Pearson correlation at 0.64.
- LS_Bi and NAS are correlated with Pearson correlation at 0.63.

Correlation in explanatory variables will introduce multicolinearity to the model, and correlation on these pairs will be closesly examined in the model selection.
Additionally, it can also be noted that only LS_US which is ultrasonic imagining of the liver is moderately correlated to the response variable. The R^2 is not expected to be high for the linear regression model.

For clarity, we recreated scatter plots pairs of variables with moderate or strong correlations, as are shown in Figure 5.

2. Transformations

We use the scatter plots and residual plots to check for possible transformation needed for our variables. Scatter plots are shown in Figure 4 and Table 5 and the residual plots are shown in Figure 6.

As is shown, there is no nonlinear relationships between the residuals and the explanatory variables that can need to be corrected using transformation. It is apparent that the residuals are not constant as a function of the explanatory variables as evident in Figure 5. This necessiates transformation of the response variable, LS_Bi. A Box-Cox procedure is applied and the best power tranformation returned from the SAS analysis is $\lambda = 0.5$, shown in Figure 7.

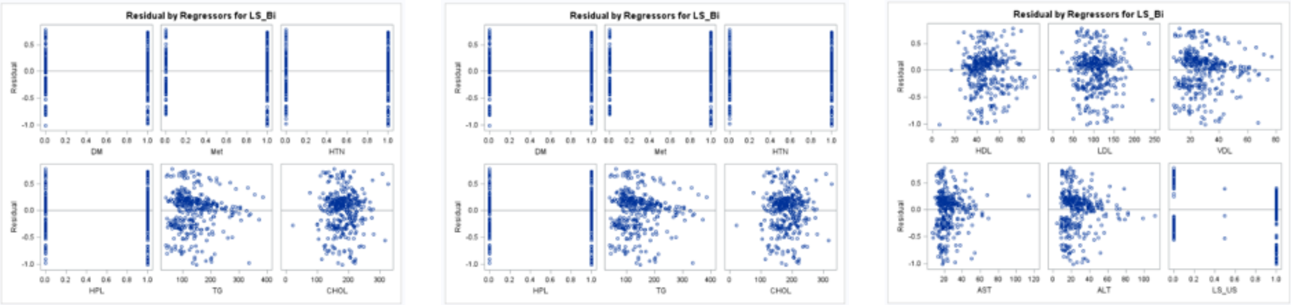


Figure 6: Scatter Plots of Correlated Variables

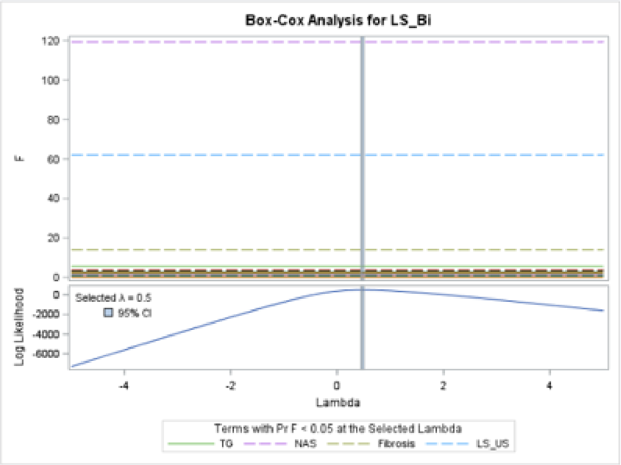


Figure 7: Box Cox Transformation of LS_Bi

We apply this power transformation of Y and test for quality of the model. Note that since the response variable LS_Bi is categorical and takes only 0 or 1, we increase this value by 1 to avoid 0 denominator in box-cox Transformation.

To determine the model quality, we use two measures: R^2 improvement and misclassification rate. The two measures for full model before and after the transformation is shown in the Table 4.

Note we use 0.5 as the threshold for classifying the model predicted $\widehat{LS\_Bi}$, i.e. if $\widehat{LS\_Bi} > 0.5$, the observation is classified as "liver steatosis positive", and otherwise, negative. This classification if contradicting with the observed LS_Bi value, the observation will be counted toward misclassification. Therefore misclassificaiton rate is the percentage of unmatched model classification with observed LS_Bi classification over the total observations. Observation with any missing values are not counted.

Table 4: R^2 and Misclassification Rate Before and After Y transformation

	Before	After
R^2	0.3644	0.3706
Misclassification Rate	0.188	0.172

The residual plots for the transformed response variable is shown in Figure 8. The plots indicate the little improves were made to the non-constant variance issue.

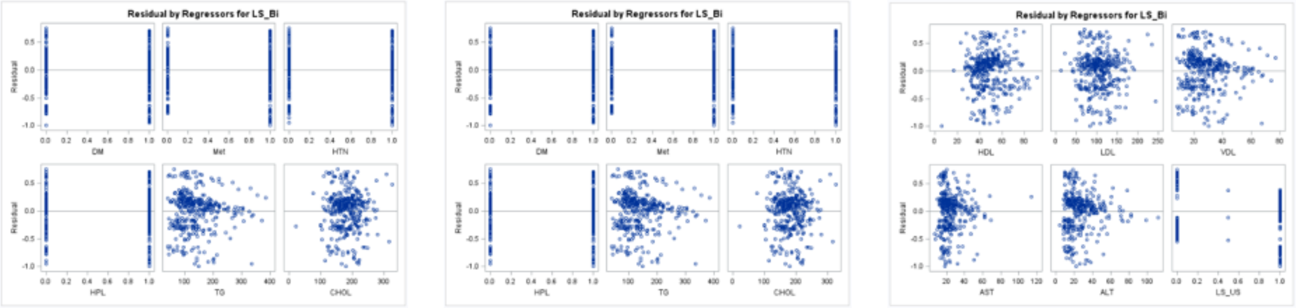


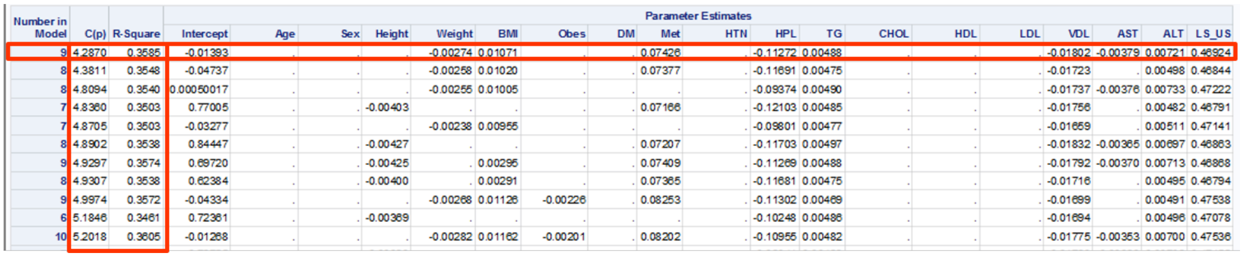
Figure 8: Scatter Plots of Correlated Variables

Based on the little improvement and in favor of simpler model, we concluded no transformatoin is to be used for the response variable LS_Bi.

3. C_p for Best Set of Variables

The Cp criterion was used to select a regression model that explains the most variance in the response variable with the fewest number of explanatory variables. We conduct a model selection based on C_p, selecting the smallest model for $C_p \leq p$, which gives the best model with 9 variables: Weight, BMI, Met, HPL, TG, VDL, AST, ALT and LS_US, as is shown in Table 5.

Table 5: Model Selection Based on the C_p Criteria



Number in Model	C(p)	R-Square	Parameter Estimates																		
			Intercept	Age	Sex	Height	Weight	BMI	Obes	DM	Met	HTN	HPL	TG	CHOL	HDL	LDL	VDL	AST	ALT	LS_US
9	4.2870	0.3585	-0.01393	.	.	.	-0.00274	0.01071	.	.	0.07426	.	-0.11272	0.00488	.	.	.	-0.01802	-0.00379	0.00721	0.46924
8	4.3811	0.3548	-0.04737	.	.	.	-0.00258	0.01020	.	.	0.07377	.	-0.11691	0.00475	.	.	.	-0.01723	.	0.00498	0.46844
8	4.8094	0.3540	0.00050017	.	.	.	-0.00255	0.01005	.	.	.	.	-0.09374	0.00490	.	.	.	-0.01737	-0.00376	0.00733	0.47222
7	4.8360	0.3503	0.77005	.	.	-0.00403	.	.	.	.	0.07166	.	-0.12103	0.00485	.	.	.	-0.01756	.	0.00482	0.46791
7	4.8705	0.3503	-0.03277	.	.	.	-0.00238	0.00955	.	.	.	.	-0.09801	0.00477	.	.	.	-0.01659	.	0.00511	0.47141
8	4.8902	0.3538	0.84447	.	.	-0.00427	.	.	.	.	0.07207	.	-0.11703	0.00497	.	.	.	-0.01832	-0.00365	0.00697	0.46863
9	4.9297	0.3574	0.69720	.	.	-0.00425	.	0.00295	.	.	0.07409	.	-0.11269	0.00488	.	.	.	-0.01792	-0.00370	0.00713	0.46868
8	4.9307	0.3538	0.62384	.	.	-0.00400	.	0.00291	.	.	0.07365	.	-0.11681	0.00475	.	.	.	-0.01716	.	0.00495	0.46794
9	4.9974	0.3572	-0.04334	.	.	.	-0.00268	0.01126	-0.00226	.	0.08253	.	-0.11302	0.00469	.	.	.	-0.01699	.	0.00491	0.47538
6	5.1846	0.3461	0.72361	.	.	-0.00369	.	.	.	.	.	.	-0.10248	0.00486	.	.	.	-0.01694	.	0.00496	0.47078
10	5.2018	0.3605	-0.01268	.	.	.	-0.00282	0.01162	-0.00201	.	0.08202	.	-0.10955	0.00482	.	.	.	-0.01775	-0.00353	0.00700	0.47536

We again measure the quality of this model with R^2 and misclassification rate. The values for the full model and this best model are listed in Table 6.

Table 6: R^2 and Misclassification Rate for Full and Cp Models

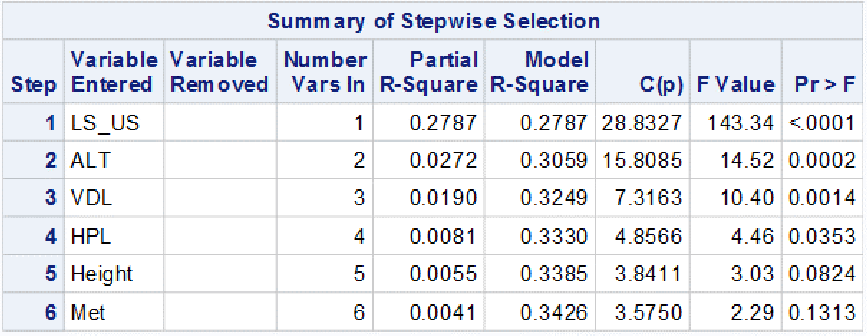
	Before	After
R^2	0.3644	0.3662
Misclassification Rate	0.188	0.190

From this comparison, R^2 again did not see significant improvement, and the misclassification rate is about the same. Yet with similar accuracy, this model is able to predict with only 9 variables down from 18.
Also note in this best model, we see that previously mentioned highly correlated pairs of variables are present. Significant multicollinearity among explanatory variables may lead to insignificant parameter estimation. It can be addressed by removing one of the highly correlated explanatory variables. This will be discussed further in the "best" model diagnostics section.

4. Stepwise for Best Set of Variables

The stepwise model selection in SAS returns a model with 6 variables: Height, Met, HPL, VDL, AST, and LS_US, as is shown in Table 7.

Table 7: Stepwise Model Selection



Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	LS_US		1	0.2787	0.2787	28.8327	143.34	<.0001
2	ALT		2	0.0272	0.3059	15.8085	14.52	0.0002
3	VDL		3	0.0190	0.3249	7.3163	10.40	0.0014
4	HPL		4	0.0081	0.3330	4.8566	4.46	0.0353
5	Height		5	0.0055	0.3385	3.8411	3.03	0.0824
6	Met		6	0.0041	0.3426	3.5750	2.29	0.1313

We again measure the quality of this model with R^2 and misclassification rate. The values for the full model and this best model are listed in Table 8.

Table 8: R^2 and Misclassification Rate for Full and Cp Models

	Before	After
R^2	0.3644	0.3260
Misclassification Rate	0.188	0.196

From this comparison, R^2 and the misclassification rate remain similar for the full model and selected best model. Not suprisingly, stepwise model remedied the multicolinearity issues the no longer have variables that are highly correlated.

5. Assumptions Check

The model used for the rest of the questions has nine explanatory variables is provided below. The values of the regression coefficients are omitted for brevity.

$$\begin{aligned} LS\_Bi_{predicted} =& b_0 + b_1 \cdot Weight + b2 \cdot BMI + b3 \cdot Metabolic\ Syndrome + b4 \cdot Hyperlipidemia + \\ & b5 \cdot Plasma\ Triglycerides + b6 \cdot Very\ Low\ Density\ Lipoprotein\ Cholesterol + \\ & b7 \cdot Aspartate\ Aminotransferase + b8 \cdot Alanine\ Aminotransferase + b9 \cdot Ultrasound \end{aligned}$$

The model assumptions to evaluate are 1) linearity, 2) normality of residuals, 3) constant residuals as function of explanatory variables.

Linearity: LS_Bi was plotted as a function of each of the explanatory variables used in the full model (Table 9). The variables Met, HPL, and LS_US do not appear to have significant linearity issues with LS_Bi, as evident by lack of curvature in the smooth lines for the plots shown in Figure 3. Curvature is observed in the plots for Weight, BMI, TG, VDL, AST and ALT. The curvature in Weight, BMI, TG, and AST appears to be due to outliers. Outliers and influential observations were examined in more detail in Part II, Problem 6. The curvature in ALT and VDL does not seem to be due to outliers. Power transformations were attempted for these variables but the not significantly improve the linearity of the relationship. Using piecewise regression for these two variables, as described in Part I, Problem I for VDL, may be a possible remedy.

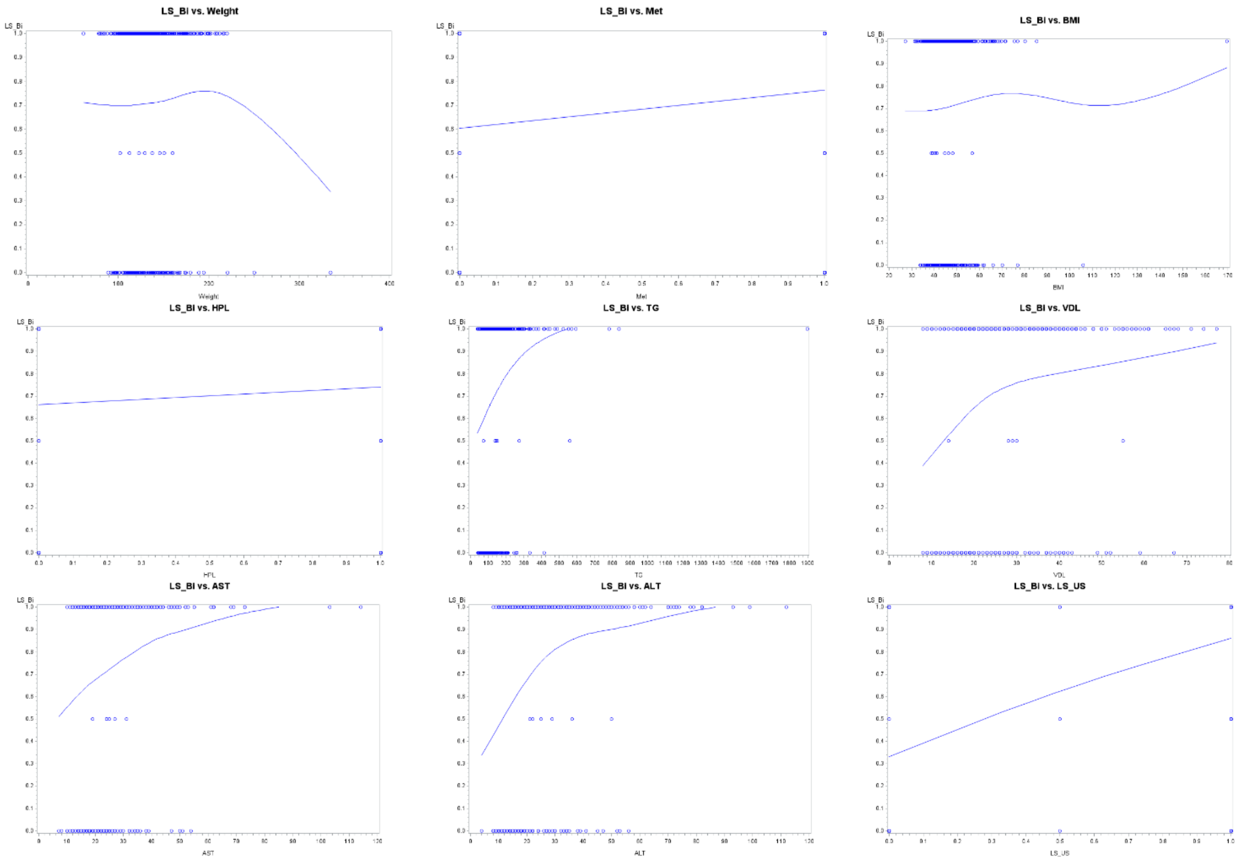
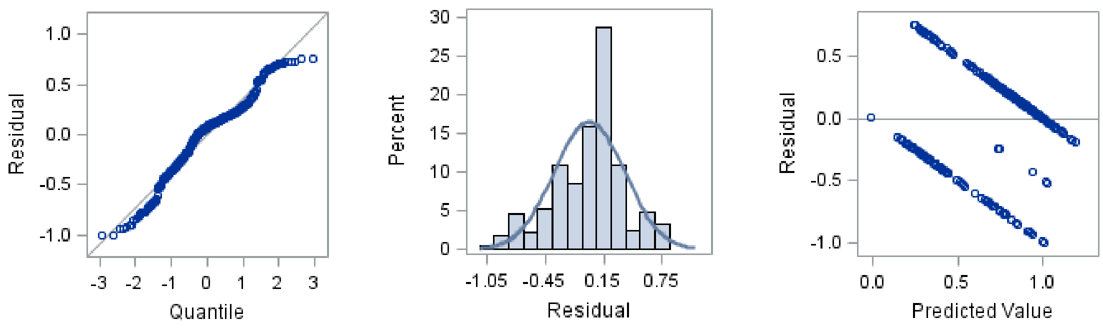


Figure 9: Scatter plots of LS_Bi vs the explanatory variables used in the model based on C_p criterion

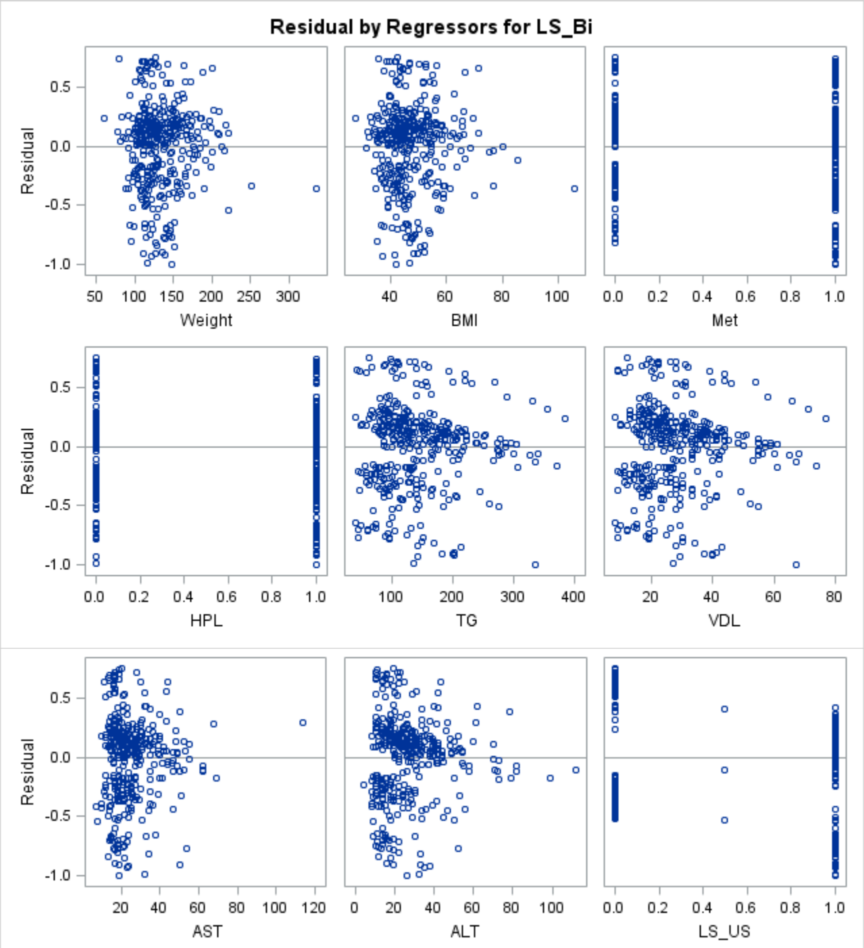
Normality and homoscedasticity of residuals for full model: Normality and homoscedasticity of residuals for full model: The QQ plot and histogram of the residuals for the best model are shown below in Figure 9, along with the plot of the residuals vs. predicted values. Slight deviations from normality are seen in the QQ plot at the lower (-3 to -1) and higher (1 to 3) quantiles. The histogram appears fairly normal, but the distribution is skewed slightly to the right. Overall, the normality of the residuals seems sufficient for use in hypothesis testing that assume normality of the residuals. The residuals as a function of predicted value does not appear homoscedastic; the residuals tend to be positive at lower values of $LS\_Bi_{Predicted}$ and negative at higher values of $LS\_Bi_{Predicted}$. The lack of homoscedasticity of residuals in the model will cause issues in hypothesis testing.

Table 9: QQ plot and histogram of residuals for the model based on C_p criterion



Constant residuals as function of explanatory variables: This assumption is evaluated by plotting the model residuals as a function of each of the explantory variables. The partial residual plots are shown below in Figure 10. The residuals for TG, VDL, AST, ALT, and LS_US appear to have patterns or heteroscedastic. It does not appear that power transformations would be useful in significantly correcting the issues presented in the residuals for these variables. The residual plots for Weight, BMI, Met, HPL appear to be sufficiently homoscedastic.

Table 10: Partial residual plots for the model based on C_p criterion



6. Further model diagnostics and prediction of LS_Bi (Outliers, influential observations, and multicollinearity)

Outliers and influential observations were evaluated using studentized residuals, Cooks D, and Hat Matrix. Visuals for these from the SAS outputs are provided in Table 10.

The cutoff for the studentized residuals is calculated using $t_{n-p-1,(1-/2n)} = t_{378-9-1,1-0.05/(2\times378)} = t_{368,0.99} \approx 3$. Values greater than three are considered outliers. None of the cases met this criterion so none are considered outliers based on studentized residuals.

Cooks D is used to detect influential observations. Cases with values greater than $4/\text{n} = 0.011$ are considered overly influential based on this criterion. Cases 136,345, 287, 132, 121, 252, 217, 346, 249, 33, 175, 383 are above the Cooks D threshold and are considered overly influential.

Finally, Hat Matrix diagonals were used to detect influential observations using the cutoff 2p/n = 0.0489. Cases 405, 379 ,94, 2, 273, 217, 368, 407, 347, 84, 230, 398, 346, 33, 383 are above the threshold and are considered overly influential. Removing these cases or giving them smaller weight in the regression model would improve the robustness of the model and may remedy some of the problems observed in the partial residual plots as discussed in problem 5.

Multicollinearity was evaluated using the variance inflation factor. A $VIF > 10$ indicates excessive multicollinearity among the explanatory variables. Based on the parameter estimate table provided below that includes the VIF, TG and VDL are highly correlated based the VIF of 36.5 and 36.8, respectively. The model would be improved by removing one of these variables. The high correlation of these variables was also noted in Part II, question 1.

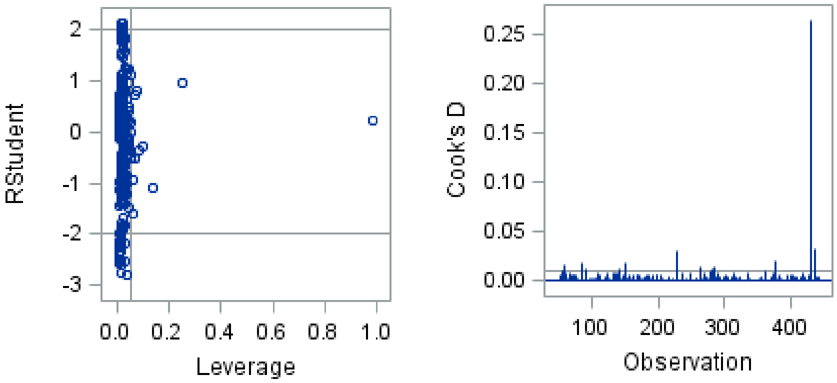
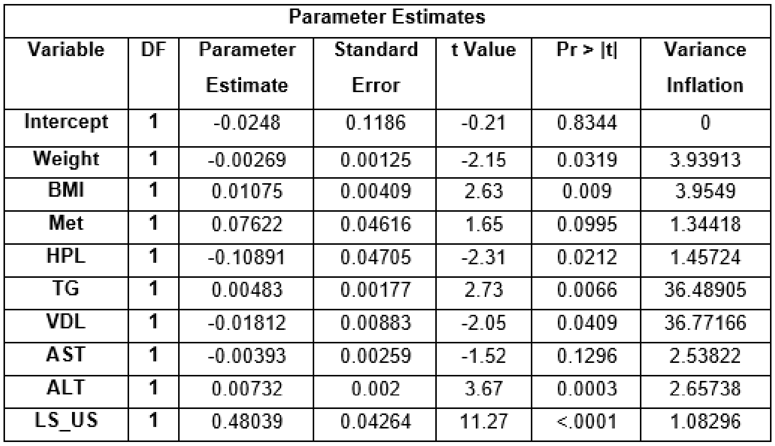


Figure 10: Studentized residual and Cookś D plots for the candidate model

Table 11: Parameter estimate table containing variance inflation factor



Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	**1**	-0.0248	0.1186	-0.21	0.8344	0
Weight	**1**	-0.00269	0.00125	-2.15	0.0319	3.93913
BMI	**1**	0.01075	0.00409	2.63	0.009	3.9549
Met	**1**	0.07622	0.04616	1.65	0.0995	1.34418
HPL	**1**	-0.10891	0.04705	-2.31	0.0212	1.45724
TG	**1**	0.00483	0.00177	2.73	0.0066	36.48905
VDL	**1**	-0.01812	0.00883	-2.05	0.0409	36.77166
AST	**1**	-0.00393	0.00259	-1.52	0.1296	2.53822
ALT	**1**	0.00732	0.002	3.67	0.0003	2.65738
LS_US	**1**	0.48039	0.04264	11.27	<.0001	1.08296

The relationship between the predicted value of LS_Bi and the observed value of LS_Bi is shown below from the SAS output was well as in box-plots. The relationship is easier to visualize using the box plots due to the large number of data points at the discrete values of LS_Bi. It is evident from the box plots that the predicted value of LS_Bi tends to be higher for the cases where LS_Bi is actually 1, indicating presence of liver steatosis. However, there is a large spread in the data, and a significant amount of cases are predicted to have values of LS_Bi that are less than 0.5 when the observed value of LS_Bi is 1.

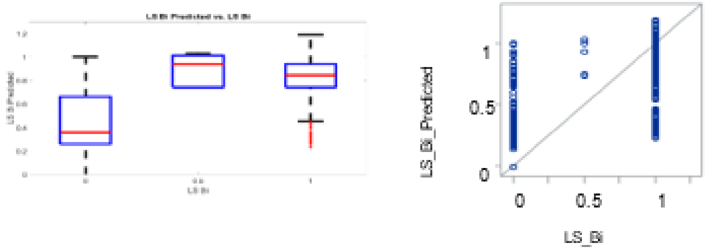


Figure 11: Plots relating the predicted value of LS_Bi with the observed value of LS_Bi

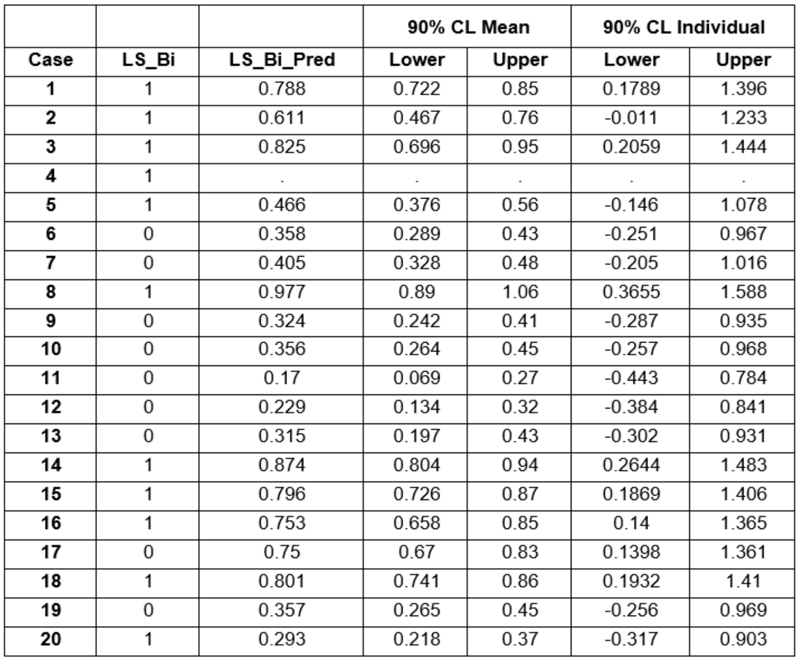
7. Equation of final model and inference on regression parameters and predicted values

Final Model: The final regression model equation is provided below

$$\begin{aligned} LS\_Bi_{predicted} = & -0.024 - 0.0027 \cdot Weight + 0.0108 \cdot BMI + 0.0762 \cdot (Metabolic\ Syndrome) - \\ & 0.1089 \cdot (Hyperlipidemia) + 0.0048 \cdot (Plasma\ Triglycerides) - \\ & 0.0181 \cdot (Very\ Low\ Density\ Lipoprotein\ Cholesterol) - \\ & 0.0039 \cdot (Aspartate\ Aminotransferase) + 0.0073 \cdot (Alanine\ Aminotransferase) + \\ & 0.4804 \cdot Ultrasound \end{aligned}$$

The table listing all of the 90% confidence intervals for the mean value of the response variable (LS_Bi) and the 90% confidence interval for the individual observations are provided in Table 12.

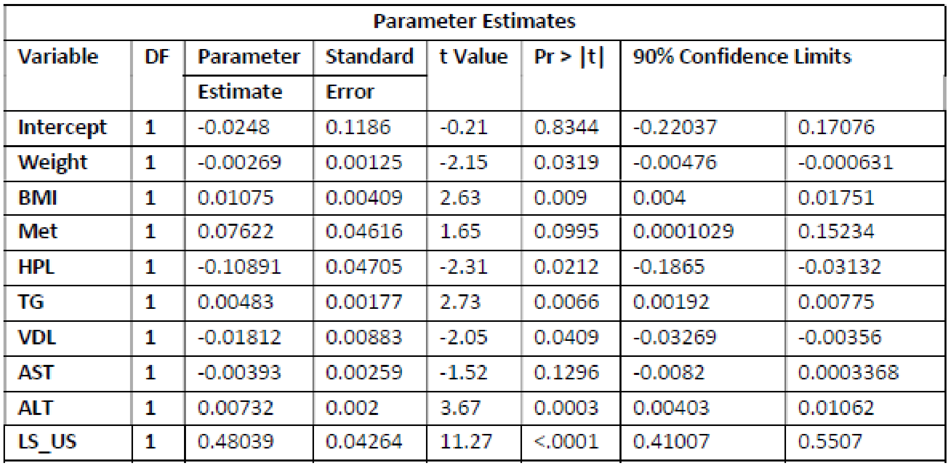
Table 12: LS_Bi predicted and 90% prediction intervals of mean and individual observations for all cases



			90% CL Mean		90% CL Individual	
Case	**LS_Bi**	**LS_Bi_Pred**	**Lower**	**Upper**	**Lower**	**Upper**
1	1	0.788	0.722	0.85	0.1789	1.396
2	1	0.611	0.467	0.76	-0.011	1.233
3	1	0.825	0.696	0.95	0.2059	1.444
4	1	.	.	.	.	.
5	1	0.466	0.376	0.56	-0.146	1.078
6	0	0.358	0.289	0.43	-0.251	0.967
7	0	0.405	0.328	0.48	-0.205	1.016
8	1	0.977	0.89	1.06	0.3655	1.588
9	0	0.324	0.242	0.41	-0.287	0.935
10	0	0.356	0.264	0.45	-0.257	0.968
11	0	0.17	0.069	0.27	-0.443	0.784
12	0	0.229	0.134	0.32	-0.384	0.841
13	0	0.315	0.197	0.43	-0.302	0.931
14	1	0.874	0.804	0.94	0.2644	1.483
15	1	0.796	0.726	0.87	0.1869	1.406
16	1	0.753	0.658	0.85	0.14	1.365
17	0	0.75	0.67	0.83	0.1398	1.361
18	1	0.801	0.741	0.86	0.1932	1.41
19	0	0.357	0.265	0.45	-0.256	0.969
20	1	0.293	0.218	0.37	-0.317	0.903

The parameter estimate table with 90% confidence intervals for the regression coeficients is provided in Table 13:

Table 13: Parameter estimate table containing 90% confidence intervals for the regression coefficients



Parameter Estimates							
Variable	**DF**	**Parameter**	**Standard**	**t Value**	**Pr > |t|**	**90% Confidence Limits**	
		Estimate	**Error**				
Intercept	**1**	-0.0248	0.1186	-0.21	0.8344	-0.22037	0.17076
Weight	**1**	-0.00269	0.00125	-2.15	0.0319	-0.00476	-0.000631
BMI	**1**	0.01075	0.00409	2.63	0.009	0.004	0.01751
Met	**1**	0.07622	0.04616	1.65	0.0995	0.0001029	0.15234
HPL	**1**	-0.10891	0.04705	-2.31	0.0212	-0.1865	-0.03132
TG	**1**	0.00483	0.00177	2.73	0.0066	0.00192	0.00775
VDL	**1**	-0.01812	0.00883	-2.05	0.0409	-0.03269	-0.00356
AST	**1**	-0.00393	0.00259	-1.52	0.1296	-0.0082	0.0003368
ALT	**1**	0.00732	0.002	3.67	0.0003	0.00403	0.01062
LS_US	**1**	0.48039	0.04264	11.27	<.0001	0.41007	0.5507

Concluding Remarks

The accuracy of the proposed model using a mix of clinical variables, blood test results, and liver ultrasound imaging to predict liver steatosis does not match that of liver biopsy. The proposed model may be useful in screening patients for biopsy. There is a large spread in the predicted values of LS_Bi, and a significant amount of cases are predicted to have values of LS_Bi that are less than 0.5 when the observed value of LS_Bi is 1. It is not clear what the parameter estimates resulting from our analyses mean since they are being used to predict a categorical variable. This may limit the amount of scientific knowledge that can be derived from the model. The parameter estimates table shows that the ultrasound imaging, plasma triglycerides, alanine aminotransferase, and BMI are highly significant in explaining the variance in LS_Bi. These variables should be explored in detail in further studies to understand if any causal relationship between them and liver steatosis exists. Further studies should perform the analysis with multiple logistic regression model which would provide more accurate parameter estimates. Finally, additional explanatory variables could be added to improve the predictive performance of the model.

References

[1] Wu J, You J, Yerian L, Shiba A, Schauer PR, Sessler DI. Prevalence of liver steatosis and fibrosis and the diagnostic accuracy of ultrasound in bariatric surgery patients. Obesity surgery. 2011;22(2):240-7.

[2] Brunt EM, Tiniakos DG. Histopathology of nonalcoholic fatty liver disease. World journal of gastroenterology. 2010;16(42):5286-96.

[3] Brunt EM, Kleiner DE, Wilson LA, Belt P, Neuschwander-Tetri BA. The NAS and The Histopathologic Diagnosis in NAFLD: Distinct Clinicopathologic Meanings. Hepatology. 2011;53(3):810-20.