**Project Proposal Stat 271** 6th Oct 2015

**Group 2**

Prash Medirattaa 300137275

Gurinder Singh 300133063

Kuldeep Kaur 300129423

**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**

N Case number.

X The number of days between registration and the earlier of

death, liver transplantation, or study analysis time in July, 1986.

D 1 if X is time to death, 0 if time to censoring

Z1 Treatment Code, 1 = D-penicillamine, 2 = placebo.

Z2 Age in years. For the first 312 cases, age was calculated by

dividing the number of days between birth and study registration by 365.

Z3 Sex, 0 = male, 1 = female.

Z4 Presence of ascites, 0 = no, 1 = yes.

Z5 Presence of hepatomegaly, 0 = no, 1 = yes.

Z6 Presence of spiders 0 = no, 1 = Yes.

Z7 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

Z8 Serum bilirubin, in mg/dl.

Z9 Serum cholesterol, in mg/dl.

Z10 Albumin, in gm/dl.

Z11 Urine copper, in mg/day.

Z12 Alkaline phosphatase, in U/liter.

Z13 SGOT, in U/ml.

Z14 Triglycerides, in mg/dl.

Z15 Platelet count; coded value is number of platelets

per-cubic-milliliter of blood divided by 1000.

Z16 Prothrombin time, in seconds.

Z17 Histologic stage of disease, graded 1, 2, 3, or 4.

|  |  |
| --- | --- |
| Methodologies used | Response Variable |
| Logistic Regression | D |
| Survival Time data | X |
| Linear Regression | X |
| One Way ANOVA | X |

|  |  |
| --- | --- |
| Quantitative variable (**10 –**Variables) | Qualitative Variable (**7 –**Variables) |
| Z2 ,Z8,Z9,Z10,Z11,Z12,Z13,Z14,Z15,Z16 | Z1,Z3,Z4,Z5,Z6,Z7,Z17 |

|  |  |
| --- | --- |
| Categorical Variable | Level |
| Z7 | 3 |
| Z17 | 4 |

**Aim of the Project**

* **Logistic Regression**

By using logistic Regression we will model the probability of censoring. We will correlate risk of death with other explanatory variables (independent variables). Using logistic regression, all diagnostic plots will be created and assumptions verified. For Z7 we will re code as 0 1 and 2 to get proper odds ratio (output of logistic regression).

* **Survival Time**

We will use survival time to model the survival time and identify the factors that are significant to hazard of death. Response variable will be X.

* **Linear Regression**

We will use linear regression model to predict which factor effected the survival time in a linear model the most. We will check the assumptions and produce a prediction model.

* **One way ANOVA Model**

We will use ANOVA model to see whether the treatment as a single factor is effective in changing the survival time. Z1 is the factor and X is the response variable.