Analysis of Common Symptoms and Relative Blood Test Index Association with Hepatitis Patients' Survival

Yiziying Chen

Abstract

Importance: Some common symptoms were known to be shared among hepatitis patients, including fatigue, malaise, abdominal pain etc. Certain types of hepatitis could develop into life-threatening liver infection/cancer. Given the presence/absence of certain symptoms combined with blood test index results, a hepatitis patient's survival chance is yet to be analyzed.

Objective: To analyze the association between blood test index and the presence/absence of certain symptoms to predict the survival chance of hepatitis patients.

Design, Setting, Participants: This hepatitis patients' data was provided by Gail Gong from Carnegie-Mellon University via Bojan Cestnik Jozef Stefan Institute on November 1st, 1988. The dataset includes 155 hepatitis patients' symptoms and numeric blood test index records relative to clinical identification of hepatitis. Binary logistic regression and multinomial logistic regression were used to assess the outcome referring as patients' survival chance.

Main Outcomes and Measures: The main outcome of the study was the categorical "class" of living status (die/live). The secondary outcome is the abnormality of a protein's amount present in blood.

Results: Among 155 hepatitis patients in the study (mean age 41.2; 139 males [89.68%]), 123 had death outcome ([79.35%], all males). At 0.1 significance level, there was no significant independency between the outcome and Boolean attributes antiviral, anorexia, liver big and liver firm. Chi-square tests suggested significantly greater odds in death for patients who felt fatigue (OR = 11.14; 95% CI, 2.60-99.35; p = 5.18e-05); had malaise (OR = 5.65; 95% CI, 2.39-13.36; p = 2.76e-05); whose spleen was palpable (OR = 3.54; 95% CI, 1.47-8.54; p = 0.003); had spider angioma (OR = 7.59; 95% CI, 3.14-18.31; p = 1.07e-06); took doses of anabolic steroids (OR = 1.96; 95% CI, 0.88-4.37, p = 0.095); had ascites (OR = 15.51;95% CI, 5.25-45.84; p = 4.83e-09); had varices (OR = 8.8, 95% CI, 3.05-25.41; p = 6.26e-06). Breslow-Day test suggests patients with/out any symptom had no significant independence with death when adjusting for another symptom (p-values < 0.05). A binary *logit* model was built taking steroid (β = 1.41, p = 0.125), ascites (β=1.83, p=0.034), varices (β=1.78, p=0.088), sgot (β=0.0108, p=0.184) and albumin (β=0.042, p=0.0284) as predictors. A *probit* model was built using age (β=-0.03, p=0.148), steroid (β=1.09, p=0.0626), ascites (β=1.21, p=0.0185), varices (β=1.01, p=0.105), alk_phosphate (β=-0.006, p=0.146), sgot (β=0.009, p=0.088) and protime (β=0.023, p=0.0317) as predictors. Both models have power (AUC(*logit*) = 0.8955, AUC(*probit*) = 0.915) in predicting log hazards of death.

Conclusions and Relevance: Hepatitis patients who had symptoms including fatigue, malaise, palpable spleen, ascites, varices had much higher hazards of death. Blood test index, blood sgot level and albumin amount, can be used for predicting survival chance of hepatitis patients. Hepatitis patients who had ascites will have increased odds of damaged liver cells. These findings point to the importance of certain symptoms and objective measurement of blood test index for existing hepatitis patients,

which may provide effective precaution strategy from worsening the existing hepatitis status into cirrhosis/cancer, which causes much higher chance of death.

Introduction

Hepatitis is an inflammation of liver commonly caused by hepatitis viruses. There are 5 main hepatitis viruses, types A, B, C, D and E.[1] According to the Centers for Disease Control and prevention (CDC), approximately 4.4 million Americans are currently living with chronic hepatitis B and C.[2] Many more people don't even know that they have hepatitis.

Hepatitis B virus is transmitted through exposure to infective blood, semen and other body fluids.[3,4,6] The causes of noninfectious hepatitis include overuse or overdose of medication. When infected, acute hepatitis appears quickly with several major symptoms as fatigue, dark urine, abdominal pain, appetite loss, etc.[5,6]

The likelihood that hepatitis infection becomes chronic is dependent on the age at which a person becomes infected.[3,6] Children and infants have much higher chance of developing the infection into chronic disease. Specifically, > 80% of infants infected during the first year of life develop chronic infections and ~40% of children infected before age 6 develop chronic infections.[3] Adults have relatively lower risk for chronic hepatitis. < 5% healthy persons infected as adults will develop chronic infection and 20%-30% of adults who are chronically infected will develop cirrhosis and/or liver cancer.[3]

When in the procedure of diagnosing hepatitis, it requires both clinic and laboratory blood tests for confirmation of hepatitis type and distinguish acute and chronic infections. No specific treatment has been proposed for treating acute hepatitis B but more aiming towards maintaining the patient's physical comfort and adequate nutritional balance. Chronic hepatitis B can be treated with medicines but mostly don't cure the disease. Treatment can slow the pression of cirrhosis, reduce incidence of liver cancer and improve long term survival.[3,5]

Methods

Study Sample Description:

A retrospective cohort study was conduced upon 155 hepatitis patients' symptom records and blood test index measurements. Each hepatitis patient was examined for common symptoms related to hepatitis and was asked whether took any treatment that could possibly affect liver functioning (including anabolic steroid, antiviral treatments). Symptoms that were examined and reported including fatigue, loss of appetite, weight loss, swollen vessels, abdominal pain. Blood test were given, and relative indices were reported including but not limited to blood bilirubin and albumin levels. The end event of the cohort study was death of the patient. This dataset was collected via Bojan Cestnik Jozef Stefan Institute and made public by Gail Gong from Carnegie-Mellon University on November 1st, 1988. The data also included demographic information, age and gender, of hepatitis patients.

Variable Information:

Binary outcome variable—Class: DIE(0), LIVE(1);

18 attributes:

Attributes	Type	Description	Symbol-binary
Age	numeric	Age of individual patients	Min = 7, max =78, no missing observation
Sex	binary	Gender of individual patients	male(0), female(1); no missing observation
Steroid	binary	whether took anabolic steroid	yes(0), no(1); 1 missing observation
Antiviral	binary	whether took antiviral medications	yes(0), no(1); 1 missing observation
Fatigue	binary	whether took anabolic steroid	yes(0), no(1); 1 missing observation
Malaise	binary	whether experienced malaise	yes(0), no(1); 1 missing observation
Anorexia	binary	whether experience anorexia	yes(0), no(1); 1 missing observation
Liver big	binary	whether the size of liver became bigger	yes(0), no(1); 10 missing observation
Liver firm	binary	whether liver became firm	yes(0), no(1); 11 missing observation
Spleen palpable	binary	Whether spleen became palpable	yes(0), no(1); 5 missing observation
Spiders	binary	Whether had swollen vessels	yes(0), no(1); 5 missing observation
Ascites	binary	Whether fluid accumulated in abdomen	yes(0), no(1); 5 missing observation
Varices	binary	Whether had enlarged veins	yes(0), no(1); 5 missing observation
Bilirubin	numeric	Bilirubin is a compound that occurs in the normal catabolic pathway that breaks down heme in vertebrates. Levels of bilirubin in the blood go up and down in patients with hepatitis C.[7]	Min=0.3 mg/dL max = 8 mg/dL; normal range: < 0.3mg/dL[8]; 6 missing records
Alk phosphate	numeric	Alkaline phosphate is a protein enzyme. A high alk phosphate level occurs when there is a blockage of flow in the biliary tact or a buildup of pressure in the liver.[7]	Min = 26 U/L, max = 295 U/L; normal range: 44-147 U/L[9]; 29 missing records
Sgot	numeric	Sgot is a protein made by liver cells. When liver cells are damaged, sgot leaks out	Min = 14 U/L, max = 648 U/L; normal range: 0-40 U/L[10]; 4 missing records

		into the blood stream and the level of sgot in the blood becomes higher than normal.[7]	
Albumin	numeric	Albumin is a protein made by liver, it prevents fluid from leaking out of blood vessels into tissues.[7]	Min = 2.1 g/dL, max = 6.4 g/dL; normal range: 3.4-5.4 g/dL[11]; 16 missing records
Protime	numeric	Prothrombin is a protein made by liver, it helps blood to make normal clots. The prothrombin time is one way of measuring how long it takes blood to form a clot.[7]	Min = 0, max = 100sec; normal range: 11-13.5sec[12]; 67 missing records

Statistical Methods:

Constructed contingency tables between individual symptoms and outcome counts. Hypothesis testing of independence was performed to analyze the association between individual symptoms and outcome using Chi-Squared tests. Fisher's Exact test was used when cell counts are small (<5). Tested conditional independence between a symptom and outcome given another symptom using Cochran-Mantel-Haenszel test and reported significant conditional independence pairs. Log-binomial regression models were built using logit and probit link. Performed stepwise selection to select for the best model (with lowest AIC value). Compared the full model with reduced model using ANOVA test and checked multicollinearity on the regression; fitted an expanded model to test whether additional interaction term was significant and assessed lack of fit; performed residual analysis to check outliers and/or high influential observations. Evaluated predictive power of the regression models using receiver operating characteristic (ROC) curve and areas under ROC curve. numeric variables, sgot and albumin were each categorized into 4 levels using the its own 25^{th}, 50^{th}, 75^{th} and 100^{th} quantile value as cutoff lines. Multicategory logit models were built for estimating the significant change in log odds of ordinal outcome compared to reference level for unit change in numeric predictor or with/out certain symptoms. Statistical analysis was performed all using 2-sided tests with a significance level <0.1.

Results

Among all patients from the study, 32 died (20.6%) and 123 remained alive. Within total 139 male patient, 32 (23%) died and all 16 female patients were alive. The mean age of patients was 41.2, with youngest patient aged 7 and the oldest aged 78. To check the association between individual symptoms and survival outcome, contingency tables are built and summarized in **Appendix Table 1**. Counts of patients with/out any symptoms are summarized in **Table 1**. At 0.1 significance level, the odds of death for patients who took antiviral treatment (OR = 0.31; 95% CI, 0.033-1.38; p = 0.168), experienced anorexia (OR = 0.31; 95% CI, 0.033-1.38; p = 0.168), had bigger liver size (OR = 0.55; 95% CI, 0.1-

Table 1. Counts of Hepatitis Patients regarding each symptoms and characteristics of ORs for death between 2 grouped counts

Symptoms	#Yes	#No	Estimated OR	95% CI OR	P Value
Steroid[a,b]	76	78	1.96	0.88-4.34	0.0946
Antiviral[c]	24	131	0.31	0.033-1.38	0.168
Fatigue[a,c]	100	54	11.14	2.60-99.36	5.18e-05
Malaise[a,b]	61	93	5.65	2.39-13.36	2.76e-05
Anorexia[a,b]	32	122	2.07	0.86-4.97	0.10
Liver Big[a,c]	25	120	0.55	0.10-2.06	0.57
Liver Firm[a,b]	60	84	1.38	0.60-3.21	0.449
Spleen Palpable[a,b]	30	120	3.54	1.47-8.55	0.00346
Spiders[a,b]	51	99	7.59	3.14-18.31	1.07e-06
Ascites[a,b]	20	130	15.51	3.05-25.41	4.84e-09
Varices[a,b]	18	132	0.16	0.065-0.40	6.26e-06

[a]with missing observations
[b]Chi-square test for independence
[c]Fisher's exact test for independence

2.06; p = 0.57), had firm liver (OR = 1.38; 95% CI, 0.6-3.21; p = 0.449) are not significantly different from patients who didn't have each of these symptoms. In this study, all female patients were free from death, so gender effect was removed from analysis.

Summaries of numeric attributes in **Appendix Table 2**.

Three-way tables are built and tested by **Breslow-Day** test for homogeneous ORs between stratum. Breslow-day test results are collectively summarized in **Appendix Table 3**. The test results show there is no significant independent (p < 0.1) association between any symptom and death outcome when adjusting for another symptom. Since there is no homogenous ORs, no need to perform additional **Mantel–Haenszel** test for testing all ORs = 1.

For predicting binary death outcome, a full generalized linear model is built with *logit* link. Using stepwise selection, a final *logit* model is selected:

$$logit[\pi(x)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 * steroid(1) + \beta_2 * ascites(1) + \beta_3 * varices(1) + \beta_4 * sgot + \beta_5 * protime.$$

ML estimates for *logit* model coefficients together with model deviance and AIC score are summarized in **Appendix Table 4**. Fitting of the model is assessed by Goodness of fit test (LR test, G^2 = 44.762, df = 74, p = 0.9972). Test result supports good fitting of the model.

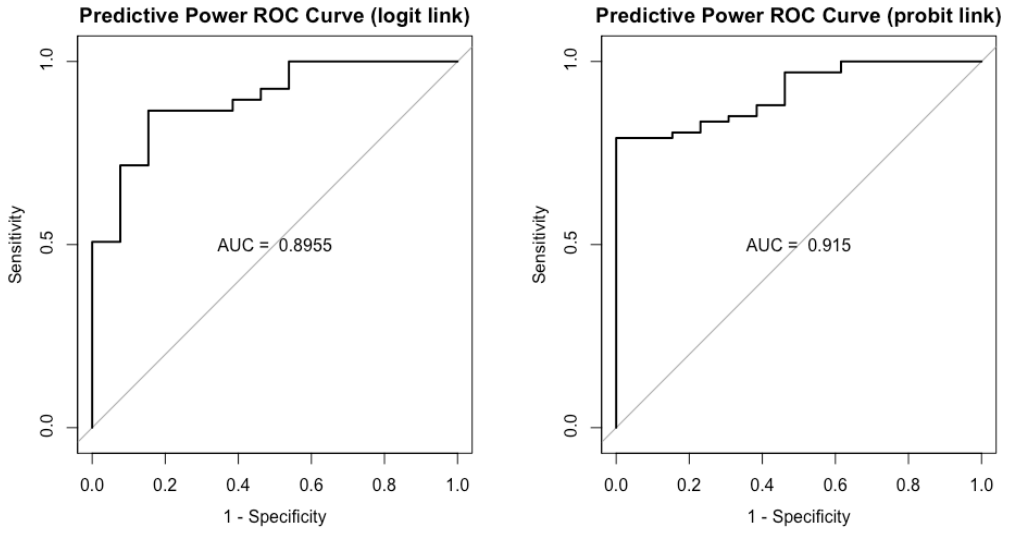
A *probit* model was built upon stepwise selection of the full model,

$$\pi(x) = \Phi(\beta_0 + \beta_1 * age + \beta_2 * steroid(1) + \beta_3 * ascites(1) + \beta_4 * varices(1) + \beta_5 * alk\_phosphate + \beta_6 * sgot + \beta_7 * protime)$$

ML estimates for *probit* model coefficients together with model deviance and AIC score are summarized in **Appendix Table 5**. The *probit* model is tested for goodness of fit using LR Chi-square test (G^2 = 40.223, df = 72, p = 0.999), which supports good fitting.

To check the predictive power of logistic models, receiver operating characteristic (ROC) curves are plotted for both *probit* model and *logit* model (See **Figure 1**). AUC score is 0.8955 for *logit* model,

Figure 1. ROC for *logit* Model and *probit* Model



0.915 for *probit* model. Collinearity was checked for two models by looking at pair-wise scatter plots between predictors (See **Appendix Figure 1** and **Figure 2**). No clear trend is detected for any predictors pair, no obvious multicollinearity existing within the two models.

Residuals are plotted to check any obvious trend/pattern in scattering of residuals and deviance for potential outlier(s) /influential point(s) within 2 models (See **Appendix Figure 3** and **Figure 4**). On *logit* model residual plots, one extreme residual appears at 135th patient's outcome record, whose standardizes residual score reaches -5. After removing this influential observation and refit the *logit* model, the predictive power increases with a higher AUC = 0.94. On *probit* model residual scatter plots, 3 individuals' absolute residual values are extreme (>3.5). These extreme observations are 99th, 135th and 95th patients, whose absolute standardizes residuals are >2. Removing these 3 observations and refit the *probit* model gives a greater predictive power of AUC = 0.954 (See **Figure 2**). In general, the residuals and deviances scatter into a wider range among latter patients (64th and after) and this scene presents when fitting both *logit* and *probit* model.

Test the significant effect of interaction term sgot*protime on logit odds. Both Wald test (p = 0.2692) and LR test (p = 0.2319) provide no strong evidence of interaction effect.

Numeric variables sgot and albumin are each categorized into 4 levels using the four quantiles as cutoff lines (see **Appendix Table 2** for detailed numeric variables' characteristic values). 2 multinomial logistic models are built for predicting baseline-category logits for the 2 ordinal responses; model *a* uses all categorical as predictors; model *b* uses the rest numeric variables as predictors. Model *a* and model *b* are compared using Likelihood-Ratio test (see **Appendix Table 6** and **Table 7** for LR test statistics). Based on LR test result (predict ordinal sgot, p = 0.3769; predict ordinal albumin, p = 0.2276), at α = 0.05, model *a* is always preferred over model *b*.

Ordinal baseline-category logit regression (reference level = 1),

$$logit\left(\frac{\pi_j}{\pi_1}\right) = \alpha_j + \beta_{1j} * bilirubin_j + \beta_{2j} * alk\_phosphate_{2j} + \beta_{3j} * albumin_j + \beta_{4j} * protime_j$$

Figure 2. Improved ROC for Refitted *logit* Model and *probit* Model (extreme observations removed)

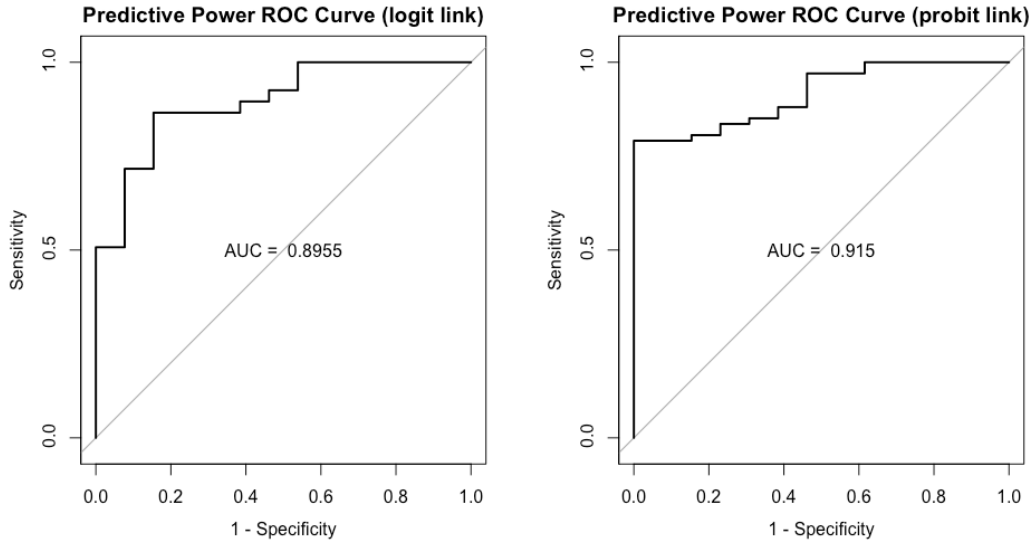


Table 2. Summary of Binary Logistic Model for Predicting Odds of Abnormal sgot				
Coefficients	Estimate	Std.Error	Z	Pr(>|z|)
Intercept	-2.51	2.71	-0.927	0.3537
bilirubin	0.39	0.50	0.787	0.4312
Alk_phosphate	0.017	0.0076	2.201	0.0277*
albumin	0.69	0.61	1.146	0.2519
protime	-0.020	0.013	-1.534	0.1250
Null deviance: 94.107, df = 79; Residual deviance: 83.766, df = 75; AIC = 93.766				
***significant predictor at α = 0.1				

Wald test of ML estimates for multinomial logistic model coefficients suggests 1 second longer of patient's blood clot formation will result in a significant increase in log odds of having level 2 (β = 0.056, p = 0.00426), level 3 (β = 0.062, p = 0.00312) and level 4 (β = 0.053, p = 0.0128) alk_phosphate.

Numeric variable sgot is further characterized for abnormality (normal = 0, abnormal = 1) based on the standard blood sgot range (see **variable information** section for details). 2 logistic models are built for predicting the logit odds of having abnormal sgot for hepatitis patients; model *c* uses all categorical predictors; model *d* uses the rest numeric variables as predictors. Model *c* and model *d* are compared using ANOVA test (see **Appendix Table 8** for ANOVA test statistics). Based on ANOVA test result (p = 0.9916), model *c* is preferred over model *d*.

Binary logistic regression,

$$logit[\pi(x)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 * bilirubin + \beta_2 * alk\_phosphate + \beta_3 * albumin + \beta_4 * protime$$

Wald test for ML estimates of binary logistic model coefficients together with model deviance and AIC score are summarized in **Table 2**.

Discussion

Study Description:

In this retrospective cohort study of hepatitis patients' survival outcomes, presence of certain symptoms was found to have significant association with log odds of death. Hepatitis patients who had fluid accumulated in abdomen will have exp(1.826) = 6.2 times hazards of death than who never had such symptom; hepatitis patients who had enlarged veins will have exp(1.774) = 5.9 times odds of death than who never had such symptom. Every one second longer for blood cells to form clot will result in exp(0.042) = 1.5 increased odds of death among hepatitis patients. A 1U/L increase in blood alkaline phosphate will result in exp(0.017) = 1.02 increased chance of abnormal blood sgot level.

No strong correlation was found between paired numeric variables (see **Appendix Table 9** for correlation matrix). Age and gender's association with survival chance was not detected among hepatitis patients.

Limitations:

1. This study only contained 32 female patients, none was dead during study period. Gender effect on survival chance for hepatitis patients can't be assessed. Should consider gender distribution/balance in future study for analyzing its effect on the outcome;

2. Many missing observations present in the dataset. After removing observations with missing records, the leftover sample size is small (n = 80). Small sample can greatly deteriorate the power of hypothesis testing. Missing values prevent hypothesis testing of symptoms lacking observations. Consider better patients follow-up and record tracking strategy in future cohort study to secure data validity;

3. No demographic factor was well studied for association to survival chance. Suggest adding additional patients' demographic information and expanding the analysis to demographic effects on survival chance.

Conclusions

Among hepatitis patients, those experienced abdominal fluid accumulations, had enlarged veins or had longer blood clots formation time tend to have substantially lower chance of survival or may develop chronic liver disease. Signs of abnormally high level of alkaline phosphate in blood or long blood clots formation time are indicative of damaged liver cells for hepatitis patients.

Dataset Information

Published: November 1st, 1988

Donor: G.Gong (Carnegie-Mellon University) via Bojan Cestnik | Jozef Stefan Institute, Jamova 39, 61000 Ljubljana; Yugoslavia (tel.: (38)(+61) 214-399 ext.287) }

Link: https://archive.ics.uci.edu/ml/datasets/Hepatitis

References

1. What is hepatitis? (2018, October 09). Retrieved from https://www.who.int/features/qa/76/en/
2. CDC - Hepatitis - Topics - Did You Know - STLT Gateway. (n.d.). Retrieved from https://www.cdc.gov/publichealthgateway/didyouknow/topic/hepatitis.html
3. Hepatitis B. (n.d.). Retrieved from https://www.who.int/news-room/fact-sheets/detail/hepatitis-b
4. Hepatitis B | HBV . (2019, February 07). Retrieved from https://medlineplus.gov/hepatitisb.html

5. Hepatitis: Types, Symptoms, and Treatment - Healthline. (n.d.). Retrieved from https://www.healthline.com/health/hepatitis
6. CCRN, R. N. (n.d.). How is hepatitis C transmitted? Retrieved from https://www.medicalnewstoday.com/articles/318888.php
7. Viral Hepatitis and Liver Disease. (2018, October 04). Retrieved from https://www.hepatitis.va.gov/HEPATITIS/hcv/patient/index.asp
8. Bilirubin blood test: MedlinePlus Medical Encyclopedia. (n.d.). Retrieved from https://medlineplus.gov/ency/article/003479.htm
9. ALP - blood test: MedlinePlus Medical Encyclopedia. (n.d.). Retrieved from https://medlineplus.gov/ency/article/003470.htm
10. Davis, C. P. (n.d.). Liver Blood Tests Abnormal Values (High, Low, Normal) Explained. Retrieved from https://www.medicinenet.com/liver_blood_tests/article.htm
11. Albumin (Blood). (n.d.). Retrieved from https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=albumin_blood
12. Prothrombin time (PT): MedlinePlus Medical Encyclopedia. (n.d.). Retrieved from https://medlineplus.gov/ency/article/003652.htm

Appendix

Table 1. Summary of Contingency Tables

Symptoms	Yes/No	Death Outcome	
		Dead	Live
steroid	Y	20	56
	N	12	66
antiviral[b]	Y	2	22
	N	30	101
fatigue[b]	Y	30	70
	N	2	52
malaise	Y	23	38
	N	9	84
anorexia	Y	10	22
	N	22	100
liver big[b]	Y	3	22
	N	24	96
liver firm	Y	13	47
	N	14	70
spleen palpable	Y	12	18
	N	19	101
spiders	Y	22	29
	N	9	90
ascites	Y	14	6
	N	17	113
varices	Y	11	7
	N	20	112

[a]All deaths were males, not meaningful for testing independence and its contingency table was not included for any further analysis.
[b]Small sampl size (containing cell counts < 5).

Table 2. Characteristics of Numeric Variables							
Attributes	Min	25th Qu.	Median	Mean	75th Qu.	Max	Missing
age	7	32	39	41.2	50	78	0
bilirubin(mg/dL)	0.3	0.7	1	1.428	1.5	8	6
alk_phosphate(U/L)	26	74.25	85	105.33	132.25	295	29
sgot(U/L)	14	31.5	58	85.89	100.5	648	4
albumin(g/dL)	2.1	3.4	4	3.817	4.2	6.4	16
protime(sec)	0	45	61	61.85	76.25	100	67

Table 3. Breslow-Day Test Statistics for Homogeneity			
Symptom1	**Symptom2**	**X-squared**	**p-value**
fatigue	malaise	NA*	NA*
fatigue	spleen_palpable	12.17	0.00049
fatigue	spiders	6.74	0.0094
fatigue	ascites	NA*	NA*
fatigue	varices	11.37	0.00074
malaise	spleen_palpable	20.93	4.76e-06
malaise	spiders	13.62	0.00022
malaise	ascites	9.45	0.0021
malaise	varices	13.77	0.00021
spleen_palpable	spiders	9.73	0.0018
spleen_palpable	ascites	10.67	0.0011
spleen_palpable	varices	9.99	0.0016
spiders	ascites	17.09	3.56e-05
spiders	varices	23.99	9.70e-07
ascites	varices	23.37	1.34e-06

*lack observations that fit into each category of 2 symptoms and death outcome

Table 4. Summary of Binary Logistic Model for Log Odds in Death (link = *LOGIT*)				
Coefficients	Estimate	Std.Error	Z	Pr(>|z|)
Intercept	-4.763098	1.673516	-2.846	0.00442
Steroid(1:0)	1.407171	0.916273	1.536	0.12460
ascites(1:0)	1.825922	0.861404	2.120	0.03403*
varices(1:0)	1.773667	1.038890	1.707	0.08777*
sgot(1:0)	0.010787	0.008121	1.328	0.18408
protime(1:0)	0.041984	0.019158	2.191	0.02842*
Null deviance: 71.07, df = 79; Residual deviance: 44.762, df = 74; AIC = 56.762				
***significant predictor p<0.1				

Table 5. Summary of Binary Logistic Model for Log Odds in Death (link = *PROBIT*)				
Coefficients	Estimate	Std.Error	Z	Pr(>|z|)
Intercept	-1.069781	1.310962	-0.816	0.4145
age	-0.029649	0.020477	-1.448	0.1476
steroid(1:0)	1.090320	0.585537	1.862	0.0626*
ascites(1:0)	1.208920	0.513199	2.356	0.0185*
varices(1:0)	1.012956	0.624374	1.622	0.1047
alk_phosphate	-0.006147	0.004229	-1.454	0.1460
Sgot	0.008871	0.005199	1.706	0.0880
protime	0.023000	0.010709	2.148	0.0317*
Null deviance: 71.007, df = 79; Residual deviance: 40.223, df = 72; AIC = 56.223				
*significant predictor p<0.1				

Table 6. ANOVA Test Statistics for Comparing 2 Multinomial Logistic Models for Predicting Ordinal sgot					
Model	#Df	LogLik	Df	Chisq	Pr(>Chi)sq
a	210	-80.628			
b	225	-88.667	15	16.078	0.3769

Table 7. ANOVA Test Statistics for Comparing 2 Multinomial Logistic Models for Predicting Ordinal albumin					
Model	#Df	LogLik	Df	Chisq	Pr(>Chi)sq
a	210	-80.628			
b	225	-89.978	15	18.7	0.2276

Table 8. ANOVA Test Statistics for Comparing 2 Multinomial Logistic Models for Predicting Binary sgot					
Model	Resid. Df	Resid.Dev	Df	Deviance	Pr(>Chi)
c	75	83.766			
d	70	83.251	5	0.5148	0.9916

Table 9. Correlation Matrix of Pair-wise Numeric Variables					
	bilirubin	**alk_phosphate**	**sgot**	**albumin**	**protime**
nilirubin	1	0.3164	0.2541	-0.3505	-0.3779
alk_phosphate	0.3164	1	0.2931	-0.4060	-0.2218
sgot	0.2541	0.2931	1	-0.2	-0.1867
albumin	-0.3505	-0.4060	-0.2	1	0.4521
protime	-0.3779	-0.2218	-0.1867	0.4521	1

Figure 1. Pair-wise Scatter Plot between Covariates of *logit* Model

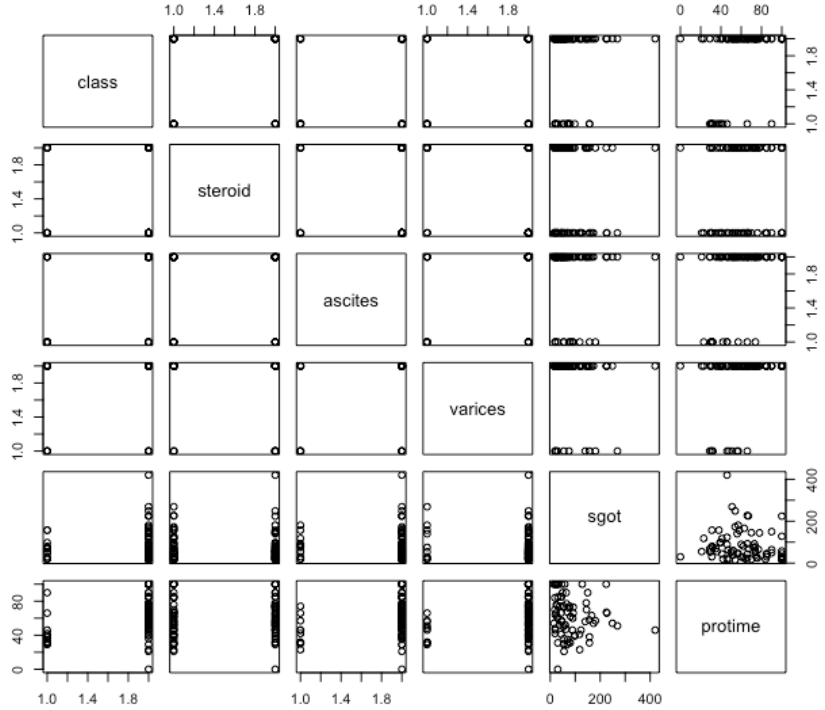


Figure 2. Pair-wise Scatter Plot between Covariates of *probit* Model

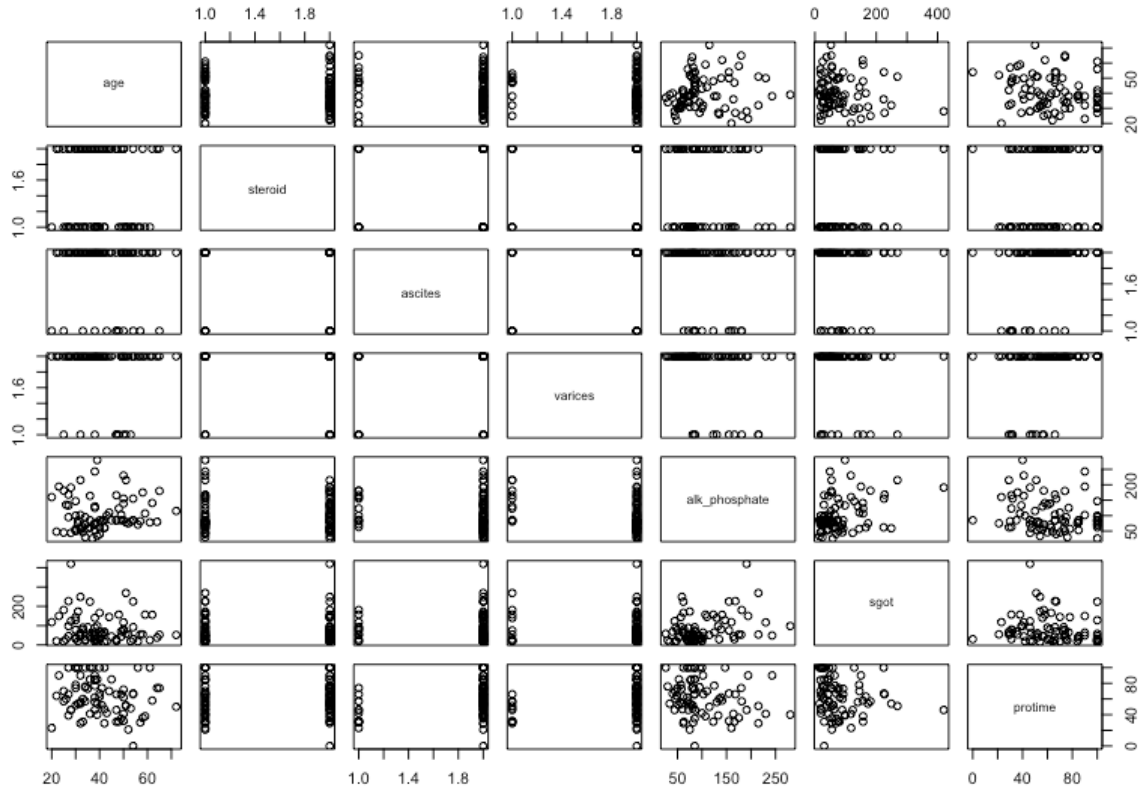


Figure 3. Scatter Plot of Residuals and Deviance for *logit* Model

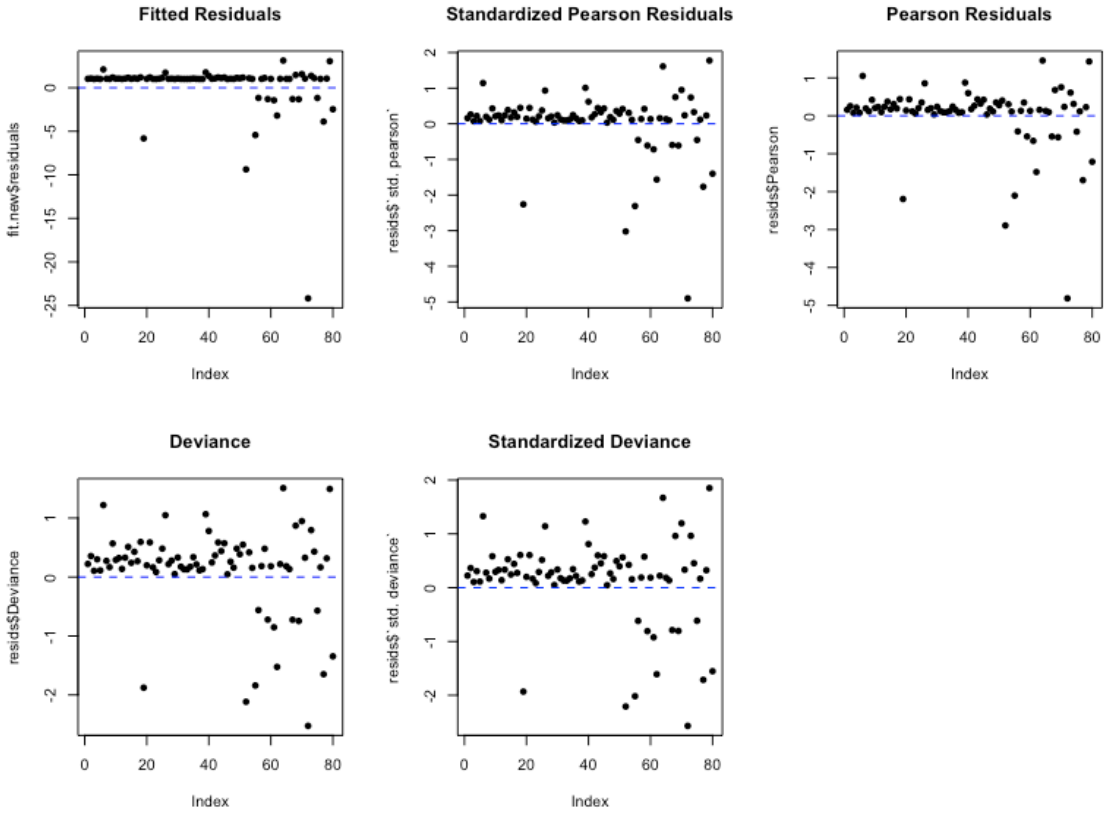
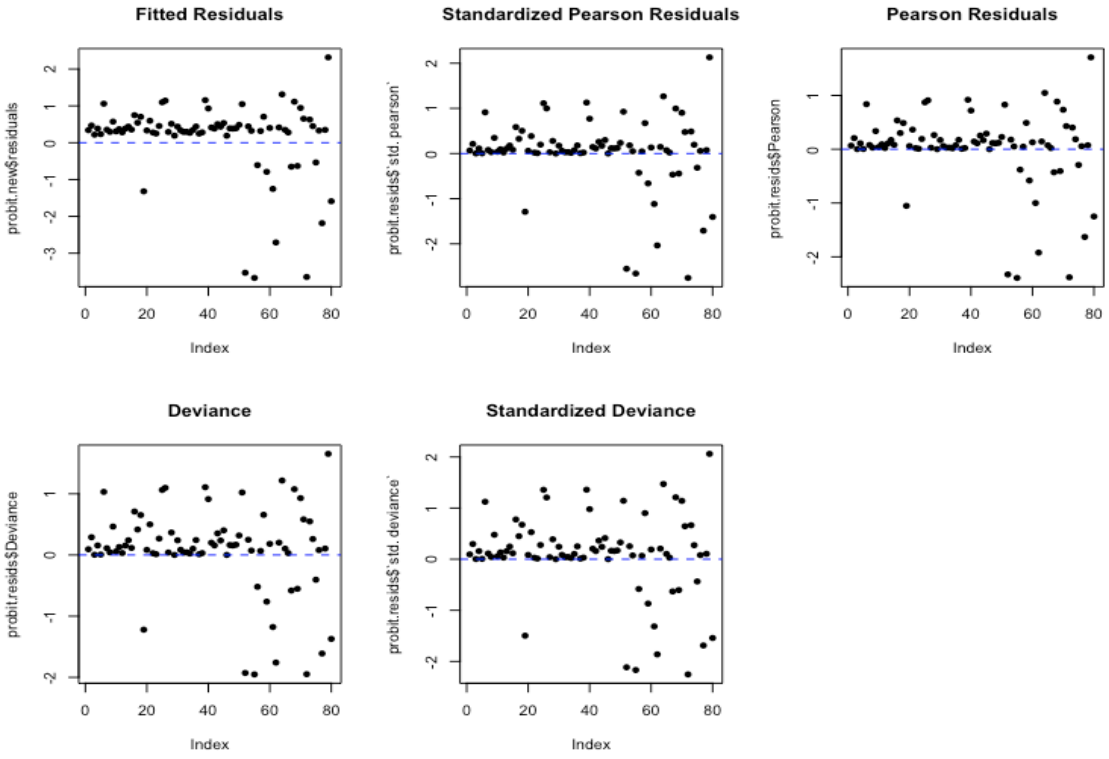


Figure 4. Scatter plot of Residuals and Deviance for *probit* model



R Code:

```
#read in dataset

data = read.table("hepatitis.data", sep = ",")[,-20]
colnames(data) = c("class", "age", "sex", "steroid", "antiviral", "fatigue",
"malaise", "anorexia","liver_big", "liver_firm", "spleen_palpable", "spiders"
, "ascites", "varices", "bilirubin", "alk_phosphate", "sgot", "albumin", "pro
time")
for (i in 1:length(colnames(data))){
  levels(data[,i])[levels(data[,i]) == "?"] <-NA
}


##data processing
data$bilirubin <- as.numeric(as.character(data$bilirubin))
data$protime <- as.numeric(as.character(data$protime))
data$alk_phosphate <- as.numeric(as.character(data$alk_phosphate))
data$sgot <- as.numeric(as.character(data$sgot))
data$albumin <- as.numeric(as.character(data$albumin))
data$age <- as.numeric(as.character(data$age))
data$sex <- as.factor(data$sex)
data$class <- as.factor(data$class)
data$antiviral <- as.factor(data$antiviral)


##check correlation
pairs(data[, c("class", "age", "sgot","albumin", "alk_phosphate", "protime",
"bilirubin", "varices")])
```

Contingency Table for OR, RR→ Test for Association

```
##create contingency tables
cat.attributes = c("sex", "steroid", "antiviral", "fatigue", "malaise", "anor
exia","liver_big", "liver_firm", "spleen_palpable", "spiders", "ascites", "va
rices")

t = list()
table = list()
for(i in 1:length(cat.attributes)){
  t[[i]] = data[complete.cases(data[ , c("class", cat.attributes[i])]),]
  table[[i]] = table(t[[i]][,cat.attributes[i]], t[[i]]$class)
  colnames(table[[i]]) = c("Death", "Live")
}
#label table row&col names
rownames(table[[1]]) = c("M", "F")
for(i in 2:length(cat.attributes)){
 rownames(table[[i]]) = c(paste(cat.attributes[i], "Yes"), "No")
}
```

Risk difference, Relative Risk and Odds ratio

Chi-square test and Exact Test for Independence

```
library('epiR')

#test for independence
OR = list()
result = list()
IR.CI = list()
OR.CI = list()
indep.test = list()
fisher = list()
for (i in 1:12){
      result[[i]] = epi.2by2(table[[i]], method = "cohort.count", conf.level
= 0.95)
  if (all(as.vector(table[[i]]) > 5)) {
      indep.test[[i]] = data.frame(test.stat = result[[i]]$massoc$chisq.strat
a$test.statistic, p.value = result[[i]]$massoc$chisq.strata$p.value, attri =
cat.attributes[i])
  }
  else{
    fisher[[i]] =fisher.test(table[[i]], alternative = 'two.sided')
    indep.test[[i]] = data.frame(est = fisher[[i]]$estimate, ci.lwr = fisher[
[i]]$conf.int[1],ci.upr = fisher[[i]]$conf.int[2], p.value = fisher[[i]]$p.va
lue, attri = cat.attributes[i])
    }
  OR[[i]] = result[[i]]$tab
  IR.CI[[i]] = result[[i]]$massoc$RR.strata.wald
  OR.CI[[i]] = data.frame(result[[i]]$massoc$OR.strata.wald, attri = cat.attr
ibutes[i])
}

#find insignificant independence
library(pipeR)
library(rlist)
indep.test %>>%
  list.filter(p.value > 0.1) %>>%
  list.mapv(attri)

## [1] antiviral  anorexia   liver_big  liver_firm
```

hypothesis testing–prop test

```
prop.test(sum(as.numeric(data$class)-1), length(data$class), p = 0.5, alterna
tive = "two.sided", conf.level = 0.95)

##
##  1-sample proportions test with continuity correction
##
## data:  sum(as.numeric(data$class) - 1) out of length(data$class), null pro
bability 0.5
```

```
## X-squared = 52.258, df = 1, p-value = 4.867e-13
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.7196073 0.8525854
## sample estimates:
##         p
## 0.7935484
```

3-way table; Breslow-Day test Homogeneous Associations; Mantel - Haenszel Test

```
options(scipen = 999)
library(DescTools)

sig.attri = c("fatigue", "malaise",  "spleen_palpable", "spiders", "ascites",
 "varices")
comb = combn(sig.attri,2)

t2 = list();table2 = list();dp = list();df = list();ftable = list();mantel =
list()
mantel.test = list();bd = list();bd.test = list()
for(i in 1:ncol(comb)){
  t2[[i]] = data[complete.cases(data[ , c("class", comb[1,i], comb[2,i])]),]
  table2[[i]] = table(t2[[i]]$class,t2[[i]][,comb[1,i]], t2[[i]][,comb[2,i]])
  #colnames(table2[[i]]) = c("Death", "Live")
  dp[[i]] = as.vector(table2[[i]])
  df[[i]] <- array(dp[[i]], dim = c(2,2,2))
  dimnames(df[[i]]) <- list(Class = c("DIE", "ALIVE"), attri1 = c("YES", "NO"
), attri2 = c("YES", "NO"))
  ftable[[i]] = ftable(df[[i]], row.vars = c("attri1", "attri2"), col.vars =
"Class")
  mantel[[i]] = mantelhaen.test(df[[i]], correct = F)
  mantel.test[[i]] = c(mantel[[i]]$statistic, p.value = mantel[[i]]$p.value,
conf.int.lwr = mantel[[i]]$conf.int[1], conf.int.upr = mantel[[i]]$conf.int[2
], est = mantel[[i]]$estimate, attri1 = comb[1,i], attri2 = comb[2,i])
  bd[[i]] = BreslowDayTest(x = table2[[i]], OR = 1)
  bd.test[[i]] = c(test.stat = bd[[i]]$statistic, p.value = bd[[i]]$p.value,
attri1 = comb[1,i], attri2= comb[2,i])
}

#check homogeneous OR among stratum
bd.test%>>%
  list.filter(p.value > 0.1) %>>%
  list.mapv(c(attri1, attri2)) %>>% matrix(ncol = 2, byrow = T)

#check whether all ORs = 1 among stratum

mantel.test%>>%
  list.filter(p.value > 0.1) %>>%
  list.mapv(c(attri1, attri2)) %>>% matrix(ncol = 2, byrow = T)
```

Residual Analysis

```
library(vcd)

death.margin = list(); odds = list(); chi = list();row.resid = list(); pearso
n.resid = list(); std.resid = list(); resid.result = list()

for(i in 1:12){
# odds ratio of margin table and chisquare test
odds[[i]] = exp(oddsratio(table[[i]])$coefficients)
chi[[i]] = chisq.test(table[[i]])
#row resid
row.resid[[i]] = chi[[i]]$observed - chi[[i]]$expected
#pearson resid
pearson.resid[[i]] = chi[[i]]$residuals
#standardized resid
std.resid[[i]] = chi[[i]]$stdres
##colnames(std.resid[[i]]) = c(paste(comb[1,i], "Yes"), paste(comb[1,i], "No"
))
##rownames(std.resid[[i]]) = c("Death_Yes", "Death_No")
}
```

Logistic Regression(logit & probit link)–stepwise model selection

```
library(dplyr)

library(MASS)

#remove missing observations
t.complete = data[complete.cases(data),]

num.complete = data[complete.cases(data$bilirubin, data$alk_phosphate, data$s
got, data$albumin, data$protime),]

cor(num.complete[,c("bilirubin", "alk_phosphate", "sgot", "albumin", "protime
")])

#check class distribution
shapiro.test(as.numeric(t.complete$class))

#logit link logistic regression
fit = glm(class ~ age + steroid + fatigue + malaise + spleen_palpable + spide
rs + ascites+ varices + bilirubin + alk_phosphate + sgot + albumin + protime,
family = binomial(link = logit), data = t.complete)


stepAIC(fit, direction = "both")

                #fit.new = glm(class ~ steroid + spiders+varices + sgot + alb
umin,family = binomial(link = logit), data = t.complete)
fit.new = glm(class ~ steroid + ascites + varices + sgot + protime,family = b
inomial(link = logit), data = t.complete)
```

```
##ANOVA→model comparison:

anova(fit.new, fit, test = "LRT")

                #fit.exp <- glm(formula = class ~ steroid + spiders + varices
 * albumin + sgot, family = binomial(link = logit), data = t.complete)

##check multicollinearity:
pairs(t.complete[,c("class","steroid" ,"ascites","varices" , "sgot", "protime
")])
```

```
#add interaction term to expand logit model
fit.exp <- glm(formula = class ~ steroid + ascites + varices + sgot * protime
, family = binomial(link = logit), data = t.complete)


#ANOVA test whether expanded model is better (significance of interaction ter
m)

anova(fit.new, fit.exp, test = "LRT")

#probit link logistic regression
probit.fit = glm(class ~ age + steroid + fatigue + malaise + spleen_palpable
+ spiders + ascites+ varices + bilirubin + alk_phosphate + sgot + albumin + p
rotime, family = binomial(link = probit), data = t.complete)


probit.new = glm(formula = class ~ age + steroid + ascites + varices + alk_ph
osphate + sgot + protime, family = binomial(link = probit), data = t.complete
)

anova(probit.new, probit.fit, test = "LRT")

#p(Y = death|X )= pnorm(predict(probit.fit))
#p(Y = live|X) = 1-p(Y = death|X)
#OR = p(Y=death|X) / (1-p(Y = death|X))


##check multicollinearity:
pairs(t.complete[,c("age" ,"steroid" , "ascites" , "varices" , "alk_phosphate
" , "sgot" , "protime")])
```

Predictive Power

```
# ROC
library(pROC)

par(mfrow = c(1,2))
rocplot.fit.new <- roc(class ~ fitted(fit.new), data = t.complete)
plot.roc(rocplot.fit.new, legacy.axes = T, main = "Predictive Power ROC Curve
 (logit link)")
txt.fit.new <- paste("AUC = ", round(auc(rocplot.fit.new),4))
```

```
text(0.5, 0.5, txt.fit.new)
#auc(rocplot)
probit.rocplot.new <- roc(class ~ fitted(probit.new), data = t.complete)
plot.roc(probit.rocplot.new, legacy.axes = T, main = "Predictive Power ROC Cu
rve (probit link)")
probit.txt.new <- paste("AUC = ", round(auc(probit.rocplot.new),4))
text(0.5, 0.5, probit.txt.new)
```

Model Checking

```
##logit model

##residual
fit.new$df.residual

##deviance
fit.new$deviance

##goodness of fit
#p-value for deviance goodness-of-fit test
1-pchisq(fit.new$deviance, fit.new$df.residual)

qchisq(fit.new$deviance, fit.new$df.residual)

logit.resids <- cbind(rstandard(fit.new,type="pearson"), residuals(fit.new,ty
pe="pearson"), residuals(fit.new,type="deviance"), rstandard(fit.new,type="de
viance"))
resids <- as.data.frame(logit.resids)
colnames(resids) = c("std. pearson", "Pearson", "Deviance",  "std. deviance")

par(mfrow = c(2,3))
plot(fit.new$residuals, pch = 16, main = "Fitted Residuals")
abline(h = 0, lty = 2, col = "blue")
plot(resids$`std. pearson` , pch = 16, main = "Standardized Pearson Residuals
")
abline(h = 0, lty = 2, col = "blue")
plot(resids$Pearson, pch = 16, main = "Pearson Residuals")
abline(h = 0, lty = 2, col = "blue")
plot(resids$Deviance, pch = 16, main = "Deviance")
abline(h = 0, lty = 2, col = "blue")
plot(resids$`std. deviance`, pch = 16, main = "Standardized Deviance")
abline(h = 0, lty = 2, col = "blue")

#remove outlier and refit model--greater predictive power
##logit model
which.max(abs(fit.new$residuals))

## 135
##  72

t.new = t.complete[-72,]

fit1 = glm(class ~ steroid + ascites + varices + sgot + protime,family = bino
```

```r
mial(link = logit), data = t.new)
rocplot.fit.new1 <- roc(class ~ fitted(fit1), data = t.new)
plot.roc(rocplot.fit.new1, legacy.axes = T, main = "Predictive Power ROC Curv
e (logit link)")
txt.fit.new1 <- paste("AUC = ", round(auc(rocplot.fit.new1),4))
text(0.5, 0.5, txt.fit.new1)

##probit model

probit.new$df.residual

## [1] 72

##deviance
probit.new$deviance

## [1] 40.22272

##goodness of fit
#p-value for deviance goodness-of-fit test
1-pchisq(probit.new$deviance, probit.new$df.residual)

## [1] 0.9991168

probit.resids <- cbind(rstandard(probit.new,type="pearson"), residuals(probit
.new,type="pearson"), residuals(probit.new,type="deviance"), rstandard(probit
.new,type="deviance"))
colnames(probit.resids) = c("std. pearson", "Pearson", "Deviance",  "std. dev
iance")
probit.resids <- as.data.frame(probit.resids)

par(mfrow = c(2,3))
plot(probit.new$residuals, pch = 16, main = "Fitted Residuals")
abline(h = 0, lty = 2, col = "blue")
plot(probit.resids$`std. pearson` , pch = 16, main = "Standardized Pearson Re
siduals")
abline(h = 0, lty = 2, col = "blue")
plot(probit.resids$Pearson, pch = 16, main = "Pearson Residuals")
abline(h = 0, lty = 2, col = "blue")
plot(probit.resids$Deviance, pch = 16, main = "Deviance")
abline(h = 0, lty = 2, col = "blue")
plot(probit.resids$`std. deviance`, pch = 16, main = "Standardized Deviance")
abline(h = 0, lty = 2, col = "blue")

#remove outlier and refit model--greater predictive power
##probit model
which.max(abs(probit.new$residuals))

## 99
## 55

order(abs(probit.new$residuals))
```

```
##  [1] 46 29  3  5 37 23 34 22 66 38 12 27 32  8 33 10 57 54 76 20 31  1 35
## [24] 65 78  7 15 11 13  4 48 42 47 49 60 41 63 44 30 14 36 53 74 24  2 50
## [47] 43 28 45 75 17  9 21 56 73 69 67 71 18 58 16 59 40 70 51  6 25 68 26
## [70] 39 61 64 19 80 77 79 62 52 72 55
```

```
probit.new$residuals[c(52,72,55)]
```

```
##        95       135        99
## -3.530288 -3.645103 -3.670577
```

```
t.new1 = t.complete[-c(52,72,55),]

prob1 = glm(formula = class ~ age + steroid + ascites + varices + alk_phospha
te + sgot + protime, family = binomial(link = probit), data = t.new1)

probit.rocplot.new1 <- roc(class ~ fitted(prob1), data = t.new1)
plot.roc(probit.rocplot.new1, legacy.axes = T, main = "Predictive Power ROC C
urve (probit link)")
probit.txt.new1 <- paste("AUC = ", round(auc(probit.rocplot.new1),4))
text(0.5, 0.5, probit.txt.new1)

par(mfrow = c(1,2))
plot.roc(rocplot.fit.new1, legacy.axes = T, main = "ROC for Refitted Model(lo
git link)")
txt.fit.new1 <- paste("AUC = ", round(auc(rocplot.fit.new1),4))
text(0.5, 0.5, txt.fit.new1)
probit.rocplot.new1 <- roc(class ~ fitted(prob1), data = t.new1)
plot.roc(probit.rocplot.new1, legacy.axes = T, main = "ROC for Refitted Model
(probit link)")
probit.txt.new1 <- paste("AUC = ", round(auc(probit.rocplot.new1),4))
text(0.5, 0.5, probit.txt.new1)
```

Multinomial Analysis

```
#create multiple levels
summary(t.complete$sgot)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   30.75   56.50   82.03  102.75  420.00
```

```
summary(t.complete$bilirubin)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.300   0.700   1.000   1.221   1.300   4.800
```

```
summary(t.complete$alk_phosphate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   68.25   85.00  102.91  133.50  280.00
```

```
summary(t.complete$albumin)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.100   3.500   4.000   3.844   4.200   5.000
```

```r
##categorize numeric variables into ordinal values

for(i in 1:length(t.complete$alk_phosphate)){

  ##categorical sgot
  if (t.complete$sgot[i] <=28){t.complete$cat.sgot[i] = 1}
  else if (t.complete$sgot[i]<=46){t.complete$cat.sgot[i] = 2}
  else if (t.complete$sgot[i]<=65.25){t.complete$cat.sgot[i] = 3}
  else if (t.complete$sgot[i] <= 84){t.complete$cat.sgot[i] = 4}
  ##normal sgot
  if (t.complete$sgot[i] <=40){t.complete$norm.sgot[i] = 0}
  else {t.complete$norm.sgot[i] = 1}

  ##categorical albumin
  if (t.complete$albumin[i] <=3.4){t.complete$cat.albumin[i] = 1}
  else if (t.complete$albumin[i]<=4){t.complete$cat.albumin[i] = 2}
  else if (t.complete$albumin[i]<=4.2){t.complete$cat.albumin[i] = 3}
  else if (t.complete$albumin[i] <= 6.4){t.complete$cat.albumin[i] = 4}
  ##binary albumin
  if (3.4 <= t.complete$albumin[i] | t.complete$albumin[i] <=5.4){t.complete$
norm.albumin[i] = 0}
  else {t.complete$norm.albumin[i] = 1}

}

#check response distribution
shapiro.test(data$sgot) #non-normal->use glm
##  Shapiro-Wilk normality test
##
## data:  data$sgot
## W = 0.67861, p-value < 2.2e-16

shapiro.test(data$albumin) #non-normal->use glm
##  Shapiro-Wilk normality test
##
## data:  data$albumin
## W = 0.94791, p-value = 4.4e-05

#check correlation
pairs(t.complete[,c("sgot","steroid","spleen_palpable" ,"antiviral","fatigue"
,"malaise",  "albumin")])

pairs(t.complete[,c("sgot","anorexia","spiders","ascites" ,"varices","albumin
")])

pairs(t.complete[,c("bilirubin" ,"albumin" , "protime" , "alk_phosphate" , "s
got")])

library(VGAM)

##build multilevel logistic models for predicting sgot
sgot.fit <- vglm(factor(cat.sgot) ~ bilirubin + alk_phosphate + albumin + pro
```

```
time, family=multinomial(refLevel = "1"), data = t.complete)


sgot.fit.cat <- vglm(factor(cat.sgot) ~ steroid + spleen_palpable +antiviral+
fatigue+malaise+anorexia+spiders +ascites + varices, family = multinomial(ref
Level = "1"), data = t.complete)


lrtest(sgot.fit.cat, sgot.fit)

##binary logistic model for predicting abnormality of sgot
bin.sgot.fit <- glm(factor(norm.sgot) ~ bilirubin + alk_phosphate + albumin +
 protime, family=binomial(logit), data = t.complete)


bin.sgot.fit.cat <- glm(factor(norm.sgot) ~ steroid + spleen_palpable +antivi
ral+fatigue+malaise+anorexia+spiders +ascites + varices, family = binomial(lo
git), data = t.complete)
summary(bin.sgot.fit.cat)

anova(bin.sgot.fit, bin.sgot.fit.cat, test = "LRT")

##multilevel logistic model for predicting albumin
albumin.fit <- vglm(factor(cat.albumin) ~ bilirubin + sgot + protime, family=
multinomial(refLevel = "1"), data = t.complete)

albumin.fit.cat <- vglm(factor(cat.sgot) ~ steroid + spleen_palpable +antivir
al+fatigue+malaise+anorexia+spiders +ascites + varices, family=multinomial(re
fLevel = "1"), data = t.complete)


lrtest(albumin.fit.cat,albumin.fit)
```