Capstone 1 Project: Liver Disease Prediction
Inferential Statistics

Nils Madsen

Approach

The goal of this statistical analysis was to find significant associations among the predictive features (age, sex, and blood test results), as well as associations between the predictive features and the presence of liver disease. A frequentist approach was used, due to its simplicity and theoretical robustness. The specific statistical tests used include Pearson correlation, t-test for means, and z-test for proportions.

Central Limit Theorem
T-tests and z-tests rely on the assumption of normally distributed sampling statistics. Therefore, it is important to first establish that the preconditions of the Central Limit Theorem are satisfied for this data. The two main conditions to ensure that the Central Limit Theorem applies are that there is a sufficiently large number of observations, and that the individual observations are independent.

Since each observation is from an individual patient, we can safely consider all observations to be independent.

The liver disease group consists of 416 observations, while the control group consists of 167 observations, so the sample size is large enough for any disease-vs-control group t-tests. To ensure that the sample size is large enough for the sex vs disease proportions z-test, there should be more than 10 observations in each combination of sex and disease group. The male segment of the dataset contains 324 liver patients and 117 controls, while the female segment contains 92 liver patients and 50 controls, so there are enough observations to perform this test.

Results

Correlation among diagnostic metrics

Pearson correlation was used to generate coefficients and p-values for every pair of numeric features (i.e. excluding sex and disease group). The null hypothesis in each case was that the two features do not covary.

The liver enzyme tests (ALP, ALT, AST) are all at least weakly associated with each other, with statistically-significant p-values ($p < 0.01$). A particularly strong association exists between AST and ALT. This may indicate that their concentration in the blood is connected by similar homeostatic processes and similar response to disease states, and therefore the two measures might have overlapping diagnostic significance. ALP has only weak associations with the other liver enzymes, indicating that it may have unique diagnostic value in this group.

Within the blood protein measures (total protein, albumin, and a-g ratio), albumin is strongly associated with the other two measures. This is not surprising, given that albumin makes up a large proportion of total blood protein, and is the numerator in albumin-globulin ratio. Total protein and a-g ratio are only weakly correlated. All correlations are statistically significant with $p < 10^{-7}$.

Total and direct bilirubin are also strongly correlated, which is not surprising given that direct bilirubin is a component of total bilirubin.

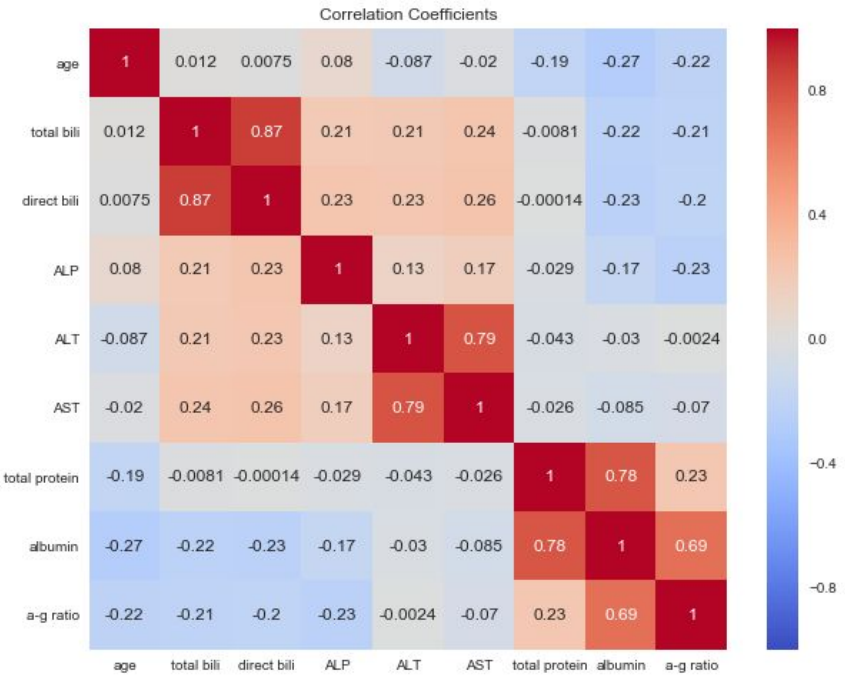

Figure 1: Correlation heatmap of diagnostic features

Association of sex with the other diagnostic metrics

To investigate whether any of the numeric features is significantly related to sex, t-tests were performed to compare men and women across each metric. The null hypothesis in each case was that there is no difference in the mean of the measure between men and women.

The results of these t-tests indicate that bilirubin, total protein, and albumin are significantly different between men and women. We cannot rule out the null hypothesis in the case of age, ALP, and a-g ratio. ALT and AST have p-values that are close to the threshold of significance ($p \sim 0.05$), with ALT having a p-value slightly below the threshold, and AST having a p-value slightly above.

diagnostic metric	male mean	female mean	t-statistic	p-value
age	45.265306	43.133803	1.365511	0.172621
total bili	3.613152	2.322535	2.160892	0.031112
direct bili	1.646032	0.989437	2.433218	0.015266
ALP	286.789116	302.338028	-0.663017	0.507583
ALT	89.238095	54.239437	1.991292	0.046917
AST	123.070295	69.042254	1.942699	0.052535
total protein	6.428345	6.653521	-2.156751	0.031435
albumin	3.099546	3.273239	-2.270944	0.023515
a-g ratio	0.946449	0.948973	-0.082046	0.934639

Table 1: Diagnostic measures by sex

Relationships of the numeric diagnostic metrics to liver disease

To investigate whether each of the numeric predictive measures is significantly related to the presence of liver disease, t-tests were performed to compare the liver disease group to the control group across each feature. The null hypothesis in each case was that there is no difference in the mean of the metric between those with liver disease and those without liver disease.

All numeric measures, with the exception of total protein, were significantly different in the disease and control groups. Age, total and direct bilirubin, ALP, ALT, and AST were all higher in the disease group, while albumin and a-g ratio were lower in the disease group. Since these tests are used in practice to diagnose liver disease, it makes sense that they would be associated with a liver disease diagnosis in a statistically-significant way.

Total protein was the exception here, and the fact that it was not significantly associated with liver disease group raises questions about its utility here. When the machine learning model is trained, it might be worthwhile to investigate the model's performance with and without this feature included.

diagnostic metric	liver patient mean	control mean	t-statistic	p-value
age	46.153846	41.239521	3.342369	8.840632e-04
total bili	4.164423	1.142515	5.441441	7.801431e-08
direct bili	1.923558	0.396407	6.118790	1.734103e-09
ALP	319.007212	219.754491	4.534141	7.027487e-06
ALT	99.605769	33.652695	3.992646	7.371781e-05
AST	137.699519	40.688623	3.705217	2.313901e-04
total protein	6.459135	6.543114	-0.844354	3.988191e-01
albumin	3.060577	3.344311	-3.941755	9.074361e-05
a-g ratio	0.914337	1.028588	-3.965115	8.251144e-05

Table 2: Diagnostic measures by disease group

Association of sex with liver disease diagnosis

A proportions z-test was used to investigate whether there is an association between sex and liver disease in this dataset. The proportion of liver patients was compared in the male and female groups, with the null hypothesis that there is no difference in proportion of liver patients between men and women.

The results of the z-test indicate that the proportion of liver patients among men is indeed different from the proportion among women (0.73 vs 0.65; $p = 0.047$). This may indicate that male sex is a risk factor for liver disease.