EXAM

1) The starter file has code for converting a couple of variables into factors, from their original type. What is the purpose behind these conversions? Base you answer in the contextual information related to the dataset. This information can be pulled up by using the help function on pbc.

Answer: The main purpose behind the conversions of the variables into factors is that the data represented by these variables are or seem like the categorical variables, that means they have groups within the variables. The snippet of summary after converting the variables is below:

```
      time         status  trt          age          sex     ascites hepato  spiders
 Min.   :  41   0:147   1:136   Min.   :26.28   m: 34   0:257   0:134   0:196
 1st Qu.:1186   1: 18   2:140   1st Qu.:41.51   f:242   1: 19   1:142   1: 80
 Median :1788   2:111           Median :49.71
 Mean   :1979                   Mean   :49.80
 3rd Qu.:2690                   3rd Qu.:56.58
 Max.   :4556                   Max.   :78.44
 edema           bili              chol           albumin          copper
 0  :234   Min.   : 0.300   Min.   : 120.0   Min.   :1.960   Min.   :  4.00
 0.5: 25   1st Qu.: 0.800   1st Qu.: 249.5   1st Qu.:3.310   1st Qu.: 42.75
 1  : 17   Median : 1.400   Median : 310.0   Median :3.545   Median : 74.00
           Mean   : 3.334   Mean   : 371.3   Mean   :3.517   Mean   :100.77
           3rd Qu.: 3.525   3rd Qu.: 401.0   3rd Qu.:3.772   3rd Qu.:129.25
           Max.   :28.000   Max.   :1775.0   Max.   :4.400   Max.   :588.00
    alk.phos            ast              trig          platelet        protime
 Min.   :  289.0   Min.   : 28.38   Min.   : 33.0   Min.   : 62.0   Min.   : 9.00
 1st Qu.:  922.5   1st Qu.: 82.46   1st Qu.: 85.0   1st Qu.:200.0   1st Qu.:10.00
 Median : 1277.5   Median :116.62   Median :108.0   Median :257.0   Median :10.60
 Mean   : 1996.6   Mean   :124.12   Mean   :125.0   Mean   :261.8   Mean   :10.74
 3rd Qu.: 2068.2   3rd Qu.:153.45   3rd Qu.:151.2   3rd Qu.:318.2   3rd Qu.:11.20
 Max.   :13862.4   Max.   :457.25   Max.   :598.0   Max.   :563.0   Max.   :17.10
 stage
 1: 12
 2: 59
 3:111
 4: 94
```

- The variables status, stage and ascites are converted to the factors at the beginning, and the variables trt, hepato, spiders, edema are converted into factors as well. The variable ascites consists of the responses 0 and 1for the presence of ascites or absences of ascites, the variable edema consists of responses 0, 0.5, and 1 for no edema, untreated edema, and treated edema respectively, the variable hepato consists of responses 0 and 1 for the presence of hepatomegaly or enlarged liver and absence respectively. The variable spiders consist of responses 0 and 1 for the presence or absence of blood vessel malformations in the skin, the variable stage consists of 4 responses of histologic stage of disease (needs biopsy) 1, 2 3 4. The variable status consists of 3 responses 0, 1 and 2 for censored, transplant, dead respectively, the variable trt consists of 2 responses 1/2/NA for D-penicillamine, placebo, and not randomized respectively. Due to the categorical nature of these variables, these are converted into factors.

2) Execute the code associated with excluding *trt*, *time*, and *status* from the data frame to produce a reduced version of the dataset (represented by the data frame object labeled *my. Pbc* in the R code file accompanying these instructions). Then, fit a model to predict *stage* using the remaining variables as predictors in the reduced version of the dataset (represented by the data frame object with the label *my.reduced.pbc*). Explain the purpose behind excluding the above listed three variables as predictors.

Answer:

- These variables are excluded from the dataset, as we are trying to predict histologic stage of disease (needs biopsy) of the and we do not need the drug given to the patient (trt), status of the patient at the

end of the clinical trial (status), and number of days between registration and the earlier of death (time) as predictors.

- Also, the drug that should be given to patient should be or might be decided after knowing the stage of the patient, the status and the time can be known after the end of the clinical trial. Hence, these variables are not of importance while predicting the significant variables for histological stage.

The variables producing the data redundancy problem are removed after identifying them through ggpairs and building the ordinal regression model on all the predictors and writing a step function to see the best model. The summary of the best ordinal regression model is attached below:

```
Call:
polr(formula = stage ~ ., data = My.Reduced.Pbc, Hess = TRUE)

Coefficients:
              Value Std. Error    t value
age       0.0179484  1.224e-02  1.466e+00
ascites1 15.6586593  3.156e-07  4.961e+07
hepato1   1.7911684  2.802e-01  6.393e+00
spiders1  0.8530970  2.890e-01  2.952e+00
chol     -0.0007905  5.214e-04 -1.516e+00
copper    0.0041810  1.663e-03  2.514e+00

Intercepts:
    Value        Std. Error   t value
1|2 -1.44810e+00  7.11800e-01 -2.03450e+00
2|3  8.20000e-01  6.74300e-01  1.21600e+00
3|4  3.22930e+00  7.06400e-01  4.57160e+00

Residual Deviance: 534.5157
AIC: 552.5157
```

3) Explain the rationale behind the specific modeling approach you have used, on the basis of the measurement type of the outcome variable.

Answer: The specific modeling approach is used to predict histological stage of the disease is ordinal regression, as the outcome variable stage has increasing ranking order between its categories 1,2,3 and 4 i.e., stage1 < stage2 < stage3 < stage4.

- The AIC of the model is 552.5157, as the value is small the model is a better model.

	Value	Std. Error	t value	p value
age	0.0179484338	1.224225e-02	1.466106e+00	1.426195e-01
ascites1	15.6586593257	3.156237e-07	4.961180e+07	0.000000e+00
hepato1	1.7911684279	2.801761e-01	6.393009e+00	1.626530e-10
spiders1	0.8530969837	2.889570e-01	2.952332e+00	3.153835e-03
chol	-0.0007905438	5.213647e-04	-1.516297e+00	1.294443e-01
copper	0.0041810064	1.663312e-03	2.513663e+00	1.194846e-02
1|2	-1.4480907996	7.117617e-01	-2.034516e+00	4.189955e-02
2|3	0.8199533632	6.743066e-01	1.215995e+00	2.239869e-01
3|4	3.2292981596	7.063865e-01	4.571574e+00	4.840739e-06

	OR	2.5 %	97.5 %
age	1.018110e+00	9.939723e-01	1.042835e+00
ascites1	6.316396e+06	6.316392e+06	6.316400e+06
hepato1	5.996455e+00	3.462647e+00	1.038439e+01
spiders1	2.346904e+00	1.332093e+00	4.134814e+00
chol	9.992098e-01	9.981892e-01	1.000231e+00
copper	1.004190e+00	1.000921e+00	1.007469e+00

- For age, when a patient's age moves 1 unit, the odds of moving from "stage 1" moving to "stage 2" or "stage 3" or "stage 4" are 1.8% higher odds.

- Patients having ascites (1) in them have 6.316396e+06 times greater odds of changing from “stage1” to “stage 2” or “stage 3” or “stage 4”.
- Patients having enlarged liver (hepato1) in them have are 5.996455e+00 times greater odds of changing from “stage1” to “stage 2” or “stage 3” or “stage 4”.
- Patients having blood malformations in skin (spider1) in them have 2.346904 times greater odds of changing from “stage1”to “stage 2” or “stage 3” or “stage 4”.
- For Chol, when a patient’s serum cholesterol (mg/dl) moves 1 unit, the odds of moving from “stage 1” moving to “stage 2” or “stage 3” or “stage 4” are 9.992098e-01 times greater.
- For copper, when a patient’s urine copper (ug/day) moves 1 unit, the odds of moving from “stage 1” moving to “stage 2” or “stage 3” or “stage 4” are 1.004190 times greater.
- The probability of t-statistic reveals that all the variables used in the model as predictors except ascites are statistically significant, as the probabilities are less than 0.05.
- The p(t) of the variables is age [1.426195e-01], hepato1[1.626530e-10], spiders1 [3.153835e-03], Chol [1.294443e-01], copper [1.194846e-02].

4) State the key assumptions underlying your choice of the modeling technique. Then, perform the necessary diagnostic tests to determine the extent to which these assumptions were met.
Summarize your results in a table with the following structure:
Answer:

Assumption	*Met?*	*Supporting Evidence*

The Dependent variable is ordered.	This assumption is met.	The histological stage of the disease (dependent variable) is ordered, i.e., [stage1<stage2<stage3<stage4].
		`  1  2  3  4`
		`12 59 111 94`

Independent variables are categorical or continuous.	This assumption is met.	`time status trt age sex ascites hepato spiders`
		`"integer" "factor" "factor" "numeric" "factor" "factor" "factor" "factor"`
		`edema bili chol albumin copper alk.phos ast trig`
		`"factor" "numeric" "integer" "numeric" "integer" "numeric" "numeric" "integer"`
		`platelet protime stage`
		`"integer" "numeric" "factor"`

No Multi-collinearity	This assumption is almost met, as the Variation inflation factor (vif) value is greater than 2 for only 2 variables that is age and ascites. However, those are not greater than 10, making them not problematic.	`age ascites hepato spiders chol copper` `4.736844 5.589378 1.070908 1.042493 1.489744 1.191856`

Proportional odds	This assumption has almost met, as the p - value > 0.05 for omnibus variable and all the other variable except the Chol variable.	``` -- Test for X2 df probability -- Omnibus 12.8 12 0.38 age 2.03 2 0.36 ascites1 0 2 1 hepato1 2.24 2 0.33 spiders1 0.26 2 0.88 chol 6.24 2 0.04 copper 1.93 2 0.38 -- H0: Parallel Regression Assumption holds ```
Goodness of fit	This assumption is met, as the p value is greater than 0.05. Therefore, the model has a good fit.	``` Lipsitz goodness of fit test for ordinal response models data: formula: stage ~ age + ascites + hepato + spiders + chol + copper LR statistic = 15.359, df = 9, p-value = 0.08154 ```

5) Next, go back to using *my.pbc* and add a new variable to represent two states of status: the original *status* values of 0 and 1 are mapped to a 0 and the original value of 2 onto a 1. In the R code, this new variable is labeled *newstatus*. What purpose would be served by combining the original 0 (censored) and 1 (transplant) into a single category, and the original 2 (dead) into another single category? (Hint: consider what you would need for status if your goal is to model survival/non-survival). What is lost when combining the original 0 and 1 categories into a single new category? How will this affect the interpretation of the results.
Answer: The code for adding a new variable is attached below:

```
## creating a new variable, combining censored and transplant into one factor, and dead as
another

my.pbc$newstatus <- as.factor(ifelse((my.pbc$status == 0),0,
                                     ifelse((my.pbc$status == 1),0,
                                            1)))

summary(my.pbc$newstatus)
```

```
  0   1
165 111
```

- The purpose of creating a new variable by combining both the censored (0) and the transplant group (1) into one category, and dead group (2) into other category is to predict the variables that are significantly affecting the survival status of the patient(dead/alive). The logic behind combining censored and transplant categories is that the patients belonging to both the groups are alive i.e., the patients in the censored group may have just started the participation in the clinical trial and the patients in the transplant group are the ones who have undergone transplantation procedure.

- On combining the original 0 and 1 categories into a single new category, the exact status of the patient and what operation procedure they have went through is lost. This might also lead to the confusion on the drug and the stage of the patients as they are correlated with each other.
- By combining the groups, we can only interpret whether the variable is contributing to the death of the patient or not, and by doing this we are leaving behind the contribution of variables to transplantation leading to the half analysis of the patients participating in the clinical trial.

6) Next, fit a CoxPH model on all of the predictors. Would it make sense to retain *status* in the set of predictors? Why/why not? Identify the variables that are indicated as being significant predictors of the outcome. What is the outcome that is being modeled?

Answer: The predictors that might produce a data redundancy problem in the model using ggpairs and stepwise CoxPH regression method. So, the model is built using the age, edema, bili, albumin, copper, ast, protime, stage as the predictors. And the outcome variable that is being modelled is newstatus, created after combining both the censored (0) and the transplant group (1) into one category, and dead group (2) into another category.
The code snippet of summary of the CoxPH model is attached below:

```
Call:
coxph(formula = Surv(time, as.numeric(newstatus)) ~ age + edema +
    bili + albumin + copper + ast + protime + stage, data = my.pbc)

  n= 276, number of events= 111

               coef  exp(coef)   se(coef)      z Pr(>|z|)
age       0.0313188  1.0318144  0.0102909  3.043  0.00234 **
edema0.5  0.1598036  1.1732804  0.3054890  0.523  0.60090
edema1    0.9121653  2.4897075  0.3545431  2.573  0.01009 *
bili      0.0869270  1.0908171  0.0198104  4.388 1.14e-05 ***
albumin  -0.7387129  0.4777284  0.2792471 -2.645  0.00816 **
copper    0.0027867  1.0027905  0.0009912  2.811  0.00493 **
ast       0.0039562  1.0039641  0.0018415  2.148  0.03168 *
protime   0.2642049  1.3023951  0.1122859  2.353  0.01862 *
stage2    1.3596258  3.8947355  1.0808743  1.258  0.20843
stage3    1.6823556  5.3782101  1.0478753  1.605  0.10839
stage4    2.0627073  7.8672396  1.0432133  1.977  0.04801 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

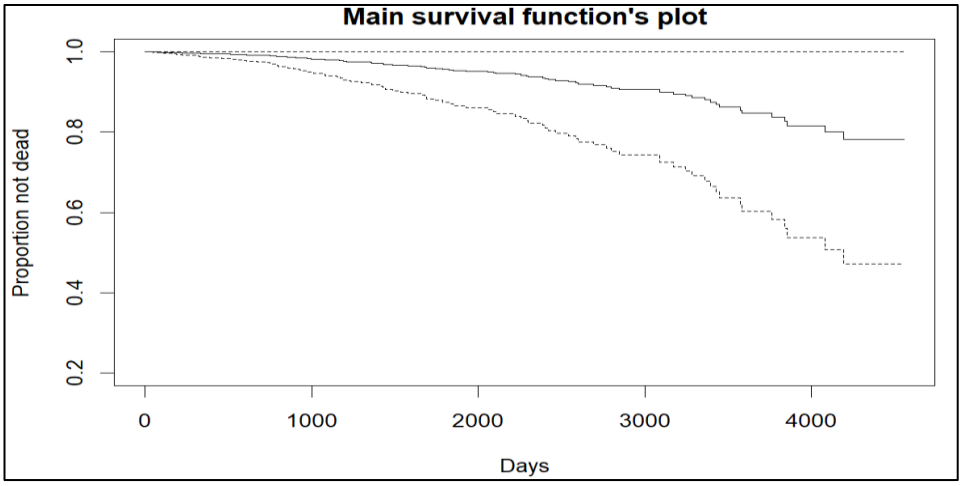
```
         exp(coef) exp(-coef) lower .95 upper .95
age         1.0318     0.9692    1.0112    1.0528
edema0.5    1.1733     0.8523    0.6447    2.1352
edema1      2.4897     0.4017    1.2427    4.9881
bili        1.0908     0.9167    1.0493    1.1340
albumin     0.4777     2.0932    0.2764    0.8258
copper      1.0028     0.9972    1.0008    1.0047
ast         1.0040     0.9961    1.0003    1.0076
protime     1.3024     0.7678    1.0451    1.6230
stage2      3.8947     0.2568    0.4682   32.3981
stage3      5.3782     0.1859    0.6897   41.9364
stage4      7.8672     0.1271    1.0182   60.7865

Concordance= 0.845  (se = 0.019 )
Likelihood ratio test= 165.7  on 11 df,   p=<2e-16
Wald test            = 176.9  on 11 df,   p=<2e-16
Score (logrank) test = 278.8  on 11 df,   p=<2e-16
```

- It would not make sense to retain the status in the predictors, as we are looking at the hazards of the death of the patients i.e., we are looking at the significance of the variables leading to the hazards of the death of the patient.
- The significant variables in the above CoxPH model with the probability of the z-statistic less than 0.05 are age [0.00234], edema1 [0.01009], bili [1.14e-05], albumin [0.00816], copper [0.00493], ast [0.03168], protime [0.01862], stage4 [0.04801].
- On observing the coefficient of age [0.0313188], it is certain that there is a chance of death in older individuals. On observing the coefficient of the edema1 [0.9121653], it is seen that the death is more in the patients suffering from edema despite diuretic therapy than people with no edema. On observing the coefficient of the bili [0.0869270], it is seen that with there is an increased probability of death in the individuals with increased levels of serum bilirubin.
- On observing the coefficient of albumin [-0.7387129], it is seen that with there is a decreased probability of death in the individuals with increased levels of serum albumin. On observing the coefficient of the

copper [0.0027867], it is seen that with there is an increased probability of death in the individuals with increased levels of urine copper. On observing the coefficient of the ast [0.0039562], it is seen that with there is an increased probability of death in the individuals with increased levels of aspartate aminotransferase.

- On observing the coefficient of the protime [0.2642049], it is seen that with there is an increased probability of death in the individuals with increased levels of standardized blood clotting time. On observing the coefficient of the stage4 [2.0627073], it is seen that the death is more in the patients suffering from stage4 of the disease than the stage1 patients.
- As Likelihood ratio test, Wald test, and Score (log rank) test have the probability of <2e-16, <2e-16, <2e-16 respectively, it can be said that the model is better model, and there is a need to fit another model.
- After 4000 weeks, only 0.8 proportion of patients have not undergone death. The plot showing this is attached below:



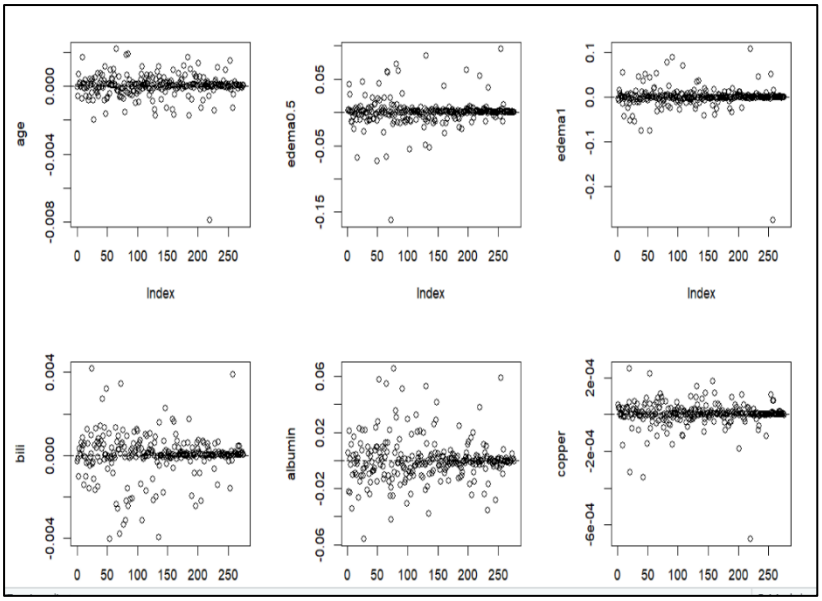
7) Identify the key assumption(s) associated with the CoxPH modeling approach, and explain whether it/they are satisfied, using appropriate evidence. Summarize your results in a table with the following structure.

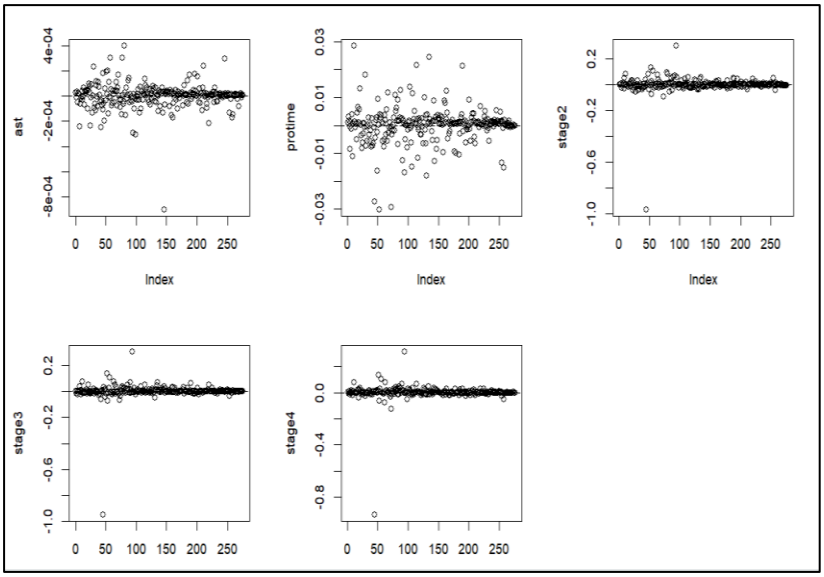
Answer:

Assumption	*Met?*	*Supporting Evidence*
Proportional Hazards	The probability of all the variables except bili, protime, stage have probabilities less than 0.05. So, this assumption is partially met.	

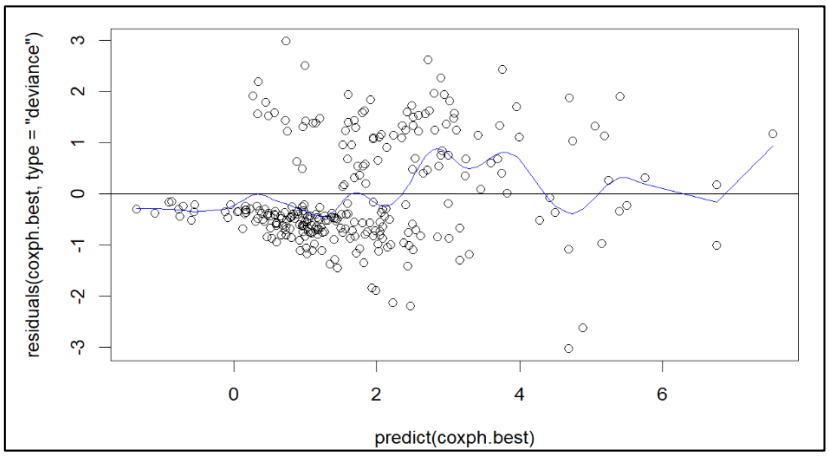
	chisq	df	p
age	2.106	1	0.1467
edema	4.435	2	0.1089
bili	9.429	1	0.0021
albumin	0.206	1	0.6495
copper	0.279	1	0.5975
ast	1.701	1	0.1922
protime	4.011	1	0.0452
stage	3.885	3	0.2742
GLOBAL	21.024	11	0.0331

Influential Observations	Since there is only a small difference between df-betas of all the variables and actual betas, this assumption maybe said to have met.





Non-linearity	This assumption is met, as all the observations are not closer to the linear line.



8) Next, fit full-parametric models, where the survival function of status is modeled using exponential and Weibull distributions to model the outcome (survival function of *newstatus*, whereas the CoxPH approach models the hazard function of *newstatus*). Which of these is a better model? Explain using appropriate evidence.
Answer: The full-parametric models, namely Weibull distributed model and the exponential models are fit to model the outcome of newstatus (alive/dead). Both the models are good fit models as the p-values were less than 0.05, Weibull distributed model has a p-value of 9.5e-31, and that of the exponential model is 3.9e-24.
The comparison of both the models with respect to AIC is attached below:

	df <dbl>	AIC <dbl>
mod.exponential	12	1995.143
mod.weibull	13	1961.954

2 rows

Between both the models, model based on Weibull distribution [1961.954] is better as the AIC value is less than the AIC of exponential model [1995.143].

The comparison of both the parametric survival models and the hazard function CoxPH model with respect to AIC is attached below.

	df <dbl>	AIC <dbl>
mod.exponential	12	1995.1428
mod.weibull	13	1961.9538
coxph.best	11	956.6894

3 rows

Between all the three models, the CoxPH model [956.6894] is the best one.

9) Taking into account the results from the three survival-related models, identify the significant predictors of the outcome. Among these, which variable is the most important one? Explain on the basis of appropriate evidence.
Answer:

- The significant variables in the above CoxPH model with the probability of the z-statistic less than 0.05 are age [0.00234], edema1 [0.01009], bili [1.14e-05], albumin [0.00816], copper [0.00493], ast [0.03168], protime [0.01862], stage4 [0.04801]. The model summary is attached below:

```
Call:
coxph(formula = Surv(time, as.numeric(newstatus)) ~ age + edema +
    bili + albumin + copper + ast + protime + stage, data = my.pbc)

  n= 276, number of events= 111

              coef  exp(coef)   se(coef)      z Pr(>|z|)
age       0.0313188  1.0318144  0.0102909  3.043  0.00234 **
edema0.5  0.1598036  1.1732804  0.3054890  0.523  0.60090
edema1    0.9121653  2.4897075  0.3545431  2.573  0.01009 *
bili      0.0869270  1.0908171  0.0198104  4.388 1.14e-05 ***
albumin  -0.7387129  0.4777284  0.2792471 -2.645  0.00816 **
copper    0.0027867  1.0027905  0.0009912  2.811  0.00493 **
ast       0.0039562  1.0039641  0.0018415  2.148  0.03168 *
protime   0.2642049  1.3023951  0.1122859  2.353  0.01862 *
stage2    1.3596258  3.8947355  1.0808743  1.258  0.20843
stage3    1.6823556  5.3782101  1.0478753  1.605  0.10839
stage4    2.0627073  7.8672396  1.0432133  1.977  0.04801 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
          exp(coef) exp(-coef) lower .95 upper .95
age          1.0318     0.9692    1.0112    1.0528
edema0.5     1.1733     0.8523    0.6447    2.1352
edema1       2.4897     0.4017    1.2427    4.9881
bili         1.0908     0.9167    1.0493    1.1340
albumin      0.4777     2.0932    0.2764    0.8258
copper       1.0028     0.9972    1.0008    1.0047
ast          1.0040     0.9961    1.0003    1.0076
protime      1.3024     0.7678    1.0451    1.6230
stage2       3.8947     0.2568    0.4682   32.3981
stage3       5.3782     0.1859    0.6897   41.9364
stage4       7.8672     0.1271    1.0182   60.7865

Concordance= 0.845  (se = 0.019 )
Likelihood ratio test= 165.7  on 11 df,   p=<2e-16
Wald test            = 176.9  on 11 df,   p=<2e-16
Score (logrank) test = 278.8  on 11 df,   p=<2e-16
```

- The significant variables in the Weibull model are age [0.0022], edema1 [0.0027], bili [5.3e-06], albumin [0.0109], copper [0.0024], ast [0.0241], protime [0.0132], stage4 [0.0491]. The model summary is attached below:

```
Call:
survreg(formula = Surv(time, as.numeric(newstatus)) ~ age + edema +
    bili + albumin + copper + ast + protime + stage, data = my.pbc)
               Value Std. Error     z       p
(Intercept) 11.367108   1.363616  8.34 < 2e-16
age         -0.019130   0.006241 -3.07  0.0022
edema0.5    -0.115817   0.184296 -0.63  0.5297
edema1      -0.620899   0.207196 -3.00  0.0027
bili        -0.051030   0.011213 -4.55 5.3e-06
albumin      0.417757   0.164181  2.54  0.0109
copper      -0.001803   0.000593 -3.04  0.0024
ast         -0.002500   0.001108 -2.26  0.0241
protime     -0.171826   0.069318 -2.48  0.0132
stage2      -0.858934   0.664119 -1.29  0.1959
stage3      -1.030796   0.645908 -1.60  0.1105
stage4      -1.265734   0.643263 -1.97  0.0491
Log(scale)  -0.501244   0.075565 -6.63 3.3e-11

Scale= 0.606

Weibull distribution
Loglik(model)= -968   Loglik(intercept only)= -1053.2
        Chisq= 170.45 on 11 degrees of freedom, p= 9.5e-31
Number of Newton-Raphson Iterations: 7
n= 276
```

- The significant variables in the exponential model are age [0.0036], bili [0.0021], copper [0.0166], ast [0.0430], protime [0.0041]. The model summary is attached below:

```
Call:
survreg(formula = Surv(time, as.numeric(newstatus)) ~ age + edema +
    bili + albumin + copper + ast + protime + stage, data = my.pbc,
    dist = "exponential")
               Value Std. Error     z       p
(Intercept) 14.330507   2.108862  6.80 1.1e-11
age         -0.028855   0.009909 -2.91  0.0036
edema0.5    -0.233554   0.295818 -0.79  0.4298
edema1      -0.653393   0.344631 -1.90  0.0580
bili        -0.057330   0.018669 -3.07  0.0021
albumin      0.468047   0.267234  1.75  0.0799
copper      -0.002291   0.000956 -2.40  0.0166
ast         -0.003690   0.001824 -2.02  0.0430
protime     -0.299489   0.104216 -2.87  0.0041
stage2      -1.477440   1.089121 -1.36  0.1749
stage3      -1.762889   1.057951 -1.67  0.0956
stage4      -2.022491   1.052469 -1.92  0.0546

Scale fixed at 1

Exponential distribution
Loglik(model)= -985.6   Loglik(intercept only)= -1054.6
        Chisq= 138.15 on 11 degrees of freedom, p= 3.9e-24
Number of Newton-Raphson Iterations: 6
n= 276
```

Among these significant variables in the survival -related models, the variables age, bili, copper, ast, protime are the most significant predictors. Since these variables have smaller p-value in all the three models.