Analysis on PBC

Final Report

STAT 271 FALL 2015

December 7, 2015

Gurinder Singh 300133063
Prash Medirattaa 300137275
Kuldeep Kaur 300129423

Analysis on PBC

Project Goal Analyze primary biliary cirrhosis (PBC) data using various methods.

Data description

Primary biliary cirrhosis (PBC) is a rare but fatal chronic liver disease of unknown cause with a prevalence of about 50-cases-per-million population. It generally strikes women between the ages of 40 and 60, but it has been diagnosed outside of this age range as well as in men. There is currently no known cure for PBC, but liver transplantation is now a common treatment. The clinical trial in Primary Biliary Cirrhosis (PBC) of the liver was conducted between 1974 and 1984 by Mayo Clinic. For that analysis, disease and survival status as of July, 1986, readings were recorded for as many patients as possible.
This is a double blinded experiment of a total of 424 PBC patients.
A randomized placebo controlled trial of the drug D-penicillamine is given to the patients.

Sources of Data

University of Massachusetts Amherst

https://www.umass.edu

Specific: https://www.umass.edu/statdata/statdata/data/pbc.txt

NAME: PBC Data (PBC.DAT)
SIZE: 418 observations, 20 variables

SOURCE: Counting Processes and Survival Analysis by T. Fleming &
D. Harrington, (1991), published by John Wiley & Sons.

Variable Description

Case number
Survival Time -The number of days between registration and the earlier of death, liver transplantation, or study analysis time in July, 1986.
Censoring - 1 if X is time to death, 0 if time to censoring
Treatment - Treatment Code, 1 = D-penicillamine, 2 = placebo.
Age - Age in years. For the first 312 cases, age was calculated by dividing the number of days between birth and study registration by 365.
Gender - 0 = male, 1 = female.
Presence of ascites - 0 = no, 1 = yes.
Presence of hepatomegaly - 0 = no, 1 = yes.
Presence of spiders - 0 = no, 1 = Yes.
Presence of edema - 0 = no edema and no diuretic therapy for edema; 0.5 = edema present for which no diuretic therapy was given, or edema resolved with diuretic therapy; 1 = edema despite diuretic therapy
Serum bilirubin - in mg/dl.
Serum cholesterol - in mg/dl.

Albumin - in gm/dl.
Urine copper - in mg/day.
Alkaline phosphatase - in U/liter.
SGOT - in U/ml.
Triglycerides - in mg/dl.
Platelet count - coded value is number of platelets per-cubic-milliliter of blood divided by 1000.
Prothrombin time - in seconds.
Histologic stage of disease - graded 1, 2, 3, or 4.

Quantitative (10)	Qualitative (7)
Age	**Two levels Categorical Variables**
Bilirubin	
Cholesterol	Treatment
Albumin	Gender
Urine copper	Ascites
Alkaline phosphatase	Hepatomegaly
SGOT	Spiders
Triglycerides	**More than two level Categorical Variables**
Platelet count	Edema – 3 levels
Prothrombin time	Histologic stage -4levels

Methodologies used	**Response Variable**
Linear Regression	Survival Time
One way ANOVA	Survival Time
Logistic Regression	Censoring
Survival Analysis	Survival Time

Multicollinearity

Aim: To check collinearity between explanatory variables.

Max |r|: Urine Copper & Bilirubin = 0.45692

Min |r|: Prothrombin Time & Platelet Count = 0.00007

We did not consider PCA since significant correlation didn't exist among our explanatory variables.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Age	Bilirubin	Cholesterol	Albumin	Urine copper	Alkaline phosphatase	SGOT	Triglycerides	Platelet count	Prothrombin time
Age	1.00000	0.00236	-0.15762	-0.18235	0.06155	-0.04725	-0.14987	0.02207	-0.14820	0.11376
		0.9616	0.0078	0.0002	0.2800	0.4056	0.0080	0.7122	0.0027	0.0203
	418	418	284	418	310	312	312	282	407	416
Bilirubin	0.00236	1.00000	0.39713	-0.31418	0.45692	0.11698	0.44173	0.43675	-0.01344	0.31489
	0.9616		<.0001	<.0001	<.0001	0.0389	<.0001	<.0001	0.7870	<.0001
	418	418	284	418	310	312	312	282	407	416
Cholesterol	-0.15762	0.39713	1.00000	-0.06973	0.12612	0.14947	0.35325	0.27683	0.19171	-0.03081
	0.0078	<.0001		0.2414	0.0343	0.0117	<.0001	<.0001	0.0013	0.6051
	284	284	284	284	282	284	284	282	280	284
Albumin	-0.18235	-0.31418	-0.06973	1.00000	-0.26477	-0.10146	-0.22005	-0.10342	0.15866	-0.20059
	0.0002	<.0001	0.2414		<.0001	0.0735	<.0001	0.0830	0.0013	<.0001
	418	418	284	418	310	312	312	282	407	416
Urine copper	0.06155	0.45692	0.12612	-0.26477	1.00000	0.18736	0.29383	0.27985	-0.06440	0.21822
	0.2800	<.0001	0.0343	<.0001		0.0009	<.0001	<.0001	0.2614	0.0001
	310	310	282	310	310	310	310	280	306	310
Alkaline phosphatase	-0.04725	0.11698	0.14947	-0.10146	0.18736	1.00000	0.11222	0.18008	0.14373	0.08938
	0.4056	0.0389	0.0117	0.0735	0.0009		0.0477	0.0024	0.0116	0.1151
	312	312	284	312	310	312	312	282	308	312
SGOT	-0.14987	0.44173	0.35325	-0.22005	0.29383	0.11222	1.00000	0.12612	-0.12015	0.11217
	0.0080	<.0001	<.0001	<.0001	<.0001	0.0477		0.0343	0.0351	0.0477
	312	312	284	312	310	312	312	282	308	312
Triglycerides	0.02207	0.43675	0.27683	-0.10342	0.27985	0.18008	0.12612	1.00000	0.10321	0.02012
	0.7122	<.0001	<.0001	0.0830	<.0001	0.0024	0.0343		0.0858	0.7365
	282	282	282	282	280	282	282	282	278	282
Platelet count	-0.14820	-0.01344	0.19171	0.15866	-0.06440	0.14373	-0.12015	0.10321	1.00000	-0.16733
	0.0027	0.7870	0.0013	0.0013	0.2614	0.0116	0.0351	0.0858		0.0007
	407	407	280	407	306	308	308	278	407	405
Prothrombin time	0.11376	0.31489	-0.03081	-0.20059	0.21822	0.08938	0.11217	0.02012	-0.16733	1.00000
	0.0203	<.0001	0.6051	<.0001	0.0001	0.1151	0.0477	0.7365	0.0007	
	416	416	284	416	310	312	312	282	405	416

Decided to keep all explanatory variables for our analysis as there is no correlation between the explanatory variables.

Linear Regression

Aim : To develop a model to describe the relationship between multiple explanatory variables and survival time while keeping treatment in the model even if it is not significant.

Technique used : Manual backward elimination
Indicator variables: Edema (0 = reference)
Histologic Stage (1 = reference)

Model 1: $\hat{y} = -180.34407 - 35.12471x_1 - 66.88416x_2 + 714.95866x_3 - 2.39953x_4 + 0.12762x_5 - 314.99749x_6$

Model 2: $\log(\hat{y}) = 6.36654 + 0.01501x_1 - 0.04634x_2 + 0.38197x_3 - 0.00163x_4 + 0.00006363x_5 - 0.23552x_6 - 0.90628x_7 - 0.26447x_8$

Model 3: $\log(\hat{y}) = 5.84251 + 0.01169x_1 - 0.04837x_2 + 0.35464x_3 - 0.23669x_6 - 0.91680x_7 - 0.26571x_8 - 0.20248x_9 + 0.20066x_{10}$

Model 4: $\log(\hat{y}) = 7.27041 + 0.01319x_1 - 0.04926x_2 + 0.35503x_3 - 0.23464x_6 - 0.91599x_7 - 0.26727x_8 - 0.53336x_9 + 0.04627x_{11}$

Model	MSE	Adjusted R^2	Constancy of variance	Normality of residuals
1	794202	0.3723	No	Yes
2	0.33909	0.5005	No	No
3	0.33390	0.5082	No	No
4	0.33222	0.5106	No	No

Where y = Survival Time
x_1 =Treatment (1 = D-penicillamine, 1 = Placebo)
x_2 = Bilirubin
x_3 = Albumin
x_4 = Urine Copper
x_5= Alkaline Phosphatase
x_6= Histologic Stage 4
x_7= Edema1
x_8= Edema.5
x_9= log(Urine Copper)
x_{10}= log(Alkaline Phosphatase)
x_{11}= log(Alkaline Phosphatase) * log(Urine Copper)

Global Hypothesis test for Model 4:

H_o: $\beta_1 = \beta_2 = \beta_3 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{11} = 0$
H_a: Regression model is overall significant

The p-value (<0.001) is small, therefore the regression model is overall significant. All the explanatory variables have a significant effect on the survival time, except for treatment which has a p-value = 0.8417.

According to MSE and Adjusted R^2 we see that Model 4 is the best fitted model, but the assumptions are still not met. **Therefore, we conclude that none of the above linear regression models are good in predicting survival time.**

ANOVA

Aim: To introduce a one way ANOVA model to see whether the treatment as a single factor is effective in changing the survival time.

Model: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$
Where y_{ij} = Survival Time
α_i = effect if i-th level of Treatment m=1, 2
ε_{ij} iidN$(0,\sigma^2)$

Hypothesis testing:

H_o: $\alpha_i = 0$ for all i
H_a: at least one $\alpha_i \neq 0$

P-value (0.8830) is large, and supports Ho. Therefore, treatment is not a significant figure to change the survival time.

Assumptions:

1) Constancy of variance
   Leven's Test reports a large p-value (0.4135) which supports Ho. Therefore, there is no violation in the constancy of variance.

2) Normality test for residuals
   By looking at the Q-Q plot, we conclude that the residuals follow a normal distribution with a few outliers.

   Goodness of for tests:
   H_o: Residuals are normally distributed
   H_a: Residuals are not normally distributed

The Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling test report p-values <.01, <0.005, <0.005 respectively. The small p-values support H_a, so we conclude that the residuals are not normally distributed.

Logistic Regression

Aim: To model the probability of death and correlate risk of death with other explanatory variables.

Technique used : Manual backward elimination

Indicator variables: Edema (0 = reference)
Histologic Stage (1 = reference)

Model: $\text{logit}(\pi) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$

$\text{logit}(\pi) = \log(\pi/(1-\pi))$

Where y = Censoring
x_1 = Age
x_2 = Treatment1
x_3 = Ascites
x_4 = Bilirubin
x_5 = SGOT
x_6 = Alkaline Phosphatase
x_7 = Urine Copper
x_8= Prothrombin Time

Fitted Model: $\text{logit}(\pi) = -13.5214 + 0.0585x_1 + 0.1614x_2 + 2.3733x_3 + 0.2077x_4 + 0.00739x_5 - 0.000212x_6 + 0.00479x_7 + 0.7065x_8$

Goodness-of-fit testing

H_o: The logistic regression model provides an adequate fit to the data

H_a: The logistic regression model provides an adequate fit to the data

The Deviance and Pearson Goodness of fit test report large p-values (0.9146 and <0.0001), which support H_o. Therefore, the logistic regression model provides an adequate fit to the data. The Hosmer Lemeshow test also report a large p-value (0.0092), also supporting H_o.

Interpretations
When age is increased by 1 year, the odds of death is multiplied by 1.060 while all the other explanatory variables are fixed.
We are 95% confident that with 1 year increase in age, the odds if death will be multiplied by between 1.028 and 1.094.

Deviance Test

Restricted Model: $\text{logit}(\pi) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_8 x_8$

Ho: Restricted Model

Ha: Full Model

G = 282.004 –268.015= 13.989

df = 2

Model	df	-2logL	AIC
Full	8	268.015	286.015
Restricted	6	282.004	296.004

P(Chi-square(df=2) > 13.989) = 0.00092

The p-value (0.00092) is small which supports H_a, so the Full model is better than the restricted model.

Survival Time

Aim: To model the survival time and identify the factors that are significant to hazard of death.

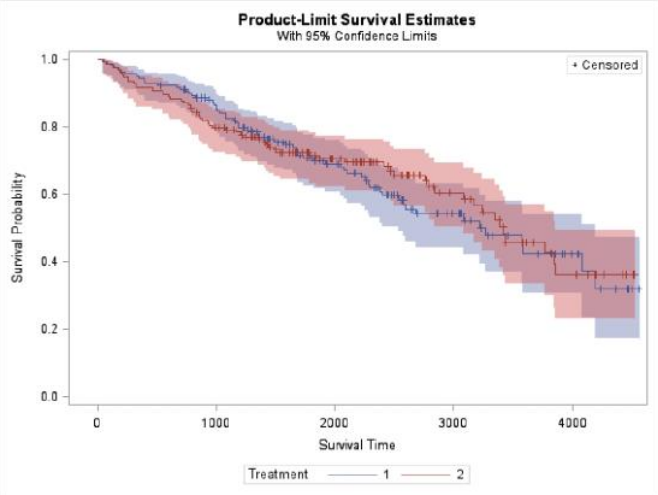
Technique used : Manual backward elimination

Indicator variables: Edema (0 = reference)
Histologic Stage (1 = reference)

Comparing survival functions:
$S_1(t)$ = survival function for D-penicillamine
$S_2(t)$ = survival function for placebo

H_o: $S_1(t) = S_2(t)$
H_a: $S_1(t) \neq S_2(t)$

Log-Rank and Wilcoxon tests, both have large p-values (.7498, .9664) so we cannot reject H_o. We conclude that there is no significant difference between the survival functions for patients given D-penicillamine and the survival function for patients given placebo.



There is a lot of overlap of the survival functions. D-penicillamine has a higher survival function until about 1700 days. After 1700 days the survival function for Placebo is higher until time 3200.

Propotional hazard model:

Global hypothesis test

H_o : The overall fitted model is not significant
H_a : The overall fitted model is significant

The small p-value(<.001) for Likelihood Ratio, Score and Wald's test support H_a. Therefore, it can be concluded that the overall fitted model is significant.
All the explanatory variables have significant effect on survival time, except treatment which has a p-value = 0.6285.

Hazard ratio

When the age is increased by 1 year, the hazard function of survival time is multiplied by 1.033 while all the other explanatory variables are fixed.

We are 95% confident, when the age is increased by 1 year the hazard function of survival time is multiplied by between 1.014 and 1.052.

Conclusion:

All four methods concluded that treatment does not have any effect on PBC patients.

Future: Study can be used for predicting survival time for PBC patients using **other** explanatory variables.