

PROJECT 11 : Australian AIDS Survival Data.

I. **Abstract.**

The immune system is your body's natural defense system. This immune system will protect your body against infection and disease. This system is made up of many different cells that work together to find and destroy viruses, bacteria, and other germs that cause infection and disease. White blood cells (or CD4 T-cells) are important immune system cells that help coordinate your immune system.

HIV (Human Immunodeficiency Virus) is a typical lymphotropic retrovirus that infects cells of the immune system destroys or damages blood cells white. As long as the infection progresses, the system immune system becomes weak and sufferers to be more susceptible to infection. HIV attacks the immune system, which is the body's defense against disease. If a person's immune system has been damaged by a virus, he will develop AIDS (Acquired Immune Deficiency Syndrome). This means they will get infections and diseases which their bodies can normally fight against.

Level HIV in the body and various occurrences of certain infections are an indicator of that HIV infection has progressed to AIDS. AIDS is a cluster symptom or disease caused by decreased immunity due to HIV.

The annual HIV data released Monday (9/24/2018) by the Kirby Institute at the University of New South Wales shows that overall, Australia has made great strides in reducing new HIV cases. The latest data released today show Australia has made significant progress in reducing HIV rates, but in some populations, HIV cases are increasing.

Health authorities say a lot of work needs to be done to ensure HIV prevention and care measures reach everyone. There were 963 new cases of HIV infection in Australia. Australia has made a strong commitment to trying to end HIV, like many countries around the world, and in the last five years, there have been many strategies in place to try and reduce HIV in communities.

Until now, no drug can eliminate HIV infection from the body. However, the symptoms of the disease can be controlled and the immune system can be improved by administering antiretroviral therapy (ARV). ARV therapy cannot completely eradicate the virus, but it can help people with HIV live longer and healthier lives. Every person with HIV can live a healthy life and carry out normal activities while undergoing antiretroviral treatment. In addition, following treatment also helps reduce the risk of transmission, especially to those closest to you. ARV therapy consists of using a group of antiviral drugs that can reduce the amount of HIV in the body by preventing the virus from reproducing. The reduction in the virus provides an opportunity for the immune system to fight off viruses that cause damage to body tissues. That way, the amount of virus in the body can be controlled and the infection will not cause symptoms. In

addition, a low number of viruses means that the risk of transmission to other people is reduced.

Hopefully, with this analysis, it is possible to predict the estimated life expectancy of patients since diagnosed infected with this virus so that treatment can be optimally performed.

II. Introduction/Motivation.

In this era of development, promiscuity is very prevalent in the community, especially adolescents. This promiscuity harms the lives of adolescents, both in terms of religion, education, psychology, health, family, and community life. One of the bad effects of promiscuity is the emergence of HIV. HIV stands for Human Immunodeficiency Virus. This virus attacks the immune system and weakens the body's ability to fight infection and disease. HIV cannot be cured yet, but some treatments can be used to slow the progression of the disease. This treatment will also make the sufferer live longer, so they can lead a normal life. With early HIV diagnosis and effective treatment, people with HIV will not turn into AIDS. AIDS is the final stage of HIV infection.

By using survival analysis, health workers can make predictions about the likelihood that a patient infected with AIDS will survive how long it has been since the patient is diagnosed. Survival analysis corresponds to a set of statistical approaches used to investigate the time it takes for an event of interest to occur.

III. Methods.

Kaplan-Meier estimate is one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment. In clinical trials or community trials, the effect of an intervention is assessed by measuring the number of subjects who survived or saved after that intervention over some time. The time starting from a defined point to the occurrence of a given event, for example, death is called survival time, and the analysis of group data in survival analysis.

The Cox proportional-hazards model (Cox, 1972) is essentially a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more predictor variables.

The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time. This rate is commonly referred to as the hazard rate. Predictor variables (or factors) are usually termed *covariates* in the survival-analysis literature.

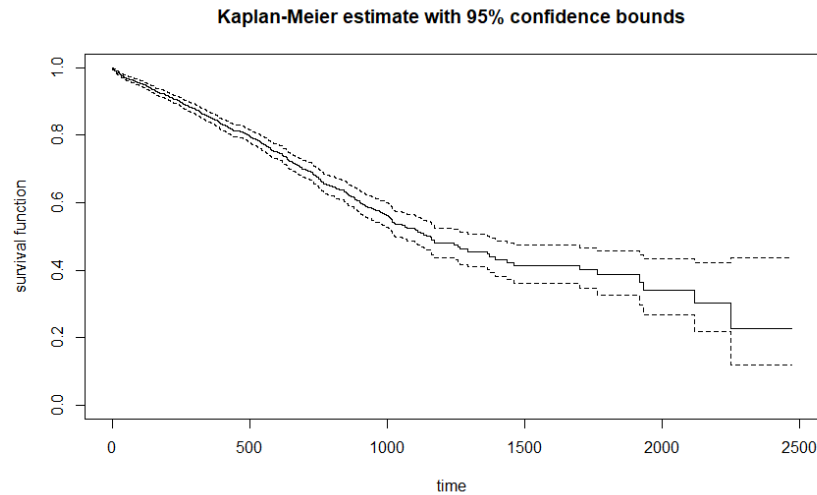
In the statistical area of survival analysis, an accelerated failure time model (AFT model) is a parametric model that provides an alternative to the commonly used proportional hazards models. Whereas a proportional hazards model assumes

that the effect of a covariate is to multiply the hazard by some constant, an AFT model assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. This is especially appealing in a technical context where the 'disease' is a result of some mechanical process with a known sequence of intermediary stages.

IV. Results.

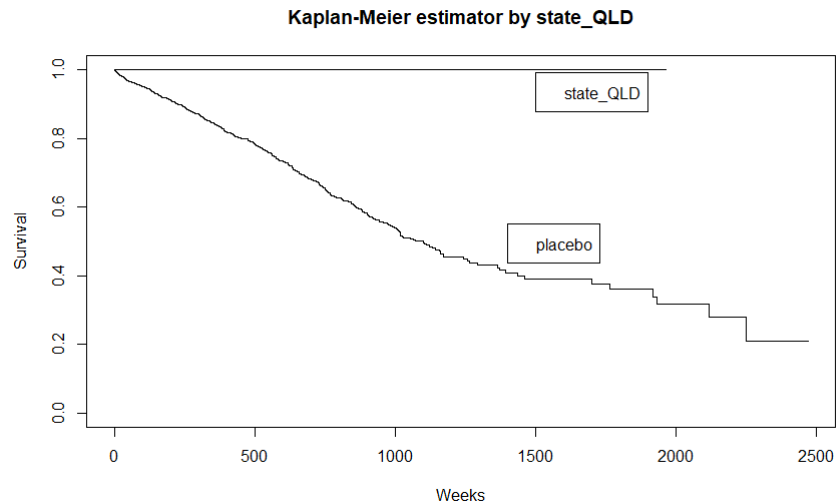
The given data is stated in picture 1. There is no null data in the dataset but since the *diag* and *death* variable is presented in the Julian calendar, I have to convert it into the Gregorian calendar system and calculate the duration of the survival time.

A. Kaplan Meier.



Picture 1. Kaplan Meier Probabilities Survival Curve

Based on picture 1, the y-axis represents the probability of the subject in the event of interest after surviving up to time t (days) represented on the x-axis. Each drop in the survival function (approximated by the Kaplan-Meier estimator with 95 % confidence interval) is caused by the event of interest happening for at least one observation. The height of the drop can also inform us about the number of observations at risk.



Picture 2. Kaplan Meier Probabilities Survival Curve Comparing state_QLD and state_VIC variable.

Based on Figure 2, in Group 1, the probability of survival for 2500 weeks is approximately 20%; conversely, if you are in Group 2, your probability of surviving the same time is 0%. The steepness of the curve is determined by the duration of survival (horizontal line length). the y-axis represents the probability of the subject in the event of interest after surviving up to time t (day), represented on the x-axis. Each drop in the survival function (approximated by the Kaplan-Meier estimator) is caused by the event of interest happening for at least one observation. The height of the drop can also inform us about the number of observations at risk.

```
> #Log-Rank test
> survdiff(Surv(duration, state_VIC) ~ state_QLD, rho=0) #log-rank or Mantel-Haenszel test
Call:
survdiff(formula = Surv(duration, state_VIC) ~ state_QLD, rho = 0)

      N Observed Expected (O-E)^2/E (O-E)^2/V
state_QLD=0 2617      588   546.3      3.18    44.9
state_QLD=1  226       0    41.7    41.69    44.9

Chisq= 44.9 on 1 degrees of freedom, p= 2e-11
> survdiff(Surv(duration, state_VIC) ~ state_QLD, rho=1) #Peto&Peto modification of the Gehan-wilcoxon test
Call:
survdiff(formula = Surv(duration, state_VIC) ~ state_QLD, rho = 1)

      N Observed Expected (O-E)^2/E (O-E)^2/V
state_QLD=0 2617      487   452.0      2.7    43.9
state_QLD=1  226       0    34.9    34.9    43.9

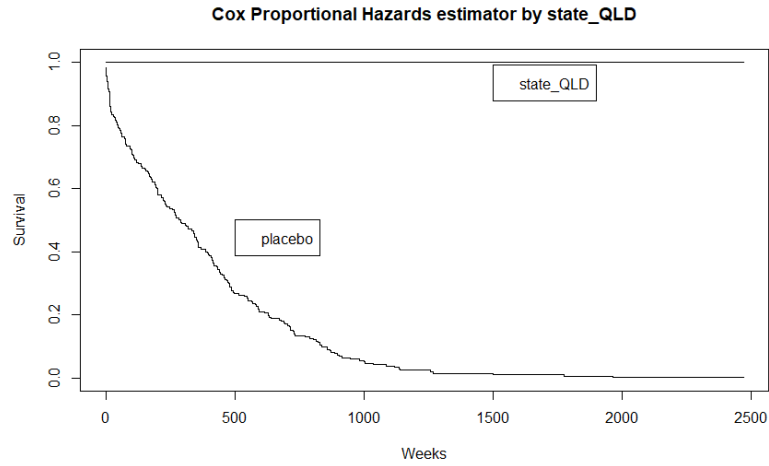
Chisq= 43.9 on 1 degrees of freedom, p= 4e-11
```

Picture 3. The Result of Log-rank Test using Kaplan Meier Method.

The log-rank test for difference in survival gives a p-value of $p = 2e^{-11}$ (Mantel-Haenszel test) and $p = 4e^{-11}$ (Peto & Peto modification of the Gehan-Wilcoxon test), indicating that the state groups differ significantly in survival. By looking at the p-value of the observation, we can see that there are no reasons to reject the null hypothesis stating that the survival functions are **not identical**. The chi-square statistic for a test of

equality which gives almost the same quality and both of the test has the same strata which are 1 degree of freedom.

B. Cox Proportional Hazards.



Picture 4. The Summary Result of Cox Proportional Hazards Model.

Based on Figure 4, in Group 1 (People who get infected in Queensland state), the probability of survival for 2500 weeks is approximately 100%; conversely, if you are in Group 2 (People who get infected in Victoria state), your probability of surviving the same time is 0%. The steepness of the curve is determined by the duration of survival (horizontal line length). the y-axis represents the probability of the subject in the event of interest after surviving up to time t (day), represented on the x-axis. Each drop in the survival function (approximated by the Cox Proportional Hazards estimator) is caused by the event of interest happening for at least one observation. The height of the drop can also inform us about the number of observations at risk.

$$HR(t, \mathbf{X}_i, \mathbf{X}_j) = \frac{h(t, \mathbf{X}_i)}{h(t, \mathbf{X}_j)}$$

	coef	exp(coef)
diag	0.00	1.00
death	-0.01	0.99
state_Other	0.06	1.06
state_QLD	0.06	1.06
state_VIC	0.01	1.01
sex_M	-0.01	0.99
status_D	-0.22	0.80
T.categ_haem	-0.36	0.70
T.categ_het	-0.42	0.66
T.categ_hs	-0.42	0.66
T.categ_hsid	-0.35	0.70
T.categ_id	-0.56	0.57
T.categ_mother	-0.54	0.58
T.categ_other	-0.22	0.80

Picture 5. Hazards Ratio calculated using Python.

```

Call:
coxph(formula = my.survival.object ~ 1 + as.factor(state_QLD),
      method = "breslow")

n= 2843, number of events= 226

              coef exp(coef)  se(coef)      z Pr(>|z|)
as.factor(state_QLD)1  2.279e+01 7.886e+09 1.648e+03 0.014    0.989

              exp(coef) exp(-coef) lower .95 upper .95
as.factor(state_QLD)1  7.886e+09  1.268e-10      0      Inf

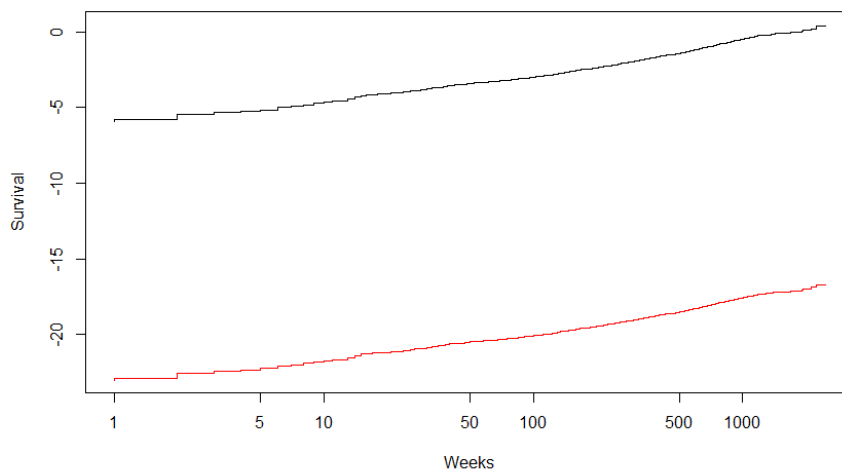
Concordance= 0.964 (se = 0.003 )
Likelihood ratio test= 1187 on 1 df,  p=<2e-16
Wald test               = 0 on 1 df,  p=1
Score (logrank) test = 2890 on 1 df,  p=<2e-16

```

Picture 5. Parameters of The Semiparametric Cox PH Model.

Based on the Wald test, it gives a conclusion that the model does not give a significant result because the p-value is greater than 5%. The hazard for the state_QLD group is 7.886 times the hazard for the state_VIC group. The value 7.886 is calculated as $\exp(\text{coefficient of the state_VIC variable})$; that is, e to the 2.279 equals 7.886. The

95% confidence interval for the treatment effect is given by the range of values 0 – Inf.

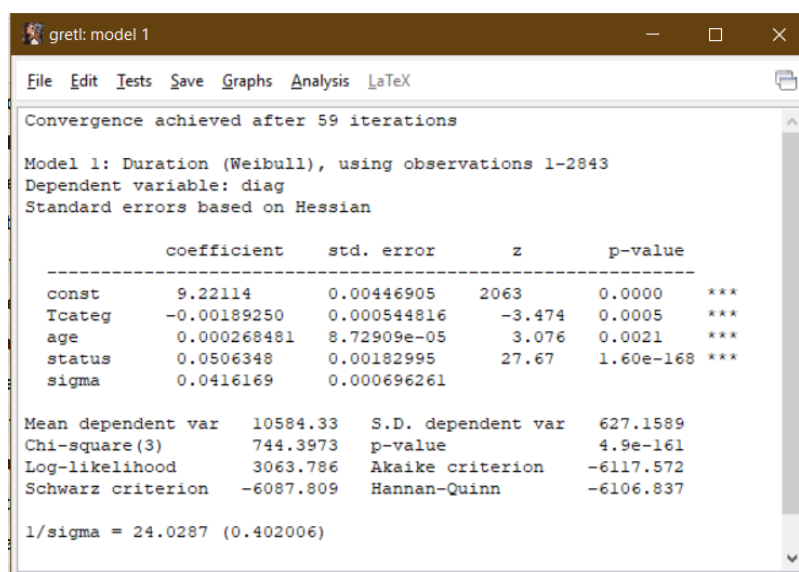


Picture 6. Proportional Hazards Model using Log-log Plots.

Since the log-log plots give the results a parallel line. It means that the Cox Proportional Hazards model is appropriate.

C. Parametric Proportional Hazards (PH) and Acceleration Failure Time (AFT).

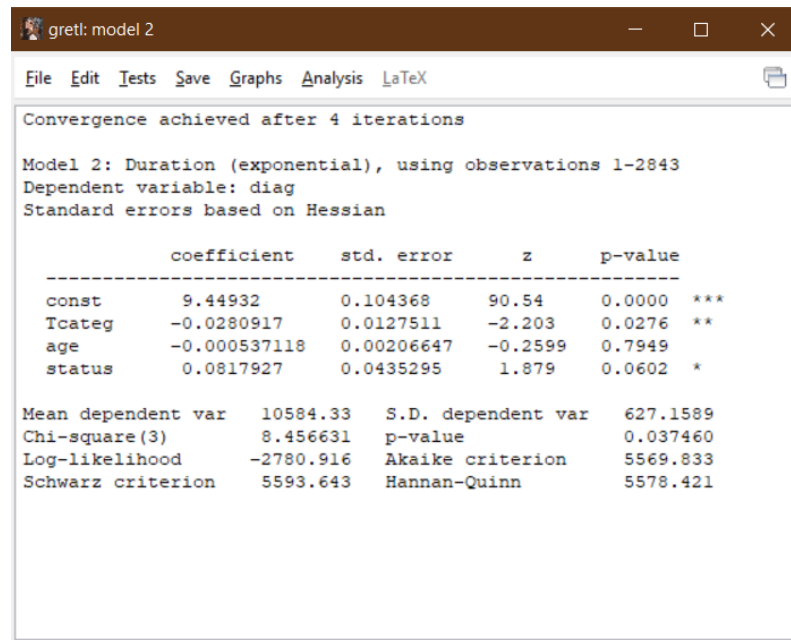
1. Weibull Proportional Hazards Model.



Picture 7. Weibull Proportional Hazards Model.

Based on picture 7, The risk got infected was 99.9% higher by transmission category (because $\exp(-0.001) = 0.999$), ceteris paribus. The shape parameter $p=24.02>1$, which means increasing hazard (as time goes on, the got infected increase).

2. Exponential Proportional Hazards Model.



gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

Convergence achieved after 4 iterations

Model 2: Duration (exponential), using observations 1-2843
Dependent variable: diag
Standard errors based on Hessian

	coefficient	std. error	z	p-value	
const	9.44932	0.104368	90.54	0.0000	***
Tcateg	-0.0280917	0.0127511	-2.203	0.0276	**
age	-0.000537118	0.00206647	-0.2599	0.7949	
status	0.0817927	0.0435295	1.879	0.0602	*
Mean dependent var	10584.33	S.D. dependent var	627.1589		
Chi-square(3)	8.456631	p-value	0.037460		
Log-likelihood	-2780.916	Akaike criterion	5569.833		
Schwarz criterion	5593.643	Hannan-Quinn	5578.421		

Picture 8. Exponential Proportional Hazards Model.

Based on picture 8, The risk got infected was 97.2% higher by transmission category (because $\exp(-0.028) = 0.972$), ceteris paribus.

Calculate the AIC value :

Weibull : $-2 * 3063.786 = -6127.57$

Exponential : $-2 * (-2780.916) = 5561.832$

Based on that calculation, we can conclude that the Weibull model is better than the Exponential model.

In Weibull model most of the variable are statistically significant and also the p-value of this model is statistically significant meanwhile in Exponential model there is one variable which is not statistically significant and we can assume this model is statistically significant because the p-value of this model is less than 5%.

V. Discussion/Conclusion

1. In the CoxPH model, hazard ratio $\exp(\beta)$ showing the percentage of increase/decrease in the hazard.
2. Kaplan-Meier will give you an estimation of the survival curve only.
3. The Kaplan Meier Curve is the visual representation of this function that shows the probability of an event at a respective time interval.
4. The log-rank test provides a statistical comparison of two groups.
5. Both of the latter two methods assume that the hazard ratio comparing two groups is constant over time.
6. The Kaplan–Meier estimator, also known as the product-limit estimator, is a non-parametric statistic used to estimate the survival function from lifetime data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment.

VI. References

Goel, Manish Kumar, Pardeep Khanna, and Jugal Kishore¹. **2010**. *“Understanding Survival Analysis: Kaplan-Meier Estimate”*. Oktober-December, 2010. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/>

Venables, W. N., and Ripley, B. D. **2002**. *Modern Applied Statistics with S*. Fourth edition. Springer.

Anonim. <http://www.sthda.com/english/wiki/cox-proportional-hazards-model>.

Anonim. <http://www.sthda.com/english/wiki/survival-analysis-basics>