Survival Analysis 'Critcare' data project

Data allocation – 037

Chapters

Chapter 1 – Overview of Data including survival probability
Chapter 2 – A comparison of the survival experience testing dependencies
Chapter 3 – A comparison of the survival experience testing invasive ventilation
Chapter 4 – A Cox proportional hazards fit, looking at age, gender and bmi
Chapter 5 – The survival probability of a future 50-year-old male
Chapter 6 – Full analysis of data testing all covariates

Chapter 1

Aim:

In this chapter I will aim to provide an overview of the data and the different covariates involved. I will also calculate a 30-day survival probability using the Kaplan-Meier estimator.

Overview of data:

The data used for this project is a subset of a complete dataset looking at the survival of 10,928 COVID-19 patients receiving critical care during the early months of the pandemic.

The variables considered in the dataset are:

Time - Days since the start of critical care
Status - Indicator of death (1) or censoring (0)
Age - Age in years
Female 1 if female, 0 otherwise
Bmi - Body mass index
Comor - 1 if comorbidities present, 0 otherwise
Invent - 1 if invasive ventilation required during first 24 hours, 0 otherwise
Depend - 1 if patient had dependency prior to hospital admission, 0 otherwise
Depriv - Deprivation index of postcode, from 1 (least deprived) to 5 (most deprived)
Apache – Apache II risk score from patient characteristics and biomarkers, with high values for higher risk

An important factor which should be explained is the censoring, in the variable **status**. Censoring occurs when we lose track of the patient before the event time. This is a problem with survival analysis as treating censored observations as if they were events will lead to biased results.

To gain a better understanding of the covariates I have provided a table of the mean and standard deviations of the non-binary variables:

	Mean	**SD**	**Min**	**Max**
Time	14.96	10.06	1.00	30.00
Age	58.61	14.52	17.00	98.00
Bmi	30.81	6.67	16.90	44.90
Apache	28.25	10.80	5.00	53.00

What can be inferred from this table is that the average number of **time** (days) since the start of critical care up to the event is 15 days. However, the standard deviation is very large indicated a large spread. Therefore, it may also be useful to note that the mode for this data is 30 days, showing that the majority of people did survive at least 30 days.

The average **age** of the patients in this dataset is 59 years old and thereby it is expected that 68% of the patients are in the age range of 44 and 73. Therefore it can be seen that the majority of people that need critical care are typically the older generation compared to the average population.

The average **bmi** of the patients is 31 which is considered overweight. However, bmi is just based on height and weight and does not take into account muscle mass / fat ratio. Therefore, I feel summarising the data as typically overweight people are in need of critical care more than others should be treated with caution.

The **apache** covariate is a risk score from patient characteristics and biomarkers, with high values for higher risk. The average apache risk score for the patients of this data is 28 and to give further information, we expect 68% of the patients to have a risk score in the range of 17 to 49.

Now looking at the binary variables:

	Proportion of 0	**Proportion of 1**
Status	0.6	0.4
Female	0.685	0.315
Comor	0.903	0.097
Invent	0.467	0.533
Depend	0.91	0.09

From this table, it can be seen that 60% of the data has censoring, meaning we have lost track of the patient before the event time. This can happen for a number of reasons, i.e. patient has moved or not available to follow-up.

Looking at the **female** covariate it can be seen that 69% of the patients are male. Therefore, based on this data, we can say that typically males are in greater need of critical care than females.

Comorbidity occurs when a patient has more than one disease/condition at the same time. From the data, it can be seen that roughly 10% of the patients have more than one disease/condition at the same time.

Invasive ventilation is when ventilation is required through an invasive airway, such as a mechanical ventilation machine. This table tells us that 53% of the patients required invasive ventilation during the first 24 hours of critical care.

We can also see that roughly 9% of patients had **dependencies** prior to hospital admission

Now looking at the final covariate **Depriv**:

Deprivation index	**1**	**2**	**3**	**4**	**5**
Proportion	0.16	0.15	0.18	0.26	0.25

Depriv is looking at a deprivation index of the patients' postcode where 1 indicates the least deprived area compared to 5 which is the most deprived area. From the data, we can see the majority of patients are from a deprived area index of 4 with an average of 3.3. This indicates that typically people that live in a poorer/more-deprived areas are more likely to need to receive critical care.

Survival probability:

To calculate a 30-day survival probability, I will be using the Kaplan-Meier estimator. This is a nonparametric estimator that does not make any assumptions about the underlying probability distribution of the survival times. Instead, it estimates the survival function based on the observed survival times in a sample.

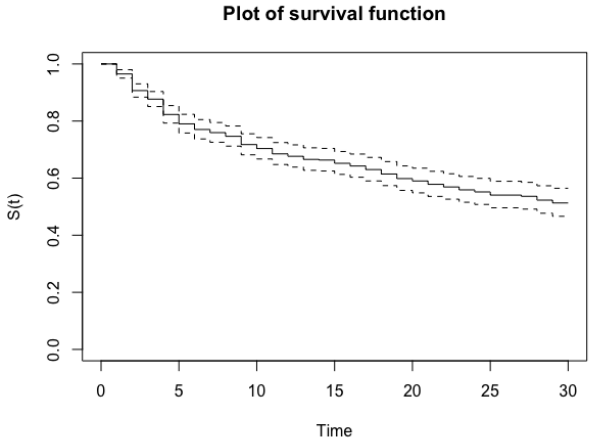
The Kaplan-Meier estimator works by calculating the probability of surviving at each point in time, taking into account the number of individuals who are still at risk of experiencing the event.

However, there are some assumptions this estimator makes;

- The survival times are independently and identically distributed
- The censored observations are non-informative

Using R, I calculated that the estimated 30-day survival probability of a patient is 0.513. Censoring in this dataset is fairly high (60%), therefore, I used a 'log-log' interval to achieve a 95% confidence interval: (0.463, 0.561).

This means that there is a 51.3% chance that a patient will survive at least 30 days after starting critical care. However, it is important to note that this probability is specific to this dataset alone.



Testing significance of covariates:

Before looking into the data further, I performed a test to investigate the different covariate effects. I did this by fitting a cox proportional hazards model using all covariates.

Using R, I obtained the following p-values;

	P-values	**Significance code**
Age	1.22e-13	***
Female	0.000126	***
Bmi	8.76e-07	***
Comor	0.394050	
Invent	0.550094	
Depend	0.362649	
Depriv	0.045750	*
Apache	0.030964	*

From this table, we can clearly see that **age**, **female** and **bmi** are highly significant while **depriv** and **apache** are also significant, albeit to a lesser degree.
The other covariates **comor**, **invent** and **depend** do not appear to be significant, however, we will investigate this further later on in this report.

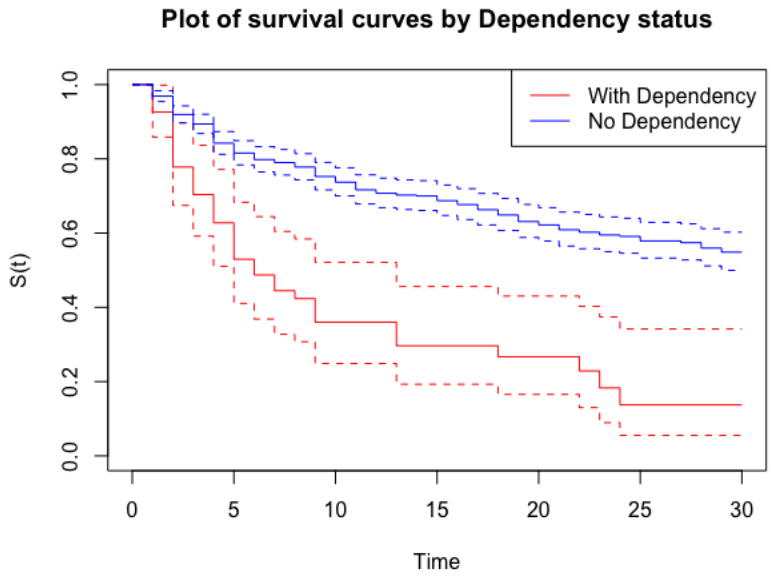
Chapter 2

Aim:

In this chapter I aim to compare the survival experience of patients with and without dependencies prior to their admission. To do this, I will be performing a log-rank test. It is important to note that for this chapter I will be ignoring other risk factors.

Log-Rank test:

The plot below shows the Kaplan-Meier estimators of patients with and without dependencies.



Performing the log-rank test:

H_0: Survival curves are the same
H_1: Survival curves are different

From R, we obtain a test-statistic of 45.8 and p-value of 1e-11 which is extremely small, therefore there is very strong evidence to reject the null hypothesis. We can conclude that there is an extremely significant difference in survival experience between patients with and without dependencies prior to critical care admission.

From the Kaplan-Meier estimators obtained, I was able to compare the 30-day survival probability.
For a patient with dependency = 0.137
For a patient without dependency = 0.549

I therefore concluded that a patient without dependencies has a 55% chance of surviving at least 30 days since critical care. Whereas, a patient with dependencies only has a 14% chance of surviving at least 30 days since critical care.

Conclusion:

The data suggests that there is strong evidence of a significant difference in survival experience between patients with and without dependencies. A patient without dependencies is four times more likely to survive at least 30 days since their admission to critical care, compared to a patient with dependencies.

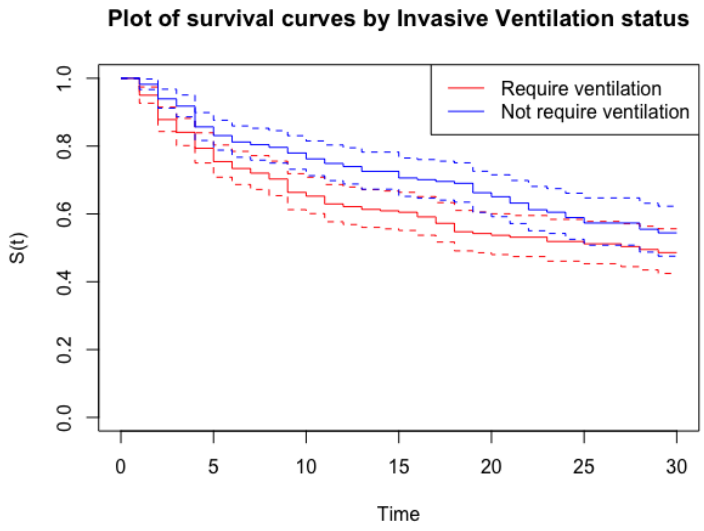
Chapter 3

Aim:

In this chapter I aim to compare the survival experience of patients who did and did not require invasive ventilation in the first 24 hours. To do this, I will be performing a log-rank test. It is important to note that for this chapter I will be ignoring other risk factors.

Log-Rank test:

The plot below shows the Kaplan-Meier estimators of patients who required and did not require invasive ventilation in the first 24 hours.



Performing the log-rank test:

H_0: Survival curves are the same
H_1: Survival curves are different

From R, I obtained a test-statistic of 5.7 and p-value of 0.02, therefore there is only some evidence to reject the null hypothesis. We can conclude that there is only a slight significant difference in survival experience between patients who did and did not require invasive ventilation within the first 24 hours of critical care.

From the Kaplan-Meier estimators obtained, I was able to compare the 30-day survival probability.
For a patient who did require invasive ventilation = 0.486
For a patient who did not require invasive ventilation = 0.544

We can therefore conclude that a patient who does not require invasive ventilation during the first 24 hours of critical has a 54% chance of surviving at least 30 days since critical care. Whereas, a patient who does require invasive ventilation during the first 24 hours has a slightly less 49% chance of surviving at least 30 days since critical care.

Conclusion:

The data suggests that there is some evidence of a significant difference in survival experience between patients with and without dependencies. A patient who did not require invasive ventilation has roughly a 5% higher chance of surviving at least 30 days since critical care.

Chapter 4

Aim:

In this chapter I aim is to investigate the covariates **age**, **female** and **bmi** using the Cox proportional hazards model.

Cox proportional hazards fit:

The Cox proportional hazards model is a semi-parametric model that allows us to estimate the effect of different covariates on the hazard rate. The model assumes that the hazard ratio is constant over time and it does not make any assumptions about the distribution of survival times. I will be confirming the hazard ratio assumption further in this document.

Using R to investigate the covariates **age**, **female** and **bmi**, I obtained the following coefficients and p-vales;

	Coefficient	**Exp(coef)**	**P-value**
Age	0.05003	1.05130	2e-16
Female	-0.54637	0.57905	0.000413
Bmi	0.06863	1.07104	1.27e-12

Inspection of this shows that all 3 covariates are highly significant.
Increasing **age** leads to an increased risk at a rate of around 5% per increasing year.
Females typically do better than males having a hazard rate more than half the value of males, (58% lower).
With each increasing unit in **bmi**, the hazard rate will increase by around 7%.

Chapter 5

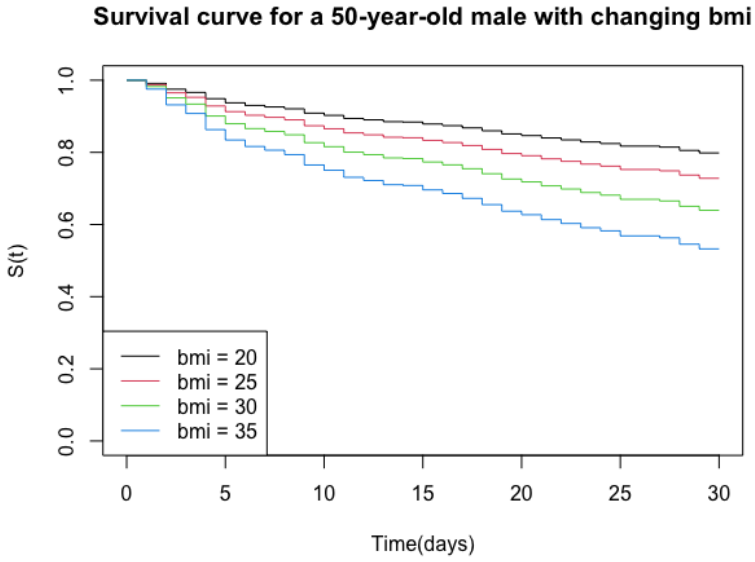
Aim:

In this chapter, my aim to discover how **bmi** affects the estimated 30-day survival probability for a future 50-year-old male.

Survival probability:

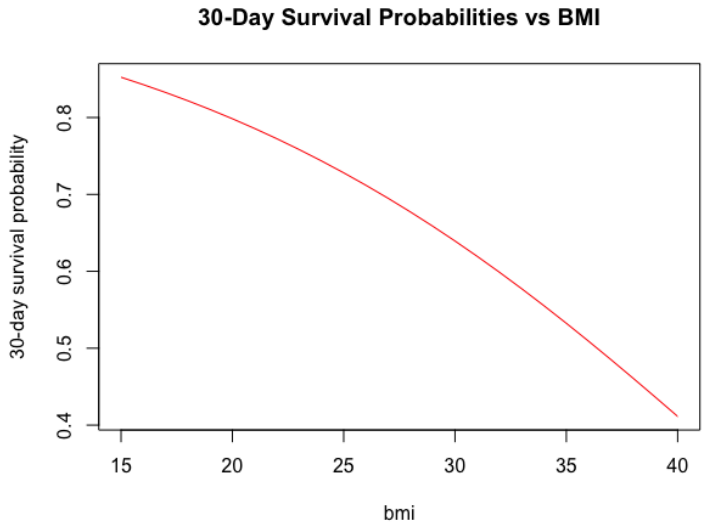
Using R, I was able to obtain the following 30-day survival probabilities for a 50-year-old male:

Bmi	**20**	**25**	**30**	**35**
30-day Survival Probability	0.798	0.728	0.639	0.532
Difference		0.07	0.089	0.107



From this we can see that an increase in bmi leads to a lower chance of survival, confirming our observations about bmi in the previous chapter.

From the 'difference' row, it appears that the relationship between bmi and survival probability is non-linear. We can infer that the higher the patients bmi, their survival probability becomes exponentially lower.
Having found this interesting relationship, I wanted to investigate further by producing a plot of 30-day survival probabilities against different bmi's



From this plot, we can clearly see that **bmi** and survival probability have a negative non-linear relationship. The general trend confirms our observation that having a high bmi leads to a lower survival probability.

Chapter 6

Aim:

In this chapter, my aim is to carry out a full analysis of the data taking into account all covariates.

Cox proportional hazards with all covariates:

Using R, I obtained the following information from a Cox proportional hazards fit.

	Coefficient	Exp(coef)	P-value	Significance
Age	0.043667	1.044635	1.22e-13	***
Female	-0.599262	0.549217	0.000126	***
Bmi	0.055439	1.057004	8.76e-07	***
Comor	0.172429	1.188188	0.394050	
Invent	0.081820	1.085260	0.550094	
Depend	0.178989	1.196007	0.362649	
Depriv	0.093925	1.098477	0.045750	*
Apache	0.015089	1.015203	0.030964	*

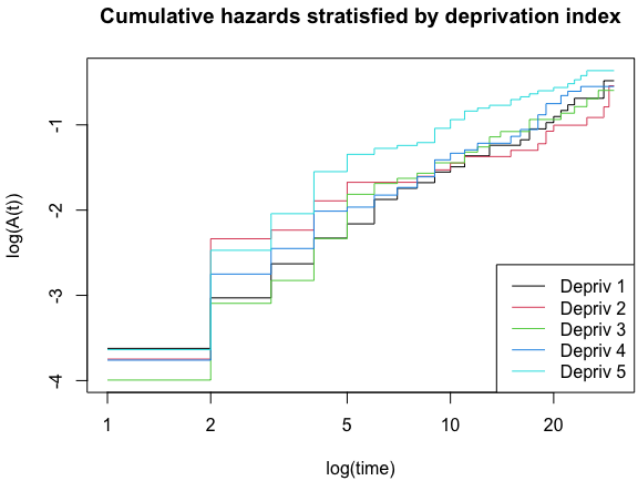
As previously mentioned in Chapter 1, we can clearly see that **age**, **female** and **bmi** are highly significant while **depriv** and **apache** are also significant, albeit to a lesser degree. The covariates **comor**, **invent** and **depend** are not significant, although they have higher coefficients. This is because their p-values are very high which suggests that the observed relationship between these covariates and time could have occurred by chance.

This shows the higher your score on every variable leads to a lower survival probability, except the gender variable where females typically do better than males.

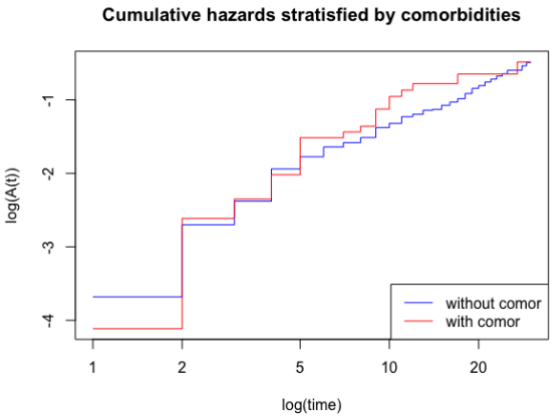
Having performed a log-rank test on the variables **depend** and **invent** in previous chapters and finding that both are significant in accepting the null hypothesis, it was interesting that they do not seem to be important in the proportional hazards fit.

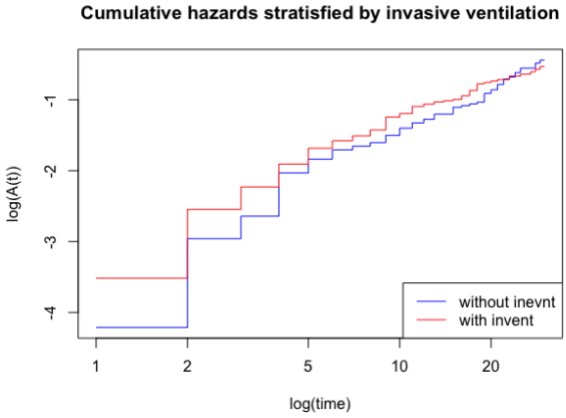
However, it is very important we understand the difference between the log-rank test and the proportional hazards model. The log rank test is testing whether there is a significant difference in the survival curves for the particular covariate. Whereas, our proportional hazard model measures the relationship between time and all of the covariates.

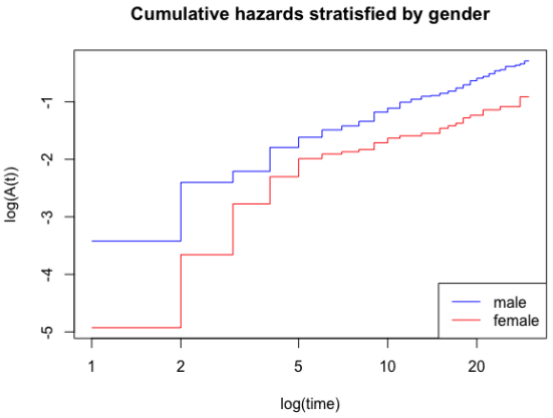
Checking the proportional hazard assumption:

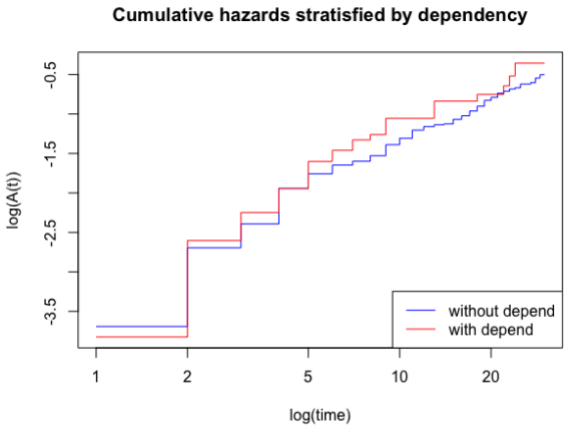


From the graph, we can see that there are five roughly parallel lines, with one of them slightly higher than the others. This suggests that the proportionality assumption of the Cox proportional hazards model holds for **depriv**. However, it may be sensible to reclassify the deprivation index into two groups: most deprived (5) and the rest (1-4) as there appears to be a difference in survival between these groups.





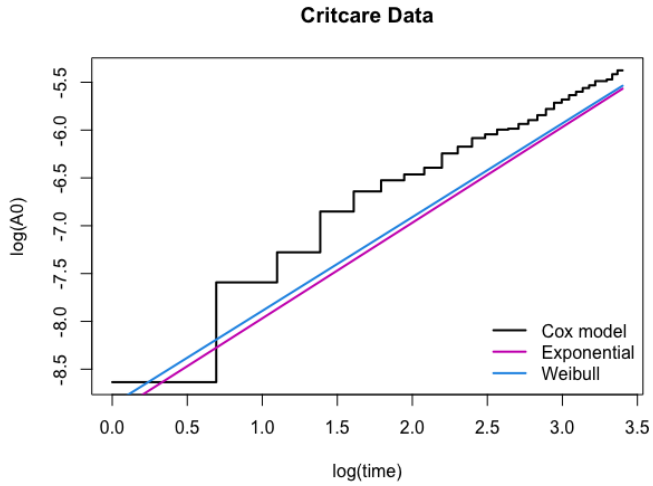




From the above graphs, we can see the lines are roughly parallel for all the covariates, indicating that the proportionality assumption does hold. For the other covariates; **age**, **bmi** and **apache**, these plots would not be appropriate as their data is continuous. These will be checked using Schoenfeld residuals further in this chapter.

Comparison with parametric models:

We can check whether a parametric model is suitable for analysing the 'critcare' survival data by plotting the log cumulative hazard function against log time.
If the Weibull model is suitable we would expect a straight line and if the Exponential is suitable, we would expect the line to have a unit slope.



We can see from this plot that the Weibull and Exponential both have straight lines close to the Cox model. The Weibull scale parameter 1.02 is very close to 1, further supporting that the Exponential is also suitable. While the parametric models seem like a good fit, both lines consistently underestimate the hazard function relative to the Cox model. Therefore, suggesting that the parametric models may not be appropriate for the data.

Reducing the model:

AIC (Akaike Information Criterion)– This is used to choose which model best balances complexity and the goodness of the fit. Smaller values are preferred as they indicate a better trade-off between complexity and accuracy. Having too many variables can lead to overfitting.

C-index – The concordance is a measure of how well the model can predict. i.e A higher risk person will die before a lower risk person. Higher values indicate a better predictive capability.

Schempers R-squared – Measures the proportion of explained variation in survival probabilities. This value is often low for survival data, even when covariates are highly statistically significant.

Here are the results of a backwards stepwise procedure on the 'critcare' data.

Model	**Covariates**	**AIC**	**C-Index**	R^2
1	age + bmi + female + comor + invent + depend + depriv + apache	2723.42	0.721	19.38
2	age + bmi + female + comor + depend + depriv + apache	2721.78	0.722	19.29
3	age + bmi + female + depend + depriv + apache	2720.50	0.722	19.22
4	age + bmi + female + depriv + apache	2719.24	0.722	19.18
5	age + bmi + female + apache	2721.56	0.718	18.61
6	age + bmi + female	2723.38	0.715	18.27
7	age + bmi	2735.06	0.703	16.80
8	age	2780.38	0.674	11.04

From Schempers R-squared, models 1-4 seem to be appropriate. Model 4 has the lowest AIC and the highest Concordance, indicating that this is the best model to use. It is worth noting that this model only contains the significant covariates from the original Cox proportional hazard model.

Likelihood ratio test on the reduced model:

Preforming the likelihood ratio test between the full model and the reduced model.

Log-likelihood on full model = -1353.711 (df=8)
Log-likelihood on reduced model = -1354.621 (df=5)

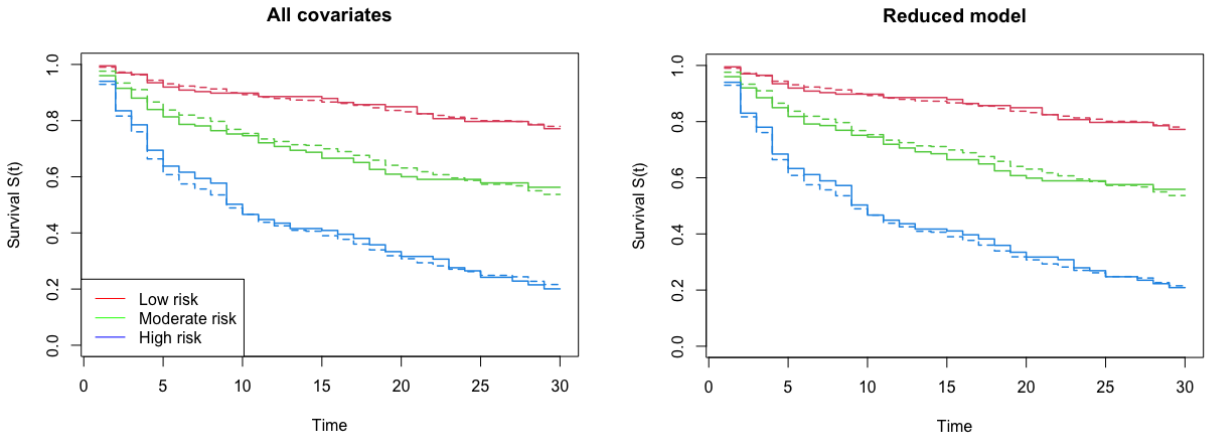
Test statistic = 1.82
Comparing with chisq distribution gives p-value = 0.61

As 0.61 > 0.05 this indicates that the reduced model is preferred, supporting our conclusion above.

Grouping by prognostic index for full and reduced model:

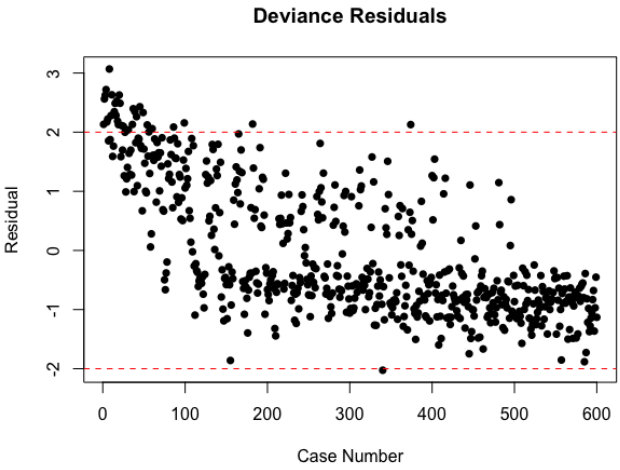
I will be using the grouping by prognostic index method to assess the fit in the full and reduced model.

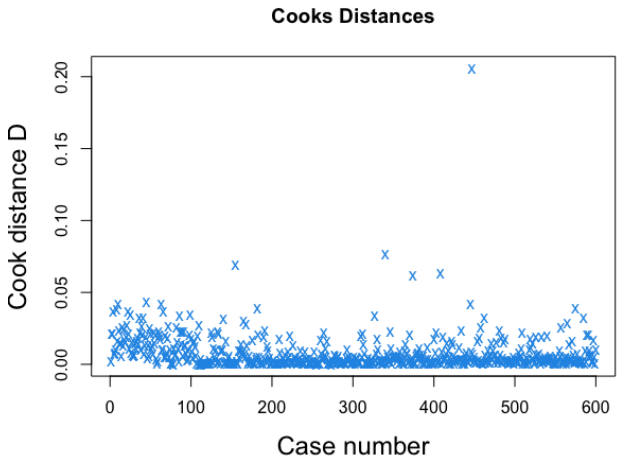
Grouping the data into high/moderate/low risk based on their prognostic index, I have plotted their survival curves.



We can see that the fit is very good in both models as the observed values are very close to the model values. Although the improvement is very minimal, the reduced model appears to fit slightly better for the higher risk group. It is worth noting that initially the fit underestimates the survival probability for the higher risk group and overestimates the survival probability for the moderate risk group in both models.

Testing for outliers/influential observations:





From the left-side plot, we can see there are clearly some outliers that are more than two standard deviations from zero. There are 33 values greater than +2 indicating that the event occurred sooner than we would expect and only 1 value less than -2, indicating that the event occurred later than expected. Overall, we can say that our model typically predicts a higher survival probability.

From the right-side plot, we can see there is one obvious influential point, case number 447. This individual has the following data:

	time	**age**	**female**	**bmi**	**comor**	**invent**	**depend**	**depriv**	**apache**
Individual	24	91	1	36.4	0	0	1	5	46

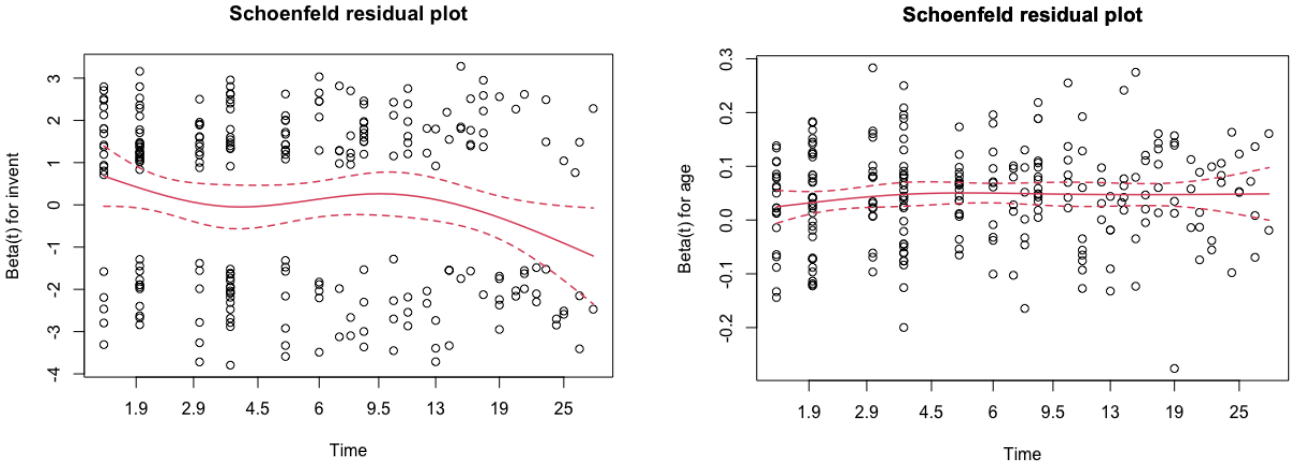
This individual has a very high **age**, **bmi**, **apache** score, had **dependency** prior to hospital admission and lives in a **deprived** area. All of which are associated with a short survival time, however, this individual survived much longer than expected (24 days).

When removing this outlier from the data, all the covariates excluding **comor** and **invent** become a little more significant, as we don't include the unusual results from this person. Removing this outlier does not lead to any change in conclusions, which is expected as the data set is fairly large (n = 600) and a single case in unlikely to dominate this.

I will now be looking at Schoenfeld residuals to test for linear trends against time.

	Test-statistic	**p-value**
age	0.590	0.441
female	0.009	0.924
bmi	1.743	0.187
comor	0.007	0.931
invent	2.917	0.088
depend	0.125	0.724
depriv	1.295	0.255
apache	0.058	0.809

This suggests that there is no significant trend for any of the covariates, except **invent,** as the p-value 0.088 < 0.1.



From a Schoenfeld residual plot we would expect the red line to be roughly constant if the model is suitable, seen in the right-side **age** plot. We can see in the left-side **invent** plot there is curvature indicating there could be a time-varying effect.
This suggests that whether a person did or did not require invasive ventilation during the first 24 hours, has less of an effect as time goes on. This could seem appropriate as the patient's condition may stabilise after essentially 'surviving' the initial ventilation phase. It appears that, from the plot, this phase is roughly 10 days.
This time-varying effect is statistically supported with at least some evidence, as the p-value 0.088 < 0.1. In survival data, we typically expect a larger spread at the end as there are fewer observations.

Final conclusion:

From the dataset, we can conclude that the covariates **age, bmi** and **female** are the most important factors when looking at the survival times of patients undergoing critical care. We should therefore pay close attention to these variables and be aware that typically older males who are not in great shape, may require more attention when testing positive for COVID-19.

It is also worth noting that the risk score covariate **apache**, which looks at variables such as blood pressure and heart rate etc, is also significant. Therefore, we should encourage people to live a healthier lifestyle to increase their chance of surviving.

The covariate **invent**, if invasive ventilation was required during first 24 hours, does not have an overall significant affect. However, it is worth noting that there is a time-varying affect. It seems that this covariate is much more important in the first 10 days. Therefore, future patients requiring invasive ventilation should be treated with higher caution during the earlier stage of their admission into critical care.