Assignment 4

Zhiwen Tan

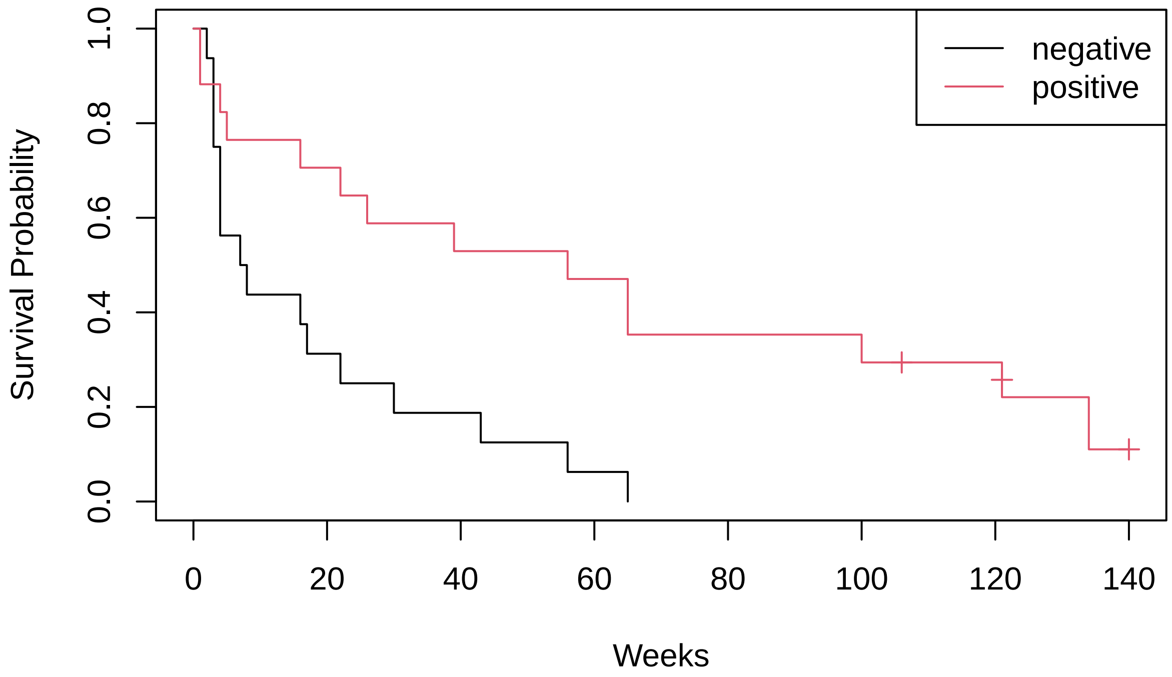
4/7/2022

Question 1

Question 2

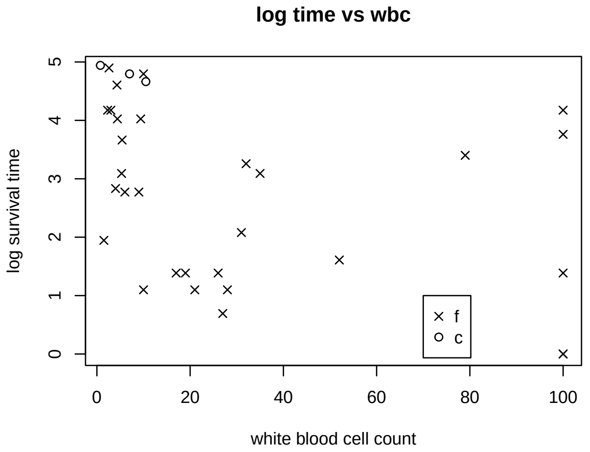
a) Because AG is binary data, so scatter plot may not give us many information compare to KM graph.

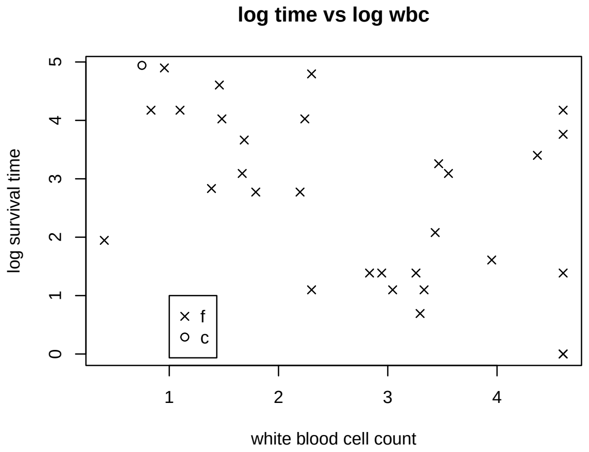
- Kaplan-Meier plot for covariate AG

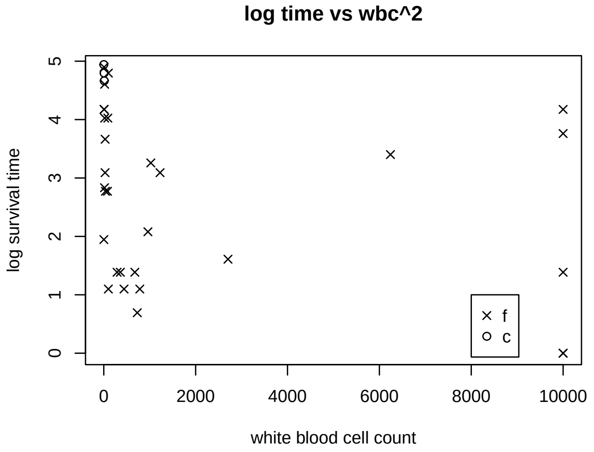


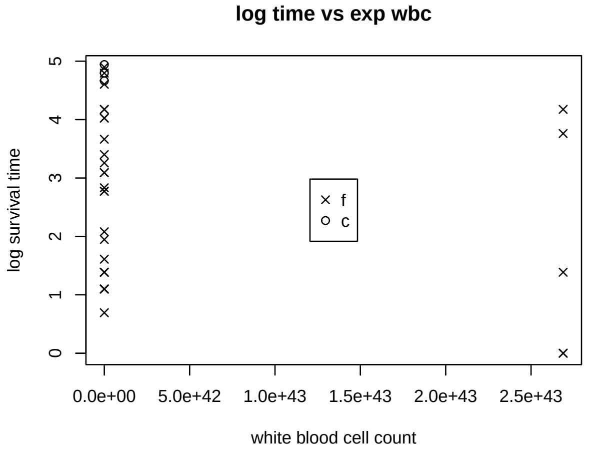
From the KM graph, we can see both positive and negative group has similar survival probability at the beginning. However, the difference in survival probability is greater after 4 weeks. Overall, the positive AG group patients has higher survival probability and longer survival time. Also, the positive 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
```

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
```

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 not 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
```

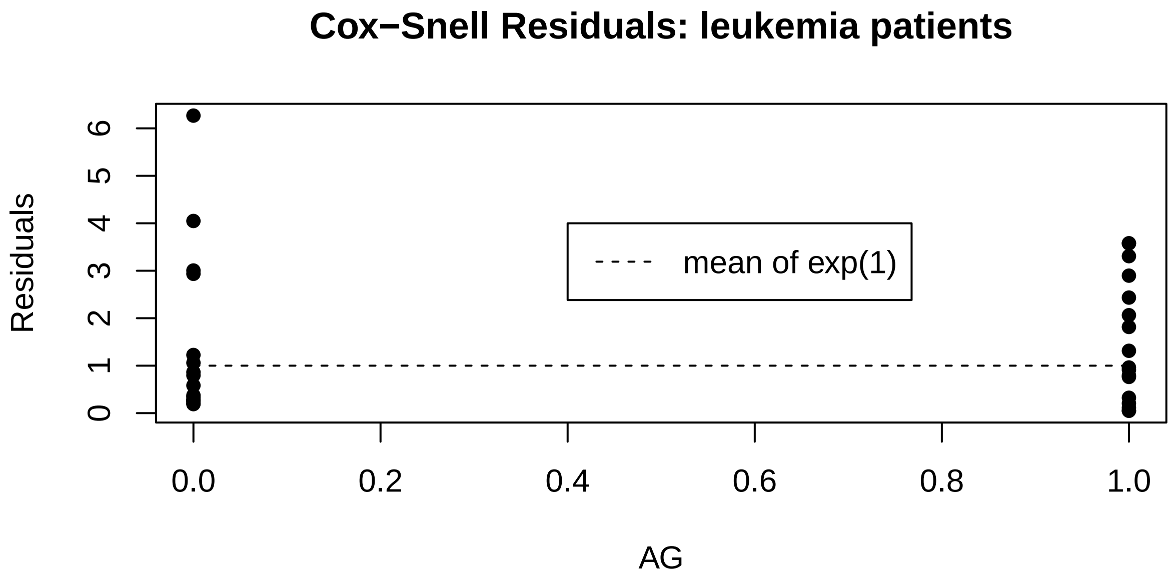
To check if this model is better than the first one, we need to do the LR test.

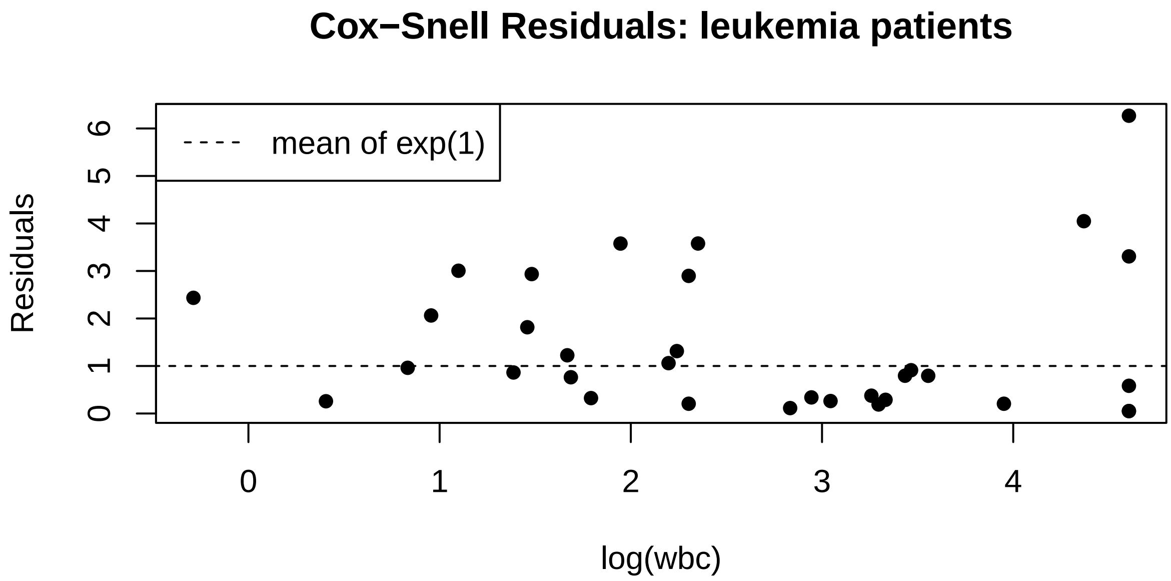
$$\begin{aligned}\lambda_{obs} &= 2 * (-132.5 + 136) \\ &= 7\end{aligned}$$

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 not reject the null hypothesis, and model 1 is better.

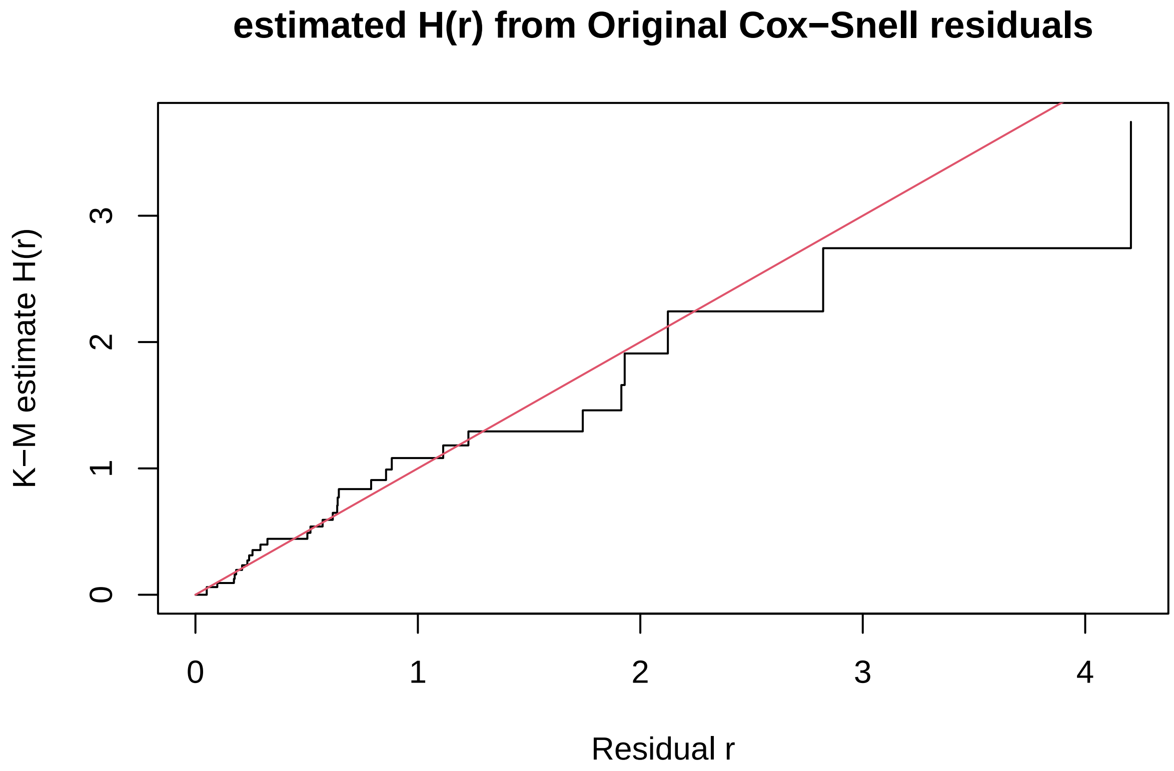
Therefore, we can conclude that model 1 with both AG and log(wbc) is the best parametric regression model for this dataset.

c) Cox-snell residual





The Cox-Snell residual plot has expected behavior, and look like exp(1) observation. therefor, model 1 fits pretty well.


estimated H(r) from Original Cox−Snell residuals

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. To clean the data, I used excel to deleted all extra recurrence time and replace empty with 0 for the patient without recurrence. Then we have a cleaned table like below.

```
##   Group futime number size recurrence_times
## 1     1      0      1    1                0
## 2     1      1      1    3                0
## 3     1      4      2    1                0
## 4     1      7      1    1                0
## 5     1     10      5    1                0
## 6     1     10      4    1                6
```

Now we can apply Cox regression model on the cleaned data frame.