Data Analytics for Business

Homework 3 – Part 1

Survival Models

Imagine that you are hired as a consultant by a local bank to help them improve customer retention. For the analysis, the management of a local bank has given you access to 2505 customers, among whom 449 (about 18%) have closed their accounts within one year. Instead of simply having a binary variable 0/1 indicating whether a customer churned or not, you are now observing the exact time (measured in months) when a customer churns since s/he opened an account. If s/he didn't churn by the end of our observation period, the data are considered right-censored and are marked by **Censored**=1.

The data file is "**Bank_Attrition.csv**". It has the following variables:

Data Description

Column Name	Variable Description
Age	The customer's age
Income	The customer's income
HomeVal	The customer's home value
Tenure	How long this person has been a customer of the bank
DirectDeposit	Indicator dummy=1 if the customer uses direct deposit and 0 otherwise
Loan	Loan indicator dummy = 1 if the customer has ever taken loans from her bank and 0 if not
NumAccounts	The total number of accounts the customer has with this bank
Dist	Distance from customer's home to the nearest bank branch
MktShare	Bank's market share in the customer's market
ChurnTime	The exact time (measured in months) when a customer churns since s/he opened an account. If the customer hasn't churned at the end of the observation period, **ChurnTime** is right-censored.

Censored	**Censored**=1 means **ChurnTime** is right-censored at the end of the observation period. **Censored**=0 means no right-censoring, so the complete spell (i.e., the exact churn time) is observed.

Import that data into R and perform the following analysis.

Tasks:

1. Estimate a proportional hazard model with the Weibull distribution using the **survreg()** function. Properly specify the dependent variable **ChurnTime** with the **Censored** indicators, and include all the other variables (except **CustomerID**) as the explanatory variables (i.e., 9 in total).

2. As we explained in class, the reported estimates from the **survreg()** function is not the β estimates directly (but the δ estimates instead), and the "scale" estimate is not directly the shape parameter a in the Weibull model either. Transform them to the β estimates and the shape parameter a as we demonstrated in class. Interpret both the reported δ estimates and the transformed β estimates. For example, what do the two types of coefficients before **Income** mean?

3. **Plot the hazard function** based on the estimation results, evaluated at the mean values of the explanatory variables from the data sample. Is it positively or negatively duration dependent? What does it mean? Also, how would you interpret the hazard rates? Try interpreting the meaning of a particular point on the plot.

4. **Plot the density function of the duration** based on the estimation results, evaluated at the mean values of the explanatory variables from the data sample. To help you better understand what this density function means, **add the histogram** of the **ChurnTime** variable from the original data (for those *uncensored* observations) to the same plot (plotting the *density* rather than the *count* for the *y*-axis). They should be on the same scale and roughly follow the same pattern. (You may set "**breaks=50**" for the **hist()** function.)