STAT5140 Final Project

Yunsheng Lu

1 Introduction

This project aims to analyze the data from Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver between 1974 and 1984. There are 418 samples in total, 276 of which have full records. The following is the list of the variables: days between the registration of the experiment and death, status, age, sex, presence of asicites/hepatomegaly/spiders/edema, amount of serum bilirubin/serum cholesterol/albumin/urine copper/alkaline phosphatase/S-GOT/triglicerides/ platelets, prothrombin time, and stage of the disease. In particular, status is coded as 0=censored, 1=censored due to liver transplant, 2=death. Among the samples with full records, the following tables demonstrates the range and the mean.

status type	number	range	mean
0	147	[788, 4556]	2391.782
1	18	[533, 3092]	1511.611
2	111	[41, 4191]	1508.55

Table 1: Basic Information for different status subgroups

The following histogram records more precise information for each status.

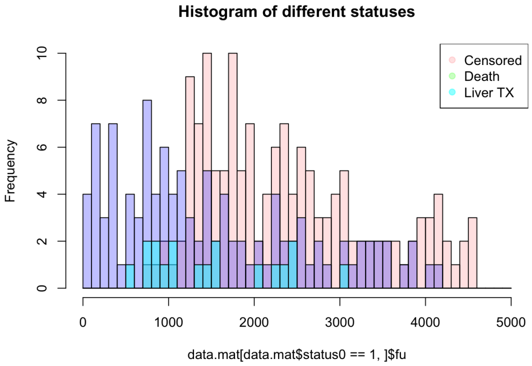


Figure 1: Histogram-status

Notice that status, drug(D-penicillamine), sex, presence of asicites/hepatomegaly/spiders/edemaAmong, stages, are all categorical variables. Converting all variables mentioned above as factors, correlation of samples with full records is calculated:

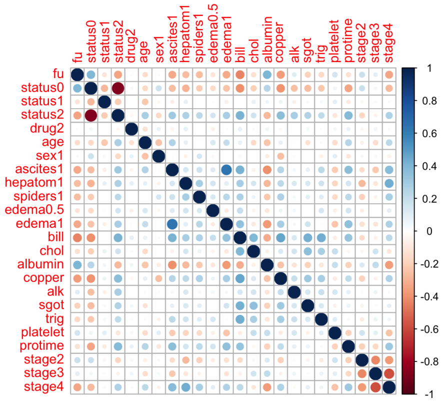


Figure 2: Correlation Plot

There is no strong correlation between different variables except the presence of ascites and second-level edema. Indeed, ascites is abdominal interstitial fluid retention, which can be viewed as serious edema, and there could be a cause-and-effect relationship.

2 Question Proposals

According to the description of the data above, we naturally formulate the following question:

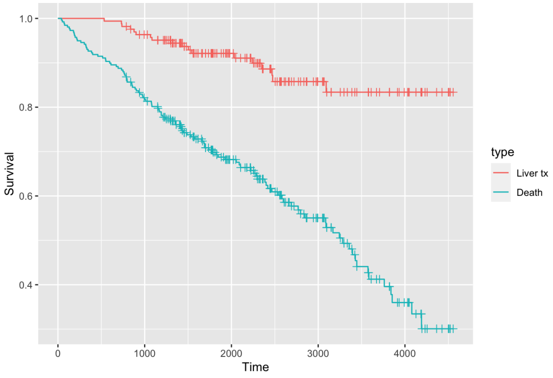
1. How to deal the case of status=1?
2. Does D-penicillamine have an effect on the survival rate of the patients?
3. How do symptoms interact with D-penicillamine and affect the hazard rate?
4. How are other predictors(sex, age, stage) related to the death of the patients?
5. How to formulate a way to see if status=1 is informative censoring or not?

3 Statistical Analysis

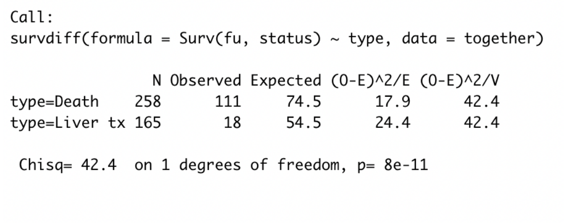
3.1 Treating Status=1

According to the rule of thumbs, for "delta" in survival data, we only could have two possible choice: censored and death, so we have three choices: first, combine status1 with status0; second, combine status1 with status2; third, treat status1 as a covariate.

First, according to table 2, as there is a significant difference between the mean of status=0 and the mean of status=1, we should not consider combine them together due to the disparity in distribution. Now, if we treat status=1 as death, then We further plot estimated survival function according to kaplan-meier estimator.



(a) Survival function for 1 and 2



```
Call:
survdiff(formula = Surv(fu, status) ~ type, data = together)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
type=Death     258      111     74.5      17.9      42.4
type=Liver tx 165       18     54.5      24.4      42.4

 Chisq= 42.4  on 1 degrees of freedom, p= 8e-11
```

(b) Log-rank test table

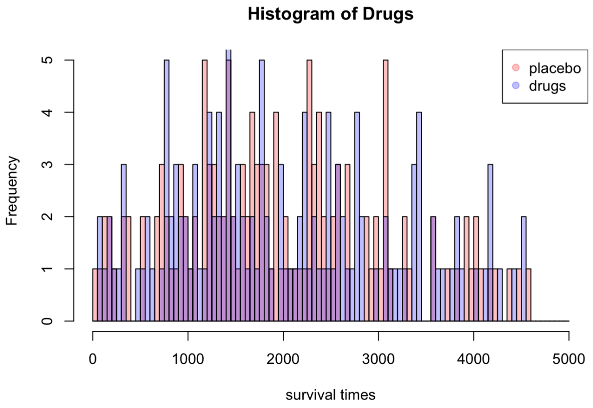
Figure 3: Comparison of 1 and 2

The divergence in estimated survival functions between two subgroups suggests that status=1 should not be considered as equivalent to death. Further supported by log-rank test in *survdiff()*, $p = 8 \times 10^{-11}$.

In the end, treating it as a covariate is also not a good idea as it's impossible to have corresponding death time, unless there is additional data provided. As a consequence, we simply discard samples receiving liver transplants.

3.2 Effectiveness of D-penicillamine

First, we should take a look of the distribution of survival time for both treatment group and control group. In fact, they are not well separate

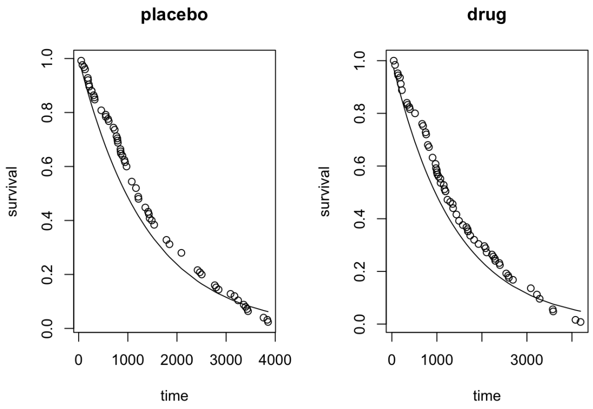


(a) drug vs placebo

To measure the effectiveness of the drug, we now use two naive parametric models: linear regression and exponential distribution. Notice that we take the subgroup with status=2, because for these two naive approaches, we don't have techniques to take censoring samples into consideration. For linear regression, we regress survival versus drugs, and time versus drug, respectively. For exponential model, we assume the hazard rate to be fixed and we regress $-\log \hat{S}$ versus t, where $\hat{S}(t) = \frac{\#\text{ of samples with time}\geq t}{\text{total }\#\text{ of samples}}$ to find the best λ, and then we compare two different lambdas from drug group and placebo group.

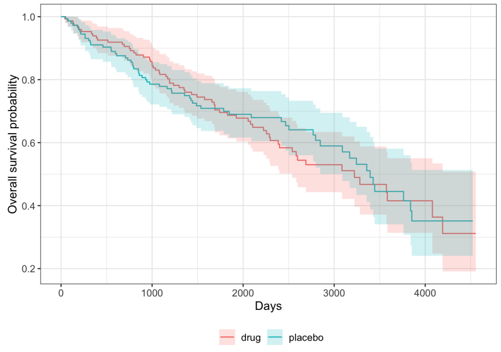
For time linear regression, $\beta = (1429.03, 89.23)$, where 89.23 corresponds to the difference between two groups. Although this might be an evidence of positive effect of the drug, the p-value is unfortunately large: $p = .643$, so we can hardly draw a determined conclusion. For survival rate linear regression, $\beta = (0.53013, -0.04989)$, with drug showing a negative effect, in constrast to time linear regression. However, again, p-value is large: $p = 0.338$, so linear regression demonstrates no significance difference for the two drug groups.

For exponential model, we have the following plots. In fact according to output in the appendix, the estimated $\lambda_d = \lambda_p = 7.204 \times 10^{-7}$, and only differ in residuals and standard errors. Thus, exponential models, again, demonstrate no significance difference for the two drug groups.

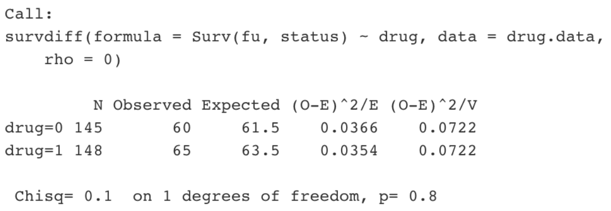


(a) drug vs placebo

Now we again use kaplan-meier estimator to take censoring into consideration. The plot shows no significant difference between the two group, which is also supported by the *survdiff()* output below.



(a) drug vs placebo



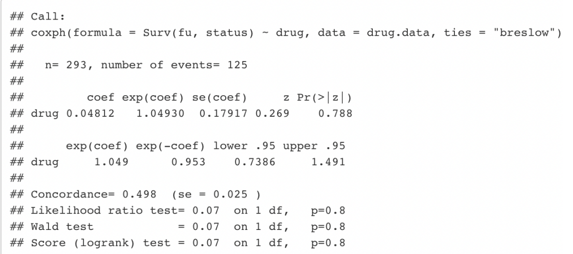
```
Call:
survdiff(formula = Surv(fu, status) ~ drug, data = drug.data,
    rho = 0)

         N Observed Expected (O-E)^2/E (O-E)^2/V
drug=0 145       60     61.5    0.0366    0.0722
drug=1 148       65     63.5    0.0354    0.0722

 Chisq= 0.1  on 1 degrees of freedom, p= 0.8
```

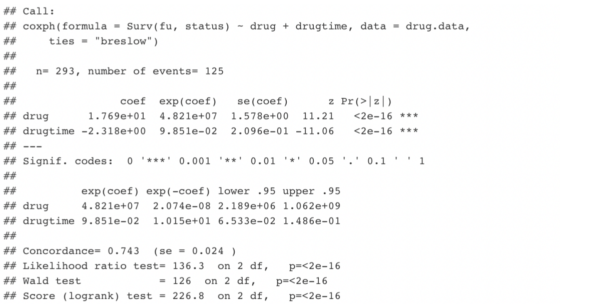
(b) drug vs placebo

The plot above shows that two curves cross each other for several times, so we suspect that the drug might be a time dependent covariate. Here we choose the linear functional, i.e., $g(t) = t$ and we multiply the time to the drug, and drugtime$= I_D * t$. Indeed, when we apply the Cox proportional hazard model, both drug and drugtime are in fact significant:



```
## Call:
## coxph(formula = Surv(fu, status) ~ drug, data = drug.data, ties = "breslow")
##
##   n= 293, number of events= 125
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## drug 0.04812   1.04930  0.17917 0.269    0.788
##
##      exp(coef) exp(-coef) lower .95 upper .95
## drug     1.049      0.953    0.7386     1.491
##
## Concordance= 0.498  (se = 0.025 )
## Likelihood ratio test= 0.07  on 1 df,   p=0.8
## Wald test            = 0.07  on 1 df,   p=0.8
## Score (logrank) test = 0.07  on 1 df,   p=0.8
```

(a) drug



```
## Call:
## coxph(formula = Surv(fu, status) ~ drug + drugtime, data = drug.data,
##     ties = "breslow")
##
##   n= 293, number of events= 125
##
##                coef  exp(coef)   se(coef)      z Pr(>|z|)
## drug      1.769e+01  4.821e+07  1.578e+00  11.21   <2e-16 ***
## drugtime -2.318e+00  9.851e-02  2.096e-01 -11.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## drug     4.821e+07  2.074e-08 2.189e+06 1.062e+09
## drugtime 9.851e-02  1.015e+01 6.533e-02 1.486e-01
##
## Concordance= 0.743  (se = 0.024 )
## Likelihood ratio test= 136.3  on 2 df,   p=<2e-16
## Wald test            = 126  on 2 df,   p=<2e-16
## Score (logrank) test = 226.8  on 2 df,   p=<2e-16
```

(b) drug and drugtime

3.3 Interaction between D-penicillamine and several symptoms

Now we are interested in if drug and drugtime have interactions with several symptoms listed above. It has been confirmed that D-penicillamine is an effective copper antidotes, so we are also interested in the interactions of the drug with copper in patients' urine. The following is an exhaustive table recording all the significant variables ($\alpha = 0.2$) we get. Notice that we treat all presences of the symptoms as categorical variable.

In order to interpret the coefficients above, we discuss different cases appeared above. If a binary covariate Z is time-independent, then the presence of Z (treatment or symtom) will increase/decrease the baseline hazard function by a factor of e^{β}. If a binary covariate Z is time-dependent, then the presense of Z will increase/decrease the baseline hazard by a factor of $e^{\beta Z(t)}$ at time t. The case of numerical variable is similar, except that we are talking about "per unit". The interpretation of interaction is much difficult. We consider a simpler case. Let S be the symptom, and T be the treatment, then we consider the case of linear model:

$$y = aS + bT + cST \implies \frac{\partial y}{\partial S} = a + cT$$

If c is negative, then it means the presence of treatment T will decrease the effect of S, and this is a very ideal situation as the treatment proves to be effective. Now in Cox regression model, if the coefficient of the interaction term between the drug and the sympotom is negative, then the treatment will reduce the effect of the symptom on the baseline hazard function by a factor of $e^{\beta T}$.

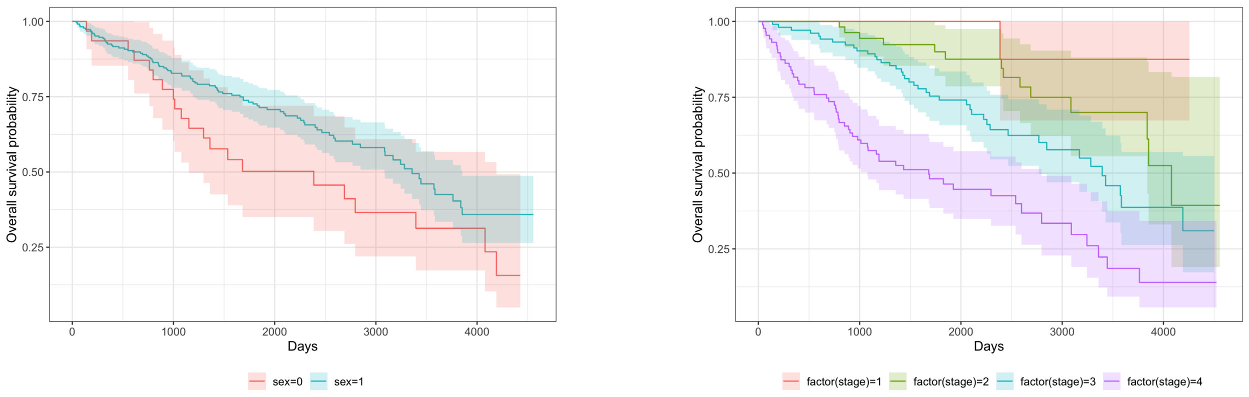
Although some of the interaction above are negative, they only reduce the baseline factor by a moderate amount. By the same logic, we only care about the symptom and the drug effect that significantly affect the hazard function. According to the principles above, we have the following conclusions: 1. the presence of the drug in all cases significantly increase the baseline hazard function. 2. the presence of ascites1(hr=13.5), edema2(7.64) appear to be dangerous, while the effect spider1(2.8) is mild. The drug appears to be effective in the following cases: ascites1(hr=0.3375), hepatom1(0.3651). Finally, it's worth noting that the effectiveness of the drug is significant in reducing the effect of copper's negative effects in the sense of p-value.

variable	est coefficient	p-value
dr	3.61	2.52e-14
ascites1	2.60	5.67e-11
drtime	-0.0015	1.43e-12
dr:asc1	-1.086	0.168
asc1:drtime	-0.001	0.137
dr	4.46	4.76e-12
hepatom1	1.43	2.14e-06
drtime	-0.0017	8.28e-09
dr:hep1	-1.0075	0.146
dr	3.60	4.76e-12
spider1	1.06	2.14e-06
drtime	-0.0014	1.40e-09
spi1:drtime	-0.007	0.07
dr	3.15	3.75e-09
edema.5	1.63	0.0008
edema1	2.03	1.06e-06
drtime	-0.001	1.91e-08
ede1:drtime	-0.0015	0.19
dr	4.247e+00	5.40e-13
copper	8.939e-03	3.86e-12
drtime	-5.672e-03	4.13e-09
dr:cop	-1.197e-07	0.0184

Table 2: Basic Information for different status subgroups

3.4 Effects of Sex, Age, and Stage

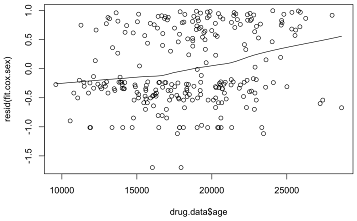
Now we are interested in the relation between sex, age, stage and survival rate. First, we plot the kaplan-meier estimators according to sex and stage, respectively.



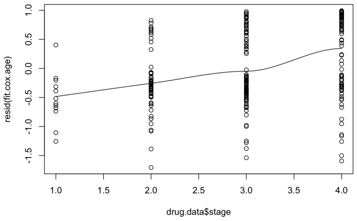
(a) sex

(b) stage

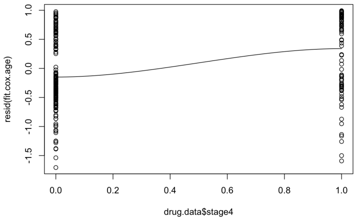
As there is obvious disparity between the male and female's curve, we choose to stratify the data according to sex first. We then plot age against the martingale residual of the first model, and the martingale residual plot suggests that we should enter a linear term of the age. Next we plot stage against the martingale residual of the previous model, and again it suggests to enter stage as a linear term. The following shows the plots:



(a) age

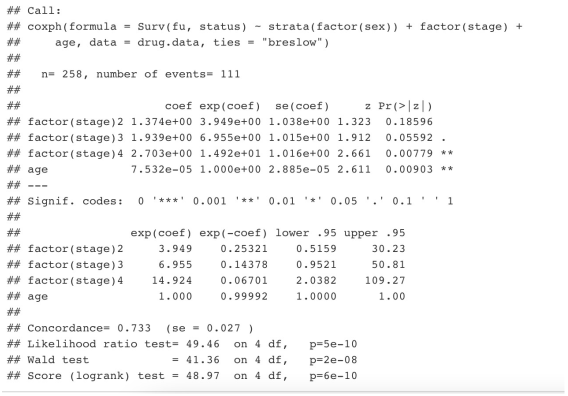


(b) stage

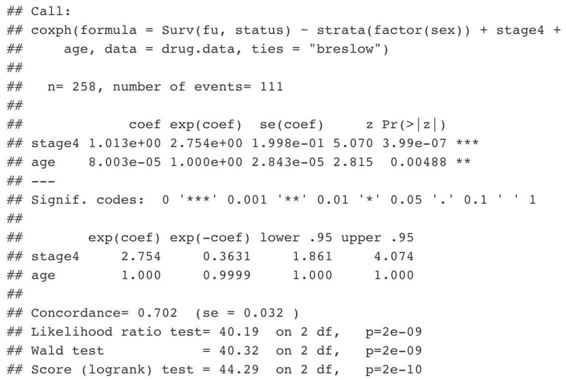


(c) stage4

However, the cox proportional hazard model suggest that only stage=4 is significant, so we now consider a new functional I(stage=4), and indeed there is a threshold between stage=4 and other stages, according to martingale residual plot.



```
## Call:
## coxph(formula = Surv(fu, status) ~ strata(factor(sex)) + factor(stage) +
##     age, data = drug.data, ties = "breslow")
##
##   n= 258, number of events= 111
##
##                     coef exp(coef)  se(coef)     z Pr(>|z|)
## factor(stage)2 1.374e+00 3.949e+00 1.038e+00 1.323  0.18596
## factor(stage)3 1.939e+00 6.955e+00 1.015e+00 1.912  0.05592 .
## factor(stage)4 2.703e+00 1.492e+01 1.016e+00 2.661  0.00779 **
## age            7.532e-05 1.000e+00 2.885e-05 2.611  0.00903 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## factor(stage)2     3.949    0.25321    0.5159     30.23
## factor(stage)3     6.955    0.14378    0.9521     50.81
## factor(stage)4    14.924    0.06701    2.0382    109.27
## age                1.000    0.99992    1.0000      1.00
##
## Concordance= 0.733  (se = 0.027 )
## Likelihood ratio test= 49.46  on 4 df,   p=5e-10
## Wald test            = 41.36  on 4 df,   p=2e-08
## Score (logrank) test = 48.97  on 4 df,   p=6e-10
```

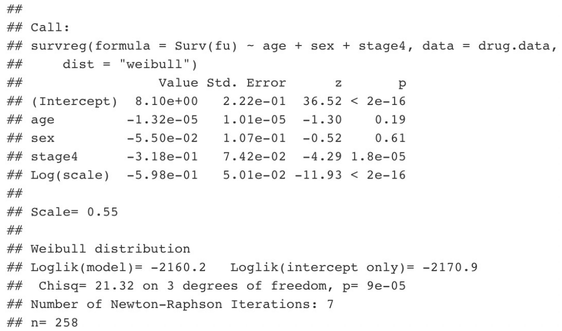


```
## Call:
## coxph(formula = Surv(fu, status) ~ strata(factor(sex)) + stage4 +
##     age, data = drug.data, ties = "breslow")
##
##   n= 258, number of events= 111
##
##             coef exp(coef)  se(coef)     z Pr(>|z|)
## stage4 1.013e+00 2.754e+00 1.998e-01 5.070 3.99e-07 ***
## age    8.003e-05 1.000e+00 2.843e-05 2.815  0.00488 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## stage4     2.754     0.3631     1.861     4.074
## age        1.000     0.9999     1.000     1.000
##
## Concordance= 0.702  (se = 0.032 )
## Likelihood ratio test= 40.19  on 2 df,   p=2e-09
## Wald test            = 40.32  on 2 df,   p=2e-09
## Score (logrank) test = 44.29  on 2 df,   p=2e-10
```

(a) Cox Regression with four stages

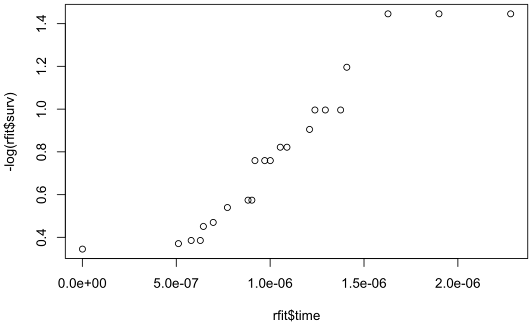
(b) Cox Regression with stage4

Finally, we see if Weibull's distribution is appropriate in our analysis. We fit time against sex, age and stage4. We get $\lambda = 4.03 \times 10^{-7}$ and $\alpha = 1.82$. Unfortunately, the Cox-Snell residual plot suggests the model is problematic.



```
##
## Call:
## survreg(formula = Surv(fu) ~ age + sex + stage4, data = drug.data,
##     dist = "weibull")
##                 Value Std. Error      z       p
## (Intercept)  8.10e+00   2.22e-01  36.52 < 2e-16
## age         -1.32e-05   1.01e-05  -1.30    0.19
## sex         -5.50e-02   1.07e-01  -0.52    0.61
## stage4      -3.18e-01   7.42e-02  -4.29 1.8e-05
## Log(scale)  -5.98e-01   5.01e-02 -11.93 < 2e-16
##
## Scale= 0.55
##
## Weibull distribution
## Loglik(model)= -2160.2   Loglik(intercept only)= -2170.9
##  Chisq= 21.32 on 3 degrees of freedom, p= 9e-05
## Number of Newton-Raphson Iterations: 7
## n= 258
```

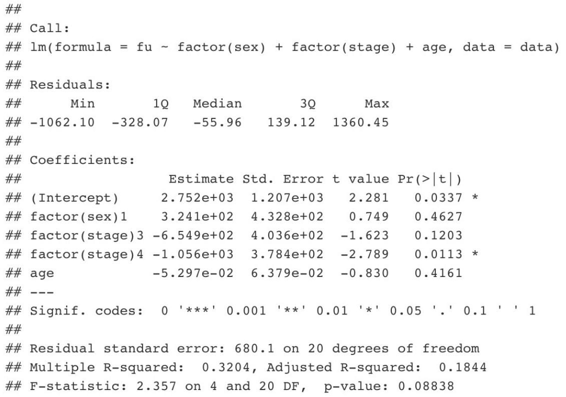
(a) Weibull distribution



(b) Cox-Snell residual

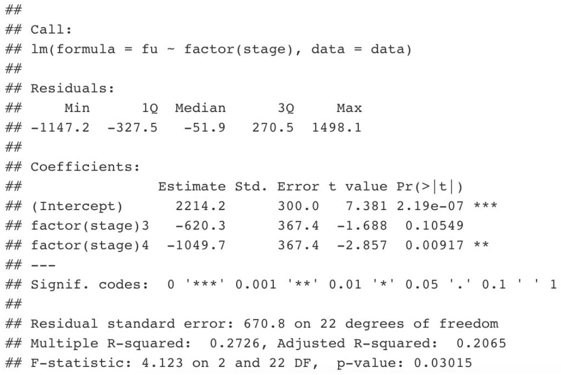
3.5 Informativeness of Status=1

As we have seen in the previous section, all three predictors are strongly correlated with the death of the patients. If they are correlated with the status=1, then we should conclude that it's informative. We simply perform a linear regression, and the output is as the following:



```
##
## Call:
## lm(formula = fu ~ factor(sex) + factor(stage) + age, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1062.10  -328.07   -55.96   139.12  1360.45
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.752e+03  1.207e+03   2.281   0.0337 *
## factor(sex)1    3.241e+02  4.328e+02   0.749   0.4627
## factor(stage)3 -6.549e+02  4.036e+02  -1.623   0.1203
## factor(stage)4 -1.056e+03  3.784e+02  -2.789   0.0113 *
## age            -5.297e-02  6.379e-02  -0.830   0.4161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 680.1 on 20 degrees of freedom
## Multiple R-squared:  0.3204, Adjusted R-squared:  0.1844
## F-statistic: 2.357 on 4 and 20 DF,  p-value: 0.08838
```

(a) full model



```
##
## Call:
## lm(formula = fu ~ factor(stage), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1147.2  -327.5   -51.9   270.5  1498.1
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2214.2      300.0   7.381 2.19e-07 ***
## factor(stage)3   -620.3      367.4  -1.688  0.10549
## factor(stage)4  -1049.7      367.4  -2.857  0.00917 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 670.8 on 22 degrees of freedom
## Multiple R-squared:  0.2726, Adjusted R-squared:  0.2065
## F-statistic: 4.123 on 2 and 22 DF,  p-value: 0.03015
```

(b) better sub-model

The summary suggests that stage=4 is very significant, together with age, which is consistent with the output in the previous section. From another perspective, there are **0.4** of samples receiving liver transplant have stage=4. It's reasonable to conclude that only when the disease develop to a critical condition (stage=4) will the patients consider a liver transplant. Thus, we should conclude that status=1 is informative.

4 Conclusion

1. First of all, we should discard the samples with liver transplants because according to our analysis, it fits neither into the first type of censoring nor into death, and it's impossible to be treated as a covariate because information about survival time is thus lost.

2. According to two parametric models, linear and exponential models, together with time-independent Cox regression semiparametric model, D-penicillamine looks ineffective to PBC. However, once we consider the drug to be time-dependent, the corresponding Cox regression model concludes that the drug is significant.

3. the presence of ascites1(hr=13.5), edema2(7.64) appear to be dangerous, while the effect of spider1(2.8) is mild. The drug appears to be effective in the following cases: ascites1(hr=0.3375), hepatom1(0.3651). being a copper antidote, D-penicillamine is confirmed to be effective in reducing the effect of copper's negative effects (p=0.0184).

4. Stages, ages, and sex are related to the hazard rate of the patients. According to kaplan-meier estimators, in particular, male has higher hazard rate than the female, and there is a natural ordering in stages, with stage 1 being the mildest and stage 4 being the most dangerous. Also, Weibull's distribution is improper if we only consider age, stage4 and sex.

5. The cox regression model confirms that age and stage(especially stage 4) are strongly correlated with liver transplants. Since they both have strong correlation with hazard rate of the patients, liver transplants is related to the death of the patients, so it's informative. Meanwhile, as there are 40 percent of patients who receive liver transplant has disease stage=4, it's likely to conclude that only when patients' disease develop to a serious condition will they consider a liver transplant.

5 Discussion

First of all, the effects of D-penicillamine is still not totally clear, because the presence of the drug, in time-dependent Cox regression, is negatively related to the survival rate, while it's counterpart, the time-dependent covariate, is positively related to the survival rate. Further research and other models beyond Cox regression might be needed. Second, when investigating the interactions between the drug and the sympotoms appeared, there is a tradeoff between the plausibility and level of influence. Namely, some of the covariates with extreme hazard ration has p-value around 0.1 to 0.2. It's hard to conclude that they are indeed significant, but directly discarding them might also be improper. Finally, our exploration of relation between sex and survival rate is limited. We choose to stratify the data according to sex due to the shape of their kaplan-meier's estimators. However, it's also possible that entering sex as a categorical variable would have good performance in model fitting. Meanwhile, as in the 4th section, our analysis is restricted to sex, stage and age, it's not clear if Weibull distribution would still remain improper if we enter other variables.