Medical Statistics Assignment 2

Stephen Brownsey

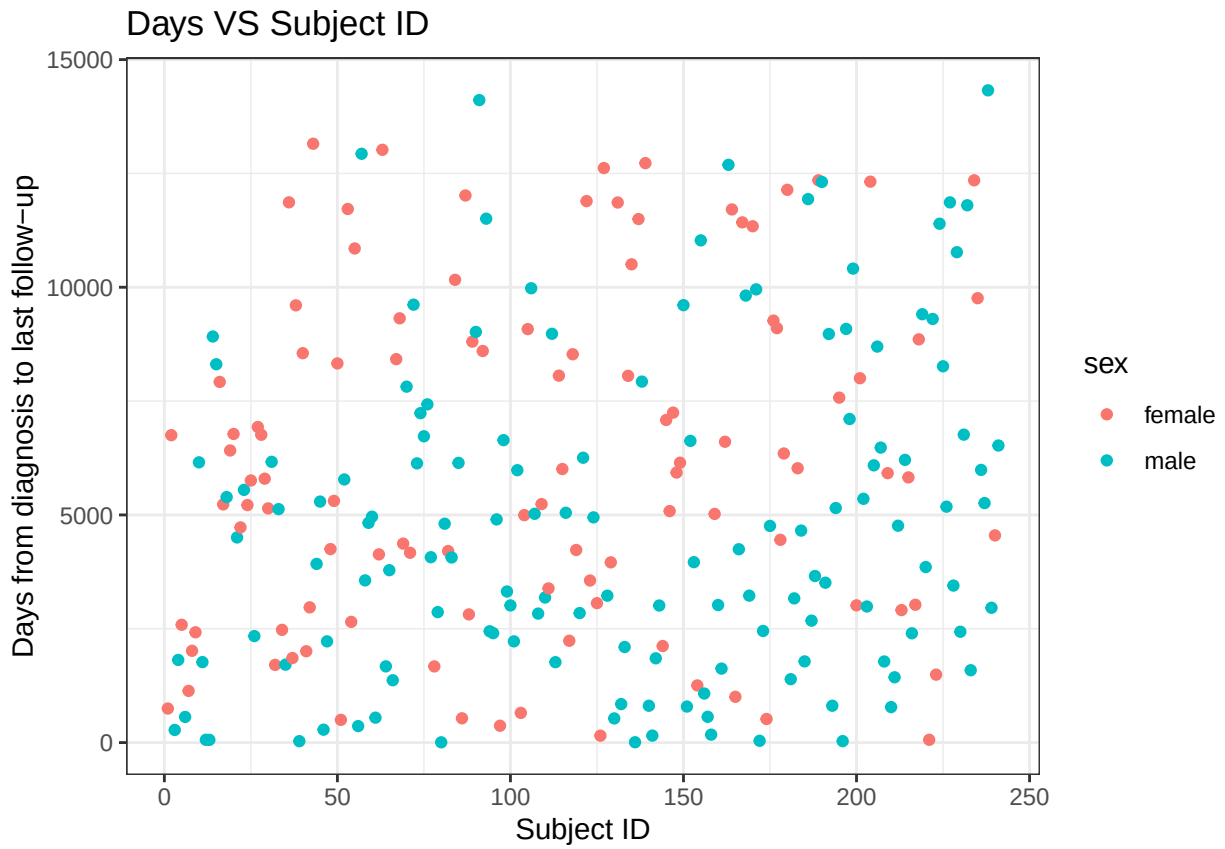
Question 1:

The plots in this section will be used to visualise the mgus data from a couple of different perspectives, each of which will be coloured by sex to demonstrate the slight differences between genders.

```
#Load the mgus data
data(mgus)
head(mgus)
```

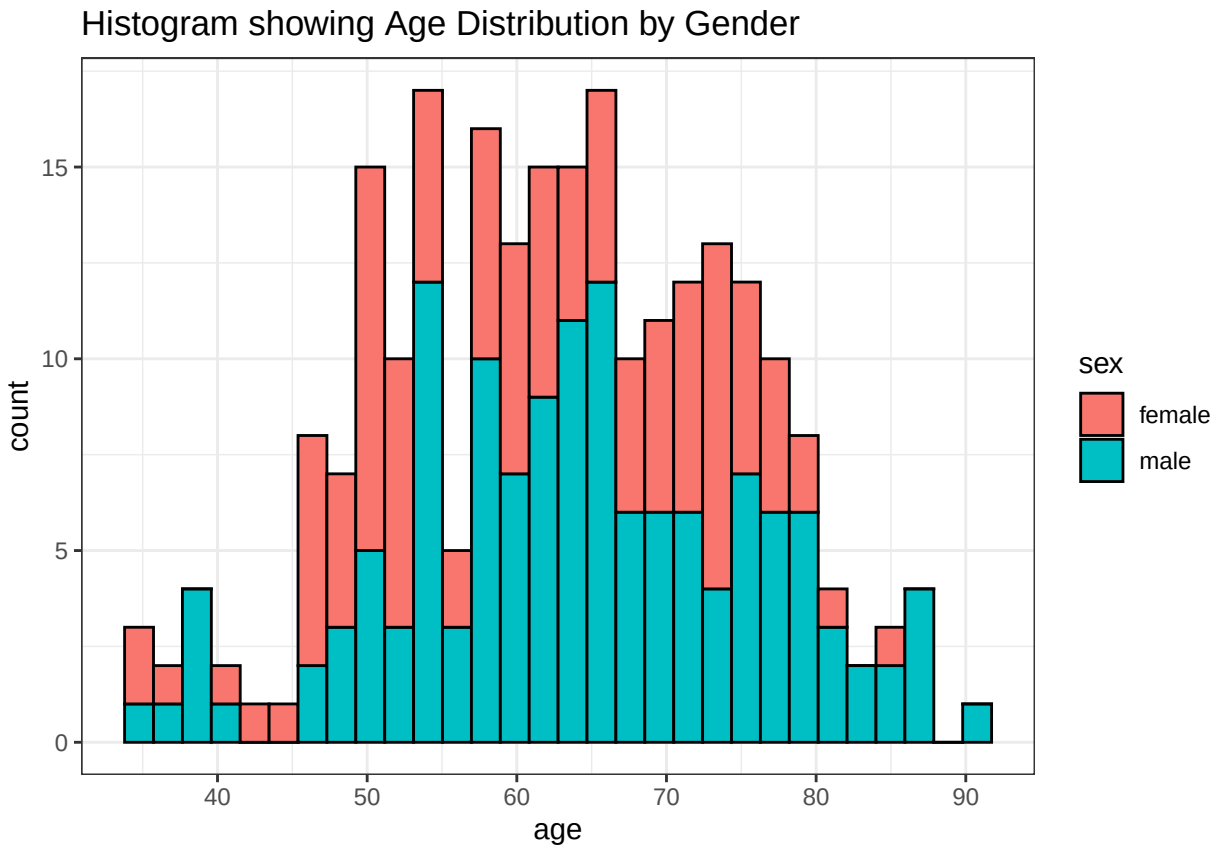
```
##   id age    sex dxyr pcdx pctime futime death alb creat  hgb mspike
## 1  1  78 female   68 <NA>     NA    748     1 2.8   1.2 11.5    2.0
## 2  2  73 female   66   LP   1310   6751     1  NA    NA   NA    1.3
## 3  3  87   male   68 <NA>     NA    277     1 2.2   1.1 11.2    1.3
## 4  4  86   male   69 <NA>     NA   1815     1 2.8   1.3 15.3    1.8
## 5  5  74 female   68 <NA>     NA   2587     1 3.0   0.8  9.8    1.4
## 6  6  81   male   68 <NA>     NA    563     1 2.9   0.9 11.5    1.8
```

```
#Data points showing Days till last follow up against subject ID, coloured by Sex
mgus %>%
  ggplot() +
  geom_point(aes(x = id, y = futime, col = sex)) +
  theme_bw() +
  labs(y = "Days from diagnosis to last follow-up",
       x = "Subject ID", title = "Days VS Subject ID")
```

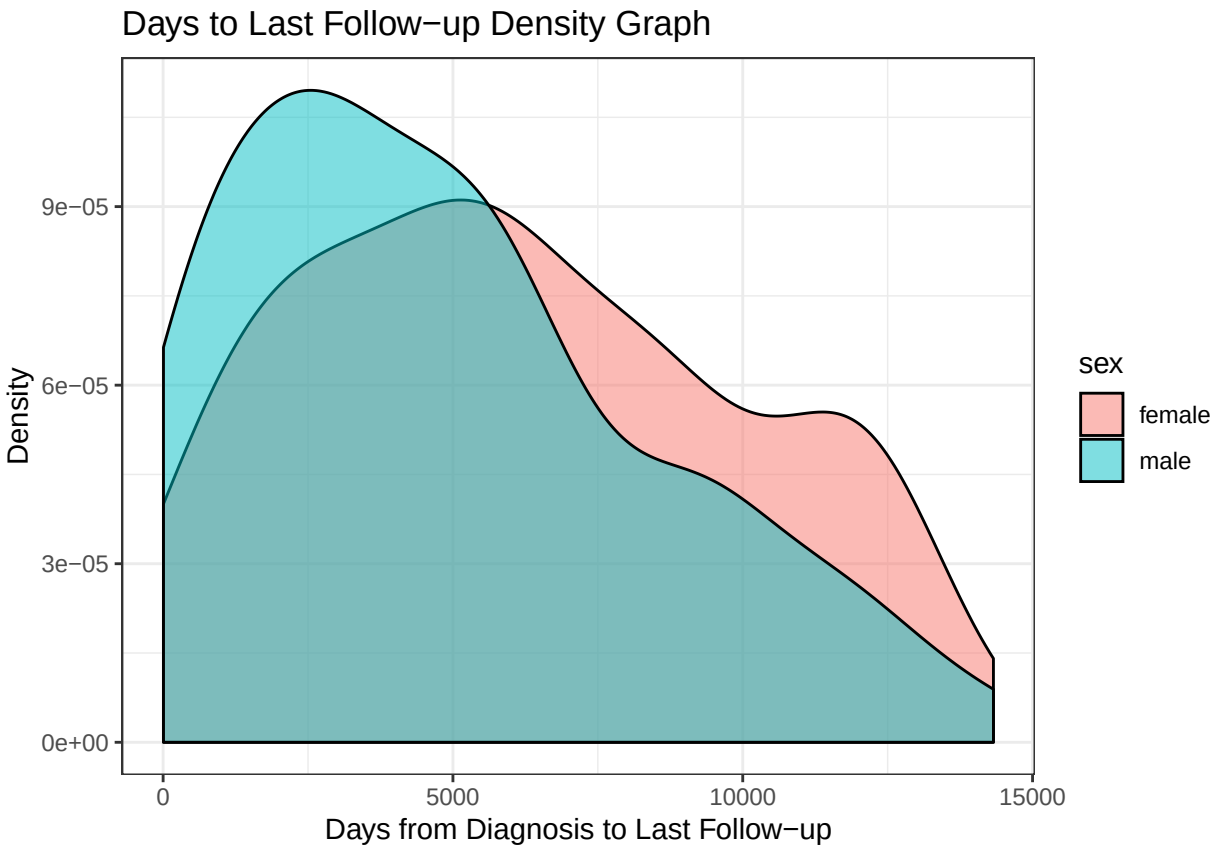



```
#Histogram demonstrating the distribution of the ages by gender.
mgus %>%
  ggplot() +
  geom_histogram(aes(age, fill = sex), col = "black") +
  theme_bw() +
  ggtitle("Histogram showing Age Distribution by Gender")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Density plot to show distributions of
mgus %>%
  ggplot() +
  geom_density(aes(futime, fill = sex), alpha = 0.5, col = "black") +
  theme_bw() +
  ggtitle("Density showing Time till Last Follow-up Distribution by Gender") +
  labs(y = "Density", x = "Days from Diagnosis to Last Follow-up",
       title = "Days to Last Follow-up Density Graph")
```



Question 2:

The Kaplan-Meier Estimator is a non-parametric statistic used to estimate the survival function from lifetime data. As seen in lectures, the estimator was given in two formats: Under the assumption that there are no ties then estimator can be denoted as:

$$\hat{s}(t) = \prod_{i=y_i \leq t} \left(\frac{n-i}{n-i+1} \right)^{\delta(i)}$$

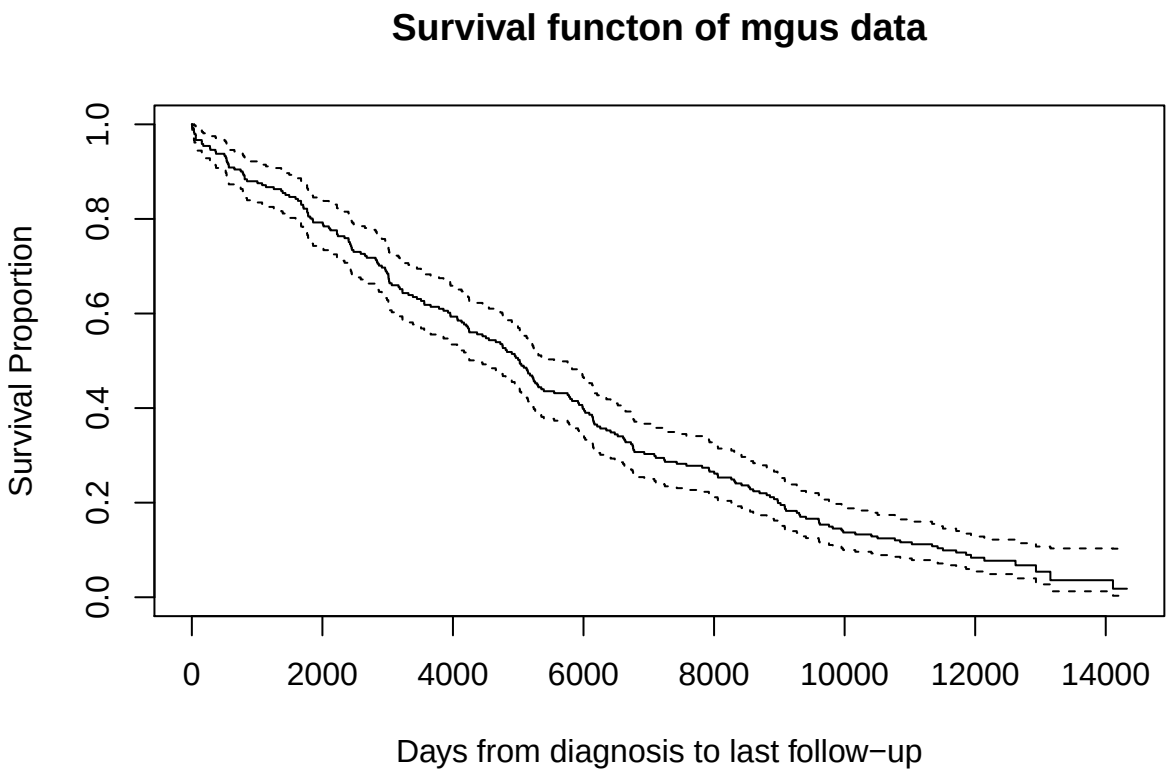
If there are ties then it can be denoted as:

$$\hat{s}(t) = \prod_{i=y_i \leq t} \left(1 - \frac{d_i}{n_i} \right)^{\delta(i)}$$

Where n_i denotes the number of subjects at beginning of time point I_i, d_i denotes the number of deaths that occur in I_i, T_i denotes the time of death for observation i, C_i denotes the censoring and δ_i is the indicator function that both T_i and C_i occur $\delta_i = 1(T_i, C_i)$

The plot, as seen below, is a series of decreasing horizontal steps which, with a sufficiently large sample size approaches the true survival rate of the population.

```
km_fit = survfit(Surv(futime, death) ~ 1, data = mgus)
plot(km_fit, main="Survival functon of mgus data",
     xlab = "Days from diagnosis to last follow-up", ylab = "Survival Proportion")
```



Question 3:

The first step to solving this problem is to see whether are ties in the data:

```
#The first step is to check whether there are any ties:
n_occur <- data.frame(table(mgus$futime)) %>%
  filter(Freq > 1) %>%
  rename(futime = Var1)

kable(n_occur, "latex", booktabs = T) %>%
kable_styling(latex_options = "striped")
```

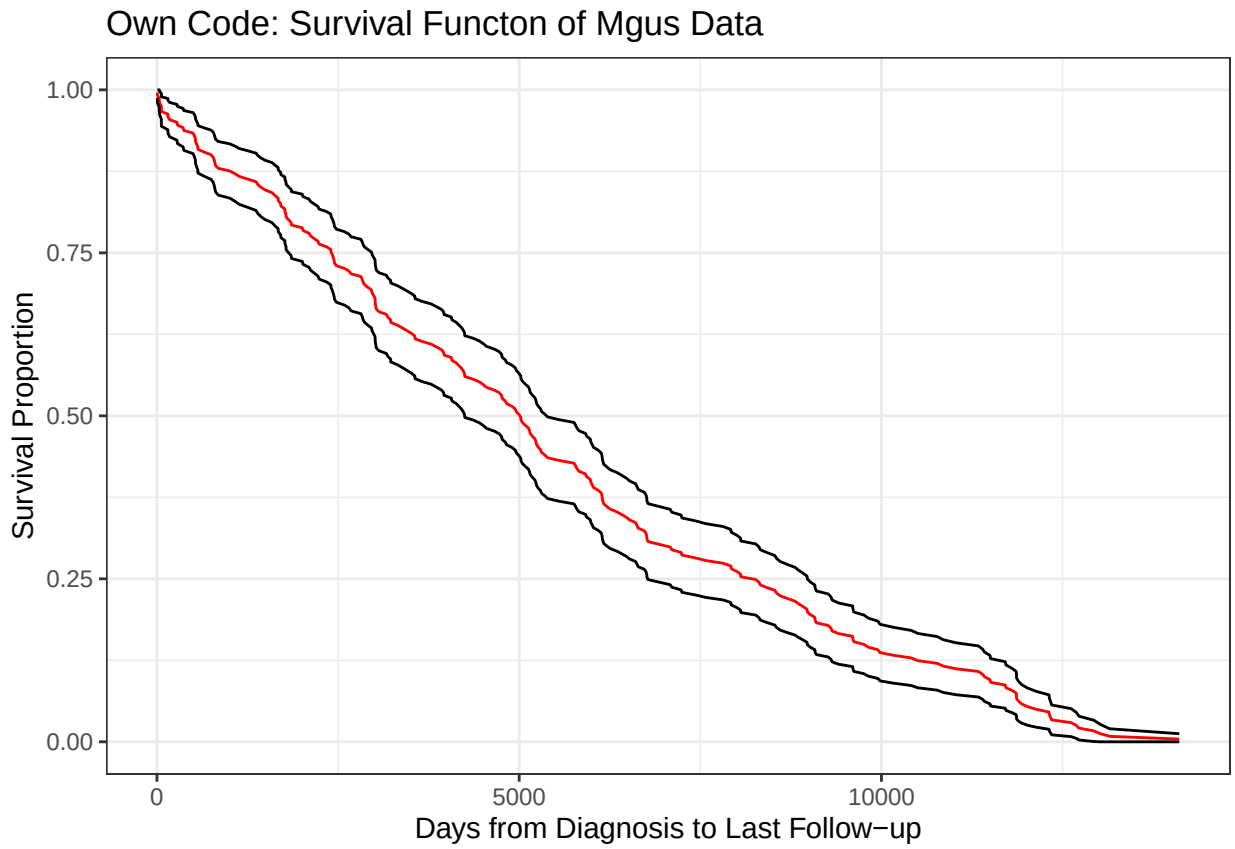
futime	Freq
61	2
809	2
2223	2
12349	2

From this table it can be clearly seen that there are ties for four futime points, as such the code below needs to take into account that the number of deaths at any time point could be greater than 1.

```
mgus3 <- mgus %>%
  select(futime, death) %>%
  arrange(futime) %>%
```

```r
  unique() %>%
  #Add in number of deaths at each futime step
  mutate(death = if_else(futime %in% n_occur$futime, 2, 1))

#Update the number of patients who are still alive at each futime step
risk <- c(241, rep(0, 236))
for(i in 1:236) {
  risk[i+ 1] <- risk[i] - mgus3[i, 2]
}
mgus3 <- cbind(mgus3, risk) %>%
  #Calculate the sum variable
  mutate(sum = death/(risk * (risk - death)))

#Define The cummulant
cummulant <- c(mgus3$sum[1], rep(0, 236))
for(i in 1:236) {
  #update cumulant with sum
  cummulant[i+1] <- cummulant[i] + mgus3$sum[i+1]
}

mgus3 <- cbind(mgus3, cummulant)  %>%
  #sqrt cummulant as required
  mutate(cummulant = sqrt(cummulant)) %>%
  #Calculate pHat value
  mutate(pHat = (1 - (1/risk))^death)

#Calculate the sHat
sHat <- c(mgus3$pHat[1], rep(0, 236))
for(i in 1: 236) {
  sHat[i+1] <- sHat[i]*mgus3$pHat[i+1]
}

mgus3 <- cbind(mgus3, sHat) %>%
  #Calculate standard error
  mutate(se = cummulant *sHat) %>%
  #defining the 95% confidence interval between lower and upper
  mutate(upper = sHat + 1.96 * se) %>%
  mutate(lower = sHat - 1.96 * se) %>%
  #Stop survival rate going over 1
  mutate(upper = if_else(upper > 1 ,1 , upper)) %>%
  #Stop the survival rate going below 0
  mutate(lower = if_else(lower < 0 ,0, lower)) %>%
  #Last observation has division by 0 so exclude
  filter(futime != 14325) %>%
  select(futime, sHat, lower, upper)

mgus3 %>%
  ggplot(aes(x = futime)) +
  geom_line(aes(y = sHat), col = "red") +
  geom_line(aes(y = upper)) +
  geom_line(aes(y = lower)) +
  theme_bw() +
  labs(x = "Days from Diagnosis to Last Follow-up",
```

```
      y = "Survival Proportion", title = "Own Code: Survival Functon of Mgus Data")
```



Question 4:

The hypothesis test undertaken will test for a difference between two survival curves using the G-rho family of tests. The idea being that if the null hypothesis is true then there is no difference between the two survival curves and therefore the two datasets contain survival data which originates from the same underlying population. On the contrary, if the alternative hypothesis is true, then there is statistically signicant evidence to suggest that since the two datasets have different survival curves then they come from different underlying populations. The test can be formulated as followed:

Null Hypothesis: $H0$: There is no difference between the two survival curves. Alternative Hypothesis: $H1$: There is a difference between the two survival curves. Test statistics: $T < 2 \times 10^{-16}$ Conclusion: Since T $<$ 0.05, there is sufficient evidence to reject the null hypothesis and accept the alternative hypothesis - there is a difference between the survival curves generated for the two datasets and the data in each dataset comes from different underlying populations.

```
#It is worth noting that both datasets contain no null values for the futime or death variables
mgus_data <- mgus %>%
  select(futime, death) %>%
  mutate(dataset = "mgus")
mgus2_data <- mgus2 %>%
  select(futime, death) %>%
  mutate(dataset = "mgus2")
combined <- bind_rows(mgus_data, mgus2_data)
```

```r
survdiff(Surv(futime, death) ~ dataset, data = combined)
```

```
## Call:
## survdiff(formula = Surv(futime, death) ~ dataset, data = combined)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## dataset=mgus    241      225      478     133.9       378
## dataset=mgus2 1384      963      710      90.1       378
##
##  Chisq= 378  on 1 degrees of freedom, p= <2e-16
```

The outcome from this test is highly significant which demonstrates the difference between the two survival curves generated from the different datasets. This makes sense as there is a large difference in survival proportion between the two datasets. In the mgus data the observed proportion of deaths = 0.934 whereas in the mgus2 data is it = 0.696.

Question 5

There are a couple of observations about the data which need to be considered. Firstly, it is useful to check the number of NA values in each column to see whether these variables should be used for analysis:

```r
#checking NAs
nulls <- apply(is.na(mgus), 2, sum)

kable(t(nulls), "latex", booktabs = T) %>%
kable_styling(latex_options = "striped")
```

id	age	sex	dxyr	pcdx	pctime	futime	death	alb	creat	hgb	mspike
0	0	0	0	177	177	0	0	31	43	1	0

A key observation here is that both pcdx and pctime contain 177 $NULL$ values out of a total of 241 observations. These variables are related and as such only one should be included in any model to reduce the risk associated with confounding variables. The pcdx variable gives the type of plasma cell malignancy if the subject progressed or NA otherwise. This variable can be changed to a 0/1 indicator variable as to whether the person progressed, where 1 denotes they did and 0 denotes they did not.

```r
#Changing the pcdx variable
cox_data <- mgus %>%
  mutate(pcdx = if_else(is.na(pcdx), 0, 1))

#Full model: Model1
cox = coxph(
  Surv(futime,death) ~ age + sex + dxyr + pcdx + alb + creat + hgb + mspike, data = cox_data)
cox
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + sex + dxyr + pcdx +
##     alb + creat + hgb + mspike, data = cox_data)
##
##              coef exp(coef)  se(coef)      z       p
## age      0.076567  1.079574  0.009173  8.347 < 2e-16
## sexmale  0.313329  1.367972  0.168638  1.858 0.06317
## dxyr     0.063163  1.065201  0.032589  1.938 0.05260
```

```
## pcdx     0.545646  1.725723  0.202875  2.690 0.00715
## alb     -0.181406  0.834097  0.207852 -0.873 0.38279
## creat    0.422114  1.525182  0.146739  2.877 0.00402
## hgb     -0.134150  0.874459  0.063469 -2.114 0.03455
## mspike   0.144567  1.155539  0.204634  0.706 0.47990
##
## Likelihood ratio test=106.2  on 8 df, p=< 2.2e-16
## n= 176, number of events= 165
##    (65 observations deleted due to missingness)
```

```
cox_fit = survfit(cox)

#Best model in terms of LRT: Model2
cox_2 = coxph(
  Surv(futime,death) ~ age + sex + dxyr + pcdx  + creat + hgb, data = cox_data)
cox_2
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + sex + dxyr + pcdx +
##     creat + hgb, data = cox_data)
##
##              coef exp(coef)  se(coef)      z       p
## age      0.079005  1.082210  0.008512  9.282 < 2e-16
## sexmale  0.294374  1.342286  0.158928  1.852 0.06399
## dxyr     0.053730  1.055199  0.029475  1.823 0.06832
## pcdx     0.472395  1.603830  0.183615  2.573 0.01009
## creat    0.442914  1.557238  0.138064  3.208 0.00134
## hgb     -0.164617  0.848219  0.052693 -3.124 0.00178
##
## Likelihood ratio test=122.4  on 6 df, p=< 2.2e-16
## n= 198, number of events= 184
##    (43 observations deleted due to missingness)
```

```
cox_fit_2 = survfit(cox_2)

#Testing to see confirm that pcdx is worth including: Model 3
cox_3 = coxph(
  Surv(futime,death) ~ age + sex + dxyr  + creat + hgb, data = cox_data)
cox_3
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + sex + dxyr + creat +
##     hgb, data = cox_data)
##
##              coef exp(coef)  se(coef)      z       p
## age      0.073564  1.076337  0.008063  9.124 < 2e-16
## sexmale  0.225459  1.252897  0.156371  1.442 0.14935
## dxyr     0.037820  1.038544  0.028697  1.318 0.18754
## creat    0.414557  1.513699  0.137991  3.004 0.00266
## hgb     -0.161486  0.850879  0.051550 -3.133 0.00173
##
## Likelihood ratio test=116.1  on 5 df, p=< 2.2e-16
## n= 198, number of events= 184
##    (43 observations deleted due to missingness)
```

```r
cox_fit_3 = survfit(cox_3)

#All variables included in this model were significant (<0.05): Model 4
cox_4 = coxph(
  Surv(futime,death) ~ age + creat + hgb, data = cox_data)
cox_4
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + creat + hgb, data = cox_data)
##
##            coef exp(coef)  se(coef)      z        p
## age    0.075096  1.077988  0.008058  9.319  < 2e-16
## creat  0.446417  1.562703  0.131407  3.397 0.000681
## hgb   -0.129269  0.878737  0.049635 -2.604 0.009203
##
## Likelihood ratio test=112.3  on 3 df, p=< 2.2e-16
## n= 198, number of events= 184
##    (43 observations deleted due to missingness)
```

```r
cox_fit_4 = survfit(cox_4)

lrt_scores <- tibble(Model = c("Model 1: Full Model", "Model 2: High p-value variables removed",
                               "Model 3: pcdx removed aswell", "Model 4: Only significant variables"),
                     LRT_Score = c(cox$score, cox_2$score, cox_3$score, cox_4$score))
kable(lrt_scores, "latex", booktabs = T) %>%
kable_styling(latex_options = "striped")
```

Model	LRT_Score
Model 1: Full Model	105.9970
Model 2: High p-value variables removed	122.2091
Model 3: pcdx removed aswell	119.5881
Model 4: Only significant variables	115.4027

All the above models have a LRT value compared to the NULL model, since they are all being compared to the same NULL model, the best model will the one returning the highest Likelihood Ratio Test (LRT) score. From the models tested above it can be seen that model 2 returns the highest value for the LRT score and as such is the most best Cox regression model. This is interesting as not all the parameters in the model are statistically significant (p.value < 0.05), but they are very close to being significant. Adding in the binary variable pcdx for whether the the subject progressed to plasma cell malignancy, did increase the LRT and was a statistically significant contributing variable. The final Model can be denoted as: $y = 0.079age + 0.294sexmale + 0.054pcdx + 0.443creat - 0.165hgb$