**Stat 271 fall 2015**

December 7, 2015

Gurinder Singh 300133063

Prash Medirattaa 300137275

Kuldeep Kaur 300129423

|  |
| --- |
|  |
| Analysis on PBC |
|  |
| Final Report |

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.

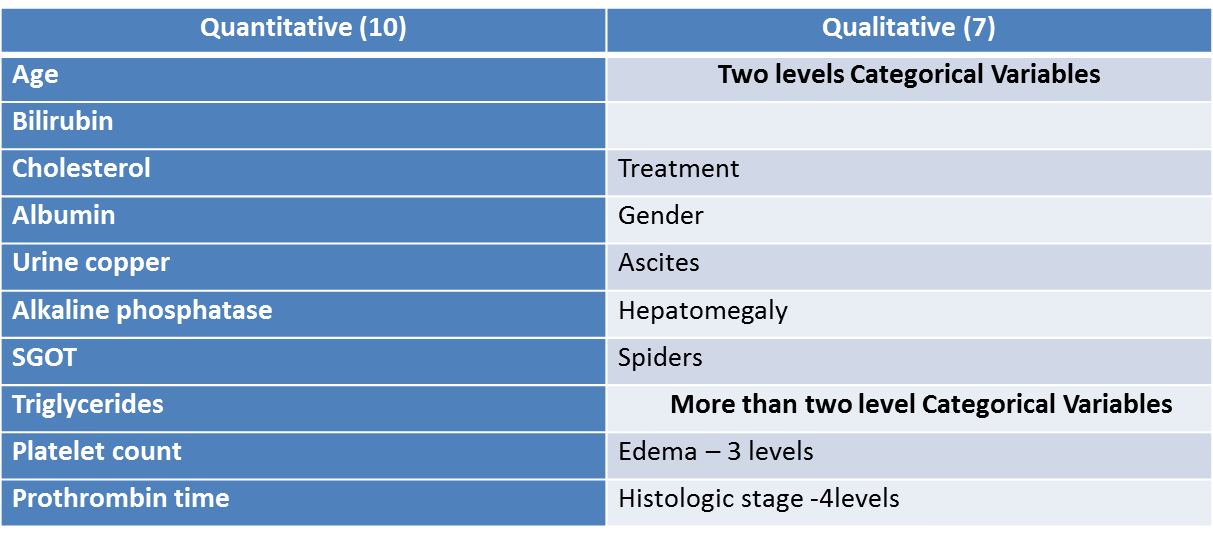
SGOT - in U/ml.

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.



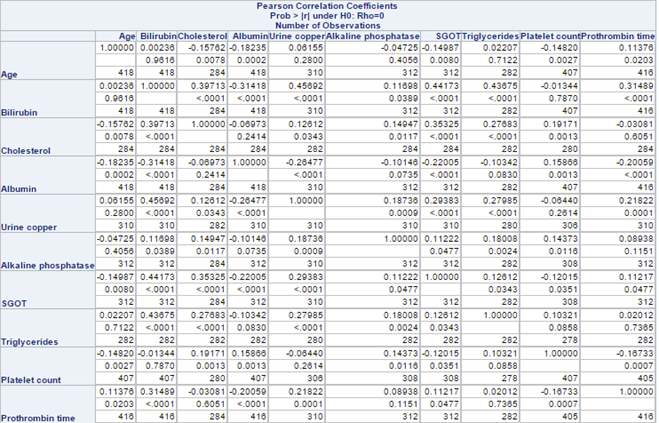
|  |  |
| --- | --- |
| **Methodologies used** | **Response Variable** |
| Linear Regression | Survival Time |
| One way ANOVA | Survival Time |
| Logistic Regression | Censoring |
| Survival Analysis | Survival Time |

**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.

**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:** ŷ =-180.34407 - 35.12471x1 - 66.88416x2+ 714.95866x3 - 2.39953x4 + 0.12762x5 - 314.99749x6

**Model 2:** log(ŷ) = 6.36654 + 0.01501x1 -0.04634x2 + 0.38197x3 - 0.00163x4 + 0.00006363x5 -0.23552x6 – 0.90628x7 – 0.26447x8

**Model 3:** log(ŷ) = 5.84251 + 0.01169x1 – 0.04837x2 + 0.35464x3 – 0.23669x6 – 0.91680x7 – 0.26571x8 – 0.20248x9 + 0.20066x10

**Model 4:** log(ŷ) = 7.27041 + 0.01319x1 – 0.04926x2 + 0.35503x3 – 0.23464x6 – 0.91599x7 – 0.26727x8 – 0.53336x9 + 0.04627x11

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MSE | Adjusted R2 | 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

x1 =Treatment (1 = D-penicillamine, 1 = Placebo)

x2 = Bilirubin

x3 = Albumin

x4 = Urine Copper

x5= Alkaline Phosphatase

x6= Histologic Stage 4

x7= Edema1

x8= Edema.5

x9= log(Urine Copper)

x10= log(Alkaline Phosphatase)

x11= log(Alkaline Phosphatase) \* log(Urine Copper)

**Global Hypothesis test for Model 4:**

Ho: β1 = β2 = β3= β6 = β7= β8 = β9 = β11 = 0

Ha: 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 R2 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:** yij= µ + αi + ɛij

Where yij = Survival Time

αi = effect if i-th level of Treatment m=1, 2

ɛij iidN(0,ơ2)

**Hypothesis testing:**

Ho: αi = 0 for all i

Ha: at least one αi ≠ 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.

1. 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:

Ho: Residuals are normally distributed

Ha: 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 Ha, 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:** logit(π) = βo + β1x1+ β2x2 + β3x3 + β4x4 + β5x5 + β6x6 + β7x7 + β8x8

logit(π) = log(π/(1-π))

Where y = Censoring

x1 = Age

x2 = Treatment1

x3 = Ascites

x4 = Bilirubin

x5 = SGOT

x6 = Alkaline Phosphatase

x7 = Urine Copper

x8= Prothrombin Time

**Fitted Model:** logit(π) =-13.5214+ 0.0585x1 + 0.1614x2 + 2.3733x3 + 0.2077x4 + 0.00739x5 0.000212x6 + 0.00479x7 + 0.7065x8

**Goodness-of-fit testing**

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

Ha: 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 Ho. 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 Ho.

**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**: logit(π) = βo + β1x1+ β2x2 + β4x4 + β5x5 + β6x6 + β8x8

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 Ha, 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:

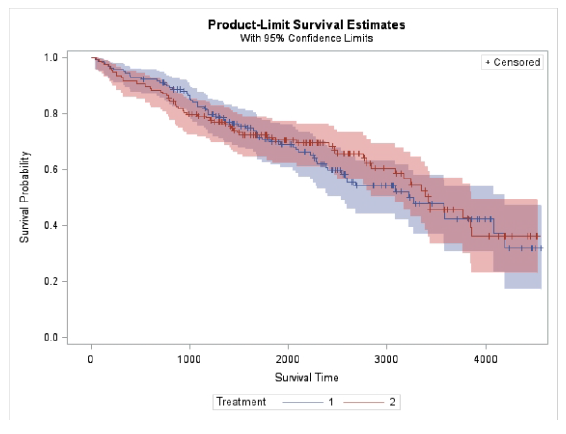
S1(t) = survival function for D-penicillamine

S2(t) = survival function for placebo

Ho: S1(t) = S2(t)

Ha: S1(t) ≠ S2(t)

Log-Rank and Wilcoxon tests, both have large p-values (.7498, .9664) so we cannot reject Ho. 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**

Ho : The overall fitted model is not significant

Ha : The overall fitted model is significant

The small p-value(<.001) for Likelihood Ratio, Score and Wald’s test support Ha. 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.