# MATH 3330 (Regression Analysis) Project

York University, Toronto

# Liver Disorder

Ammar Mughal, Kevin Cao, Giuliano Iervasi

# Contents

# Introduction

The liver disorder dataset on the UCI Machine Learning Repository provides data for studying liver disorders. It includes 344 observations of male individuals which measure seven variables:

1. Mean Corpuscular Volume (mcv - $x_1$): this measures the average volume of red blood cells. It is important for assessing anemia and other blood-related conditions.
2. Alkaline Phosphatase (alkphos - $x_2$): this enzyme is found in the liver, bones, and other tissues. Elevated levels can indicate liver disease. The normal ranges for this measure are 44-147 U/L (Haldeman-Englert et al., 2024).
3. Alanine Aminotransferase (sgpt - $x_3$): also known as serum glutamate-pyruvate transaminase (SGPT), is an enzyme primarily found in the liver. The normal ranges for this measure are below 40 U/L (Haldeman-Englert et al., 2024). Levels over 1,000 U/L are often related to liver inflammation or damage.
4. Aspartate Aminotransferase (sgot - y): also known as serum glutamate-oxaloacetate transaminase (SGOT), is an enzyme found in the liver, heart, muscles, and other tissues. Elevated levels are related to liver damage or disease. The healthy range for this measure is 8-33 U/L (Cleveland Clinic, 2021).
5. Gamma-Glutamyl Transpeptidase (gamma gt - $x_4$): this enzyme is involved in the metabolism of glutathione and is present in the liver and other organs. Elevated levels can indicate liver disease or alcohol abuse. The normal ranges for this measure are 4-47 U/L (Finke et al., 2024).
6. Drinks (drinks - $x_5$): the number of half-pint equivalents of alcohol consumed by an individual in a day. The average number of drinks in the USA was around 603 standard drinks per year in 2013, or ~1.652 half-pints per day (Schaeffer & DeSilver, 2024).
7. Train/Test Set Categories (selector): This is a categorical variable which the researchers used to arbitrarily split the two groups into two sets. The variable was not included as it is statistically irrelevant to our investigation.

In this project, we are going to use the fourth variable (sgot) as our dependent variable (y), with the remaining four blood test variables and number of drinks as the independent variables ($x_1$ through $x_5$). This will allow us to use a variety of different factors to predict an individual's measure of aspartate aminotransferase, which is connected to a risk of liver damage or disease. Our goal is to be able to use our model to accurately predict healthy and unhealthy amounts of aspartate aminotransferase in their body. The outline of our analysis will be as follows:

1. **Fit a model (Chapter 7).** First, we will determine a model that is most appropriate for this dataset by comparing it with all possible first-order linear models. By doing so, we can remove insignificant variables to produce a more accurate model.
2. **Perform a test for multicollinearity (Chapter 5).** We will determine if any of the independent variables (blood test measurements) are related by multicollinearity. We will then take necessary measures to fit a better model, such as removing variables if required or performing a Ridge Regression.
3. **Interpret results (Chapter 3/4).** We will discuss the results of our findings in more detail and its statistical significance. We will include analysis of our model's effectiveness and precision before discussing limitations and possible improvements.

# Part I: Fitting a Model

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 5 | 20555 | 4110.95678 | 97.63 | <.0001 |
| Error | 338 | 14232 | 42.10583 | | |
| Corrected Total | 343 | 34787 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 6.48890 | R-Square | 0.5909 |
| Dependent Mean | 24.66570 | Adj R-Sq | 0.5848 |
| Coeff Var | 26.30739 | | |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| x1 | 1 | 27.99687 | 27.99687 | 0.66 | 0.4154 |
| x2 | 1 | 148.05275 | 148.05275 | 3.52 | 0.0616 |
| x3 | 1 | 10253.02499 | 10253.02499 | 243.51 | <.0001 |
| x4 | 1 | 671.81949 | 671.81949 | 15.96 | <.0001 |
| x5 | 1 | 166.22552 | 166.22552 | 3.95 | 0.0477 |

We began this investigation with a model that included all five predictor variables. The model produced a moderately strong fit with the key statistics from the Analysis of Variance table, with $R^2 = 0.5909$ and $\underline{R^2} = 0.5848$. The model is statistically significant because it carries a F value of 97.63 with a P value of <0.0001. By the Type III Sum of Squares chart, $x_3$ (sgpt) is the most significant predictor (F(model) = 243.51 and $P(T > |t|) < 0.0001$). The variables that are not considered to be significant in this model are variables $x_1$ and $x_2$ as they both have a P value above our significance value of $\alpha=0.05$ ($x_1$ has P value of 0.4154 and $x_2$ has P value of 0.0616). Thus, we see that $x_1$ is the least significant variable by a considerable margin, so we will remove it moving forward.

We can see that the $\underline{R^2}$ value slightly increases when we remove $x_1$ from the model, from 0.5848 to 0.5852, which shows a marginally greater proportion of the variance in the predictor variable. Additionally, when $x_1$ is removed from the data, the mean square value increases while the mean square error decreases. The higher mean square model displays more of a variance in the dependent variable (5131.69 to 4110.95) and the lower mean square error (42.06 to 42.10) shows that the prediction accuracy is slightly better. These three key statistics demonstrate that the model that excludes $x_1$ has a better overall fit to the data then the base model.

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 4 | 20527 | 5131.69676 | 122.00 | <.0001 |
| Error | 339 | 14260 | 42.06421 | | |
| Corrected Total | 343 | 34787 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 6.48569 | R-Square | 0.5901 |
| Dependent Mean | 24.66570 | Adj R-Sq | 0.5852 |
| Coeff Var | 26.29438 | | |

Attempting to further improve the model, we will try to remove $x_2$, the only other variable that is considered not significant in the Type III Sum of Squares table. However, when the $x_2$ variable is also removed from the dataset, the model becomes less accurate with the adjusted $R^2$ dropping (0.5852 to 0.5822) along with the mean square error increasing (42.06 to 42.37). This shows that we should remain with the second model for the purpose of this project due to its accuracy being the highest.

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 3 | 20379 | 6793.00161 | 160.31 | <.0001 |
| Error | 340 | 14408 | 42.37515 | | |
| Corrected Total | 343 | 34787 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 6.50962 | R-Square | 0.5858 |
| Dependent Mean | 24.66570 | Adj R-Sq | 0.5822 |
| Coeff Var | 26.39139 | | |

As shown circled in **black**, the preceding findings can be confirmed upon seeing the results of the r-square selection method (on the right): all possible first-order models with their $R^2$ and C

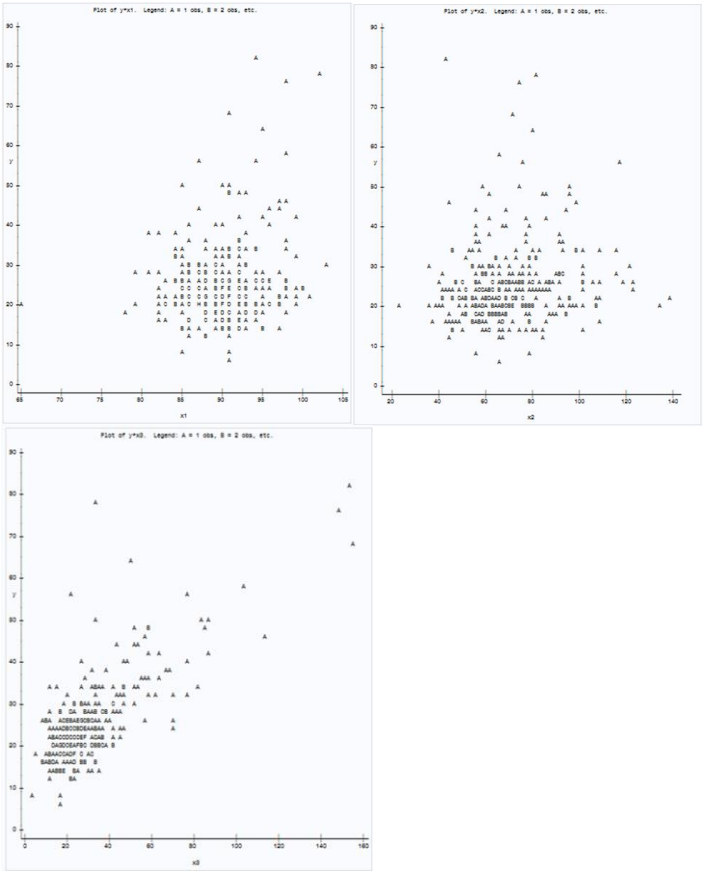
statistic values are shown. For example, another reason why we consider the model without $x_1$ to be better than the model without $x_1$ and $x_2$ is the C statistic: the former is smaller than k (as 4.6649 < k=6) while the latter is larger (although to a small effect, with 6.1747 > 6).

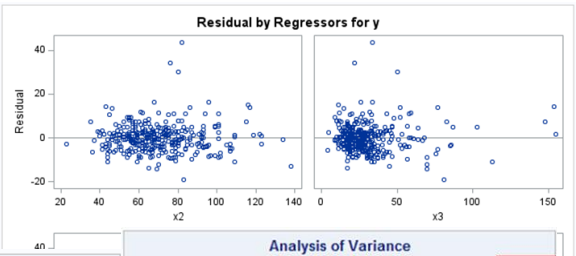
There are several key takeaways from the first-order models from the r-square test. As discussed, the model without the $x_1$ term has the highest $R^2$ value. We consider this model to be superior to the original model with all five predictor variables, as a small difference between $R^2$ values (0.5901 vs 0.5908) is compensated by a larger difference in the C statistic (4.6649 vs 6).

Thus, moving forward, we can consider the model without the $x_1$ variable, as a desirable $R^2$ value and C statistic stands out among the other models available.

| Number in Model | R-Square | C(p) | Variables in Model |
|---|---|---|---|
| 1 | 0.5467 | 34.5284 | x3 |
| 1 | 0.2777 | 256.7241 | x4 |
| 1 | 0.0794 | 420.5849 | x5 |
| 1 | 0.0349 | 457.3065 | x1 |
| 1 | 0.0209 | 468.9136 | x2 |
| 2 | 0.5789 | 9.9068 | x3 x4 |
| 2 | 0.5637 | 22.4558 | x3 x5 |
| 2 | 0.5547 | 29.9301 | x2 x3 |
| 2 | 0.5529 | 31.3606 | x1 x3 |
| 2 | 0.2893 | 249.1815 | x4 x5 |
| 2 | 0.2834 | 253.9918 | x2 x4 |
| 2 | 0.2829 | 254.4534 | x1 x4 |
| 2 | 0.0929 | 411.4338 | x2 x5 |
| 2 | 0.0901 | 413.7089 | x1 x5 |
| 2 | 0.0536 | 443.8885 | x1 x2 |
| 3 | 0.5858 | 6.1747 | x3 x4 x5 |
| 3 | 0.5838 | 7.8388 | x2 x3 x4 |
| 3 | 0.5813 | 9.9386 | x1 x3 x4 |
| 3 | 0.5698 | 19.4273 | x2 x3 x5 |
| 3 | 0.5655 | 22.9481 | x1 x3 x5 |
| 3 | 0.5605 | 27.1291 | x1 x2 x3 |
| 3 | 0.2941 | 247.2281 | x2 x4 x5 |
| 3 | 0.2913 | 249.4692 | x1 x4 x5 |
| 3 | 0.2885 | 251.8475 | x1 x2 x4 |
| 3 | 0.1034 | 404.7812 | x1 x2 x5 |
| 4 | 0.5901 | 4.6649 | x2 x3 x4 x5 |
| 4 | 0.5866 | 7.5162 | x1 x3 x4 x5 |
| 4 | 0.5861 | 7.9478 | x1 x2 x3 x4 |
| 4 | 0.5716 | 19.9555 | x1 x2 x3 x5 |
| 4 | 0.2961 | 247.5061 | x1 x2 x4 x5 |
| 5 | 0.5909 | 6.0000 | x1 x2 x3 x4 x5 |

# Part II: Multicollinearity Test






**Pearson Correlation Coefficients, N = 344**
**Prob > |r| under H0: Rho=0**

| | x1 | x2 | x3 | y | x4 | x5 |
|---|---|---|---|---|---|---|
| x1 | 1.00000 | 0.04301<br>0.4265 | 0.14699<br>0.0063 | 0.18691<br>0.0005 | 0.22164<br><.0001 | 0.31413<br><.0001 |
| x2 | 0.04301<br>0.4265 | 1.00000 | 0.07494<br>0.1655 | 0.14452<br>0.0073 | 0.13188<br>0.0144 | 0.10278<br>0.0569 |
| x3 | 0.14699<br>0.0063 | 0.07494<br>0.1655 | 1.00000 | 0.73937<br><.0001 | 0.50291<br><.0001 | 0.20843<br><.0001 |
| y | 0.18691<br>0.0005 | 0.14452<br>0.0073 | 0.73937<br><.0001 | 1.00000 | 0.52699<br><.0001 | 0.28175<br><.0001 |
| x4 | 0.22164<br><.0001 | 0.13188<br>0.0144 | 0.50291<br><.0001 | 0.52699<br><.0001 | 1.00000 | 0.34308<br><.0001 |
| x5 | 0.31413<br><.0001 | 0.10278<br>0.0569 | 0.20843<br><.0001 | 0.28175<br><.0001 | 0.34308<br><.0001 | 1.00000 |

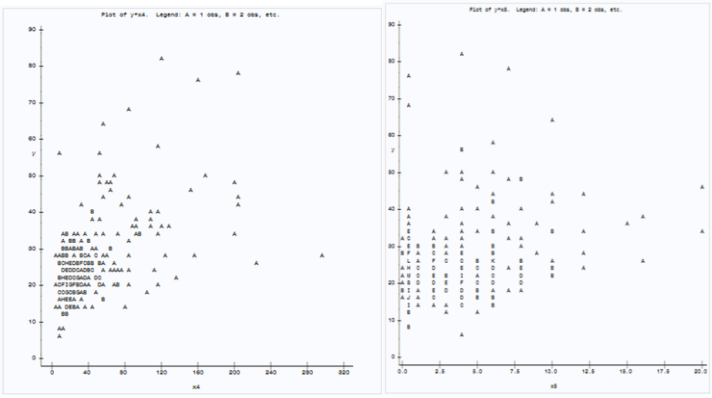
**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
|---|---|---|---|---|---|---|
| Intercept | 1 | 9.67579 | 1.45990 | 6.63 | <.0001 | 0 |
| x2 | 1 | 0.03614 | 0.01928 | 1.87 | 0.0617 | 1.02166 |
| x3 | 1 | 0.32490 | 0.02077 | 15.65 | <.0001 | 1.34124 |
| x4 | 1 | 0.04420 | 0.01079 | 4.10 | <.0001 | 1.46540 |
| x5 | 1 | 0.25477 | 0.11195 | 2.28 | 0.0235 | 1.13993 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 20527 | 5131.69676 | 122.00 | <.0001 |
| Error | 339 | 14260 | 42.06421 | | |
| Corrected Total | 343 | 34787 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 6.48569 | R-Square | 0.5901 |
| Dependent Mean | 24.66570 | Adj R-Sq | 0.5852 |
| Coeff Var | 26.29438 | | |



By performing a test for multicollinearity, we will determine if there is evidence of variables being redundant in the model (without $x_1$). Throughout this course, we discussed five different methods to detect multicollinearity. According to these methods, we were able to conclude the following:

1. The scatterplots (first five images above) which illustrate the relationships between $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$ with y do not appear to have any obvious linear relationships. This suggests that multicollinearity does not exist so far.
2. All values of the correlation matrix (circled in **black**) are significantly smaller than the threshold value of r = 0.90. This also suggests there is no evidence for multicollinearity.
3. All Variance Inflation values ($VIF_j$ for j = 1-5, circled in **blue**) are significantly smaller than the threshold value of $VIF_j = 10$. Additionally, the mean of the $VIF_j$ values (1.239) is not substantially greater than 1. This further supports that there is no evidence for multicollinearity.
4. The F-value for the overall test (circled in **red**) has a very small ($P < 0.0001$) suggests that the model with all five variables has a good fit and that at least one of the variables is significant. Additionally, most of the p-values from the t-test (circled in **purple**) are small, which results in the t-test being smaller in value compared to the t-value $t_{0.05}^{(342)}$

= 1.967. This results in the $\hat{\beta}_j$ values being insignificant. This also suggests there is multicollinearity between the variables in this model.

5. Since values of the correlation matrix are all low (ranging from 0.04301 to 0.73937), there is little to no evidence of a linear effect of $x_j$ on y. Therefore the effect of the t-test is minimal, which does not suggest that multicollinearity exists.

We are able to discern that there is little evidence of multicollinearity, as 4 out of the 5 methods discussed in this course support this. Therefore, it is not necessary to remove any variables or perform a Durbin-Watson Test with the results of this multicollinearity test.

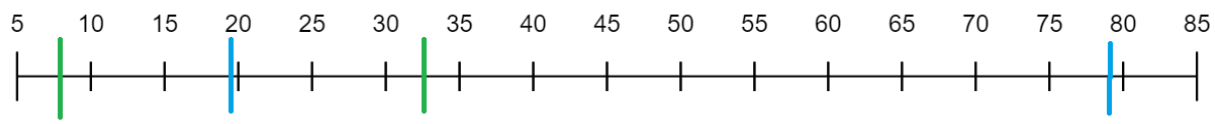
## Part III: Interpretation of Results

Based on the parameter estimates derived from the previous section, we may interpret our four-variable model for a given individual as follows.

- Our model is expressed by the equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_2 + \hat{\beta}_2 x_3 + \hat{\beta}_3 x_4 + \hat{\beta}_4 x_5$.
- $\hat{\beta}_0 = 9.67579$. This is the level of aspartate aminotransferase in an individual when all other measures ($x_2$ = alkaline phosphatase, $x_3$ = alanine aminotransferase, $x_4$ = gamma-glutamyl transpeptidase, $x_5$ = number of drinks) are 0.
- $\hat{\beta}_1 = 0.03614$. This is the change in aspartate aminotransferase associated with an increase in 1 level of alkaline phosphatase when all other measures are 0. Note that the normal ranges for this measure are 44-147 U/L (Haldeman-Englert et al., 2024), and range of this dataset is $x_2 \in [65, 103]$. Thus, $\hat{\beta}_1 x_2 \in [2.3491, 3.7224]$.
- $\hat{\beta}_2 = 0.32490$. This is the change in alanine aminotransferase associated with an increase in 1 level of alkaline phosphatase when all other measures are 0. Note that the normal ranges for this measure are below 40 iU/L, although levels above 1,000 iU/L are related to liver damage (Haldeman-Englert et al., 2024). The range of this dataset for this variable is $x_3 \in [22, 138]$. Then, $\hat{\beta}_2 x_3 \in [7.5438, 47.3202]$.
- $\hat{\beta}_3 = 0.04420$. This is the change in gamma-glutamyl transpeptidase associated with an increase in 1 level of alkaline phosphatase when all other measures are 0. The normal levels for this measure are 4-47 U/L (Finke et al., 2024). In this dataset $x_4 \in [5, 297]$ and thus $\hat{\beta}_3 x_4 \in [0.221, 13.1274]$.
- $\hat{\beta}_4 = 0.25477$. This is the change in the number of half-pint drinks consumed in a day associated with an increase in 1 level of alkaline phosphatase, when all other measures are 0. Note that the average number of drinks consumed by Americans is ~1.652/day (Schaeffer & DeSilver, 2024). The data we studied had a range of $x_5 \in [0, 20]$. Therefore, $\hat{\beta}_4 x_5 \in [0, 5.0954]$.

By using the ranges of each variable in the dataset, we are able to calculate the range of our predictor variable $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_2 + \hat{\beta}_2 x_3 + \hat{\beta}_3 x_4 + \hat{\beta}_4 x_5$.

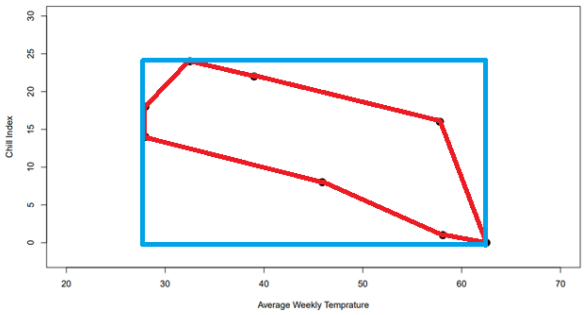
- The minimum contribution from each term is 9.67579, 2.3491, 7.5438, 0.221 and 0. Thus, the smallest possible value for $\hat{y}$ is 19.7869.
- The maximum contribution from each term is 9.67579, 3.7224, 47.3202, 13.1274, and 5.0954. Thus, the largest possible value for $\hat{y}$ is 78.9412.
- Thus, $\hat{y} \in [19.7869, 78.9412]$. This suggests that this model can predict an individual's measure of aspartate aminotransferase within the ranges of each independent variable, yielding a value within 19.7869–78.9412 U/L (see the **blue** interval above). This falls




within the healthy range of 8-33 U/L of aspartate aminotransferase for an individual (the **green** interval above), and includes a sufficient range of unhealthy amounts exceeding

33 U/L. Thus, this model is able to predict an individual's measure of this enzyme in a realistic range and determine if their liver is considered healthy.

## Part III.I: Limitations & Improvements




It is important to remember that the predictors in this investigation are only valid within the context of the experimental region. This is effectively the range of the different combinations of the independent variables. For example, in the fuel consumption example (see Example 3.3 in class on the right), this refers to the area the two variables $x_1$ (average hourly temperature) and $x_2$ (chill index) form. Notice, however, that the correct experimental region is only the region the variables enclose (the **red** area), as opposed to the area of the maximum and minimum of the variables (the **blue** area).

In our model, it is not possible to visualize a four-dimensional area. Thus, this represents a limitation with creating a model with four predictor variables - we can determine the area the experimental region is *enclosed* in for our model (the **blue** area in Example 3.3) by the minima and maxima of the four variables, but it is beyond the scope of this course to see the true experimental region (the **red** area), as we have not discussed how to generalize this procedure past four. This may result in this model being less reliable in its ability to predict values.

Another possible limitation of this model is the relatively high unexplained variation, with $R^2 = 0.5901$ and $\underline{R}^2 = 0.5852$. This suggests that this model fits moderately well with the dataset. Since none of the assumptions (zero mean, constant variance, independence, normality) are likely to be violated, some the unexplained error can be attributed to the following:

- Randomness in the data. By looking at the residual plots, there does not appear to be any relation between any of the variables and the residuals. This random pattern may suggest that this model is appropriate, even though there is a considerable amount of unexplained error.
- Limited variables in the dataset. There could be many more variables that were not measured in this dataset that could influence an individual's measure of aspartate aminotransferase. For future investigations, numerical variables that could be measured include weight and drugs (Cleveland Clinic, 2021), while other categorical variables such as sex, exposure to hepatitis, diabetes, or family history (MedlinePlus, 2022). By including more variables in the dataset, more of the variance could be explained and produce a more effective model.
- Influential points. In each scatter plot, there are a considerable number of points that are suspected to be outliers with respect to x and y. By removing these, it could be possible to produce a model that is more accurate. However, the effect of this may not change too significantly, as the sample size of n = 344 should be sufficient in minimizing the effect of influential points in the first place.

# Bibliography


Cleveland Clinic. (2021) Aspartate Transferase (AST) Blood Test: What It Is, Procedure & Results' Cleveland Clinic. Available at: https://my.clevelandclinic.org/health/diagnostics/22147-aspartate-transferase-ast (Accessed: 19 Aug 2024).

Finke, A., Haldeman-Englert, C., and Novick, T. (2024) 'Gamma-Glutamyl Transpeptidase - Health Encyclopedia', *Health Encyclopedia.* University of Rochester Medical Center. Available at: https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=gamma_glutamyl_transpeptidase (Accessed: 18 August 2024).

Haldeman-Englert, C., Foley, M. and Turley Jr, R. (2024) 'ALT - Health Encyclopedia', *Health Encyclopedia*. University of Rochester Medical Center. Available at: https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=alt_sgpt (Accessed: 18 August 2024).

Haldeman-Englert, C., Turley Jr, R. and Novick, T. (2024) 'Alkaline Phosphatase - Health Encyclopedia', *Health Encyclopedia*. University of Rochester Medical Center. Available at: https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=alkaline_phosphatase. (Accessed: 19 August 2024).

Liver Disorders. (1990). UCI Machine Learning Repository. https://doi.org/10.24432/C54G67 (Accessed: 15 August 2024).

MedlinePlus. (2022). AST Test: MedlinePlus Lab Test Information. MedlinePlus. https://medlineplus.gov/lab-tests/ast-test/ (Accessed: 18 August 2024).

Ritchie, H. and Roser, M. (2018) Alcohol Consumption, Our World in Data. Available at: https://ourworldindata.org/alcohol-consumption (Accessed: 19 August 2024).

Schaeffer, K., & DeSilver, D. (2024). 10 facts about Americans and alcohol as "Dry January" begins. Pew Research Center. Available at: https://www.pewresearch.org/short-reads/2024/01/03/10-facts-about-americans-and-alcohol-as-dry-january-begins/#:~:text=In%202021%2C%20the%20most%20 recent (Accessed: 18 August 2024)

# Appendices

### Appendix 1: Raw Data

mcv, alkphos, sgpt, sgot, gammagt, drinks, selector
85,92,45,27,31,0.0,1
85,64,59,32,23,0.0,2
86,54,33,16,54,0.0,2
91,78,34,24,36,0.0,2
87,70,12,28,10,0.0,2
98,55,13,17,17,0.0,2
88,62,20,17,9,0.5,1
88,67,21,11,11,0.5,1
92,54,22,20,7,0.5,1
90,60,25,19,5,0.5,1
89,52,13,24,15,0.5,1
82,62,17,17,15,0.5,1
90,64,61,32,13,0.5,1
86,77,25,19,18,0.5,1
96,67,29,20,11,0.5,1
91,78,20,31,18,0.5,1
89,67,23,16,10,0.5,1
89,79,17,17,16,0.5,1
91,107,20,20,56,0.5,1
94,116,11,33,11,0.5,1
92,59,35,13,19,0.5,1
93,23,35,20,20,0.5,1
90,60,23,27,5,0.5,1
96,68,18,19,19,0.5,1
84,80,47,33,97,0.5,1
92,70,24,13,26,0.5,1
90,47,28,15,18,0.5,1
88,66,20,21,10,0.5,1
91,102,17,13,19,0.5,1
87,41,31,19,16,0.5,1
86,79,28,16,17,0.5,1
91,57,31,23,42,0.5,1
93,77,32,18,29,0.5,1
88,96,28,21,40,0.5,1
94,65,22,18,11,0.5,1
91,72,155,68,82,0.5,2
85,54,47,33,22,0.5,2
79,39,14,19,9,0.5,2
85,85,25,26,30,0.5,2
89,63,24,20,38,0.5,2

84,92,68,37,44,0.5,2
89,68,26,39,42,0.5,2
89,101,18,25,13,0.5,2
86,84,18,14,16,0.5,2
85,65,25,14,18,0.5,2
88,61,19,21,13,0.5,2
92,56,14,16,10,0.5,2
95,50,29,25,50,0.5,2
91,75,24,22,11,0.5,2
83,40,29,25,38,0.5,2
89,74,19,23,16,0.5,2
85,64,24,22,11,0.5,2
92,57,64,36,90,0.5,2
94,48,11,23,43,0.5,2
87,52,21,19,30,0.5,2
85,65,23,29,15,0.5,2
84,82,21,21,19,0.5,2
88,49,20,22,19,0.5,2
96,67,26,26,36,0.5,2
90,63,24,24,24,0.5,2
90,45,33,34,27,0.5,2
90,72,14,15,18,0.5,2
91,55,4,8,13,0.5,2
91,52,15,22,11,0.5,2
87,71,32,19,27,1.0,1
89,77,26,20,19,1.0,1
89,67,5,17,14,1.0,2
85,51,26,24,23,1.0,2
103,75,19,30,13,1.0,2
90,63,16,21,14,1.0,2
90,63,29,23,57,2.0,1
90,67,35,19,35,2.0,1
87,66,27,22,9,2.0,1
90,73,34,21,22,2.0,1
86,54,20,21,16,2.0,1
90,80,19,14,42,2.0,1
87,90,43,28,156,2.0,2
96,72,28,19,30,2.0,2
91,55,9,25,16,2.0,2
95,78,27,25,30,2.0,2
92,101,34,30,64,2.0,2
89,51,41,22,48,2.0,2
91,99,42,33,16,2.0,2
94,58,21,18,26,2.0,2
92,60,30,27,297,2.0,2
94,58,21,18,26,2.0,2

88,47,33,26,29,2.0,2
92,65,17,25,9,2.0,2
92,79,22,20,11,3.0,1
84,83,20,25,7,3.0,1
88,68,27,21,26,3.0,1
86,48,20,20,6,3.0,1
99,69,45,32,30,3.0,1
88,66,23,12,15,3.0,1
89,62,42,30,20,3.0,1
90,51,23,17,27,3.0,1
81,61,32,37,53,3.0,2
89,89,23,18,104,3.0,2
89,65,26,18,36,3.0,2
92,75,26,26,24,3.0,2
85,59,25,20,25,3.0,2
92,61,18,13,81,3.0,2
89,63,22,27,10,4.0,1
90,84,18,23,13,4.0,1
88,95,25,19,14,4.0,1
89,35,27,29,17,4.0,1
91,80,37,23,27,4.0,1
91,109,33,15,18,4.0,1
91,65,17,5,7,4.0,1
88,107,29,20,50,4.0,2
87,76,22,55,9,4.0,2
87,86,28,23,21,4.0,2
87,42,26,23,17,4.0,2
88,80,24,25,17,4.0,2
90,96,34,49,169,4.0,2
86,67,11,15,8,4.0,2
92,40,19,20,21,4.0,2
85,60,17,21,14,4.0,2
89,90,15,17,25,4.0,2
91,57,15,16,16,4.0,2
96,55,48,39,42,4.0,2
79,101,17,27,23,4.0,2
90,134,14,20,14,4.0,2
89,76,14,21,24,4.0,2
88,93,29,27,31,4.0,2
90,67,10,16,16,4.0,2
92,73,24,21,48,4.0,2
91,55,28,28,82,4.0,2
83,45,19,21,13,4.0,2
90,74,19,14,22,4.0,2
92,66,21,16,33,5.0,1
93,63,26,18,18,5.0,1

86,78,47,39,107,5.0,2
97,44,113,45,150,5.0,2
87,59,15,19,12,5.0,2
86,44,21,11,15,5.0,2
87,64,16,20,24,5.0,2
92,57,21,23,22,5.0,2
90,70,25,23,112,5.0,2
99,59,17,19,11,5.0,2
92,80,10,26,20,6.0,1
95,60,26,22,28,6.0,1
91,63,25,26,15,6.0,1
92,62,37,21,36,6.0,1
95,50,13,14,15,6.0,1
90,76,37,19,50,6.0,1
96,70,70,26,36,6.0,1
95,62,64,42,76,6.0,1
92,62,20,23,20,6.0,1
91,63,25,26,15,6.0,1
82,56,67,38,92,6.0,2
92,82,27,24,37,6.0,2
90,63,12,26,21,6.0,2
88,37,9,15,16,6.0,2
100,60,29,23,76,6.0,2
98,43,35,23,69,6.0,2
91,74,87,50,67,6.0,2
92,87,57,25,44,6.0,2
93,99,36,34,48,6.0,2
90,72,17,19,19,6.0,2
97,93,21,20,68,6.0,2
93,50,18,25,17,6.0,2
90,57,20,26,33,6.0,2
92,76,31,28,41,6.0,2
88,55,19,17,14,6.0,2
89,63,24,29,29,6.0,2
92,79,70,32,84,7.0,1
92,93,58,35,120,7.0,1
93,84,58,47,62,7.0,2
97,71,29,22,52,8.0,1
84,99,33,19,26,8.0,1
96,44,42,23,73,8.0,1
90,62,22,21,21,8.0,1
92,94,18,17,6,8.0,1
90,67,77,39,114,8.0,1
97,71,29,22,52,8.0,1
91,69,25,25,66,8.0,2
93,59,17,20,14,8.0,2

92,95,85,48,200,8.0,2
90,50,26,22,53,8.0,2
91,62,59,47,60,8.0,2
92,93,22,28,123,9.0,1
92,77,86,41,31,10.0,1
86,66,22,24,26,10.0,2
98,57,31,34,73,10.0,2
95,80,50,64,55,10.0,2
92,108,53,33,94,12.0,2
97,92,22,28,49,12.0,2
93,77,39,37,108,16.0,1
94,83,81,34,201,20.0,1
87,75,25,21,14,0.0,1
88,56,23,18,12,0.0,1
84,97,41,20,32,0.0,2
94,91,27,20,15,0.5,1
97,62,17,13,5,0.5,1
92,85,25,20,12,0.5,1
82,48,27,15,12,0.5,1
88,74,31,25,15,0.5,1
95,77,30,14,21,0.5,1
88,94,26,18,8,0.5,1
91,70,19,19,22,0.5,1
83,54,27,15,12,0.5,1
91,105,40,26,56,0.5,1
86,79,37,28,14,0.5,1
91,96,35,22,135,0.5,1
89,82,23,14,35,0.5,1
90,73,24,23,11,0.5,1
90,87,19,25,19,0.5,1
89,82,33,32,18,0.5,1
85,79,17,8,9,0.5,1
85,119,30,26,17,0.5,1
78,69,24,18,31,0.5,1
88,107,34,21,27,0.5,1
89,115,17,27,7,0.5,1
92,67,23,15,12,0.5,1
89,101,27,34,14,0.5,1
91,84,11,12,10,0.5,1
94,101,41,20,53,0.5,2
88,46,29,22,18,0.5,2
88,122,35,29,42,0.5,2
84,88,28,25,35,0.5,2
90,79,18,15,24,0.5,2
87,69,22,26,11,0.5,2
65,63,19,20,14,0.5,2

90,64,12,17,14,0.5,2
85,58,18,24,16,0.5,2
88,81,41,27,36,0.5,2
86,78,52,29,62,0.5,2
82,74,38,28,48,0.5,2
86,58,36,27,59,0.5,2
94,56,30,18,27,0.5,2
87,57,30,30,22,0.5,2
98,74,148,75,159,0.5,2
94,75,20,25,38,0.5,2
83,68,17,20,71,0.5,2
93,56,25,21,33,0.5,2
101,65,18,21,22,0.5,2
92,65,25,20,31,0.5,2
92,58,14,16,13,0.5,2
86,58,16,23,23,0.5,2
85,62,15,13,22,0.5,2
86,57,13,20,13,0.5,2
86,54,26,30,13,0.5,2
81,41,33,27,34,1.0,1
91,67,32,26,13,1.0,1
91,80,21,19,14,1.0,1
92,60,23,15,19,1.0,1
91,60,32,14,8,1.0,1
93,65,28,22,10,1.0,1
90,63,45,24,85,1.0,2
87,92,21,22,37,1.0,2
83,78,31,19,115,1.0,2
95,62,24,23,14,1.0,2
93,59,41,30,48,1.0,2
84,82,43,32,38,2.0,1
87,71,33,20,22,2.0,1
86,44,24,15,18,2.0,1
86,66,28,24,21,2.0,1
88,58,31,17,17,2.0,1
90,61,28,29,31,2.0,1
88,69,70,24,64,2.0,1
93,87,18,17,26,2.0,1
98,58,33,21,28,2.0,1
91,44,18,18,23,2.0,2
87,75,37,19,70,2.0,2
94,91,30,26,25,2.0,2
88,85,14,15,10,2.0,2
89,109,26,25,27,2.0,2
87,59,37,27,34,2.0,2
93,58,20,23,18,2.0,2

88,57,9,15,16,2.0,2
94,65,38,27,17,3.0,1
91,71,12,22,11,3.0,1
90,55,20,20,16,3.0,1
91,64,21,17,26,3.0,2
88,47,35,26,33,3.0,2
82,72,31,20,84,3.0,2
85,58,83,49,51,3.0,2
91,54,25,22,35,4.0,1
98,50,27,25,53,4.0,2
86,62,29,21,26,4.0,2
89,48,32,22,14,4.0,2
82,68,20,22,9,4.0,2
83,70,17,19,23,4.0,2
96,70,21,26,21,4.0,2
94,117,77,56,52,4.0,2
93,45,11,14,21,4.0,2
93,49,27,21,29,4.0,2
84,73,46,32,39,4.0,2
91,63,17,17,46,4.0,2
90,57,31,18,37,4.0,2
87,45,19,13,16,4.0,2
91,68,14,20,19,4.0,2
86,55,29,35,108,4.0,2
91,86,52,47,52,4.0,2
88,46,15,33,55,4.0,2
85,52,22,23,34,4.0,2
89,72,33,27,55,4.0,2
95,59,23,18,19,4.0,2
94,43,154,82,121,4.0,2
96,56,38,26,23,5.0,2
90,52,10,17,12,5.0,2
94,45,20,16,12,5.0,2
99,42,14,21,49,5.0,2
93,102,47,23,37,5.0,2
94,71,25,26,31,5.0,2
92,73,33,34,115,5.0,2
87,54,41,29,23,6.0,1
92,67,15,14,14,6.0,1
98,101,31,26,32,6.0,1
92,53,51,33,92,6.0,1
97,94,43,43,82,6.0,1
93,43,11,16,54,6.0,1
93,68,24,18,19,6.0,1
95,36,38,19,15,6.0,1
99,86,58,42,203,6.0,1

98,66,103,57,114,6.0,1
92,80,10,26,20,6.0,1
96,74,27,25,43,6.0,2
95,93,21,27,47,6.0,2
86,109,16,22,28,6.0,2
91,46,30,24,39,7.0,2
102,82,34,78,203,7.0,2
85,50,12,18,14,7.0,2
91,57,33,23,12,8.0,1
91,52,76,32,24,8.0,1
93,70,46,30,33,8.0,1
87,55,36,19,25,8.0,1
98,123,28,24,31,8.0,1
82,55,18,23,44,8.0,2
95,73,20,25,225,8.0,2
97,80,17,20,53,8.0,2
100,83,25,24,28,8.0,2
88,91,56,35,126,9.0,2
91,138,45,21,48,10.0,1
92,41,37,22,37,10.0,1
86,123,20,25,23,10.0,2
91,93,35,34,37,10.0,2
87,87,15,23,11,10.0,2
87,56,52,43,55,10.0,2
99,75,26,24,41,12.0,1
96,69,53,43,203,12.0,2
98,77,55,35,89,15.0,1
91,68,27,26,14,16.0,1
98,99,57,45,65,20.0,1

**Appendix 2: SAS Code**

```
data d; /* Assigns name d to file*/
input x1 x2 x3 y x4 x5; /* assigns variable names:*/
cards; /* start to input data */
85 92 45 27 31 0.0
85 64 59 32 23 0.0
86 54 33 16 54 0.0
91 78 34 24 36 0.0
87 70 12 28 10 0.0
98 55 13 17 17 0.0
88 62 20 17 9 0.5
88 67 21 11 11 0.5
92 54 22 20 7 0.5
90 60 25 19 5 0.5
89 52 13 24 15 0.5
```

82 62 17 17 15 0.5
90 64 61 32 13 0.5
86 77 25 19 18 0.5
96 67 29 20 11 0.5
91 78 20 31 18 0.5
89 67 23 16 10 0.5
89 79 17 17 16 0.5
91 107 20 20 56 0.5
94 116 11 33 11 0.5
92 59 35 13 19 0.5
93 23 35 20 20 0.5
90 60 23 27 5 0.5
96 68 18 19 19 0.5
84 80 47 33 97 0.5
92 70 24 13 26 0.5
90 47 28 15 18 0.5
88 66 20 21 10 0.5
91 102 17 13 19 0.5
87 41 31 19 16 0.5
86 79 28 16 17 0.5
91 57 31 23 42 0.5
93 77 32 18 29 0.5
88 96 28 21 40 0.5
94 65 22 18 11 0.5
91 72 155 68 82 0.5
85 54 47 33 22 0.5
79 39 14 19 9 0.5
85 85 25 26 30 0.5
89 63 24 20 38 0.5
84 92 68 37 44 0.5
89 68 26 39 42 0.5
89 101 18 25 13 0.5
86 84 18 14 16 0.5
85 65 25 14 18 0.5
88 61 19 21 13 0.5
92 56 14 16 10 0.5
95 50 29 25 50 0.5
91 75 24 22 11 0.5
83 40 29 25 38 0.5
89 74 19 23 16 0.5
85 64 24 22 11 0.5
92 57 64 36 90 0.5
94 48 11 23 43 0.5
87 52 21 19 30 0.5
85 65 23 29 15 0.5
84 82 21 21 19 0.5

88 49 20 22 19 0.5
96 67 26 26 36 0.5
90 63 24 24 24 0.5
90 45 33 34 27 0.5
90 72 14 15 18 0.5
91 55 4 8 13 0.5
91 52 15 22 11 0.5
87 71 32 19 27 1.0
89 77 26 20 19 1.0
89 67 5 17 14 1.0
85 51 26 24 23 1.0
103 75 19 30 13 1.0
90 63 16 21 14 1.0
90 63 29 23 57 2.0
90 67 35 19 35 2.0
87 66 27 22 9 2.0
90 73 34 21 22 2.0
86 54 20 21 16 2.0
90 80 19 14 42 2.0
87 90 43 28 156 2.0
96 72 28 19 30 2.0
91 55 9 25 16 2.0
95 78 27 25 30 2.0
92 101 34 30 64 2.0
89 51 41 22 48 2.0
91 99 42 33 16 2.0
94 58 21 18 26 2.0
92 60 30 27 297 2.0
94 58 21 18 26 2.0
88 47 33 26 29 2.0
92 65 17 25 9 2.0
92 79 22 20 11 3.0
84 83 20 25 7 3.0
88 68 27 21 26 3.0
86 48 20 20 6 3.0
99 69 45 32 30 3.0
88 66 23 12 15 3.0
89 62 42 30 20 3.0
90 51 23 17 27 3.0
81 61 32 37 53 3.0
89 89 23 18 104 3.0
89 65 26 18 36 3.0
92 75 26 26 24 3.0
85 59 25 20 25 3.0
92 61 18 13 81 3.0
89 63 22 27 10 4.0

90 84 18 23 13 4.0
88 95 25 19 14 4.0
89 35 27 29 17 4.0
91 80 37 23 27 4.0
91 109 33 15 18 4.0
91 65 17 5 7 4.0
88 107 29 20 50 4.0
87 76 22 55 9 4.0
87 86 28 23 21 4.0
87 42 26 23 17 4.0
88 80 24 25 17 4.0
90 96 34 49 169 4.0
86 67 11 15 8 4.0
92 40 19 20 21 4.0
85 60 17 21 14 4.0
89 90 15 17 25 4.0
91 57 15 16 16 4.0
96 55 48 39 42 4.0
79 101 17 27 23 4.0
90 134 14 20 14 4.0
89 76 14 21 24 4.0
88 93 29 27 31 4.0
90 67 10 16 16 4.0
92 73 24 21 48 4.0
91 55 28 28 82 4.0
83 45 19 21 13 4.0
90 74 19 14 22 4.0
92 66 21 16 33 5.0
93 63 26 18 18 5.0
86 78 47 39 107 5.0
97 44 113 45 150 5.0
87 59 15 19 12 5.0
86 44 21 11 15 5.0
87 64 16 20 24 5.0
92 57 21 23 22 5.0
90 70 25 23 112 5.0
99 59 17 19 11 5.0
92 80 10 26 20 6.0
95 60 26 22 28 6.0
91 63 25 26 15 6.0
92 62 37 21 36 6.0
95 50 13 14 15 6.0
90 76 37 19 50 6.0
96 70 70 26 36 6.0
95 62 64 42 76 6.0
92 62 20 23 20 6.0

91 63 25 26 15 6.0
82 56 67 38 92 6.0
92 82 27 24 37 6.0
90 63 12 26 21 6.0
88 37 9 15 16 6.0
100 60 29 23 76 6.0
98 43 35 23 69 6.0
91 74 87 50 67 6.0
92 87 57 25 44 6.0
93 99 36 34 48 6.0
90 72 17 19 19 6.0
97 93 21 20 68 6.0
93 50 18 25 17 6.0
90 57 20 26 33 6.0
92 76 31 28 41 6.0
89 63 24 29 29 6.0
92 79 70 32 84 7.0
92 93 58 35 120 7.0
93 84 58 47 62 7.0
97 71 29 22 52 8.0
84 99 33 19 26 8.0
96 44 42 23 73 8.0
90 62 22 21 21 8.0
92 94 18 17 6 8.0
90 67 77 39 114 8.0
97 71 29 22 52 8.0
91 69 25 25 66 8.0
93 59 17 20 14 8.0
92 95 85 48 200 8.0
90 50 26 22 53 8.0
91 62 59 47 60 8.0
92 93 22 28 123 9.0
92 77 86 41 31 10.0
86 66 22 24 26 10.0
98 57 31 34 73 10.0
95 80 50 64 55 10.0
92 108 53 33 94 12.0
97 92 22 28 49 12.0
93 77 39 37 108 16.0
94 83 81 34 201 20.0
87 75 25 21 14 0.0
88 56 23 18 12 0.0
84 97 41 20 32 0.0
94 91 27 20 15 0.5
97 62 17 13 5 0.5
92 85 25 20 12 0.5

82 48 27 15 12 0.5
88 74 31 25 15 0.5
95 77 30 14 21 0.5
88 94 26 18 8 0.5
91 70 19 19 22 0.5
83 54 27 15 12 0.5
91 105 40 26 56 0.5
86 79 37 28 14 0.5
91 96 35 22 135 0.5
89 82 23 14 35 0.5
90 73 24 23 11 0.5
90 87 19 25 19 0.5
89 82 33 32 18 0.5
85 79 17 8 9 0.5
85 119 30 26 17 0.5
78 69 24 18 31 0.5
88 107 34 21 27 0.5
89 115 17 27 7 0.5
92 67 23 15 12 0.5
89 101 27 34 14 0.5
91 84 11 12 10 0.5
94 101 41 20 53 0.5
88 46 29 22 18 0.5
88 122 35 29 42 0.5
84 88 28 25 35 0.5
90 79 18 15 24 0.5
87 69 22 26 11 0.5
65 63 19 20 14 0.5
90 64 12 17 14 0.5
85 58 18 24 16 0.5
88 81 41 27 36 0.5
86 78 52 29 62 0.5
82 74 38 28 48 0.5
86 58 36 27 59 0.5
94 56 30 18 27 0.5
87 57 30 30 22 0.5
98 74 148 75 159 0.5
94 75 20 25 38 0.5
83 68 17 20 71 0.5
93 56 25 21 33 0.5
101 65 18 21 22 0.5
92 65 25 20 31 0.5
92 58 14 16 13 0.5
86 58 16 23 23 0.5
85 62 15 13 22 0.5
86 57 13 20 13 0.5

86 54 26 30 13 0.5
81 41 33 27 34 1.0
91 67 32 26 13 1.0
91 80 21 19 14 1.0
92 60 23 15 19 1.0
91 60 32 14 8 1.0
93 65 28 22 10 1.0
90 63 45 24 85 1.0
87 92 21 22 37 1.0
83 78 31 19 115 1.0
95 62 24 23 14 1.0
93 59 41 30 48 1.0
84 82 43 32 38 2.0
87 71 33 20 22 2.0
86 44 24 15 18 2.0
86 66 28 24 21 2.0
88 58 31 17 17 2.0
90 61 28 29 31 2.0
88 69 70 24 64 2.0
93 87 18 17 26 2.0
98 58 33 21 28 2.0
91 44 18 18 23 2.0
87 75 37 19 70 2.0
94 91 30 26 25 2.0
88 85 14 15 10 2.0
89 109 26 25 27 2.0
87 59 37 27 34 2.0
93 58 20 23 18 2.0
88 57 9 15 16 2.0
94 65 38 27 17 3.0
91 71 12 22 11 3.0
90 55 20 20 16 3.0
91 64 21 17 26 3.0
88 47 35 26 33 3.0
82 72 31 20 84 3.0
85 58 83 49 51 3.0
91 54 25 22 35 4.0
98 50 27 25 53 4.0
86 62 29 21 26 4.0
89 48 32 22 14 4.0
82 68 20 22 9 4.0
83 70 17 19 23 4.0
96 70 21 26 21 4.0
94 117 77 56 52 4.0
93 45 11 14 21 4.0
93 49 27 21 29 4.0

84 73 46 32 39 4.0
91 63 17 17 46 4.0
90 57 31 18 37 4.0
87 45 19 13 16 4.0
91 68 14 20 19 4.0
86 55 29 35 108 4.0
91 86 52 47 52 4.0
88 46 15 33 55 4.0
85 52 22 23 34 4.0
89 72 33 27 55 4.0
95 59 23 18 19 4.0
94 43 154 82 121 4.0
96 56 38 26 23 5.0
90 52 10 17 12 5.0
94 45 20 16 12 5.0
99 42 14 21 49 5.0
93 102 47 23 37 5.0
94 71 25 26 31 5.0
92 73 33 34 115 5.0
87 54 41 29 23 6.0
92 67 15 14 14 6.0
98 101 31 26 32 6.0
92 53 51 33 92 6.0
97 94 43 43 82 6.0
93 43 11 16 54 6.0
93 68 24 18 19 6.0
95 36 38 19 15 6.0
99 86 58 42 203 6.0
98 66 103 57 114 6.0
92 80 10 26 20 6.0
96 74 27 25 43 6.0
95 93 21 27 47 6.0
86 109 16 22 28 6.0
91 46 30 24 39 7.0
102 82 34 78 203 7.0
85 50 12 18 14 7.0
91 57 33 23 12 8.0
91 52 76 32 24 8.0
93 70 46 30 33 8.0
87 55 36 19 25 8.0
98 123 28 24 31 8.0
82 55 18 23 44 8.0
95 73 20 25 225 8.0
97 80 17 20 53 8.0
100 83 25 24 28 8.0
88 91 56 35 126 9.0

```
91 138 45 21 48 10.0
92 41 37 22 37 10.0
86 123 20 25 23 10.0
91 93 35 34 37 10.0
87 87 15 23 11 10.0
87 56 52 43 55 10.0
99 75 26 24 41 12.0
96 69 53 43 203 12.0
98 77 55 35 89 15.0
91 68 27 26 14 16.0
98 99 57 45 65 20.0
;

/*(Multicollinearity test)*/

proc plot data = d;
plot y*(x1 x2 x3 x4 x5); /* plot y vs x's */
proc corr data=d; /* correlation matrix */
proc reg data = d; /* specifies regression procedure */
model y = x1 x2 x3 x4 x5 / p vif; /* VIF*/
proc reg data = d; /* specifies regression procedure */
model y = x2 x3 x4 x5 ; /* Data without x1*/

proc reg data = d; /* specifies regression procedure */
model y = x3 x4 x5 /; /* Data without x1 and x2*/

/*(Finding model comparison for all 5 different variables x1 through x5)*/

proc rsquare cp data = d;
model y = x1 x2 x3 x4 x5;
run;

proc reg data=d;
model y=x1 x2 x3 x4 x5/selection=rsquare adjrsq cp mse;
run;

proc glm data = d;
model y = x1 x2 x3 x4 x5/ p clm;

/*(Residual Plots)*/
proc reg data = d; /* specifies regression procedure */
model y = x2 x3 x4 x5 /alpha=0.05 p clm cli clb;
output out=res p = yhat r = redid ; /*output predict value and residual to a new data set "res" */
run;
proc plot data = res; plot redid*(x2 x3 x4 x5);
run;
```