Assignment 4

Zhiwen Tan

4/7/2022

Question 1

a) Since $Y = \log(T)$ has a log location-scale distribution and

$$Y = \beta_0 + \beta_1 z + \sigma W$$

since $Z = 0$ at T_0

Then $Y_0 = \beta_0 + \sigma W$

$$T = e^{Y} = e^{\beta_0 + \beta_1 z + \sigma W}$$

$$= e^{\beta_0 + \sigma W} \cdot e^{\beta_1 z}$$

$$= e^{Y_0} \cdot e^{\beta_1 z}$$

$$= T_0 \cdot e^{\beta_1 z}$$

Then $S_T(t|Z) = P(T > t \mid Z = z)$

$$= P(T_0 e^{\beta_1 z} > t \mid Z = z)$$

$$= P(T_0 > t e^{-\beta_1 z} \mid Z = z)$$

$$= S_0(t e^{-\beta_1 z})$$

The effect of covariate Z is to accelerate / decelerate ~~the~~ time scale.

compare to a subject with baseline covariate ($Z = 0$). failure time for a subject covariate Z is accelerate / decelerate by a factor $e^{-\beta_1 z}$

as Z increases, $e^{-\beta_1 z}$ decreases, so failure time for a subject covariate Z is decelerate.

as Z decrease, $e^{-\beta_1 z}$ increases, so failure time for a subject covariate Z is accelerate.

b)

i) we can take $z=0$ into $S_T(t|z)$, then $S_0(t) = \frac{1}{1+\lambda t^\alpha}$, then

$$S_{T_0}(te^{-\beta_1 z}) = \frac{1}{1+\lambda(te^{-\beta_1 z})^\alpha}$$

$$= \frac{1}{1+\lambda t^\alpha e^{-\alpha\beta_1 z}}$$

Let $-\alpha\beta_1 = \beta$, then we have

$$= \frac{1}{1+\lambda t^\alpha e^{\beta z}}$$

$$= S_T(t|z)$$

ii) Since $T \sim$ loglogistic (μ, σ), then $Y = \log T \sim$ logistic (μ, σ)

$$F_Y(y) = \frac{e^{\frac{y-\mu}{\sigma}}}{1+e^{\frac{y-\mu}{\sigma}}} \quad -\infty < y < \infty$$

Let $W \sim$ logistc $(0,1)$, we have CDF for w, $G(w) = \frac{e^w}{1+e^w}$ $-\infty < w < \infty$

Then $F_w(w) = P(W \leq w) = P(\frac{y-\mu}{\sigma} \leq w)$

$$= P(Y \leq \mu + \sigma w)$$

$$= F_Y(\mu + \sigma w)$$

$$= \frac{e^{\frac{\mu+\sigma w - \mu}{\sigma}}}{1+e^{\frac{\mu+\sigma w-\mu}{\sigma}}}$$

$$= \frac{e^w}{1+e^w}$$

Then $\frac{y-\mu}{\sigma} \sim$ logistic $(0,1)$. Then $F_Y(y) = G(\frac{y-\mu}{\sigma})$

Thus Y has location scale distribution and T has log-location scale distribution.

Then $Y = \mu(z) + \sigma w = \beta^T z + \sigma w = \beta_0 + \beta_1 z + \sigma w$

where $\beta_0 = \mu(0)$, $\beta_0 + \beta_1 z = \mu(z)$

then $F_T(t) = P(T \geqslant t) = P(Y \geqslant \log(t)) = F_Y(\log t)$

now we have $F_T(t) = \dfrac{e^{\frac{\log t - \mu_z}{\sigma}}}{1 + e^{\frac{\log t - \mu_z}{\sigma}}}$

Since $S_T(t) = 1 - F_T(t)$, so we have

$$S_T(t) = 1 - \frac{e^{\frac{\log t - \mu_z}{\sigma}}}{1 + e^{\frac{\log t - \mu_z}{\sigma}}}$$

$$= \frac{1}{1 + e^{\frac{\log t - \mu_z}{\sigma}}}$$

$$= \frac{1}{1 + t^{-\sigma} \cdot e^{-\frac{\mu_z}{\sigma}}}$$

$$= \frac{1}{1 + t^{-\sigma} \cdot e^{-\frac{(\beta_0 + \beta_1 z)}{\sigma}}} = \frac{1}{1 + t^{-\sigma} \cdot e^{-\frac{\beta_0}{\sigma}} \cdot e^{-\frac{\beta_1 z}{\sigma}}}$$

now let $\alpha = \frac{1}{\sigma}$, $\lambda = e^{-\frac{\beta_0}{\sigma}}$, $\beta = -\frac{\beta_1}{\sigma}$. then we have

$$S_T(t) = \frac{1}{1 + t^{\alpha} \cdot \lambda \cdot e^{\beta z}} = \frac{1}{1 + \lambda t^{\alpha} e^{\beta z}}$$

c) now we want to show T is PH model, that is $h(t|z) = h_0(t)\,\psi(z)$

where $\psi(z) = e^{\beta z}$

$$h(t|z) = -\frac{d}{dt} \log S_T(t|z)$$

$$= -\frac{d}{dt} \log (1 + \lambda t^{\alpha} e^{\beta z})^{-1}$$

$$= -\frac{d}{dt} - \log (1 + \lambda t^{\alpha} e^{\beta z})$$

$$= \frac{\alpha \lambda t^{\alpha - 1} e^{\beta z}}{1 + \lambda t^{\alpha} e^{\beta z}}$$

$$h_0(t) = -\frac{d}{dt} \log S_0(t)$$

$$= -\frac{d}{dt} - \log (1 + \lambda t^{\alpha})$$

$$= \frac{\alpha \lambda t^{\alpha - 1}}{1 + \lambda t^{\alpha}}$$

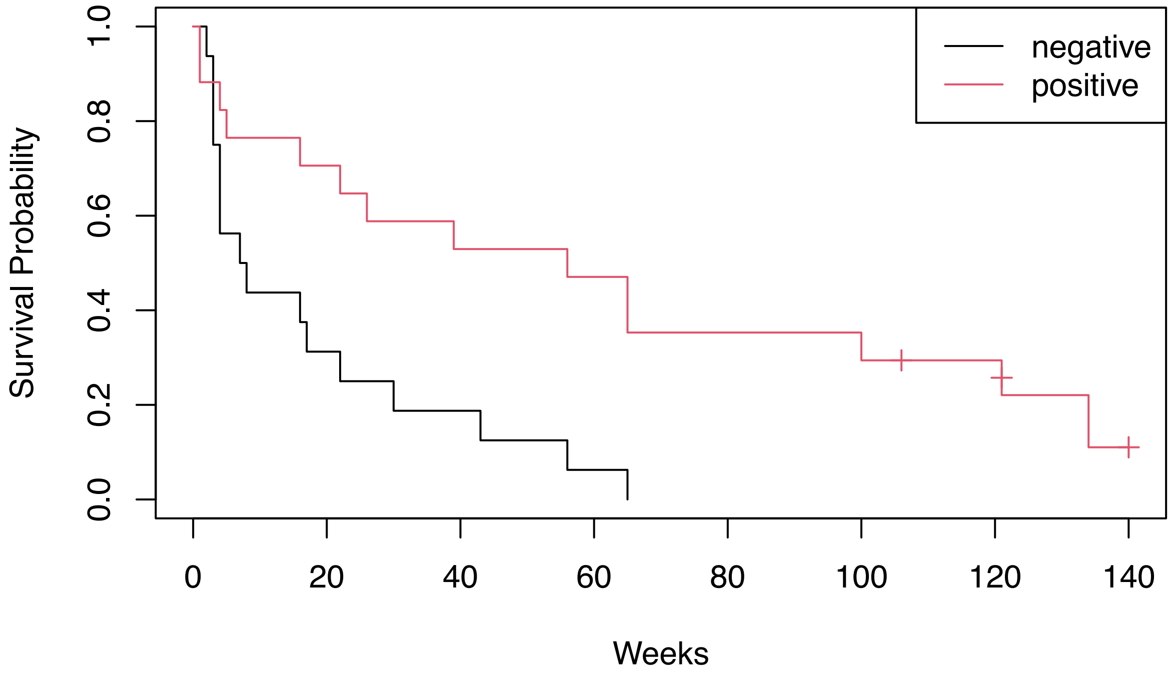
Then $h_0(t) \cdot \psi(z) = \dfrac{\alpha \lambda t^{\alpha - 1}}{1 + \lambda t^{\alpha}} e^{\beta z} \neq h(t|z)$

Therefore , T is not a proportional hazard function.

Question 2

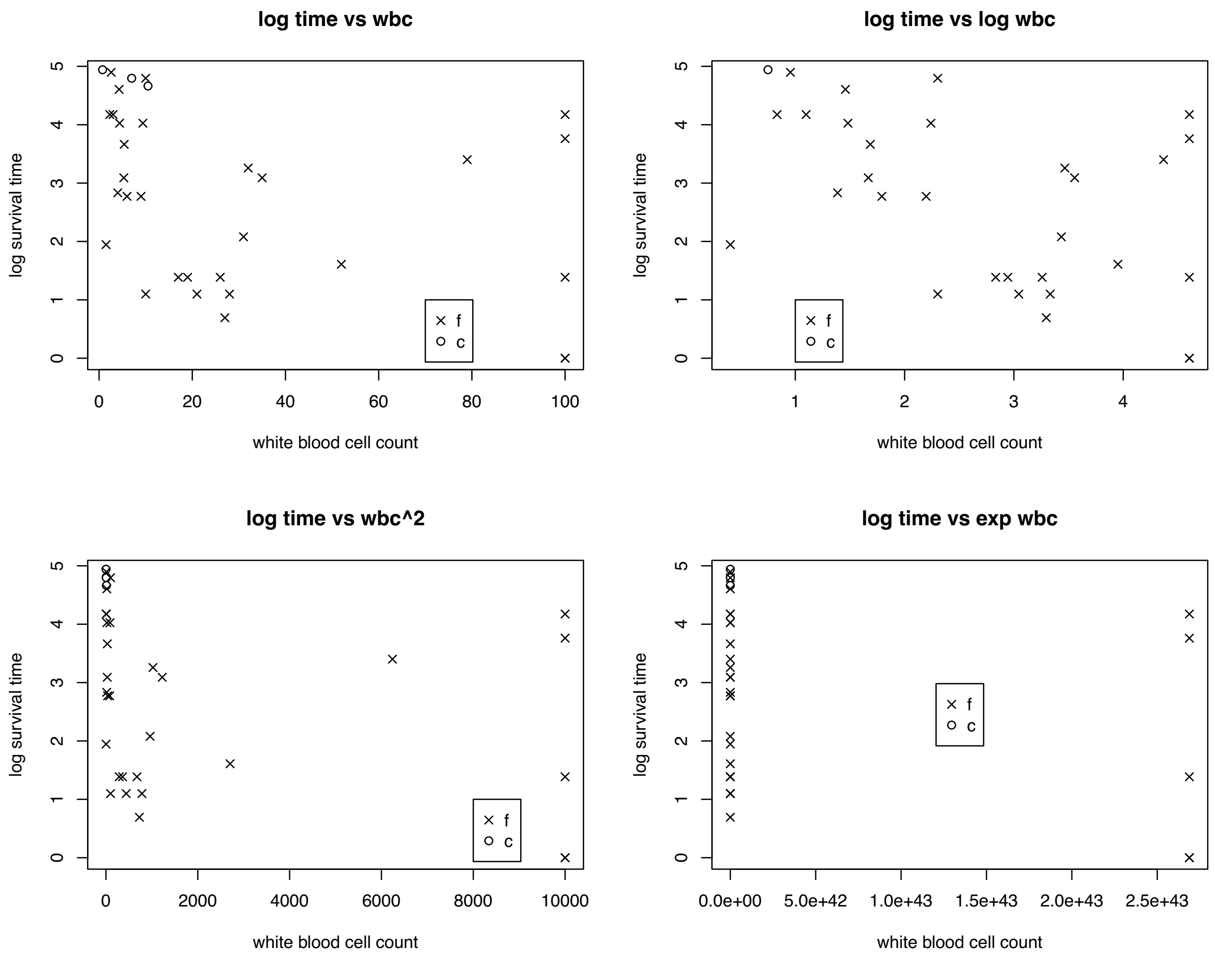
a) Because AG is binary data, so scatter plot may not give us many information compare to KM graph.

- Kaplan-Meier plot for covariate AG



From the KM graph, we can see both positive and negative group has similar survival probability at the beginning. However, the difference in survival probability is greater after 4 weeks. Overall, the positive AG group patients has higher survival probability and longer survival time. Also, the positive group has 3 censored points where the negative group has no censored data.

- The wbc is a continuous variable, so a scatter plot would be more appropriate. Also, this p42 dataset is lightly censored data because there are only 3 censored data. Therefore, we can check the covariate with scatter plot.



We have tried log time vs wbc, log wbc, wbcˆ2, and exp(wbc). Among those four scatter plots, "log time vs log wbc" shows some linear decreasing pattern, other plots does not have a clear pattern. Therefore, the survival time might have some relationship with log(wbc).

b) Here, we will build three models and check which one is the best

The first model will include both AG and log(wbc), the fitted results are:

```
##
## Call:
## survreg(formula = Surv(time, status) ~ as.factor(AG) + log(wbc),
##     data = data)
##                  Value Std. Error     z       p
## (Intercept)      3.841      0.534  7.19 6.6e-13
## as.factor(AG)1  1.177      0.427  2.76  0.0058
## log(wbc)        -0.366      0.150 -2.45  0.0143
## Log(scale)       0.112      0.147  0.77  0.4442
##
## Scale= 1.12
##
## Weibull distribution
## Loglik(model)= -132.5   Loglik(intercept only)= -140.3
##  Chisq= 15.69 on 2 degrees of freedom, p= 0.00039
## Number of Newton-Raphson Iterations: 6
## n= 33
```

model 1 has formula of

$$Y = 3.841 + 1.177 * I(AG = 1) - 0.366 * log(wbc)$$

The p value for AG and log(wbc) are 0.0058 and 0.0143, both less than 0.05. the overall p-value is 0.00039, which is way less than 0.05, so we can conclude that our model fits the data well.

The second model would be AG only, the results are shown as below:

```
##
## Call:
## survreg(formula = Surv(time, status) ~ as.factor(AG), data = data)
##                 Value Std. Error    z       p
## (Intercept)    2.800      0.305 9.17 < 2e-16
## as.factor(AG)1 1.459      0.437 3.34 0.00085
## Log(scale)     0.172      0.149 1.16 0.24650
##
## Scale= 1.19
##
## Weibull distribution
## Loglik(model)= -135.5   Loglik(intercept only)= -140.3
##  Chisq= 9.63 on 1 degrees of freedom, p= 0.0019
## Number of Newton-Raphson Iterations: 5
## n= 33
```

model 2 has formula of

$$Y = 2.800 + 1.459 * I(AG = 1)$$

To check if this model is better than the first one, we need to do the LR test.

$$\begin{aligned}\lambda_{obs} &= 2 * (-132.5 + 135.5) \\ &= 6\end{aligned}$$

This follows a chisq distribution with df=1, we can use $pchisq(6, 1, lower.tal = F)$ to get the p value, which is $0.01430588 < 0.05$. This means we can reject the null hypothesis, and model 1 is better.

The third model is log(wbc) only, the results are shown as below:

```
##
## Call:
## survreg(formula = Surv(time, status) ~ log(wbc), data = data)
##              Value Std. Error     z      p
## (Intercept)  4.854      0.500  9.71 <2e-16
## log(wbc)    -0.500      0.165 -3.03 0.0024
## Log(scale)   0.222      0.146  1.52 0.1277
##
## Scale= 1.25
##
## Weibull distribution
## Loglik(model)= -136   Loglik(intercept only)= -140.3
##  Chisq= 8.77 on 1 degrees of freedom, p= 0.0031
## Number of Newton-Raphson Iterations: 5
## n= 33
```

model 3 has formula of

$$Y = 4.854 - 0.500 * log(wbc)$$

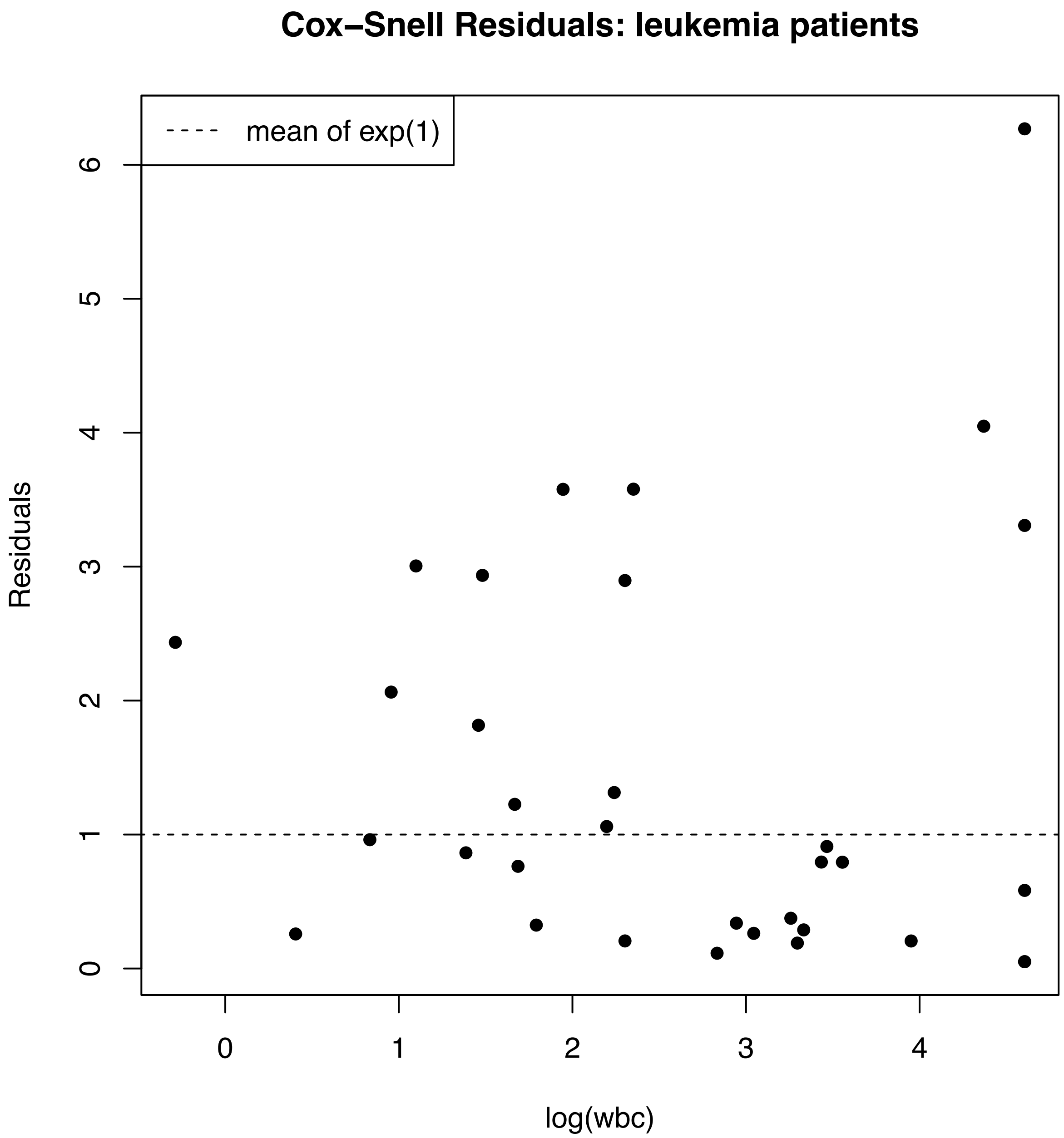
To check if this model is better than the first one, we need to do the LR test.

$$\begin{aligned}\lambda_{obs} &= 2 * (-132.5 + 136) \\ &= 7\end{aligned}$$

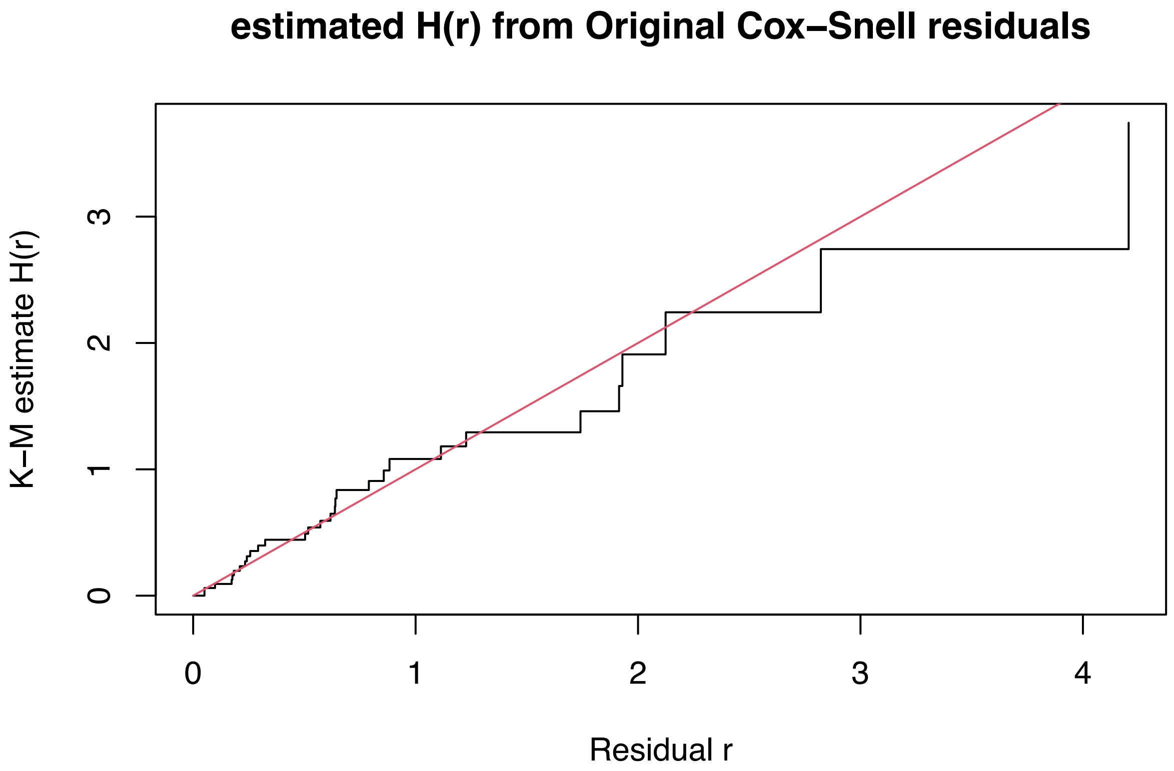
This follows a chisq distribution with df=1, we can use $pchisq(7, 1, lower.tal = F)$ to get the p value, which is $0.008150972 < 0.05$. This means we can reject the null hypothesis, and model 1 is better.

Therefore, we can conclude that model 1 with both AG and log(wbc) is the best parametric regression model for this dataset.

c) Cox-snell residual



The Cox-Snell residual plot has expected behavior. the dots are denser between 0 to 1 and look like exp(1) observation. Therefore, model 1 fits pretty well.



From the K-M estimated H(r) from original Cox-Snell residual plot, we can see the trend is similar to y = x. This suggest that our model 1 fits the data pretty well.

d) Our final model is model 1, the summary is shown as below:

```
##
## Call:
## survreg(formula = Surv(time, status) ~ as.factor(AG) + log(wbc),
##     data = data)
##                 Value Std. Error     z       p
## (Intercept)     3.841      0.534  7.19 6.6e-13
## as.factor(AG)1  1.177      0.427  2.76  0.0058
## log(wbc)       -0.366      0.150 -2.45  0.0143
## Log(scale)      0.112      0.147  0.77  0.4442
##
## Scale= 1.12
##
## Weibull distribution
## Loglik(model)= -132.5   Loglik(intercept only)= -140.3
##  Chisq= 15.69 on 2 degrees of freedom, p= 0.00039
## Number of Newton-Raphson Iterations: 6
## n= 33
```

The final model has formula of

$$log(time) = 3.841 + 1.177 * I(AG = 1) - 0.366 * log(wbc)$$

In this model, we have $\hat{\beta_0}, \hat{\beta_1}, \hat{\beta_2}$.

$\hat{\beta_0}$ = 3.841 when AG is negative and wbc cell count at diagnosis is 0. The average failure time is 3.841 - E[EV(0,1)] = 3.841 - 0.577 = 3.264. In here, the $\hat{\beta_0}$ contains both expected log failure time and the expectation of EV(0,1).

$\hat{\beta_1} = 1.177$, this means if we assume log(wbc) stays the same, then, when AG test is positive (AG = 1), the log failure time will increase by 1.177 unit compare to negative AG test.

$\hat{\beta_2} = -0.366$, this means if we assume AG stays the same, then, 1 unit log(wbc) increase will decrease the log failure time by 0.366 unit.

Question 3

The original file has multiple recurrence for each patients, and some patients has no recurrence, so we need to clean the file first, then we can read the data into data frame. After cleaning the data, we have the data looks like below

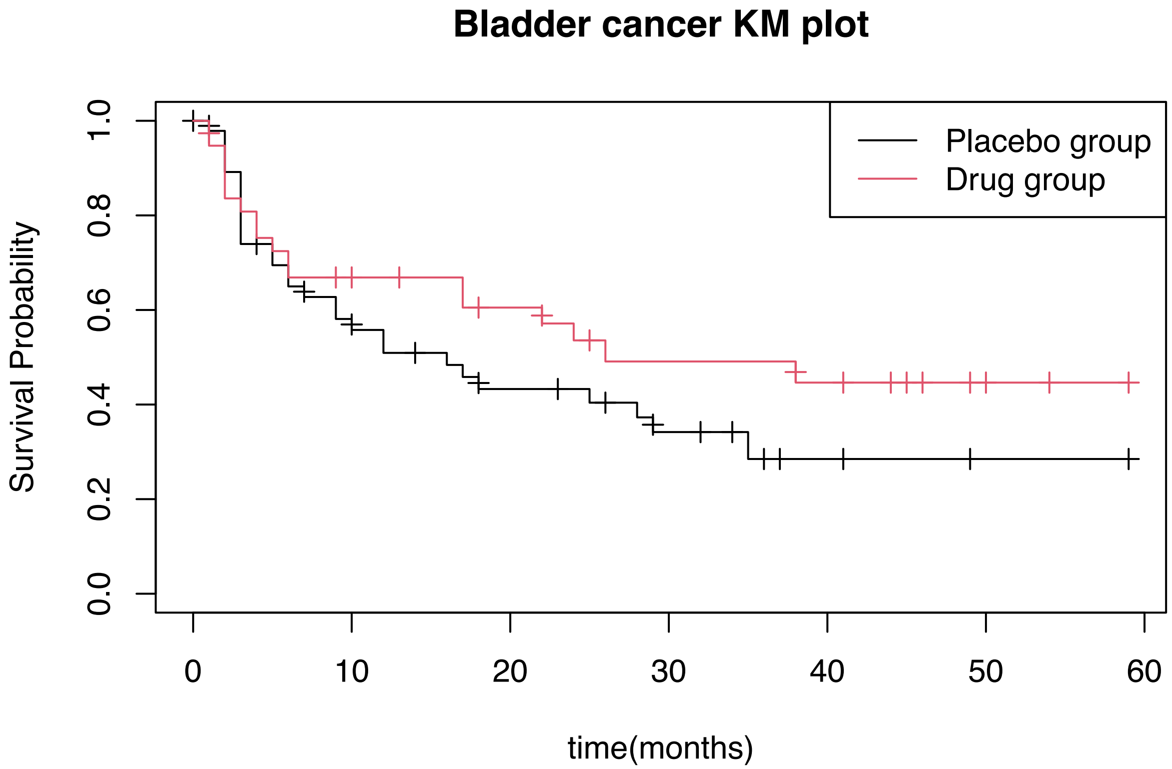
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##   Group fultime number size time status
## 1     1       0      1    1    0      0
## 2     1       1      1    3    1      0
## 3     1       4      2    1    4      0
## 4     1       7      1    1    7      0
## 5     1      10      5    1   10      0
## 6     1      10      4    1    6      1
```

Now we can draw a KM plot for our data to make compare between placebo and drug group.



From the KM plot, we can see the drug group and placebo group has similar survival probability in the first 6 months. After 6 month, there are greater difference between drug group and placebo group, the survival probability is higher in drug group.

After graph exploration, we want to fit cox model to the bladder data.

Model 1: Cox model with Group, number and size

```
## Call:
## coxph(formula = Surv(time, status) ~ as.factor(Group) + number +
##     size, data = blad)
##
##   n= 86, number of events= 47
##
##                         coef exp(coef) se(coef)      z Pr(>|z|)
## as.factor(Group)2 -0.52598   0.59097  0.31583 -1.665   0.0958 .
## number             0.23818   1.26894  0.07588  3.139   0.0017 **
## size               0.06961   1.07209  0.10156  0.685   0.4931
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## as.factor(Group)2     0.591     1.6921    0.3182     1.097
## number                1.269     0.7881    1.0936     1.472
## size                  1.072     0.9328    0.8786     1.308
##
## Concordance= 0.631  (se = 0.044 )
## Likelihood ratio test= 9.92  on 3 df,   p=0.02
## Wald test            = 10.53  on 3 df,   p=0.01
## Score (logrank) test = 11.12  on 3 df,   p=0.01
```

```
## [1] -185.1376 -180.1783
```

The hazard function of failure time T_i has the form

$$h(t|z_i) = h_0(t)exp(0.23818 * number + 0.06961 * size - 0.52598 * I(Group = 2))$$

We can see the p value for number is $0.0017 < 0.05$, which means number variable is an important covariate. the p-value for both group and size are greater than 0.05. This indicate that both variable are not statistically significant, but the p-value for group 0.0958 is just slightly over 0.05, so there might still have some relationship. However, the p-value for size is far away from 0.05, so we can delete this variable in the model.

Now we know size variable is not important, so we want to try our second model which include only Group and number to see if this make any difference.

Model 2: Cox model with Group and number

```
## Call:
## coxph(formula = Surv(time, status) ~ as.factor(Group) + number,
##     data = blad)
##
##   n= 86, number of events= 47
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## as.factor(Group)2 -0.51218   0.59919  0.31299 -1.636  0.10176
## number             0.23079   1.25960  0.07542  3.060  0.00221 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## as.factor(Group)2    0.5992     1.6689    0.3244     1.107
## number               1.2596     0.7939    1.0865     1.460
##
## Concordance= 0.633  (se = 0.043 )
## Likelihood ratio test= 9.47  on 2 df,   p=0.009
## Wald test            = 10.15  on 2 df,   p=0.006
## Score (logrank) test = 10.73  on 2 df,   p=0.005
```

```
## [1] -185.1376 -180.4023
```

```
## [1] 0.5032863
```

The hazard function of failure time T_i has the form

$$h(t|z_i) = h_0(t)exp(0.23079 * number - 0.51218 * I(Group = 2))$$

To check if this model is better than the first, we need to do a LR test first.

$$\begin{aligned}\lambda_{obs} &= 2 * (180.4023 - 180.1783) \\ &= 0.448\end{aligned}$$

This follows a chisq distribution with df=1, we can use $pchisq(0.448, 1, lower.tail = F)$ to get the p value, which is $0.5032863 > 0.05$. This means we can not reject the null hypothesis, and model 1 is not better than model 2. We can also see, the p value for number is $0.00221 < 0.05$, which means the number varaible is significant. However, the p value for Group variable is $0.10176 > 0.05$ which indicates this variable is not statistically significant and we could remove this variable out of our model.

Now we can try our third model with number variable only.

Model 3: Cox model with number only

```
## Call:
## coxph(formula = Surv(time, status) ~ number, data = blad)
##
##   n= 86, number of events= 47
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## number 0.20120   1.22287  0.07068 2.846  0.00442 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## number     1.223     0.8178     1.065     1.405
##
## Concordance= 0.604  (se = 0.041 )
## Likelihood ratio test= 6.7  on 1 df,   p=0.01
## Wald test            = 8.1  on 1 df,   p=0.004
## Score (logrank) test = 8.46  on 1 df,   p=0.004
```

```
## [1] -185.1376 -181.7882
```

```
## [1] 0.09593822
```

The hazard function of failure time T_i has the form

$$h(t|z_i) = h_0(t)exp(0.20120 * number)$$

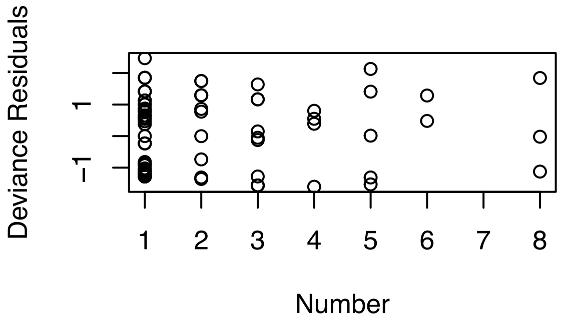
To check if this model is better than the second model, we need to do a LR test first.

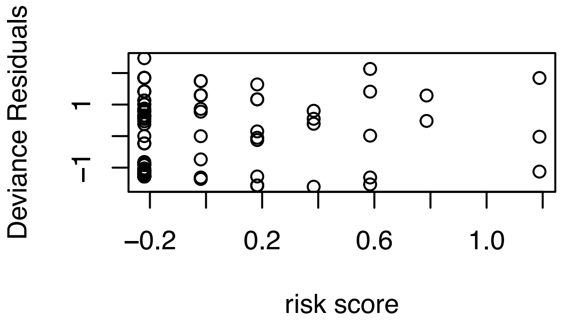
$$\begin{aligned}\lambda_{obs} &= 2 * (181.7882 - 180.4023)\\ &= 2.7718\end{aligned}$$

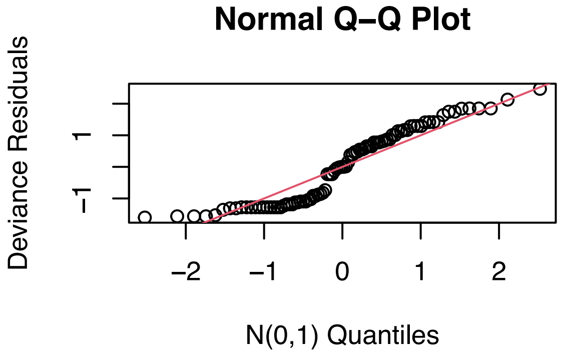
This follows a chisq distribution with df=1, we can use $pchisq(2.7718, 1, lower.tail = F)$ to get the p value, which is $0.09593822 > 0.05$. This means we can not reject the null hypothesis, and model 2 is not better than model 3. We can also see, the p value for number is $0.00442 < 0.05$, which means the number varaible is significant and the overall p value is also less than 0.05. This means model 3 is the best model.

Therefore, Model 3 is our final model.

Now, we can do the deviance residual analysis







From the model above, we know n= 86, number of events= 47, so we have 39 censoring here, the censoring rate is 39/86 = 0.4534884 > 0.4. This means a large number of residuals are close to 0, which distorts the N (0,1) distribution shape. But residuals should still mostly fall between [-2,2], and be roughly symmetric in range. From our first two plot, we can see although most of data are between [-2,2], but it is not symmetric. Therefore, both plot does not roughly follow a N (0,1) shaped distribution. From the Q-Q plot, we can also see the data do not have a linear pattern. Therefore, the residual are not normally distributed.

To conclude, the model with number variable only (model 3) have the best performance, but the Deviance Residuals does not shows a N(0,1) pattern. Therefore, the Cox model may not fit the data very well. From the model coefficient, we can see the coefficient for number is 0.20120, this represent log hazard ratio of patient gets 1 more tumor at the initial time. exp(0.20120) = 1.222869. this represent hazard ratio for patients get 1 more tumor at the initial time. In this dataset, there are 51 patients with 1 tumor initially, 11 with 2 tumors, 10 with 3 tumors, 4 with 4 tumors, 5 with 5 tumors, 2 with 6 tumors and 3 with 8 tumors. This means we need to increase the study size if number is the variable we are interested in. we could draw the KM plot for numbers to see the difference, but without a adequate sample size, the plot may not give use the right information.

Question 4

a) In this question, we have more than 2 groups, so we can set the null hypothesis as there are no difference at all among the groups, that is $H_0 : S_1(t) = S_2(t) = S_3(t)$. Then the alternative hypothesis is H_A: there is at least one group has different S(t).

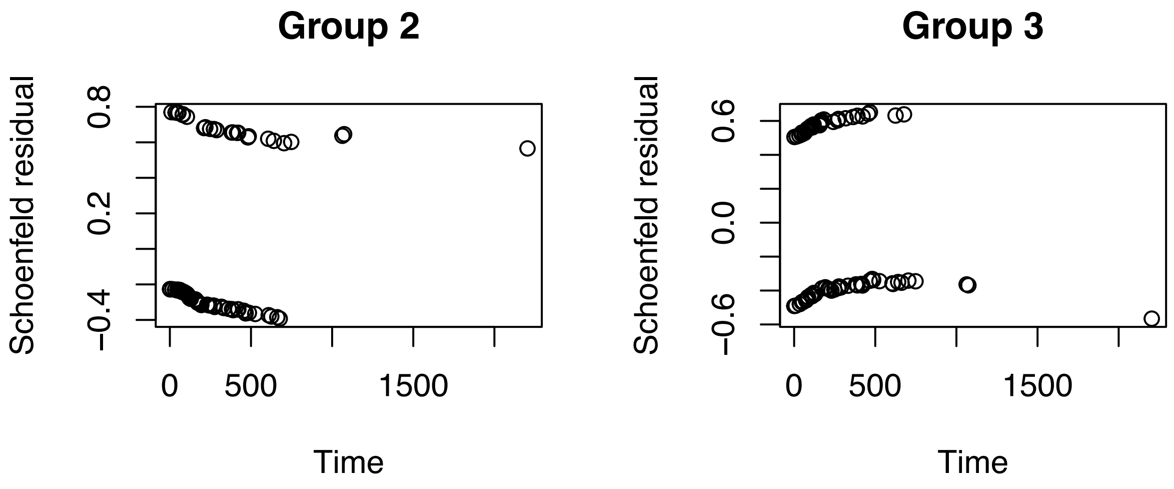
```
## Call:
## survdiff(formula = Surv(DFtime, DFstatus) ~ as.factor(Group),
##     data = BM)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## as.factor(Group)=1 38       24     21.9     0.211     0.289
## as.factor(Group)=2 54       25     40.0     5.604    11.012
## as.factor(Group)=3 45       34     21.2     7.756    10.529
##
##  Chisq= 13.8  on 2 degrees of freedom, p= 0.001
```

From the above result, we can see the $W^2_{obs} \approx \chi^2_2 = 13.8$ that is W^2_{obs} follows a chisq distribution with degree of freedom equals 2. The p-vale is $0.001 < 0.05$ which means we can reject the null hypothesis and there is at least 1 disease group has different S(t).

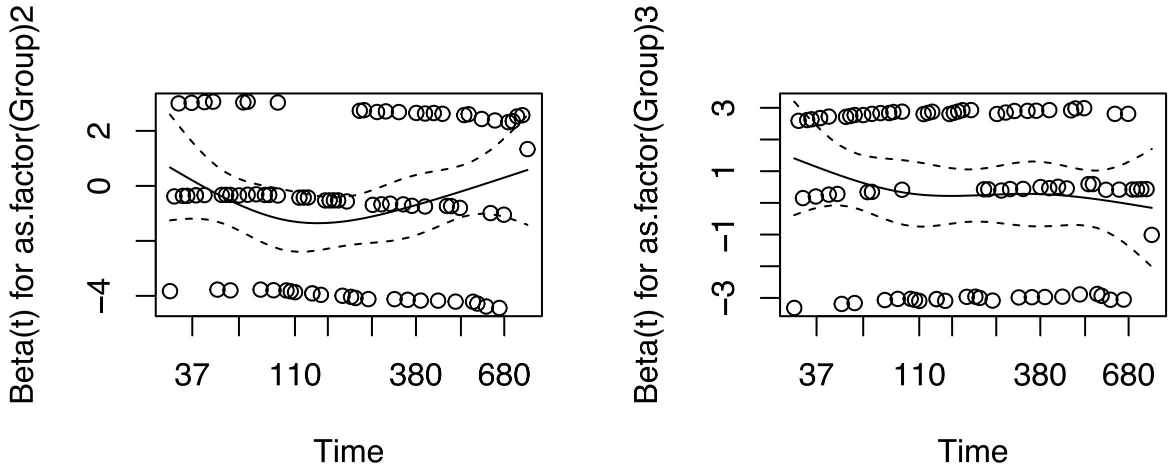
b) we can now fit cox model to our dataset, group will be the test variable

```
## Call:
## coxph(formula = Surv(DFtime, DFstatus) ~ as.factor(Group), data = BM)
##
##   n= 137, number of events= 83
##
##                      coef exp(coef) se(coef)      z Pr(>|z|)
## as.factor(Group)2 -0.5742    0.5632   0.2873 -1.999   0.0457 *
## as.factor(Group)3  0.3834    1.4673   0.2674  1.434   0.1516
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## as.factor(Group)2    0.5632     1.7757    0.3207     0.989
## as.factor(Group)3    1.4673     0.6815    0.8688     2.478
##
## Concordance= 0.625  (se = 0.03 )
## Likelihood ratio test= 13.45  on 2 df,   p=0.001
## Wald test            = 13.03  on 2 df,   p=0.001
## Score (logrank) test = 13.81  on 2 df,   p=0.001
```

From the above summary, we can see that group 2 has a p-value $0.0457 < 0.05$ which means we can rehect the null hypothesis. We can also see, the wald test is less than 0.05, so group 2 is statistical different from group1. Group 2 has a p-value 0.1516 which means we can not reject the null hypothesis and group 3 is same as group1. Therefore, the result from Cox regression model suggest that the 3 group are not the same. Next, we need to perform a residual analysis. In this case, we are using Schoenfeld residual.



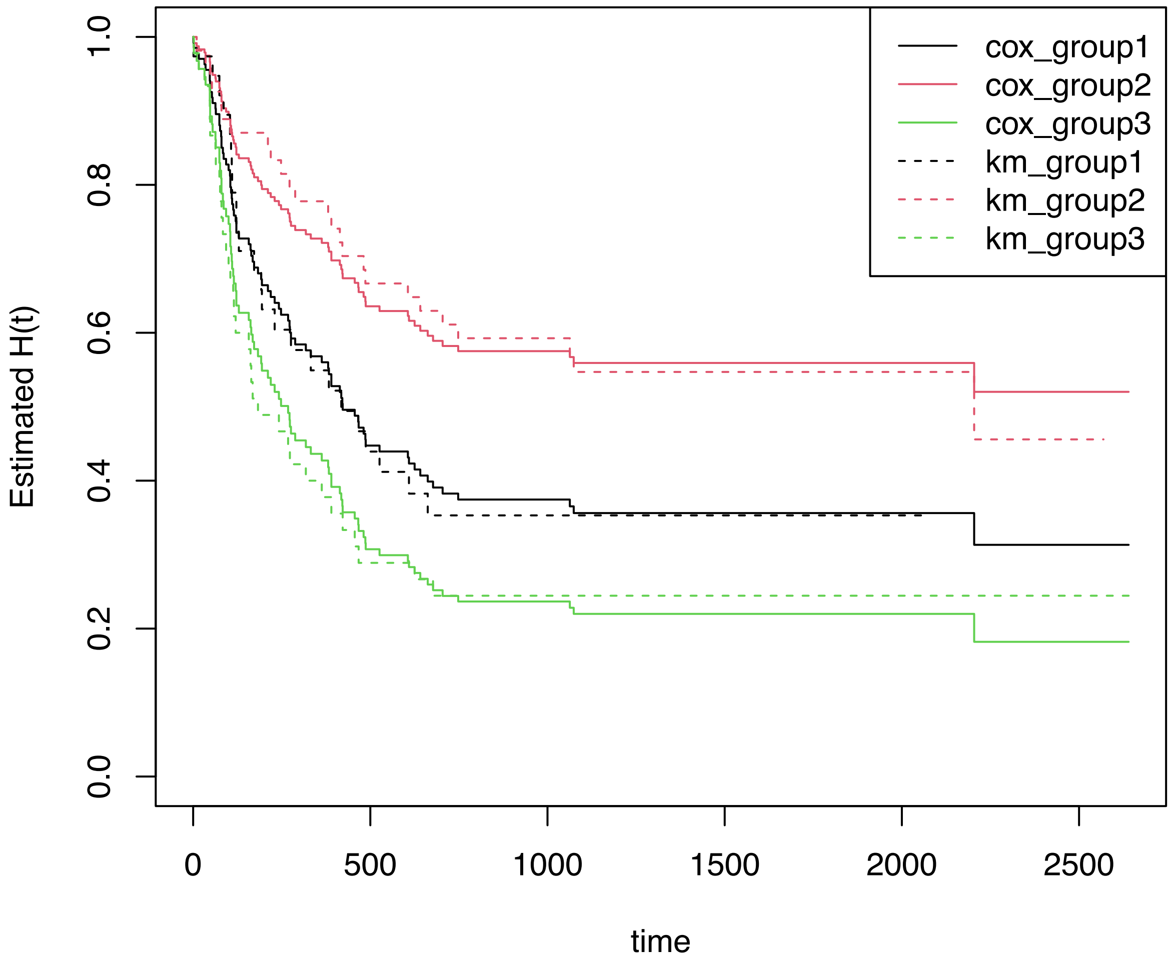
From the above two plot, we can see the Schoenfeld residual satisfies the PH assumption



From the sacled Schoenfeld residual plot, we can see the residuals does not show a time trend and this satisfied the PH assumption.

```
##                    chisq df    p
## as.factor(Group)2  1.44  1 0.23
## as.factor(Group)3  2.66  1 0.10
## GLOBAL             2.78  2 0.25
```

This part shows the p-value for each groups, all p-value in this individual test are greater than 0.05, therefore, we can not reject the null hypothesis ($H_0 : \theta_2 = 0$ and $H_0 : \theta_3 = 0$), and the Schoenfeld residual is not correlated with time which satisfied the PH assumption. The global p-value is $0.25 > 0.05$, this means we can not reject the null hypothesis ($H_0 : \theta_2 = \theta_3 = 0$) also suggest the Schoenfeld residual is not correlated with time, and this satisfied the PH assumption. In conclusion, we can say the cox model fits well becasue the residual test satisfied the PH assumption.



From the KM vs Cox plot, we can see the patterns are very similar between two methods. In group 1, all the pattern are similar. In group 2, there are some small difference between time 200-800 and after 2200. In group 3, the km method has high H(t) at the end and lower H(t) at beginning. however, all the difference are small. Therefore, the Cox model agree with the KM curve, and we can conclude the Cox model fits well for this data.

c) We can get all the beta value from the previous Cox model we build.

```
## Call:
## coxph(formula = Surv(DFtime, DFstatus) ~ as.factor(Group), data = BM)
##
##   n= 137, number of events= 83
##
##                      coef exp(coef) se(coef)      z Pr(>|z|)
## as.factor(Group)2 -0.5742    0.5632   0.2873 -1.999   0.0457 *
## as.factor(Group)3  0.3834    1.4673   0.2674  1.434   0.1516
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## as.factor(Group)2    0.5632     1.7757    0.3207     0.989
## as.factor(Group)3    1.4673     0.6815    0.8688     2.478
##
## Concordance= 0.625  (se = 0.03 )
## Likelihood ratio test= 13.45  on 2 df,   p=0.001
## Wald test            = 13.03  on 2 df,   p=0.001
## Score (logrank) test = 13.81  on 2 df,   p=0.001
```

As we know from the lecture, $\widehat{S(t|z)} = [\widehat{S_0(t)}^{e^{(\hat{\beta}^T Z)}}]$, and from the question, we know for group 1, then we have $S(t) = \hat{S_0}(t)$. For group 2 and 3, we can simply plug the β value in the equation, then we have:

$$
\begin{aligned}
Group1 &: S(t) = \hat{S_0}(t) \\
Group2 &: S(t) = \hat{S_0}(t)^{exp(-0.5742*I(Group=2))} \\
Group3 &: S(t) = \hat{S_0}(t)^{exp(0.3834*I(Group=3))}
\end{aligned}
$$

Appendix

Q2

a)

```r
library(survival)
data <- read.table("p42.txt", header = T)
fit <- survfit(Surv(time, status)~as.factor(AG), data = data)
plot(fit, conf.int = F, mark.time = T, xlab = "Weeks",
     ylab = "Survival Probability", col = 1:2)
legend("topright", c("negative", "positive"), col = 1:2, lty=1:1)

delta<-data$status
x<-data$time
wbc<-data$wbc
par(mfrow = c(2,2))
#scatter plot for wbc
plot(wbc[delta==1],log(x[delta==1]),pch=4,xlab="white blood cell count",
     ylab="log survival time",main="log time vs wbc")
points(wbc[delta==0],log(x[delta==0]),pch=1)
legend(70,1,c("f","c"),pch=c(4,1))

plot(log(wbc[delta==1]),log(x[delta==1]),pch=4,xlab="white blood cell count",
     ylab="log survival time",main="log time vs log wbc")
points(wbc[delta==0],log(x[delta==0]),pch=1)
legend(1,1,c("f","c"),pch=c(4,1))

plot(wbc[delta==1]^2,log(x[delta==1]),pch=4,xlab="white blood cell count",
     ylab="log survival time",main="log time vs wbc^2")
points(wbc[delta==0],log(x[delta==0]),pch=1)
legend(8000,1,c("f","c"),pch=c(4,1))

plot(exp(wbc[delta==1]),log(x[delta==1]),pch=4,xlab="white blood cell count",
     ylab="log survival time",main="log time vs exp wbc")
points(wbc[delta==0],log(x[delta==0]),pch=1)
legend("center",c("f","c"),pch=c(4,1))
```

b)

```r
fit1 <- survreg(Surv(time, status)~as.factor(AG) + log(wbc), data = data)
summary(fit1)

fit2 <- survreg(Surv(time, status)~as.factor(AG), data = data)
summary(fit2)

fit3 <- survreg(Surv(time, status)~log(wbc), data = data)
summary(fit3)
```

c)

```
AG <-data$AG
wbc <- data$wbc
x <- data$time
b<-fit1$coeff
mu<-b[1]+b[2]*AG+b[3]*log(wbc)
sig<-fit1$scale
delta<-data$status
r<-exp(log(x)-mu/sig)+1-delta
r1<-exp((log(x)-mu)/sig)

par(mfrow = c(2,1))
plot(AG,r,pch=16,xlab="AG"
     ,ylab="Residuals",main="Cox-Snell Residuals: leukemia patients")
lines(seq(0,1),rep(1,2),lty=2)
legend(0.4,4,"mean of exp(1)",lty = 2)


plot(log(wbc),r,pch=16,xlab="log(wbc)"
     ,ylab="Residuals",main="Cox-Snell Residuals: leukemia patients")
lines(seq(-1,6),rep(1,8),lty=2)
legend("topleft","mean of exp(1)",lty = 2)

fitr <- survfit(Surv(r1,delta)~1)
plot(fitr, fun="cumhaz", conf.int = F, xlab = "Residual r",
     ylab = "K-M estimate H(r)",
     main = "estimated H(r) from Original Cox-Snell residuals")
lines(0:4, 0:4, col = 2)
```

d)

```
summary(fit1)
```

Q3

```
library(dplyr)
blad <- read.table("bladder.txt", col.names = c("Group", "fultime",
                                                 "number", "size", "time",
                                                 "r2", "r3", "r4"),
                   fill = T, na.strings = "", header = F)
blad <- blad[,1:5]
blad$status <- ifelse(is.na(blad$time), 0,1)
blad <- blad %>%
    mutate(time = coalesce(time,fultime))
head(blad)

fit_km <- survfit(Surv(time, status)~as.factor(Group), data = blad)
plot(fit_km, conf.int = F, mark.time = T, xlab = "time(months)",
     ylab = "Survival Probability", col = 1:2, main = "Bladder cancer KM plot")
legend("topright", c("Placebo group", "Drug group"), lty = 1:1, col = 1:2)

fit_co1 <- coxph(Surv(time, status)~as.factor(Group)+number+size, data = blad)
```

```r
summary(fit_co1)
fit_co1$loglik

fit_co2 <- coxph(Surv(time, status)~as.factor(Group)+number, data = blad)
summary(fit_co2)
fit_co2$loglik
pchisq(2*(180.4023 - 180.1783), 1, lower.tail = F)

fit_co3 <- coxph(Surv(time, status)~number, data = blad)
summary(fit_co3)
fit_co3$loglik
pchisq(2*(181.7882 - 180.4023), 1, lower.tail = F)

fit_res <- resid(fit_co3, type = "deviance")
par(mfrow = c(2,2))
plot(blad$number, fit_res, xlab = "Number", ylab = "Deviance Residuals")
plot(predict(fit_co3), fit_res, xlab = "risk score", ylab = "Deviance Residuals")
qqnorm(fit_res, xlab = "N(0,1) Quantiles", ylab = "Deviance Residuals")
abline(0,1,col = 2)
```

Q4

a)

```r
BM <- read.table("BoneMarrowDFTime.txt", header = T)
fit_lr <- survdiff(Surv(DFtime, DFstatus) ~ as.factor(Group), data = BM)
fit_lr
```

b)

```r
cox_bm <- coxph(Surv(DFtime, DFstatus) ~ as.factor(Group), data = BM)
cox_bm
summary(cox_bm)

resid.sch=residuals(cox_bm,type="schoenfeld")
t=sort(BM$DFtime[BM$DFstatus==1])
par(mfrow=c(1,2))
plot(t,resid.sch[,1],ylab="Schoenfeld residual", xlab="Time")
plot(t,resid.sch[,2],ylab="Schoenfeld residual", xlab="Time")

resid.scaledsch=residuals(cox_bm, type="scaledsch")
zph=cox.zph(cox_bm,terms=F)
par(mfrow=c(1,2))
plot(zph)
print(zph)

Group <- BM$Group
g1 <- survfit(cox_bm, newdata = data.frame(Group=1))
g2 <- survfit(cox_bm, newdata = data.frame(Group=2))
g3 <- survfit(cox_bm, newdata = data.frame(Group=3))
plot(g1, conf.int = F, xlab = "time", ylab = "Estimated H(t)", col = 1)
lines(g2, conf.int = F, lty = 1, col = 2)
lines(g3, conf.int = F, lty =1, col = 3)

fit_kmb <- survfit(Surv(DFtime, DFstatus) ~ as.factor(Group), data = BM)
lines(fit_kmb, lty = c(2,2,2), col = 1:3)

legend("topright", c("cox_group1", "cox_group2",
                     "cox_group3", "km_group1",  "km_group2",  "km_group3"),
       lty = c(1,1,1,2,2,2), col = 1:3)
```