Statistical Prognoses of Primary Biliary Cholangitis with Survival Analysis.

Final Project

Richard Paul Yim

21 November 2022

Contents

1 Introduction

Primary biliary cholangitis (PBC) is an autoimmune disease that can develop into primary sclerosing cholangitis and ultimately primary biliary cirrhosis where the small bile ducts in the liver become damaged, scarred and destroyed, respectively. This is a serious condition that can certainly lead to death if left untreated and mismanaged. Between January 1974 and May 1984, 424 patients were seen at the Mayo Clinic with follow-ups extended to April 30, 1988. The report presented aims to rigorously identify factors of increased risk of mortality for those afflicted with PBC.

This report is organized as follows: Section 2 describes the data in detail along with elementary statistical descriptions; Section 3 performs survival analysis and develops Cox proportional hazards models; Section 4 identifies influential outliers from the models developed via residual analysis; the report is concluded in Section 5 with a discussion on findings as well as other considerations regarding the study and treatment; Section 6 includes additional figures, references and code.

2 Data

Two separate datasets are associated with this study: `pbc`, data consisting of first visits by patients diagnosed with PBC at the Mayo clinic; and `pbcseq`, a subset of patients from `pbc` with multiple visits.

2.1 PBC Dataset and Covariates

The `pbc` dataset includes first visit data of individuals diagnosed with at least stage 1 PBC histology (requiring biopsy), results, as well as urine and blood tests. Table 1 presents the data dictionary of the `pbc` dataset. These results were observed between January 1974 and May 1984. The final cleaned variation of this dataset includes observations of 399 patients suffering from PBC.

2.2 PBCSEQ Dataset and Covariates

The `pbcseq` dataset is an extension of the original PBC study conducted by the Mayo clinic. The accrual of patients recorded in `pbcseq` began in January 1974 and ended in May 1984. Among the original 424 patients, 312 patients enrolled in a randomized double-blind clinical trial to study the effectiveness of D-penicillamine (used today as a treatment for Wilson's disease). Multiple follow-ups on patients in this clinical trial was reported up until patient death or end of study, which was April 30, 1988. After data cleaning and wrangling 1803 different visits with biopsy, blood and urine tests were recorded between 312 patients. Table 2 shows additional features that were produced from the `pbcseq` dataset after intervaling subsequent events and observations.

2.3 Sample Statistics

It's important to note that PBC disproportionately affects women. It's found that the efficacy of D-penicillamine is nearly absent with survival being statistically unaffected by adminstration of D-pencillamine. Beyond these important facts we make note of some elementary statistical descriptions of the `pbc` data for both continuously distributed variables in Table 3 as well as categorical variables, both ordinal and nonordinal, in Table 4. (Again, these statistical descriptions are of first visits of all patients in the PBC Mayo Clinic study between 1974 and 1984.)

It's been noted in the literature that low albumin and high bilirubin serum levels are associated with increased risk for mortality for those afflicted with PBC. In this dataset, we see that the average first-visit serum bilirubin level is 5.62 mg/dl compared to censored observations at 1.80 mg/dl. The average serum albumin level for that are censored is higher at 3.59 g/dl and lower for those that die during the study at 3.35 mg/dl.

Another covariate of particular interest is histologic stage of PBC, where stage 1 indicates immune response of biliary cholangitis; stage 2 is chronic cholestasis and inflammation developing into stage 3 fibrosis and sclerosis; and finally, cirrhosis and end-stage liver disease/failure. Table 6 shows the counts of each stage and mortality rates in the dataset. (It's important to note that PBC itself affects middle and older aged individuals, so death within groups may be due to other circumstances beyond PBC complications.) We report in later sections the low effect and lack of statistical significance with respect to histologic stage as related to mortality by PBC.

Table 1: Dataset encodings and variables.

Var	Definition
age	in years
albumin	serum albumin (g/dl)
alk.phos	alkaline phosphotase (U/liter)
ascites	presence of ascites
ast	aspartate aminotransferase, once called SGOT (U/ml)
bili	serum bilirunbin (mg/dl)
chol	serum cholesterol (mg/dl)
copper	urine copper (ug/day)
edema	0 no edema, 0.5 untreated or successfully treated, 1 edema despite diuretic therapy
hepato	presence of hepatomegaly or enlarged liver
id	case number
platelet	platelet count
protime	standardised blood clotting time
sex	m/f
spiders	blood vessel malformations in the skin
stage	histologic stage of disease (needs biopsy)
status	status at endpoint, 0/1/2 for censored, transplant, dead
time	number of days between registration and the earlier of death

Table 2: Additional Time-Dependent Covariates after intervaling appointment periods.

New Var	Definition and Factor Levels
tstart	appointment day
tstop	next appointment day
endpt	event death indicator (0=censored, 1=dead)

Table 3: Statistical descriptions of continuously distributed variables.

Statistical Description	Values
Event Time (mean, median)	(1895, 1690) days
Age (mean, median)	(50.58, 51.00) years
Serum Bilirubin (mean, median)	(3.234, 1.400) mg/dl
Serum Albumin (mean, median)	(3.497, 3.520) years mg/dl
Platelet (mean, median)	(255.4, 249.0) platelet count
Prothrombin Time (mean, median)	(10.72, 10.60) standardized blood clotting time

Table 4: Statistical descriptions of categorically distributed variables.

Statistical Description	Value
Mortality Rate within Study Time	37%
(Male, Female) Count	(43, 356)
Histologic Stage (one, two, three, four)	(19, 86, 153, 141)

Table 5: PBC Stage Counts and Mortality Rate.

Measure	Stage 1	Stage 2	Stage 3	Stage 4
Counts	19	86	153	141
Mortality Rate	10.55%	23.3%	30.7%	57.45%

3 Survival Analysis and Regression

We perform survival analysis and Cox proportional hazards modeling in this section. Commentary is made on D-penicillamine and its effectiveness as studied in the PBC treatment clinical trial, concluding that with respect to understanding PBC prognoses the treatment results are merely tangential to the primary focus of analysis in this report.

3.1 D-Penicillamine

Of the 424 patients, 312 went through the randomized clinical trial of D-penicillamine to study its effectiveness on treating copper accumulation and the associated complications of copper toxicity. The efficacy of the treatment is very minimal and not found to be effective at a statistically significant level. Figure 1 shows hazard ratios between the placebo and treatment groups; there is no indication of proportional hazards and effectively modeling survival with a Cox proportional hazards model would be very inappropriate. Furthermore, there is evidence in lack of difference of survival between the placebo group and the treatment group as seen in the complementary log-log plot where the curves cross over through the entire study. Testing for the difference in survival modeled via Kaplan-Meier curves, which is how the hazard ratios were produced we keep the null hypothesis that survival rates between placebo and treatment groups are essentially same at an α of 0.05 (`p=0.7`).

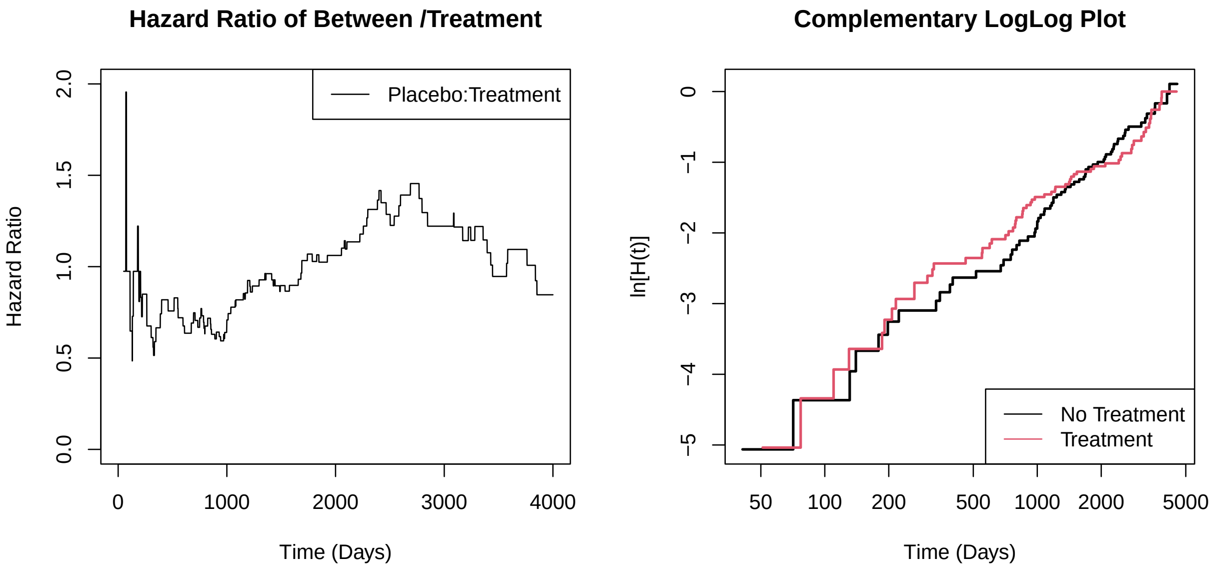


Figure 1: Hazard Ratios between placebo and treatment groups; complementary log-log plot between groups. There is a complete lack of constant hazard variance measured between both groups. Additionally, there is serious crossing over between complementary log-log curves indicating lack of distributional time-varying differences in survival between those receiving treatment and the placebo group (i.e., D-penicillamine is ineffective). The placebo group and treatment group have 41.7% mortality and 38.9% mortality rates, respectively.

3.2 Proportional Hazards and Survival

The primary concern of this report is discovering the most effective prognosis for those afflicted with PBC. Among all the covariates, edema, or swelling in the legs, proves to be the best indicator of survival—as opposed to histologic stage which requires biopsy and is not externally visible. Modeling survival of PBC patients across edema statuses requires visually verifying some assumptions required for the Cox proportional hazards model, which is what will be used. We check these assumptions across both individuals studied in the whole PBC dataset as well as separately looking at the 312 patients involved in the clinical trial.

1. Figure 2 shows Kaplan-Meier survival curve estimates between the three edema groups: no edema, edema but no diuretic therapy, edema with failed diuretic therapy. These curves model the probabilities of survival over time depending on which group a patient lies.

2. Figure 3 shows cumulative hazard curves from the Kaplan-Meier survival curves. The cumulative hazards effectively model the growing risk of mortality given the patient's initial edema status from the first clinic vist.

3. Figure 4 shows the hazard ratio curve computed by taking the ratio of the cumulative hazards of the three groups. This figure is important because it validates the use of the Cox proportional hazards model where we require fairly constant proportional variation of hazards over time.

4. Figure 5 is another variation for validating proportional hazards using complementary log-log curves, and also used for determining whether the survival processes are distributionally different. The fairly parallel curves indicate proportional hazards between the three groups, and there spread indicates different survival processes occurring between groups..

In addition to these visual procedures for checking the proportional hazards assumption, we find that for an $\alpha = 0.05$, the survival curves between different edema groups, for both the original PBC data and the clinical trial PBC participants, are different at a statistically significant level (`p<2e-16`) indicating that these groups' survival processes are truly different, yet proportional.

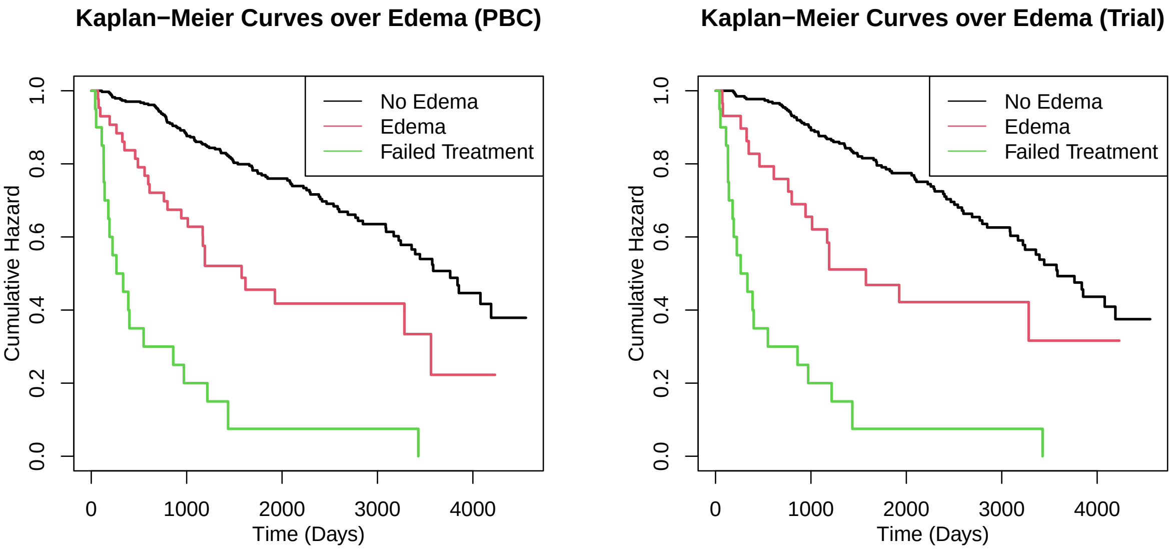


Figure 2: Kaplan-Meier estimates of survival functions between three groups over edema status. The survival probability is logically ordered between no edema with the highest survival probability, followed by edema yet to be treated diuretically, and failed diuretic treatment with lowest probability of survival.

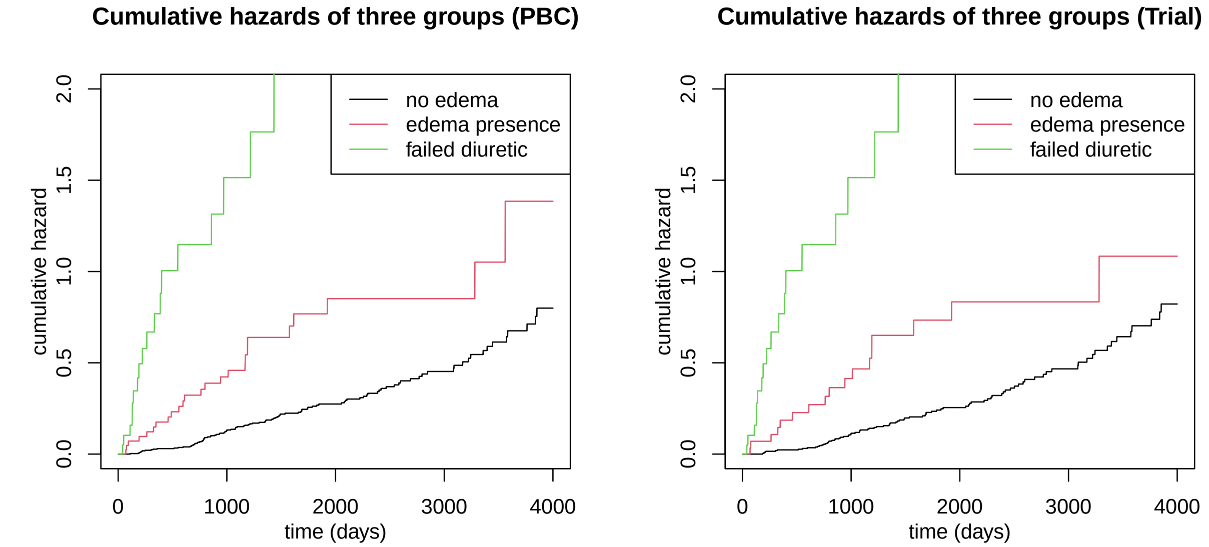


Figure 3: Cumulative hazard plot between three groups. The hazards between these groups appear to be proportional especially between patients with edema and failed edema treatment.

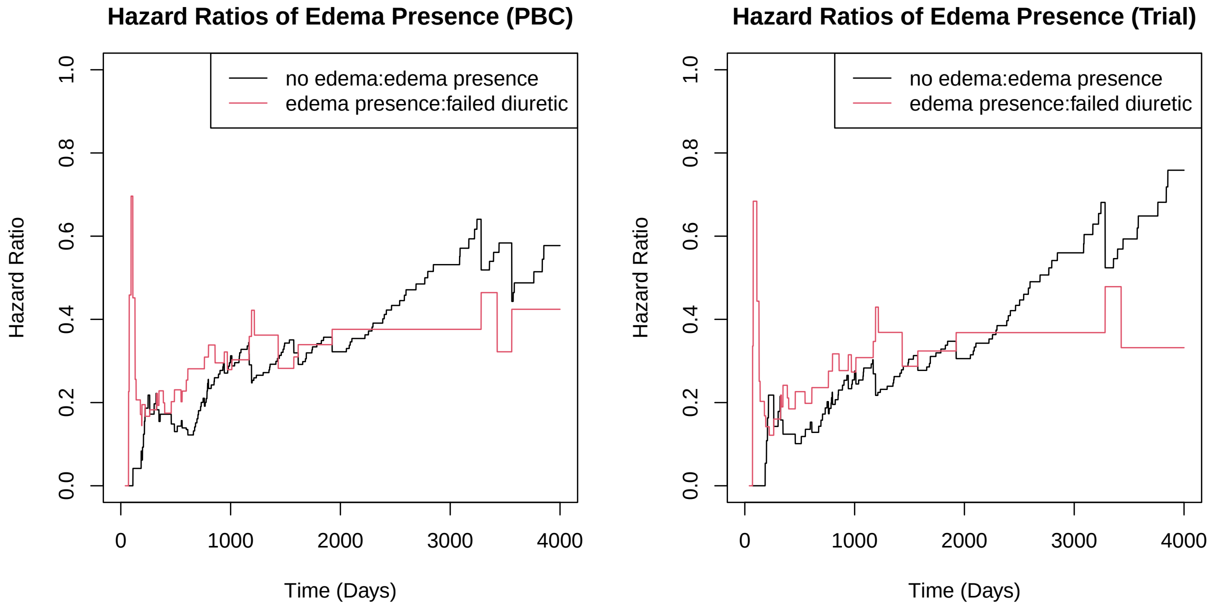


Figure 4: Hazard Ratios between groups. Hazard ratios between those with edema are constant, there is a bit of monotonicity and lack of proportionality between patients with no edema and patients with edema. The hazard ratio between the no edema and failed diuretic therapy group are excluded due to extreme lack of proportionality; this lack of proportionality if further elaborated in later sections.

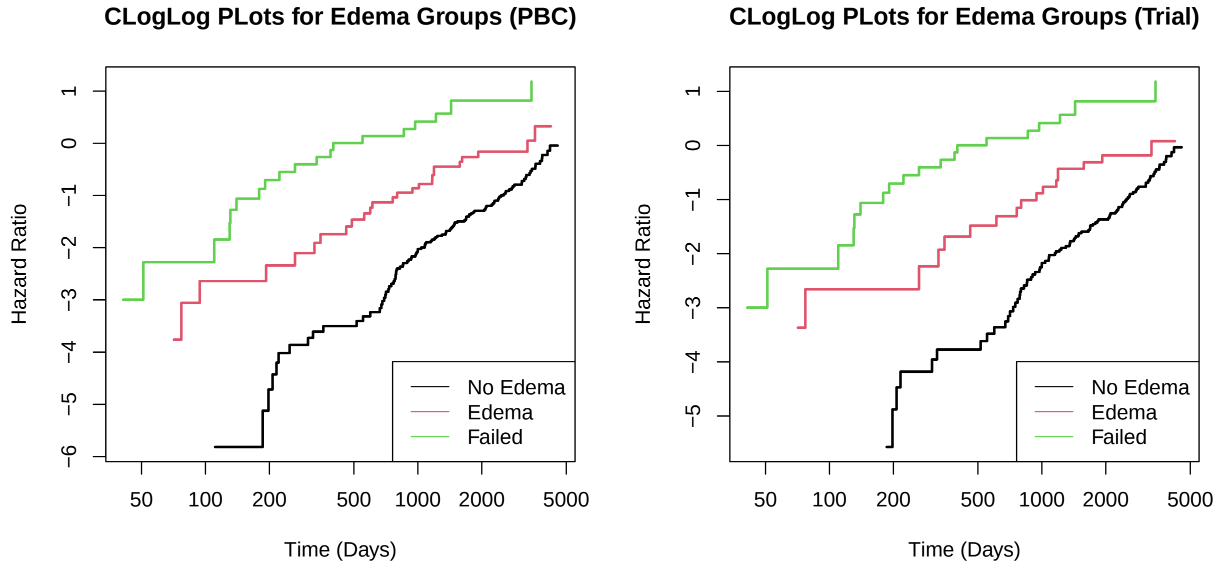


Figure 5: Complementary Log-Log plots between Edema groups. There is no overlap between curves and they appear to be relatively proportional, indicating different survival process occurring between groups as well. We can conclude proportional hazards and existence of variation between groups.

3.3 Cox Model

We build two separate Cox proportional hazards models to study the two different intersections of the pbc data. Model building was done while keeping three main points in mind:

1. We want the most elegant model to explain as much of the variance as possible. While fitting various combinations of covariates into different models, when it came to removing or adding covariates, the Akaike Information Criterion (AIC) helped to decide whether a certain feature added any additional value at a statistically significant level.

2. In addition to this principle of parsimony with respect to efficient and effective survival representation, goodness-of-fit was an important characteristic in *keeping* certain features as they improved the goodness-of-fit criterion with respect to the cumulative hazard of Cox-Snell residuals. In particular, this criterion leads to keeping the `log(prothrombin time)` as an important covariate despite statistically violating proportional hazards to be discussed in the next section. (Figure 6 shows the cumulative hazard plots for both models with sufficiently good fit.)

3. Practical significance dictates which covariates make actual clinical sense when associating prognoses with a patient's combination of risk factors. Keeping edema in the model was an obvious decision to be made given the agreement in proportional hazards across this covariate as well as its general effectiveness in determining PBC survivability due to its clear physical manifestation in a patient.

Altogether the final models that were built are listed in Table 6. With these final models we are able to conclude likelihoods of survival based off of the exponentials of the Cox model regression coefficients that are found to be statistically significant at an $\alpha = 0.05$.

For the PBC first-visits model, we find that age, edema with failed diuretic therapy, log serum albumin, log serum bilirubin, log prothrombin time are all statistically significant with `p-values` less than 0.005. For this model, platelet count was included as it helped with a slightly lower AIC score and with goodness-of-fit with the cumulative hazards of the Cox-Snell residuals. The following can be concluded from this model (all other covariates fixed):

1. An individual with PBC is 3.5% more likely to die if they are a year older than another PBC afflicted person.
2. An individual with PBC is 2.3 times more likely to die if they failed diuretic therapy with edema versus a person with no edema history.
3. An individual with PBC is 94% less likely to die if they have one unit greater amount in log serum albumin (g/dl) compared to another PBC afflicted individual.
4. An individual with PBC is 2.3 times more likely likely to die if they have one unit greater amount in log serum bilirubin (mg/dl) compared to another PBC afflicted individual.
5. An individual with PBC is 32 times more likely to die if their log prothrombin time is one log standard prothrombin time unit greater than another PBC afflicted individual.

For the PBC trial, multiple-visits model, we find that all covariates in the model are all statistically significant with `p-values` less than 0.005. The following can be concluded from this model (all other covariates fixed and appointment-to-appointment):

1. An individual with PBC is 5.6% more likely to die if they are a year older than another PBC afflicted person.
2. An individual with PBC is 2.2 times more likely to die if they failed diuretic therapy with edema versus a person with no edema history.
3. An individual with PBC is 99.15% less likely likely to die if they have one unit greater amount in log serum albumin (g/dl) compared to another PBC afflicted individual.
4. An individual with PBC is 1.17 times more likely likely to die if they have one unit greater amount in serum bilirubin (mg/dl) compared to another PBC afflicted individual.
5. An individual with PBC is 1.18 times more likely to die if their prothrombin time is one standard prothrombin time unit greater than another PBC afflicted individual.

Both variations of models and datasets appear to agree with each other very well with respect to increased/decreased likelihoods of mortality.

4 Residual Analysis

With the models developed in the previous section, we need to validate the proportional hazards assumptions across each covariate while checking for outliers in the data using other visual heuristics for studying residuals.

4.1 PBC (Leverage Points)

Across all the covariates for the PBC model, the proportional hazards assumptions appear to be well justified. In particular, Figure 13 shows the distribution of the residuals such that they are uniformly and randomly distributed with no strong nonzero linear pattern indicating independence over time of each covariate throughout the study. It's important to note the for `log(protime)` we see some nonzero linear association early in the study, but throughout the remainder of the study there is constant nonzero variation, visually. We allow this covariate to remain in the model.

Table 6: Models Studied.

Model	Covariates Used

`PBC (First Visit)`	`age + edema + log(albumin) + log(bili)`
	`+ log(protime) + platelet`

`Trial (Multiple Visits)`	`age + edema + log(albumin) + bili + protime`

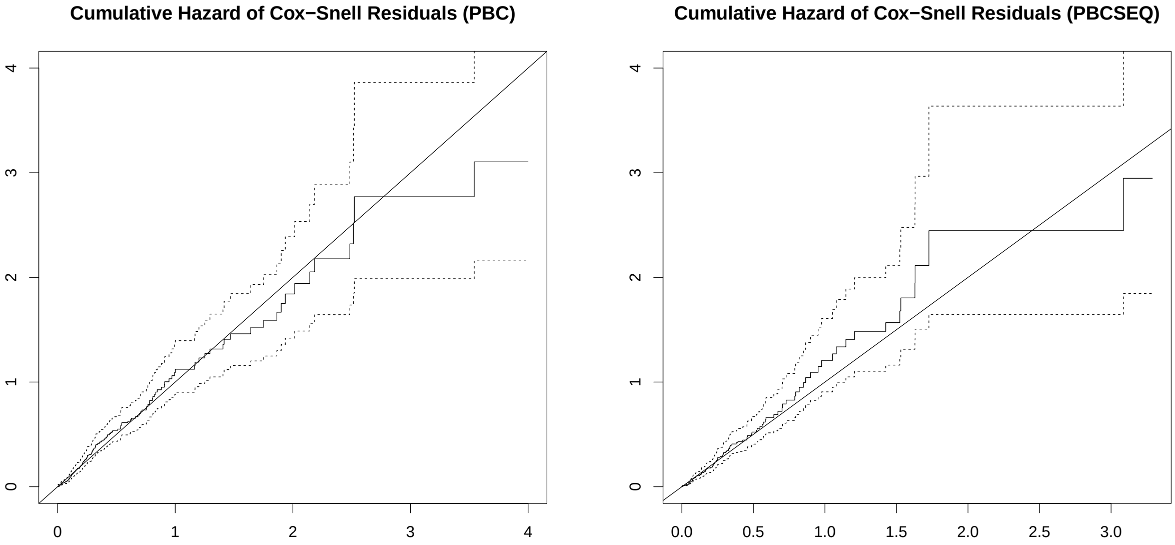


Figure 6: Cumulative hazard plot of Cox-Snell residuals indicating sufficient goodndess-of-fit across both the original PBC data and the time-dependent PBC data with multiple visits.

Figure 7 shows the Martingale and deviance residual plots, and Figure 8 and Figure 9 shows the corresponding dfbeta plots across each covariate. The Martingale plots serve as a visual heuristic for deciding if the linear model is poorly fit by any particular covariate, and in this case, there is no strong lack of fit since the linear trend appears to be relative flat in the distribution of the Martingale residuals. The deviance residuals are used to identify outliers in the model by utilizing the Martingale residuals—in particular, observations with large deviance residuals are not well predicted by the model. The `dfbeta` values model the residual change in the coefficients of each covariate of the model fit for each observation having been dropped. Data points with greater coefficient residuals indicate greater influence on the corresponding covariate with respect to the model. Among these residual plots, we find the following patients to be the most influential observations in the model: 77, 83, 85, 257, 297, 381, 412.

We comment on the influential points identified from the PBC first-visits data:

- Patient 77: This patient passed fairly early into the study within one year. This person is also a middle-aged female presenting symptoms of edema, high serum albumin levels at 6.6 mg/dl and low serum bilirubin levels at 3.41 with considerably long prothrombin time at a stage 3 progression of PBC. This patient presents to be a very normative influential point in the model.
- Patient 83: This female patient lasted long through the study at 4,050 days and is right censored. This patient appears to have been diagnosed with stage four PBC presenting with edema and a long prothrombin time—the longest among these influential points. This outlier appears to be particularly unusual considering that this patient despite the histologic stage of PBC, presents with low serum bilirubin levels at 1.3 g/dl. The survivability of this patient appears to be discordant with respect to expectation of survival of those with low serum bilirubin at stage 4 PBC.
- Patient 85: This female patient presents high bilirubin levels at 2.1 mg/dl and low albumin levels at 3.48 g/dl. Furthermore, the liver is at stage four PBC. This patient passed, but surprisingly far into the study at nearly ten years since these first measurements.
- Patient 257: This patient is 58 year old female with very low bilirubin levels at 0.5 mg/dl and high albumin levels at 4.23 g/dl. This patient presents to be a good influential observation fairly concordant with expectations of survival and in particular is seen to have only stage 2 PBC and presents to signs of edema.
- Patient 297: This patient is an age 55 years old male with no symptoms of edema at the first appointment,

but at stage 4 PBC. Furthermore, the patient has low serum bilirubin levels with relatively normal prothrombin time. This data point appears to be have been censored however within two years—a considerably low duration of observation.
- Patient 381: This is a male patient that is right censored at around five years. This patient had stage 4 PBC, but did not show any signs of edema. Yet, this patient seems to have low serum bilirubin and normal serum albumin levels. This patient appears to be a fairly concordant data point.
- Patient 412: This is a female patient censored at 1097 days and is the oldest among these patients at 67 years old. This patient had no signs of edema and had healthy levels of serum bilirubin and albumin levels despite being at stage three PBC. Interestingly this patient has the second lowest prothrombin time amongst these observations at 10.8 units.

4.2 PBC Clinical Trial (Leverage Points)

We apply the same techniques for identifying influential data points in our PBC clinical trial model for repeated visits. Figure 10 show the Martingale and deviance residual against linear predictor plots; Figure 11 and Figure 12 shows the dfbeta plots for each observation across each covariates. We list the most influential observations of visits in the model: patient 42, day 0; patient 70, day 0; patient 88, day 188; patient 104, day 195; patient 240, day 717. Table 7 provides a short summary of statistics for each of these patients over the course of the whole observation period as well as the particular visit of interest. In our discussion we discuss the results for these visits with respect to the individual patient and their appointment sequence as well as within all visits across all patients.

Table 7: Data of influential observations. Sample mean across visits and corresponding "influential" visits for covariates of interest. Edema is coded as (A, P, F), where ’A’ indicates absence of edema, ’P’ indicates presence, and ’F’ indicates failed therapy; ’N’ denotes the appointment number.

ID	Age	(A, P, F)	(N, Time, Status)	Bilirubin	Albumin	Pro. Time
42	33	(12, 2, 0)	Mean : (14, 4453, 0)	6.064 mg/dl	3.446 g/dl	11.14
		(1, 0, 0)	Visit : (1, 0, 0)	2.1 mg/dl	3.54 g/dl	11.0
70	56	(0, 11, 1)	Mean : (12, 3458, 0)	2.3 mg/dl	3.328 g/dl	10.88
		(0, 1, 0)	Visit : (1, 0, 0)	0.6 mg/dl	4.64 g/dl	10.6
88	41	(2, 2, 0)	Mean : (3, 2452, 0)	0.575 mg/dl	3.64 g/dl	11.2
		(1, 1, 0)	Visit : (2, 188, 0)	0.4 mg/dl	3.58 g/dl	10.4
104	43	(7, 0, 3)	Mean : (10, 3086, 1)	5.29 mg/dl	3.012 g/dl	11.10
		(2, 0, 0)	Visit : (2, 195, 0)	0.9 mg/dl	3.65 g/dl	11.1
240	56	(6, 0, 0)	Mean : (6, 1831, 0)	0.35 mg/dl	3.715g/dl	10.42
		(2, 0, 0)	Visit : (2, 717, 0)	0.3 mg/dl	3.79 g/dl	10.6

We comment on the influential points identified from the PBC multiple-visits data: the patients presented below are all females at least 41 years of age. Among all of these visits across these patients they record having at a serum bilirubin level of 1.0 mg/dl (the following are all females).

- Patient 42 (Appointment 1, Day 0): This first appoointment seems to be a departure from the patients average bilirubin levels measured over 10 years and 12 appointments. The average bilirubin is quite high at 6.064 mg/dl compared to the first visit at 2.1 mg/dl, but this patient appears to be censored by the end of the study. Additionally, this patient is fairly young at 33 years with no edema present in only 2/14 appointments. Furthermore, it’s worth noting the change in serum albumin levels as the average is lower over 14 appointments compared to the first appointment.
- Patient 70 (Appointment 1, Day 0): This patient is an older individual at age 56 with 12 total appointments, one of which resulted in failed diuretic therapy. Despite this, this patient seems to be right censored. Furthermore, this patient presented with fairly low bilirubin levels at a average over 12 appointments and 10 years at 2.3 mg/dl. The particular first visit is interesting since the patient begins with fairly low bilirubin levels, but the average over 12 appointments is nearly four times greater, and the serum albumin average is lower as well indicating deteriorating biomarker scores.

- Patient 88 (Appointment 2, Day 188): This patient is 41 years old, with very low average bilirubin levels over the course of 6 years and four appointments. This patient is right censored, and also reports with fairly high levels of albumin. This data point appears to be concordant with typical expectation of survival. It's also worth noting that the particular visit data, the second visit, is consistent with the average measurements over the course of this patient's observation.
- Patient 104 (Appointment 2, Day 195): This patient died during the study with three failed instances of diuretic therapy, very high serum bilirubin and very low serum albumin on average. This particular visit is consequently surprising considering the low bilirubin levels for a patient that eventually dies with high average serum bilirubin.
- Patient 240 (Appointment 2, Day 717): This patient survives over the course of 7 years and is right censored. The serum bilirubin levels are very low as are the serum albumin levels. This is a fairly concordant data point. In particular the visit of this patient at the second appointment is fairly stable over the average of 6 clinic appointments.

5 Conclusion

There are multiple courses of disease progression for those suffering from PBC across various risk factors. Prognoses of PBC symptoms vary from individual to indvidual. We have two main nonparametric models of first-visits over 400 patients and multiple visits of 312 patients participating in the clinical trials. In addition to modeling and observing likelihoods of mortality across covariates, we studied influential points in the data driving survival models of those with PBC.

5.1 Discussion

From both models we can immediately come to some general results of likelihood for survival for those with PBC:

1. Individuals that are older are more likely to die from PBC than younger individuals with PBC.
2. Individuals with edema that have failed diuretic therapy are more likely to die than those showing no edema.
3. Individuals with lower serum albumin levels have greater risk for mortality.
4. Individuals with greater serum bilirubin levels have greater risk for mortality.
5. Individuals with longer prothrombin time have greater risk for mortality.

Older age is a general risk factor for mortality across all diseases, and is the easiest to explain with respect to PBC mortality risk. For individuals with failed diuretic therapy, these individuals prove to have greater risk since the physical symptoms of PBC present themselves to be untreatable; such patients seem to fail edema therapy or have recurrent edema symptoms. It can be understood that failing to treat PBC extrinsic symptoms corresponds to extreme liver failure and cirrhosis where reversing damage done from liver fibrosis is irreversible.

Beyond age and edema, factors that are clinically accessible for both physicians and patients, blood tests focusing on serum albumin, serum bilirubin levels and prothrombin time are important biomarkers of liver health. Albumin is a protein in the body that helps transport small molecules through blood of which particular interest is bilirubin. Bilirubin is another protein that is produced when red blood cells are broken down from which it passes through the liver and is excreted from the body. It's clear then that individuals suffering from low albumin are perhaps failing to correctly transport bilirubin. Individuals with high bilirubin levels are then failing to breakdown bilirubin in the liver and consequently failing to expel it from the body leaving it to accumulate in the blood.

Across both models it was a commonly observed prognosis that individuals with PBC indicating high levels of serum bilirubin and low serum albumin levels were likely to die earlier. However, a few alternative courses of the disease presented itself among influential outliers:

1. One patient with stage four liver damage had a long period of survival presenting with edema at 4,050 days with low serum bilirubin levels.

2. Another patient with stage four liver damage, high bilirubin and low albumin levels also survived for a long time at nearly ten years.
3. A 67 year old patient, greater than the 3rd quartile, with stage three liver damage presented with no signs of edema and health levels of serum bilirubin and serum albumin levels, with healthy and normal prothrombin time.

These prognoses are particularly important because it indicates that patients that are older, patients with higher grader liver damage, and patients with edema can still manage to excede survival expectations and that the disease is manageable over the course of the patients lifetime. The patient retains control over her health.

From the multiple-visits PBC data we find the same concordant scenarios for influential observations in our multiple-visits model—it's important to notice that patients with failed diuretic therapy are at much greater risk for mortality. Additionally, a prognosis emerges where a patient with healthy serum albumin and serum bilirubin levels can eventually develop poorer scores over their lifetime.

5.2 Other Considerations

An auxiliary result of the PBC Mayo Clinic study is that D-penicillamine was found to not be an effective treatment for PBC not at all improving survival rates. A variable included in the dataset was copper, and copper accumulation can result from PBC—D-penicillamine effectively helps the body to excrete copper where copper accumulation common for those suffering from PBC. Yet, the lack of its efficicay suggests that copper accumulation is one of the many symptoms of PBC. Of greater importance is blood protein serum levels for both albumin and bilirubin.

In addition to these protein biomarkers, although not studied in our models, serum cholesterol levels are also of interest. Just looking at mortality rates for individuals with cholesterol levels greater than the sample serum cholesterol level average, there is a 55% mortality rate compared to a 32% mortality rate for those below the sample average. In other words managing cholesterol seems to correlated to liver health and survivability for those affected by PBC. Of even greater importance is the positive correlation of body mass index (BMI) and cholesterol levels: we are suggesting that one of the best therapies for managing PBC is through maintaining healthy lifestyle habits such as proper diet and nutrition, sleep, exercise and stress management. This is particularly important to make note of when considering patients in our model found to have great survivability despite being older and presenting edema.

This long-term study has certain limitations of which future studies could possible track fairly easily. The main content of this data is the histological and serological measurements of each patient. What is perhaps lacking is the patients own survey data at each visit. Questions such as other prescription drugs, lifestyle habits and the like can possibly be integrated into this data. Other physical measurements such as body mass index can also be of help as good physical fitness generally decrease risk for all-cause mortality. These factors can further inform a patient's prognosis with respect to PBC.

5.3 Current Therapies

There is currently no cure for this liver disease. The best individuals suffering from PBC can do is to, again, practice and maintain healthy lifestyle habits. In addition to maintaining health lifestyle habits, other bad habits such as smoking, drinking, consumptions of certain non-steroidal anti-inflammatory drugs (NSAIDs) such as aspirin and ibuprofen should potentially be avoided.

Although PBC is uncurable, there are drugs today that can increase liver longevity and health in a sufferer of PBC. Of particular interest and often prescribed for those with PBC is ursodeoxycholic acid (UDCA, ursodiol), which is particularly effective for those in the early stages of PBC, although a patient will have to likely rely on this drug for the rest her of her life as it becomes an external source of bile acid important for reducing cholesterol saturation.

6 Appendix

6.1 Additional figures

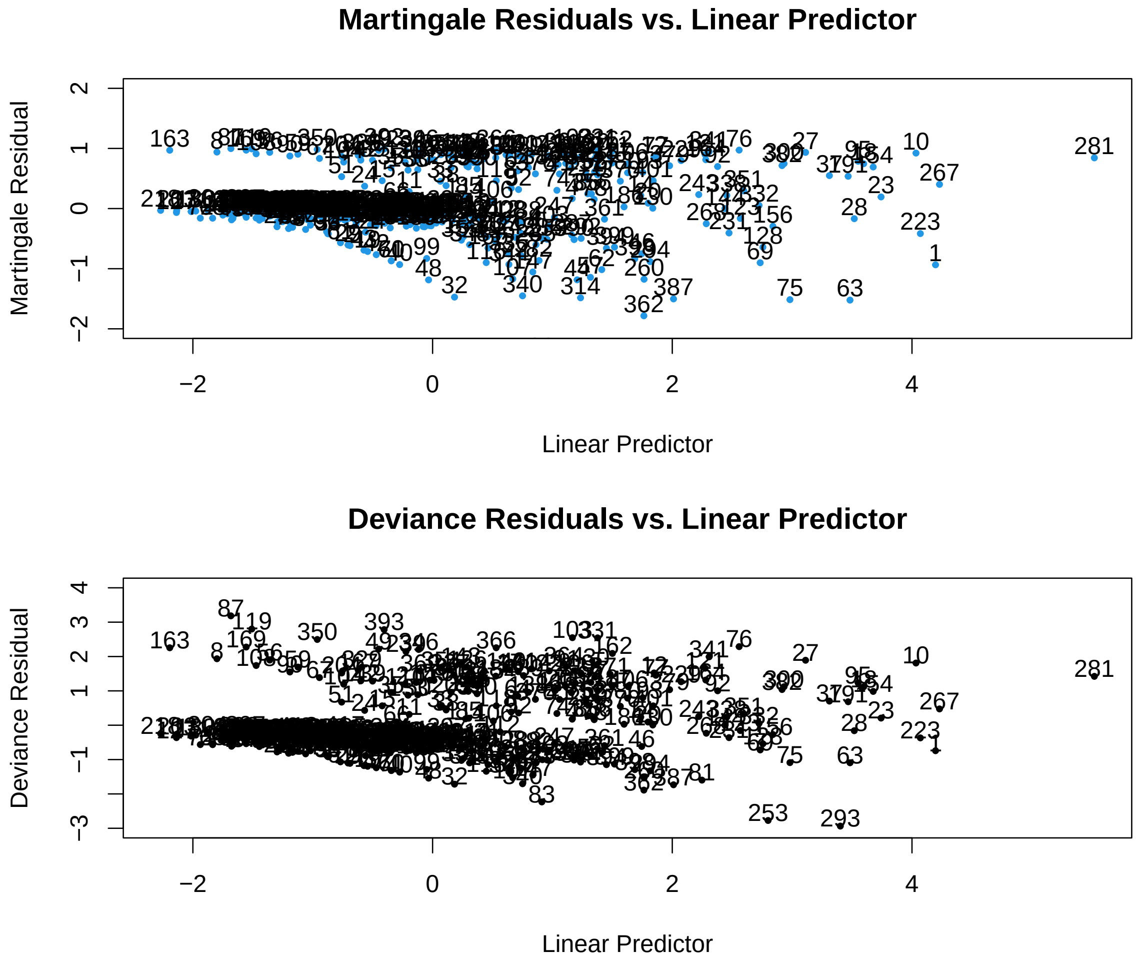
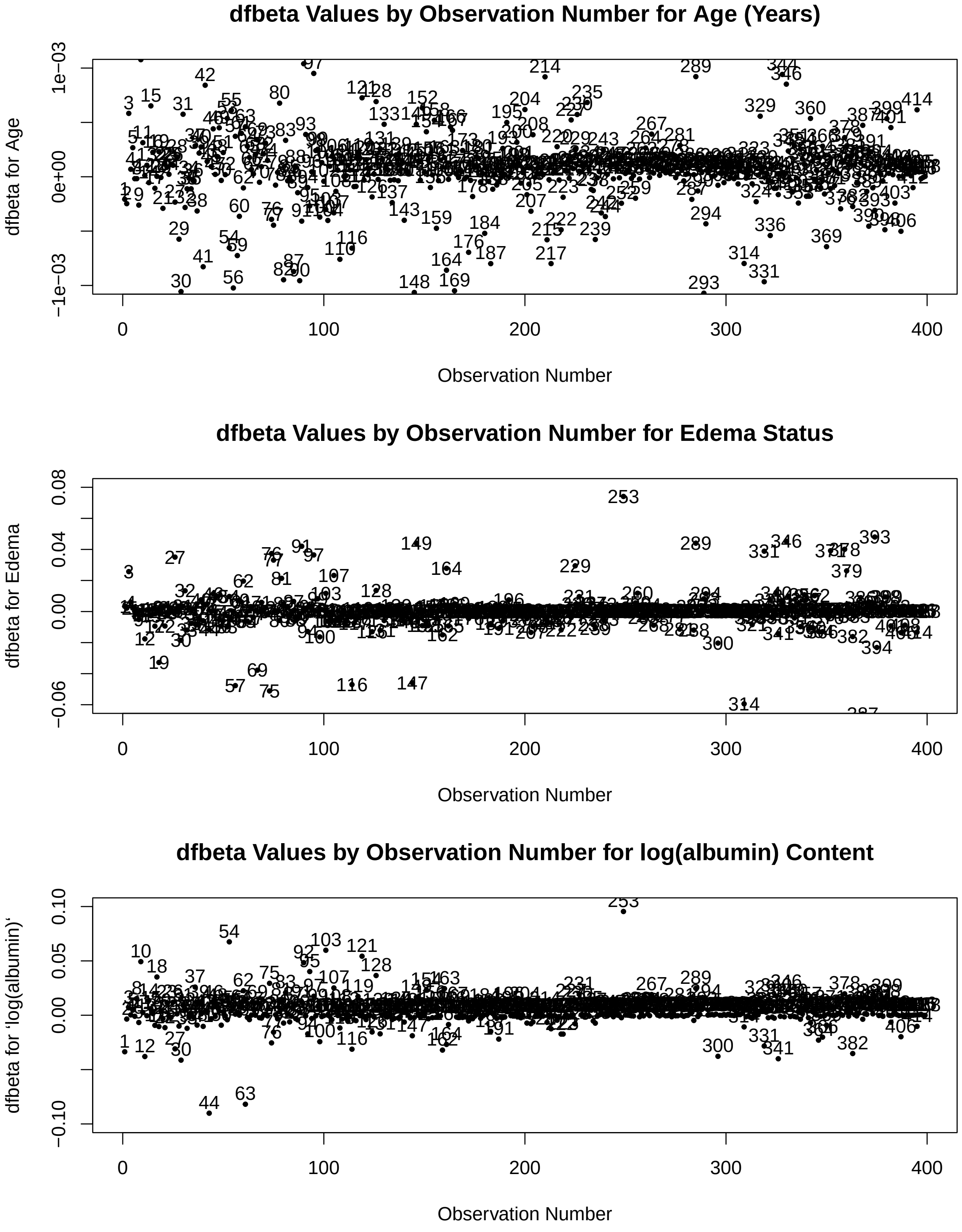


Figure 7: Martingale and deviance residuals versus linear predictor plots for PBC data model.


dfbeta Values by Observation Number for Age (Years)

dfbeta Values by Observation Number for Edema Status

dfbeta Values by Observation Number for log(albumin) Content

Figure 8: `dfbeta` value plots for each data point across age, edema status and log serum albumin levels.

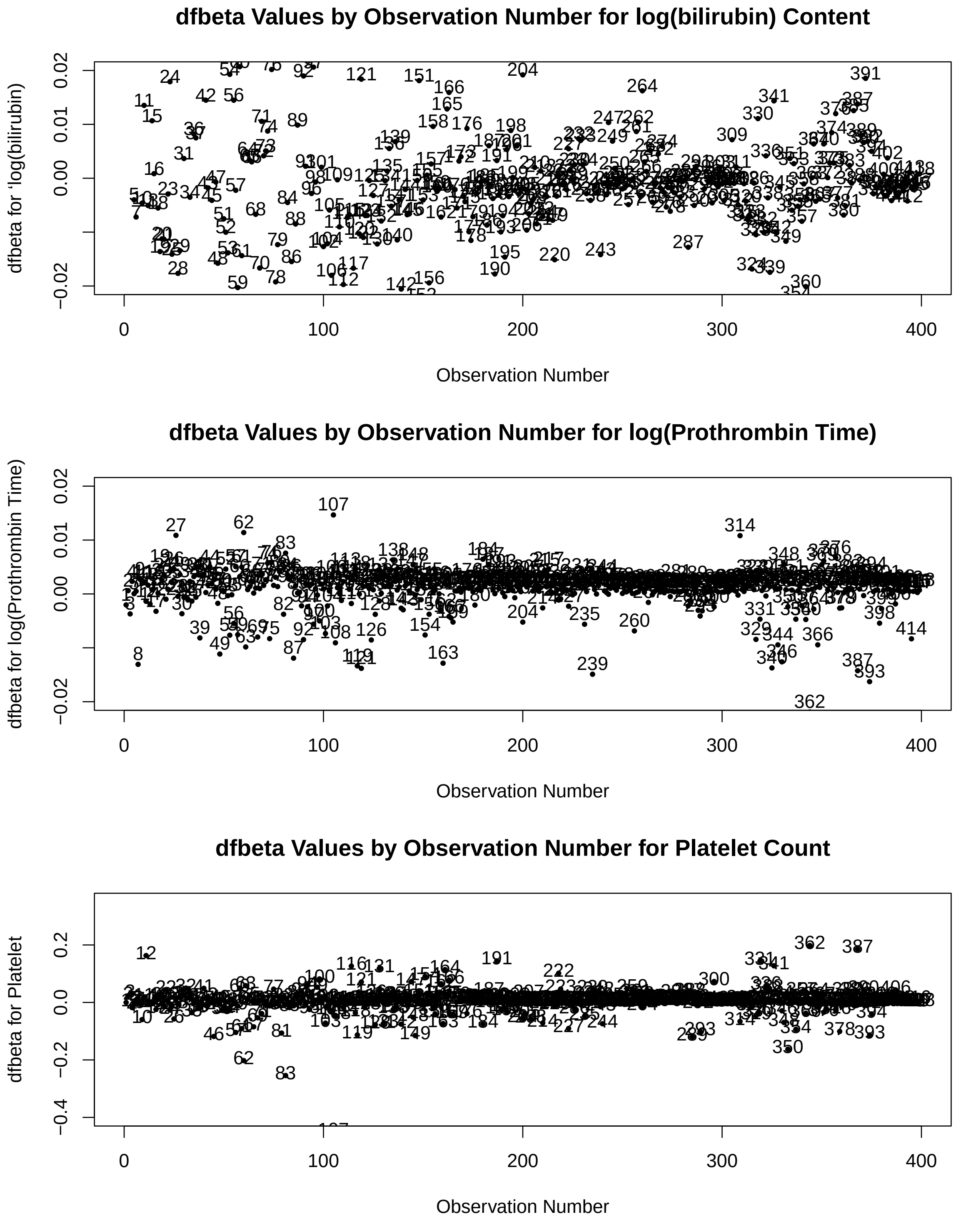


Figure 9: `dfbeta` value plots for each data point across log serum bilirubin levels, log prothrombin time and platelet count.

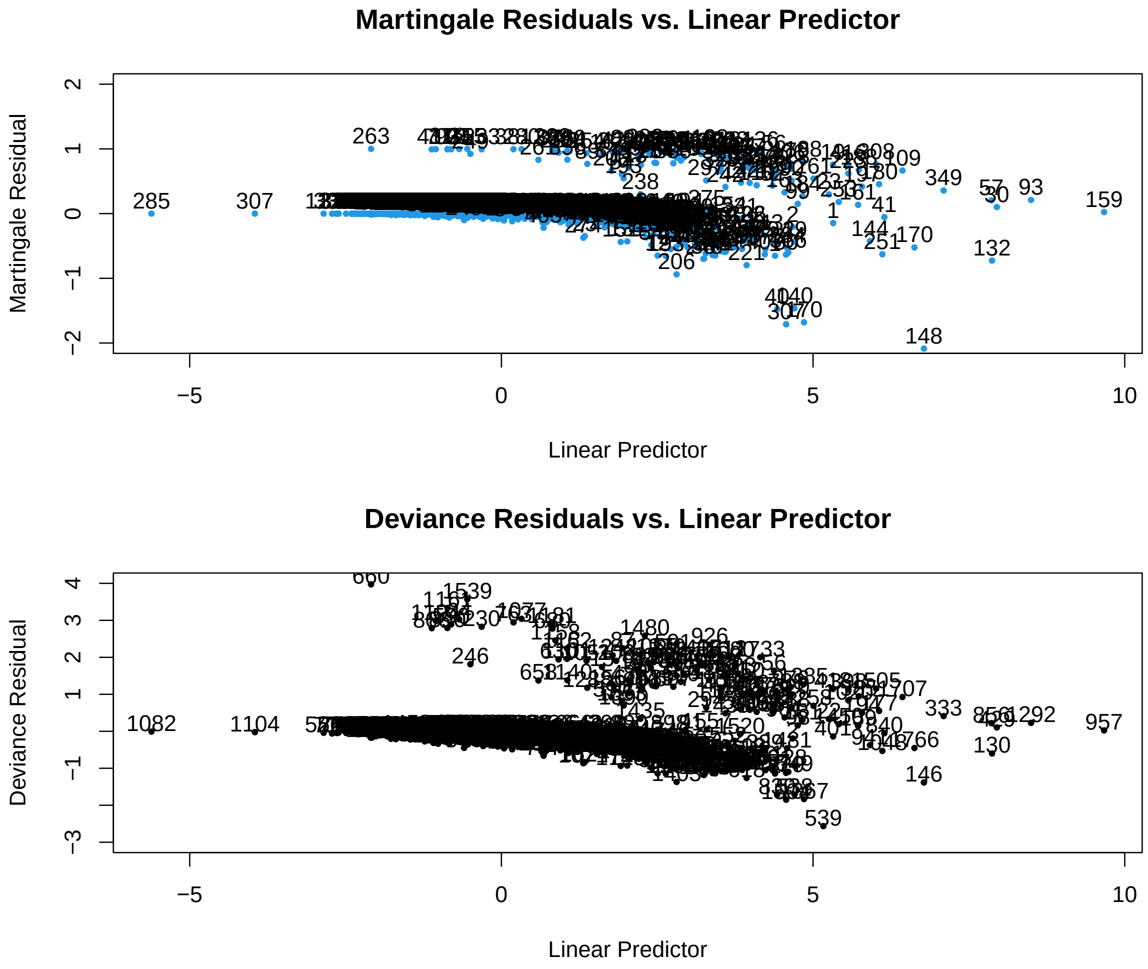


Figure 10: Martingale and deviance residual versus linear plots for PBC clinical trial participants.

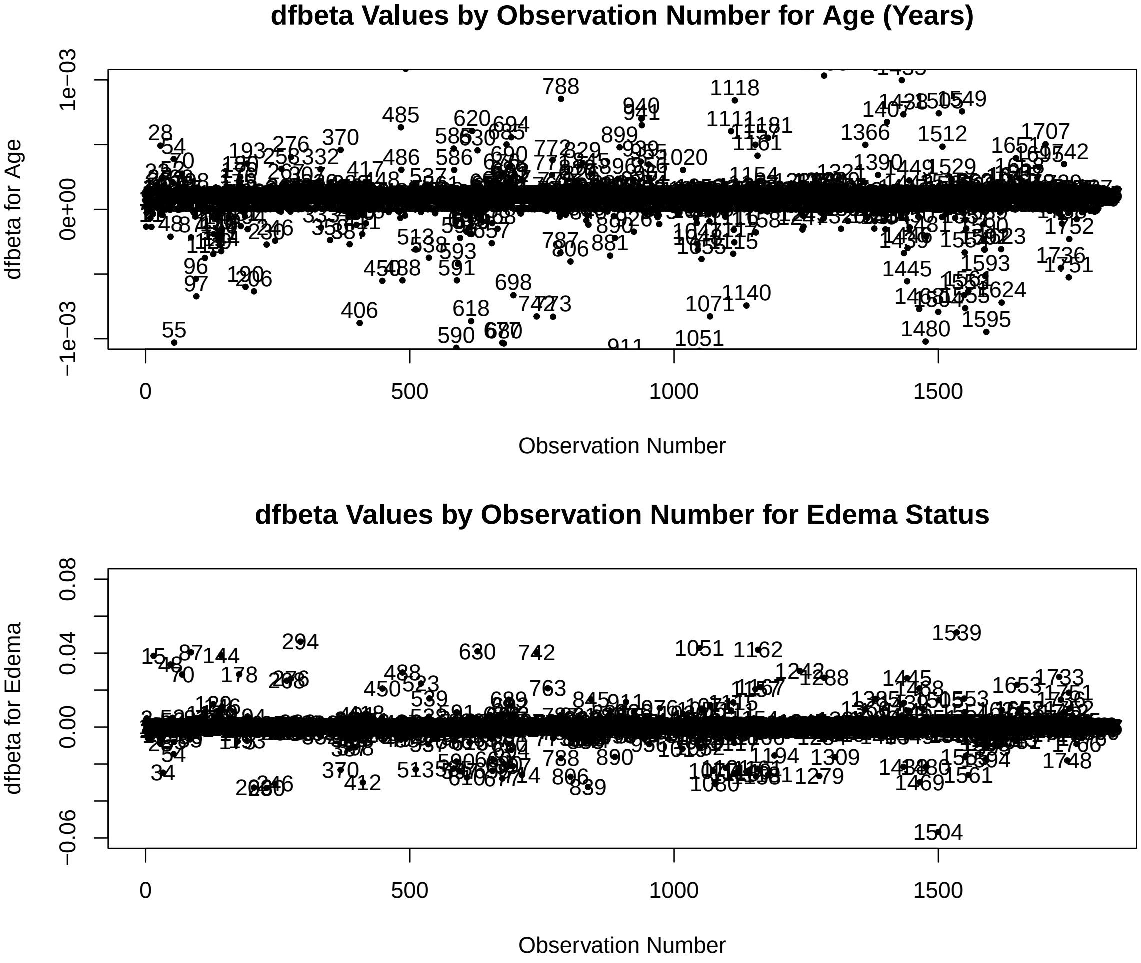
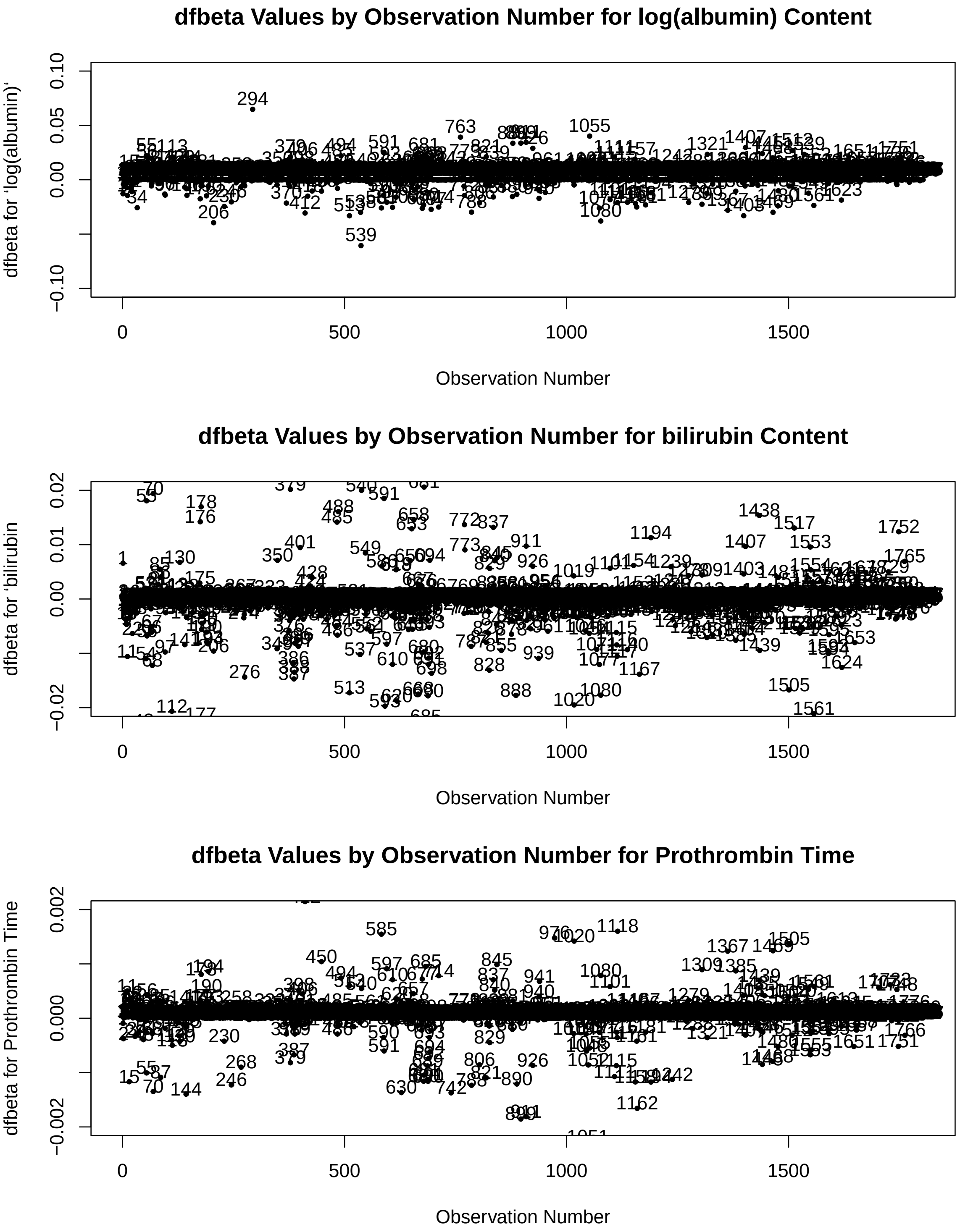


Figure 11: `dfbeta` value plots for each data point across age and edema status.


dfbeta Values by Observation Number for log(albumin) Content

dfbeta Values by Observation Number for bilirubin Content

dfbeta Values by Observation Number for Prothrombin Time

Figure 12: `dfbeta` value plots for each data point across log serum albumin levels, serum bilirubin levels and prothrombin time.

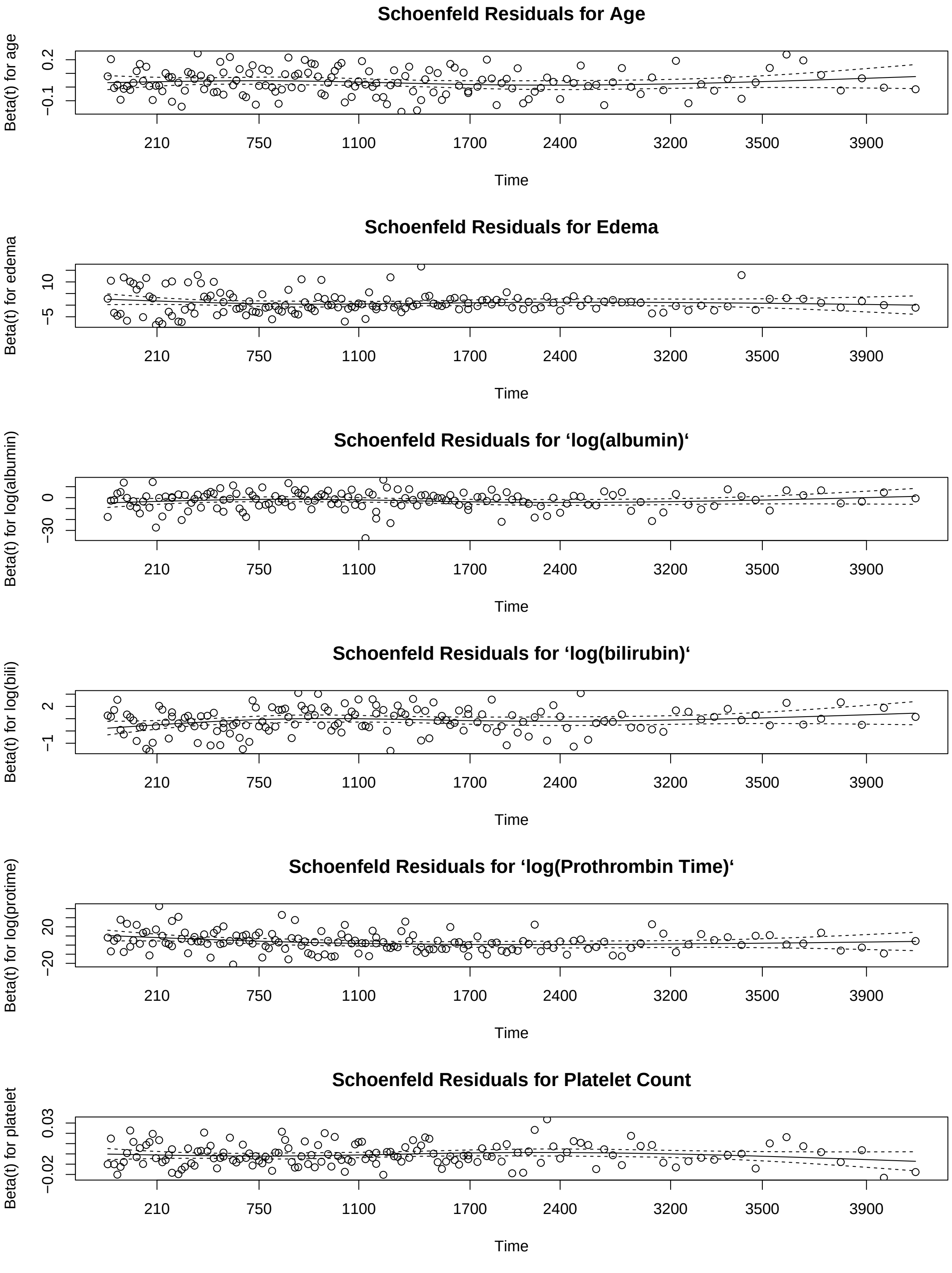


Figure 13: Schoenfeld residual plots showing fairly constant residuals for first-visits PBC model.

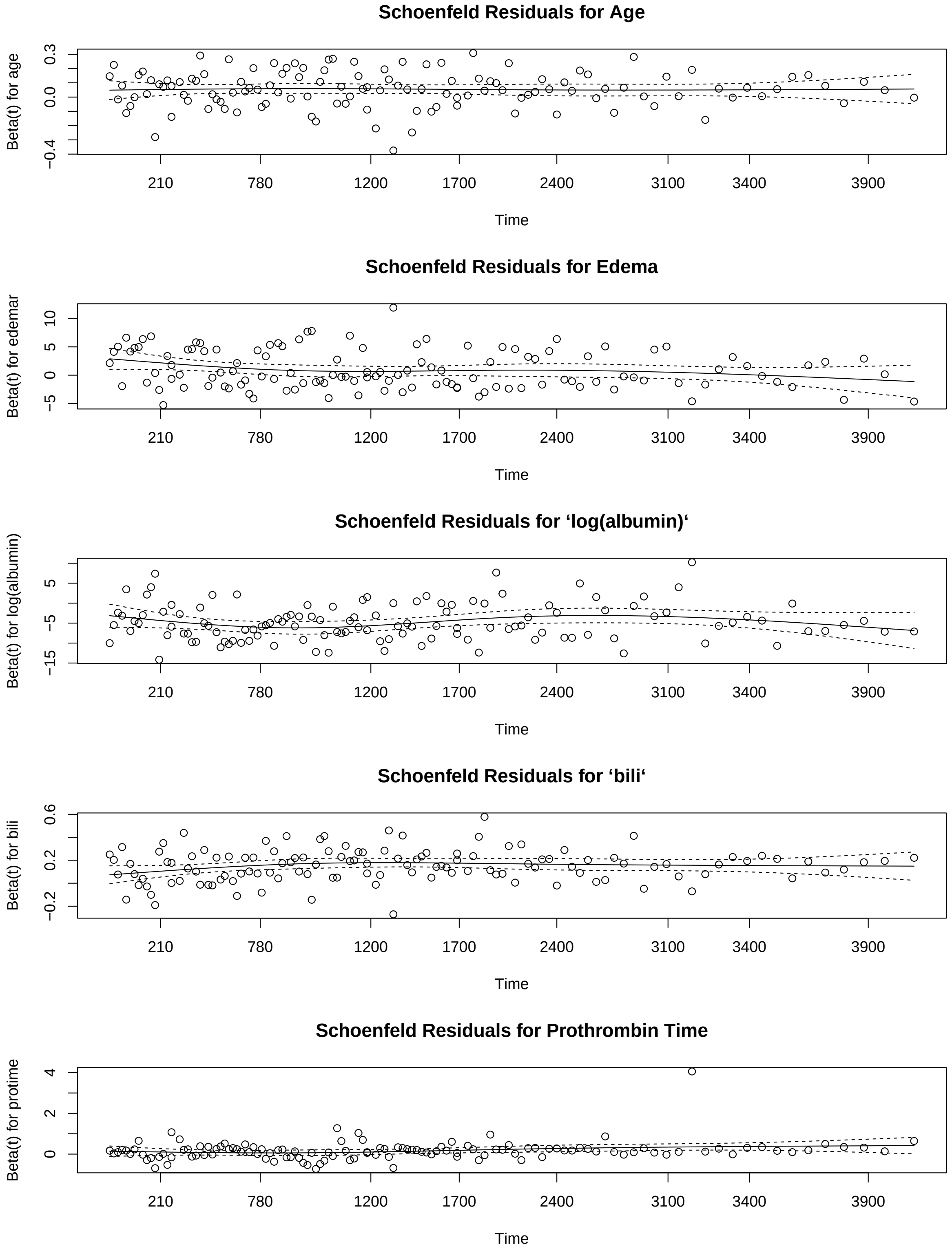


Figure 14: Schoenfeld residual plots showing fairly constant residuals for the clinical trial participants.

6.2 Code Chunks

```
knitr::opts_chunk$set(echo = TRUE)
# ============================================================================ #
# Loading Libraries and Reading Data
# ============================================================================ #
require(KMsurv)
require(survival)
data(pbc, package="survival")

# ==============================================================================
# Data Wrangling and EDA (PBC)
# ==============================================================================
# reading in the data and libraries
require(KMsurv)
require(survival)
data(pbc, package="survival")
colnames(pbc)
# na check
apply(apply(pbc,2,is.na),2,sum)
# recoded
nona <- na.omit(subset(pbc, select=
           c(id,time,status,age,sex,edema,bili,albumin,platelet,protime,stage)))
pbcDecode <- nona
pbcDecode$edema <- factor(pbcDecode$edema, labels=c("asymp", "edema", "failed"))
pbcDecode$sex<- factor(pbcDecode$sex, labels=c("male","female"))
pbcDecode$status <- as.numeric(pbcDecode$status==2)
pbcDecode$stage <- factor(pbcDecode$stage, labels=c("one","two","three","four"))
head(pbcDecode)
# log transform (EDA)
if(0){
  hist(pbcDecode$bili)
  hist(log(pbcDecode$bili))
  hist(pbcDecode$albumin)
  hist(log(pbcDecode$albumin))
  hist(pbcDecode$platelet)
  hist(log(pbcDecode$platelet))
# might want to check model difference
hist(pbcDecode$protime,freq = TRUE, breaks=10)
hist(log(pbcDecode$protime),freq = TRUE, breaks = 10)
}
# pbcDecode rename and na check
pbc1 <- pbcDecode
apply(apply(pbc1,2,is.na),2,sum)
# pbc1 simple stat descriptions
summary(pbcDecode)

# ==============================================================================
# Data Wrangling and EDA (PBCSEQ)
# ==============================================================================
# reading in the data and libraries
require(KMsurv)
require(survival)
data(pbc, package="survival")
```

```r
colnames(pbc)
# na check
apply(apply(pbcseq,2,is.na),2,sum)
# tmerge
temp <- subset(pbc, id <= 312, select=c(id:sex, stage, edema), status = tdc(day, status))
pbcseq.tdc <- tmerge(temp, temp, id=id, endpt = event(time, status))
pbcseq.tdc <- na.omit(tmerge(pbcseq.tdc, pbcseq,
               id=id, ascites = tdc(day, ascites), hepato = tdc(day, hepato),
               stager = tdc(day, stage), edemar = tdc(day, edema),
               spiders = tdc(day, spiders), bili = tdc(day, bili),
               platelet = tdc(day,platelet), albumin = tdc(day, albumin),
               protime = tdc(day, protime), alk.phos = tdc(day, alk.phos)))
head(pbcseq.tdc)
# recoded
pbcseq.tdc$spiders <- factor(pbcseq.tdc$spiders, labels=c("absent","present"))
pbcseq.tdc$hepato<- factor(pbcseq.tdc$hepato, labels=c("absent","present"))
pbcseq.tdc$ascites<- factor(pbcseq.tdc$ascites, labels=c("absent","present"))
pbcseq.tdc$stager <- factor(pbcseq.tdc$stager, labels=c("one","two","three","four"))
pbcseq.tdc$edemar <- factor(pbcseq.tdc$edemar, labels=c("asymp", "edema", "failed"))
pbcseq.tdc$sex<- factor(pbcseq.tdc$sex, labels=c("male","female"))
pbcseq.tdc$endpt<- as.numeric(pbcseq.tdc$endpt==2)
head(pbcseq.tdc)
# na check
apply(apply(pbcseq.tdc,2,is.na),2,sum)
# temp simple stat descriptions
summary(pbcseq)
# ============================================================================
# Survival Analysis (TREATMENT)
# ============================================================================
par(mfrow=c(1,2))
# km curves
pbcseq.surv <- Surv(time=temp$time, event=temp$status==2)
if(0){plot(survfit(pbcseq.surv~factor(trt),data=temp),col=1:3,lwd=2)}
# hazards (EDEMA)
NAcurves <- survfit(pbcseq.surv~factor(trt), type="fleming-harrington", data=temp)
timevec <- 1:4000
sf1 <- stepfun(NAcurves[1]$time,c(1,NAcurves[1]$surv))
sf2 <- stepfun(NAcurves[2]$time,c(1,NAcurves[2]$surv))
#now we can find the cumulative hazards
cumhaz1 <- -log(sf1(timevec))
cumhaz2 <- -log(sf2(timevec))
if(0){
  plot(timevec,cumhaz1,,type="l",ylim=c(0,2), col=1)
  lines(timevec,cumhaz2,type="l",ylim=c(0,6), col=2)
  legend("topright",c("No Treatment", "Treatment"),col=1:2,lwd=1)
  title("Cumulative Hazards of Two Groups",xlab="Time (Days)",
        ylab="Cumulative Hazard", line=2)
}
# Hazard Ratios
plot(timevec, cumhaz1/cumhaz2, type="l", ylim=c(0,2), ylab="Hazard Ratio",
     xlab="Time (Days)")
title("Hazard Ratio of Between /Treatment")
# checking survdiff
```

```r
pbc.survdiff.trt<- survdiff(pbcseq.surv~factor(trt),data=temp)
legend("topright",c("Placebo:Treatment"),col=1:2,lwd=1)
print(pbc.survdiff.trt)
print(summary(pbc.survdiff.trt))
# complementary log log
plot(NAcurves, col=1:3, lwd=2, fun="cloglog", ylab="ln[H(t)]", xlab="Time (Days)")
legend("bottomright",c("No Treatment", "Treatment"),col=1:2,lwd=1)
title("Complementary LogLog Plot")
# ===============================================================================
# KM survival curves
# ===============================================================================
par(mfrow=c(1,2))
# km curves (fixed)
pbc.surv <- Surv(time=pbc1$time, event=pbc1$status)
plot(survfit(pbc.surv~edema,data=pbc1),col=1:3,lwd=2)
title("Kaplan-Meier Curves over Edema (PBC)",xlab="Time (Days)",
      ylab="Cumulative Hazard", line=2)
legend("topright",c("No Edema", "Edema", "Failed Treatment"),col=1:3,lwd=1)
# km curves (varied)
pbcseq.surv <- Surv(time=temp$time, event=temp$status==2)
plot(survfit(pbcseq.surv~edema,data=temp),col=1:3,lwd=2,xlab="",ylab="")
title("Kaplan-Meier Curves over Edema (Trial)",xlab="Time (Days)",
      ylab="Cumulative Hazard", line=2)
legend("topright",c("No Edema", "Edema", "Failed Treatment"),col=1:3,lwd=1)
# ===============================================================================
# Cumulative Hazards
# ===============================================================================
par(mfrow=c(1,2))
# km curves (fixed)
# hazards (EDEMA)
NAcurves <- survfit(pbc.surv~edema, type="fleming-harrington", data=pbc1)
timevec <- 1:4000
sf1 <- stepfun(NAcurves[1]$time,c(1,NAcurves[1]$surv))
sf2 <- stepfun(NAcurves[2]$time,c(1,NAcurves[2]$surv))
sf3 <- stepfun(NAcurves[3]$time,c(1,NAcurves[3]$surv))
#now we can find the cumulative hazards
cumhaz1 <- -log(sf1(timevec))
cumhaz2 <- -log(sf2(timevec))
cumhaz3 <- -log(sf3(timevec))
plot(timevec,cumhaz1,type="l",ylim=c(0,2), col=1,xlab="",ylab="")
lines(timevec,cumhaz2,type="l",ylim=c(0,6), col=2)
lines(timevec,cumhaz3,type="l",ylim=c(0,6), col=3)
legend("topright",c("no edema", "edema presence", "failed diuretic"),col=1:3,lwd=1)
title("Cumulative hazards of three groups (PBC)",xlab="time (days)",
      ylab="cumulative hazard", line=2)
# hazards (EDEMA, varied)
NAcurvesr <- survfit(pbcseq.surv~edema, type="fleming-harrington", data=temp)
timevec <- 1:4000
sf1r <- stepfun(NAcurvesr[1]$time,c(1,NAcurvesr[1]$surv))
sf2r <- stepfun(NAcurvesr[2]$time,c(1,NAcurvesr[2]$surv))
sf3r <- stepfun(NAcurvesr[3]$time,c(1,NAcurvesr[3]$surv))
#now we can find the cumulative hazards
cumhaz1r <- -log(sf1r(timevec))
```

```
cumhaz2r <- -log(sf2r(timevec))
cumhaz3r <- -log(sf3r(timevec))
plot(timevec,cumhaz1r,type="l",ylim=c(0,2), col=1,xlab="",ylab="")
lines(timevec,cumhaz2r,type="l",ylim=c(0,6), col=2)
lines(timevec,cumhaz3r,type="l",ylim=c(0,6), col=3)
legend("topright",c("no edema", "edema presence", "failed diuretic"),col=1:3,lwd=1)
title("Cumulative hazards of three groups (Trial)",xlab="time (days)",
      ylab="cumulative hazard", line=2)
# ==============================================================================
# Hazard Ratio Plot
# ==============================================================================
par(mfrow=c(1,2))
# Hazard Ratios
par(mar = c(4, 4, 3, 2))
plot(timevec, cumhaz1/cumhaz2, type="l", ylim=c(0,1),col=1, ylab="", xlab="")
lines(timevec,cumhaz2/cumhaz3,type="l",ylim=c(0,6), col=2)
legend("topright",c("no edema:edema presence", "edema presence:failed diuretic"),
       col=1:3,lwd=1)
title("Hazard Ratios of Edema Presence (PBC)", ylab="Hazard Ratio", xlab="Time (Days)")
# Hazard Ratios
par(mar = c(4, 4, 3, 2))
plot(timevec, cumhaz1r/cumhaz2r, type="l", ylim=c(0,1),col=1, ylab="", xlab="")
lines(timevec,cumhaz2r/cumhaz3r,type="l",ylim=c(0,6), col=2)
legend("topright",c("no edema:edema presence", "edema presence:failed diuretic"),
       col=1:3,lwd=1)
title("Hazard Ratios of Edema Presence (Trial)", ylab="Hazard Ratio", xlab="Time (Days)")
# ==============================================================================
# Hazard Ratio Plot
# ==============================================================================
par(mfrow=c(1,2))
# complementary log log
plot(NAcurves, col=1:3, lwd=2, fun="cloglog", ylab="Hazard Ratio", xlab="Time (Days)")
legend("bottomright",c("No Edema", "Edema", "Failed"),col=1:3,lwd=1)
title("CLogLog PLots for Edema Groups (PBC)")
# complementary log log
plot(NAcurvesr, col=1:3, lwd=2, fun="cloglog", ylab="Hazard Ratio", xlab="Time (Days)")
legend("bottomright",c("No Edema", "Edema", "Failed"),col=1:3,lwd=1)
title("CLogLog PLots for Edema Groups (Trial)")
# run `fixedData.R` first
# ==============================================================================
# Model Building
# ==============================================================================
# survival object
pbc.surv <- Surv(time=pbc1$time, event=pbc1$status)
# cox model 1 (low albumin, high bilirubin is important)
cox1 <- coxph(pbc.surv ~ strata(sex) + age + edema + log(albumin) + log(bili) +
                protime + stage + platelet, data = pbc1)
# model dropping
drop1(cox1)
# cox final model (no platelet)
coxfinal.pbc <- coxph(pbc.surv ~ age + edema + log(albumin) + log(bili)
                      + log(protime) + platelet, data = pbc1)
drop1(coxfinal.pbc)
```

```r
summary(coxfinal.pbc)
# proportional hazards (all of it looks good)
coxfinal.pbc.zph <- cox.zph(coxfinal.pbc)
# run `timeData.R` first
# ==============================================================================
# Model Building
# ==============================================================================
# tdc surv object
pbc.tdc.surv <- Surv(time=pbcseq.tdc$tstart,time2=pbcseq.tdc$tstop,
                     event=pbcseq.tdc$endpt,type="counting")
# initial model
pbc.tdc.cox <- coxph(pbc.tdc.surv ~ strata(sex) + age + edemar + log(albumin) +
                       log(bili) + protime + stager + platelet, data=pbcseq.tdc)
drop1(pbc.tdc.cox)
summary(pbc.tdc.cox)
# refined model
pbc.tdc.coxr <- coxph(pbc.tdc.surv ~ age + edemar + log(albumin) + bili +
                      protime, data=pbcseq.tdc)
drop1(pbc.tdc.coxr)
summary(pbc.tdc.coxr)
# proportional hazards
coxtdc.zph <- cox.zph(pbc.tdc.coxr)
par(mfrow=c(1,2),cex=1.5)
# ==============================================================================
# goodness-of-fit testing
# ==============================================================================
#fit martingale for full model
cox1.mart <- residuals(coxfinal.pbc,type="martingale")
#find cox-snell residuals: martingales subtracted from event indicator
cox.cs <- pbc1$status - cox1.mart
#cumulative hazard of CS residuals
surv.csr <- survfit(Surv(cox.cs, pbc1$status)~1,type="fleming-harrington")
# plotting cumulative hazard of cs residuals
plot(surv.csr,fun="cumhaz",ylim=c(0,4),xlim=c(0,4))
abline(0,1)
title("Cumulative Hazard of Cox-Snell Residuals (PBC)")
# ==============================================================================
# goodness-of-fit testing (PBCSEQ)
# ==============================================================================
#fit martingale for full model
coxtdc.mart <- residuals(pbc.tdc.coxr,type="martingale")
#find cox-snell residuals: martingales subtracted from event indicator
coxtdc.cs <- pbcseq.tdc$endpt - coxtdc.mart
#cumulative hazard of CS residuals
surv.csr <- survfit(Surv(coxtdc.cs, pbcseq.tdc$endpt)~1,type="fleming-harrington")
# plotting cumulative hazard of cs residuals
plot(surv.csr,fun="cumhaz",ylim=c(0,4))
abline(0,1)
title("Cumulative Hazard of Cox-Snell Residuals (PBCSEQ)")
# run `fixedData_Modeling.R` first
# ==============================================================================
# Leverage Analysis
# ==============================================================================
#fit residuals: martingale, deviance, and df beta
cox.mart <- residuals(coxfinal.pbc,type="martingale")
cox.dev <- residuals( coxfinal.pbc,type="deviance")
cox.dfb <- residuals( coxfinal.pbc,type="dfbeta")
#find linear predictor
cox.preds <- predict(coxfinal.pbc)
n <- length(sort(abs(cox.preds)))
# ============================================================================ #
# plotting ordered important observations
# ============================================================================ #
important <- c(names(sort(abs(cox.mart))[(n-5):n]),
names(sort(abs(cox.dev))[(n-5):n]),
names(sort(abs(cox.dfb[,1]))[(n-5):n]),
names(sort(abs(cox.dfb[,2]))[(n-5):n]),
names(sort(abs(cox.dfb[,3]))[(n-5):n]),
names(sort(abs(cox.dfb[,4]))[(n-5):n]),
names(sort(abs(cox.dfb[,5]))[(n-5):n]),
names(sort(abs(cox.dfb[,6]))[(n-5):n]))
# unusuals
unusuals.cox <- sort(table(important))
unn <- length(unusuals.cox)
pbc1[as.numeric(names(unusuals.cox[(unn-6):unn])),]
# run `timeData_Modeling.R` first
# ==============================================================================
# Leverage Analysis
# ==============================================================================
#fit residuals: martingale, deviance, and df beta
coxr.mart <- residuals(pbc.tdc.coxr,type="martingale")
coxr.dev <- residuals( pbc.tdc.coxr,type="deviance")
coxr.dfb <- residuals( pbc.tdc.coxr,type="dfbeta")
#find linear predictor
coxr.preds <- predict(pbc.tdc.coxr)
n <- length(sort(abs(coxr.preds)))
# ============================================================================ #
# plotting ordered important observations
# ============================================================================ #
k=22
importantr <- c(names(sort(abs(coxr.mart))[(n-k):n]),
names(sort(abs(coxr.dev))[(n-k):n]),
names(sort(abs(coxr.dfb[,1]))[(n-k):n]),
names(sort(abs(coxr.dfb[,2]))[(n-k):n]),
names(sort(abs(coxr.dfb[,3]))[(n-k):n]),
names(sort(abs(coxr.dfb[,4]))[(n-k):n]),
names(sort(abs(coxr.dfb[,5]))[(n-k):n]))
# unusuals
unusuals.coxr <- sort(table(importantr))
unnr <- length(unusuals.coxr)
pbcseq.tdc[as.numeric(names(unusuals.coxr[(unnr-4):unnr])),]
# run `fixedData_Modeling.R` first
# ==============================================================================
# Leverage Analysis
# ==============================================================================
#fit residuals: martingale, deviance, and df beta
```

```
cox.mart <- residuals(coxfinal.pbc,type="martingale")
cox.dev <- residuals( coxfinal.pbc,type="deviance")
cox.dfb <- residuals( coxfinal.pbc,type="dfbeta")
#find linear predictor
cox.preds <- predict(coxfinal.pbc)
n <- length(sort(abs(cox.preds)))
# ============================================================================= #
# plotting ordered important observations
# ============================================================================= #
important <- c(names(sort(abs(cox.mart))[(n-5):n]),
names(sort(abs(cox.dev))[(n-5):n]),
names(sort(abs(cox.dfb[,1]))[(n-5):n]),
names(sort(abs(cox.dfb[,2]))[(n-5):n]),
names(sort(abs(cox.dfb[,3]))[(n-5):n]),
names(sort(abs(cox.dfb[,4]))[(n-5):n]),
names(sort(abs(cox.dfb[,5]))[(n-5):n]),
names(sort(abs(cox.dfb[,6]))[(n-5):n]))
# unusuals
unusuals.cox <- sort(table(important))
unn <- length(unusuals.cox)
pbc1[as.numeric(names(unusuals.cox[(unn-6):unn])),]

# ============================================================================= #
# plot deviance residuals, martingale residuals
# ============================================================================= #
par(mfrow=c(2,1), mar = c(4, 4, 4, 4), cex=1.2)
plot(cox.preds,cox.mart,xlab="Linear Predictor",
     ylab="Martingale Residual", ylim = c(-2,2), pch = 19, cex = 0.5, col=12)
text(cox.preds,cox.mart+0.2, labels = rownames(pbc1))
title("Martingale Residuals vs. Linear Predictor")
plot(cox.preds, cox.dev, xlab="Linear Predictor",ylab="Deviance Residual",
     ylim = c(-3,4), pch = 19, cex = 0.5)
text(cox.preds,cox.dev+0.23, labels = rownames(pbc1))
title("Deviance Residuals vs. Linear Predictor")
# ============================================================================= #
# plotting dfbeta values for time-independent covariates
# ============================================================================= #
par(mfrow=c(3,1), mar = c(4, 4, 4, 4), cex=1.1)
plot(cox.dfb[,1],xlab="Observation Number",ylab="dfbeta for Age",
     ylim=c(-.001,.001), pch = 19, cex = 0.5)
text(cox.dfb[,1]+0.0001, labels = rownames(pbc1))
title("dfbeta Values by Observation Number for Age (Years)")

plot(cox.dfb[,2],xlab="Observation Number",ylab="dfbeta for Edema",
     ylim=c(-.06,.08), pch = 19, cex = 0.5)
text(cox.dfb[,2], labels = rownames(pbc1))
title("dfbeta Values by Observation Number for Edema Status")

plot(cox.dfb[,3],xlab="Observation Number",ylab="dfbeta for `log(albumin)`",
     ylim=c(-.1,.1), pch = 19, cex = 0.5)
text(cox.dfb[,3]+.01, labels = rownames(pbc1))
title("dfbeta Values by Observation Number for log(albumin) Content")
par(mfrow=c(3,1), mar = c(4, 4, 4, 4),cex=1.1)
```

```r
plot(cox.dfb[,4],xlab="Observation Number",ylab="dfbeta for `log(bilirubin)",
     ylim=c(-.02,.02), pch = 19, cex = 0.5)
text(cox.dfb[,4]+.001, labels = rownames(pbc1))
title("dfbeta Values by Observation Number for log(bilirubin) Content")

plot(cox.dfb[,5],xlab="Observation Number",ylab="dfbeta for log(Prothrombin Time)",
     ylim=c(-.02,.02), pch = 19, cex = 0.5)
text(cox.dfb[,5]+.002, labels = rownames(pbc1))
title("dfbeta Values by Observation Number for log(Prothrombin Time)")

plot(cox.dfb[,6],xlab="Observation Number",ylab="dfbeta for Platelet",
     ylim=c(-.4,.35), pch = 19, cex = 0.5)
text(cox.dfb[,6]+.01, labels = rownames(pbc1))
title("dfbeta Values by Observation Number for Platelet Count")
# run `timeData_Modeling.R` first
# ==============================================================================
# Leverage Analysis
# ==============================================================================
#fit residuals: martingale, deviance, and df beta
coxr.mart <- residuals(pbc.tdc.coxr,type="martingale")
coxr.dev <- residuals( pbc.tdc.coxr,type="deviance")
coxr.dfb <- residuals( pbc.tdc.coxr,type="dfbeta")
#find linear predictor
coxr.preds <- predict(pbc.tdc.coxr)
n <- length(sort(abs(coxr.preds)))
# ============================================================================ #
# plotting ordered important observations
# ============================================================================ #
k=22
importantr <- c(names(sort(abs(coxr.mart))[(n-k):n]),
names(sort(abs(coxr.dev))[(n-k):n]),
names(sort(abs(coxr.dfb[,1]))[(n-k):n]),
names(sort(abs(coxr.dfb[,2]))[(n-k):n]),
names(sort(abs(coxr.dfb[,3]))[(n-k):n]),
names(sort(abs(coxr.dfb[,4]))[(n-k):n]),
names(sort(abs(coxr.dfb[,5]))[(n-k):n]))
# unusuals
unusuals.coxr <- sort(table(importantr))
unnr <- length(unusuals.coxr)
pbcseq.tdc[as.numeric(names(unusuals.coxr[(unnr-4):unnr])),]

# ============================================================================ #
# plot deviance residuals, martingale residuals
# ============================================================================ #
par(mfrow=c(2,1), mar = c(4, 4, 4, 4),cex=1.1)
plot(coxr.preds,coxr.mart,xlab="Linear Predictor",
     ylab="Martingale Residual", ylim = c(-2,2), pch = 19, cex = 0.5, col=12)
text(coxr.preds,coxr.mart+0.2, labels = rownames(pbc1))
title("Martingale Residuals vs. Linear Predictor")
plot(coxr.preds, coxr.dev, xlab="Linear Predictor",ylab="Deviance Residual",
     ylim = c(-3,4), pch = 19, cex = 0.5)
text(coxr.preds,coxr.dev+0.23, labels = rownames(pbcseq.tdc))
title("Deviance Residuals vs. Linear Predictor")
```

```
# ========================================================================= #
# plotting dfbeta values for time-independent covariates
# ========================================================================= #
par(mfrow=c(2,1), mar = c(4, 4, 4, 4),cex=1.1)
plot(coxr.dfb[,1],xlab="Observation Number",ylab="dfbeta for Age",
     ylim=c(-.001,.001), pch = 19, cex = 0.5)
text(coxr.dfb[,1]+0.0001, labels = rownames(pbcseq.tdc))
title("dfbeta Values by Observation Number for Age (Years)")

plot(coxr.dfb[,2],xlab="Observation Number",ylab="dfbeta for Edema",
     ylim=c(-.06,.08), pch = 19, cex = 0.5)
text(coxr.dfb[,2], labels = rownames(pbcseq.tdc))
title("dfbeta Values by Observation Number for Edema Status")
par(mfrow=c(3,1), mar = c(4, 4, 4, 4),cex=1.1)
plot(coxr.dfb[,3],xlab="Observation Number",ylab="dfbeta for `log(albumin)`",
     ylim=c(-.1,.1), pch = 19, cex = 0.5)
text(coxr.dfb[,3]+.01, labels = rownames(pbcseq.tdc))
title("dfbeta Values by Observation Number for log(albumin) Content")

plot(coxr.dfb[,4],xlab="Observation Number",ylab="dfbeta for `bilirubin",
     ylim=c(-.02,.02), pch = 19, cex = 0.5)
text(coxr.dfb[,4]+.001, labels = rownames(pbcseq.tdc))
title("dfbeta Values by Observation Number for bilirubin Content")

plot(coxr.dfb[,5],xlab="Observation Number",ylab="dfbeta for Prothrombin Time",
     ylim=c(-.002,.002), pch = 19, cex = 0.5)
text(coxr.dfb[,5]+.0001, labels = rownames(pbcseq.tdc))
title("dfbeta Values by Observation Number for Prothrombin Time")

pbcseq.tdc[pbcseq.tdc$id%in%c(7,15,104,107,88),]
if(1){
  coxfinal.pbc.zph
  par(mfrow=c(6,1), mar = c(4, 4, 4, 4),cex=1.1)
  plot(coxfinal.pbc.zph[1], main = "Schoenfeld Residuals for Age",df=4)
  plot(coxfinal.pbc.zph[2], main = "Schoenfeld Residuals for Edema",df=4)
  plot(coxfinal.pbc.zph[3], main = "Schoenfeld Residuals for `log(albumin)`",df=4)
  plot(coxfinal.pbc.zph[4], main = "Schoenfeld Residuals for `log(bilirubin)`",df=4)
  plot(coxfinal.pbc.zph[5], main = "Schoenfeld Residuals for `log(Prothrombin Time)`",df=4)
  plot(coxfinal.pbc.zph[6], main = "Schoenfeld Residuals for Platelet Count",df=4)
  correlations <- apply(X = coxfinal.pbc.zph$y, MARGIN = 2, FUN = function(x)
    {cor.test(x, coxfinal.pbc.zph$x, method = "pearson",exact=FALSE)})
  correlations
}
if(1){
  coxtdc.zph
  par(mfrow=c(5,1), mar = c(4, 4, 4, 4),cex=1.1)
  plot(coxtdc.zph[1], main = "Schoenfeld Residuals for Age",df=4)
  plot(coxtdc.zph[2], main = "Schoenfeld Residuals for Edema",df=4)
  plot(coxtdc.zph[3], main = "Schoenfeld Residuals for `log(albumin)`",df=4)
  plot(coxtdc.zph[4], main = "Schoenfeld Residuals for `bili`",df=4)
  plot(coxtdc.zph[5], main = "Schoenfeld Residuals for Prothrombin Time",df=4)
  correlations <- apply(X = coxtdc.zph$y, MARGIN = 2, FUN = function(x)
  {cor.test(x, coxtdc.zph$x, method = "spearman",exact=FALSE)})
  correlations
}
```

6.3 References

1. https://cran.r-project.org/web/packages/survival/survival.pdf
2. *Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits.* PA Murtangh, ER Dickson, GM Van Dam, M Malinchoc, PM Grambsch, AL Langworthy, CH Gips. Hepatology Vol. 20, 1994.
3. *Prognosis in Primary Biliary Cirrhosis: Model for Decision Making.* ER Dickson, PM Grambsch, TR Fleming, LD Fisher, A Langworth. Hepatology Vol 10., 1989.
4. *Modeling Survival Data: Extending the Cox Model.* Terry M. Therneau and Patricia M. Grambsch.
5. https://www.nhs.uk/conditions/primary-biliary-cirrhosis-pbc/treatment/
6. https://my.clevelandclinic.org/health/diagnostics/22390-albumin-blood-test
7. https://www.mayoclinic.org/tests-procedures/bilirubin/about/pac-20393041
8. https://www.mayoclinic.org/tests-procedures/prothrombin-time/about/pac-20384661