Statistical Learning Assignment 2

Michael J Jones

24/04/2015

Question 1

Define a Cox Proportional Hazard Model (M1) for the covariates: clinic, prison, dose.

```
library(survival)
library(knitr)

load("addicts.rda")
attach(dat)

M1 <- coxph(formula = Surv(time = survt, event = status) ~ clinic + prison + dose)

summary(M1)
```

```
## Call:
## coxph(formula = Surv(time = survt, event = status) ~ clinic +
##     prison + dose)
##
##   n= 238, number of events= 150
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## clinic -1.009896  0.364257  0.214889 -4.700 2.61e-06 ***
## prison  0.326555  1.386184  0.167225  1.953   0.0508 .
## dose   -0.035369  0.965249  0.006379 -5.545 2.94e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## clinic    0.3643     2.7453    0.2391    0.5550
## prison    1.3862     0.7214    0.9988    1.9238
## dose      0.9652     1.0360    0.9533    0.9774
##
## Concordance= 0.665  (se = 0.026 )
## Rsquare= 0.238   (max possible= 0.997 )
## Likelihood ratio test= 64.56  on 3 df,   p=6.228e-14
## Wald test            = 54.12  on 3 df,   p=1.056e-11
## Score (logrank) test = 56.32  on 3 df,   p=3.598e-12
```

Question 2

Perform a regression analysis for the model M1 and provide a discussion of the results. Remark: Follow the instructions given in Tutorial 8.

In our regression analysis, we're going to use binomial logistic regression. This model is useful when our dependent variable is restricted to the values 1 and 0. In our case, our dependent variable is whether a patient drops out a clinic or not. In our data, this is encoded as the value (1) if the patient drops out of the clinic and (0) if they do not. In other words, logistic regression is useful when our dependent variable is binary and our explanatory variables are either continuous or categorical.

To perform this a logistic regression in R, we must use the `stats::glm` function and ensure we use the `family = binomial()` argument.

```
binary.linear.regression <- glm(status ~ clinic + prison + dose, family = binomial())
summary(binary.linear.regression)
```

```
##
## Call:
## glm(formula = status ~ clinic + prison + dose, family = binomial())
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9040  -0.9449   0.6851   0.7994   1.8381
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.22797    0.78182   5.408 6.38e-08 ***
## clinic      -1.54175    0.30493  -5.056 4.28e-07 ***
## prison      -0.04155    0.29257  -0.142   0.8871
## dose        -0.02630    0.01048  -2.509   0.0121 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 313.60  on 237  degrees of freedom
## Residual deviance: 276.33  on 234  degrees of freedom
## AIC: 284.33
##
## Number of Fisher Scoring iterations: 4
```

Looking at the summary table above, we can observe the coefficients of the explanatory variables (also known as the predictor variables). The coefficients determine to what degree each of the explanatory variables contribute to the value of the dependent variables.

Firstly, considering the coefficient for which clinic the patient went to shows that the value is negatively correlated with whether a patient will drop out of a clinic. We can see that there are two clinics:

```
unique(dat$clinic)
```

```
## [1] 1 2
```

This model implies that *Clinic 1* is more likely to have patients dropping out than *Clinic 2*. We can actually verify this by looking at the data for Clinic 1 and Clinic 2:

```
dat.clinic1 <- dat[dat$clinic == 1,]
dat.clinic2 <- dat[dat$clinic == 2,]

table(dat.clinic1$status)
```

```
##
##   0   1
##  41 122
```

```
table(dat.clinic2$status)
```

```
##
##  0  1
## 47 28
```

Clinic 1 has a much higher dropout rate than Clinic 2. We can also see that this coefficient has a highly significant p value of 4.28e-04, meaning that it we have a very strong case to refute the null hypothesis that clinic doesn't contribute to our outcome that a patient will drop out of a clinic.

The coefficient for whether a patient has been to *Prison* implies that there is a very small negative correlation between having been to prison and dropping out of a clinic. This implies that those which have been to prison have a marginally lower chance of dropping out of the clinic. However, when we look at the p value for this coefficient, we can see plainly that, given this sample, we are unable to refute the null hypothesis. In other words, it appears that having been in prison probably doesn't contribute to whether a patient is likely to drop out of a clinic or not.

Unlike clinic and prison, the *dose* of methadone given to a patient is interesting in that it is a continuous variable as opposed to being categorical. The correlation of the regression with dose is -0.02630, indicating that a higher dose is negatively correlated with a patient dropping out of the clinic.

In summary then, observing the results of our linear regression, we can say that it is highly likely that a patient's risk of dropping out decreased when they go to Clinic 1 instead of Clinic 2 and two and if they are prescribed a higher dosage of methadone. However, it is very unlikely that the risk of a patient dropping out of a clinic is affected given they have been to prison or not.

Question 3

Part A

Check the proportional hazard assumption of M1 and adjust the model if necessary

---

R provides a way for us to calculate the proportional hazard assumption using `cox.zph`. This function allows us to measure proportionality with regard to log(time).

```
proportionality.test <- cox.zph(fit = M1, transform = "log")
knitr::kable(proportionality.test$table)
```

	rho	chisq	p
clinic	-0.2140030	7.7056532	0.0055048
prison	-0.0462436	0.3218268	0.5705119
dose	0.1260549	2.1238151	0.1450249
GLOBAL	NA	10.4499011	0.0151046

In the first column of the results table, we can see `rho`, which is the Pearson product-moment correlation between the scaled Schoenfeld residual and log(time) for each covariate.

The other column of interest to us is the right-most p-value column which shows the p-value given a null hypothesis that the proportionality violation has been violated. In other words, if the p-value of this column is *less than* 0.05, then we must refute the null hypothesis and assume that this particular covariate does in fact violate the proportionality assumption.

The `GLOBAL` p-value shows that the whole model violates the proportionality assumption. To remedy this, we might remove the clinic explanatory variable from our model and re-run the cox.zph command.

```
M1.improved <- coxph(formula = Surv(time = survt, event = status) ~ prison + dose)

proportionality.test.improved <- cox.zph(fit = M1.improved, transform = "log")
knitr::kable(proportionality.test.improved$table)
```

	rho	chisq	p
prison	-0.1225006	2.213461	0.1368115
dose	0.1074185	1.401130	0.2365345
GLOBAL	NA	3.526046	0.1715255

As you can see from the results of removing the clinic variable from M1, all explanatory variables as well as the GLOBAL p-value no longer are small enough to refute the null hypothesis and thus, we can conclude that the model is unlikely to refute the proportionality model.

Part B and C

Visualize and discuss the Schoenfeld residues for the covariates (and) Provide a discussion of the results.

Cox's proportional hazards assumes that there is a constant relationship between dependent variables and the explanatory variable. This means that the hazard function for any two individuals at any time are proportional. If the model assumes this then we need to test this assumption.

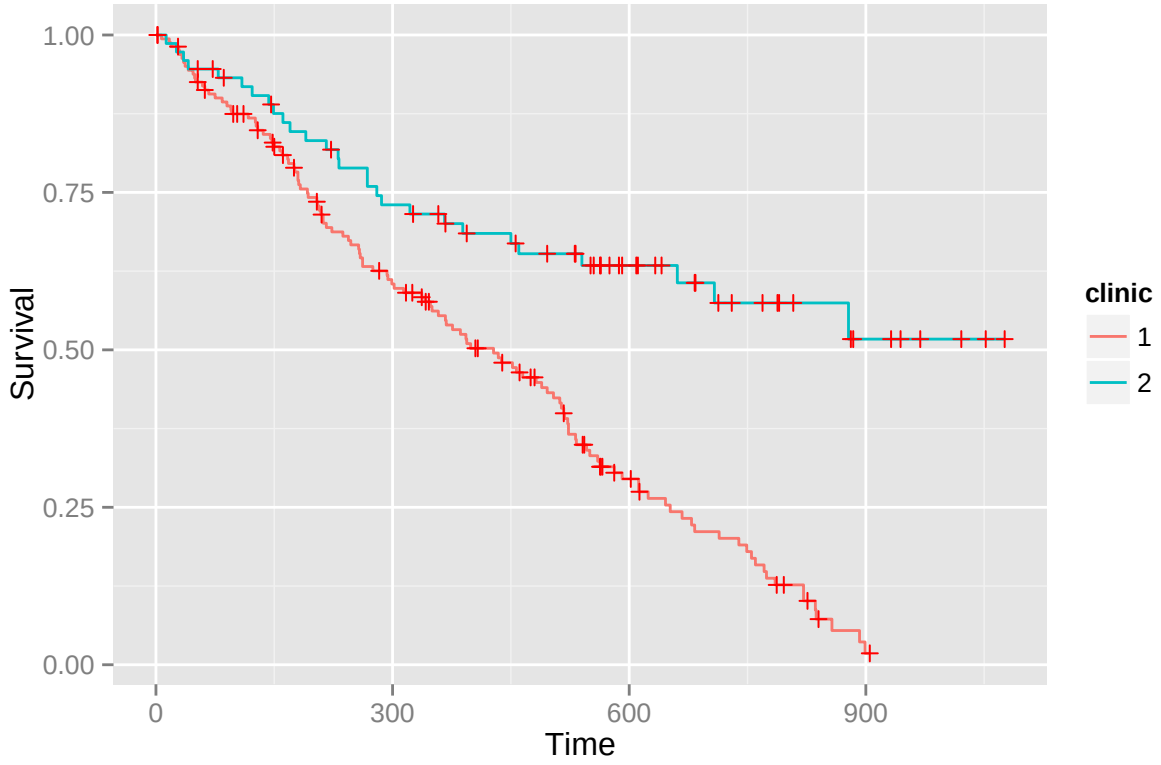
Kaplan-Meier curves

There are a number of methods for confirming that our model (M1) conforms to the *proportional hazards assumption*. A simple, graphical way for doing this is to plot Kaplan-Meier curves for the survival function of each explanatory variable against survival time of individuals. If the lines cross for different values of our explanatory variables then it is most probable that the two are not proportional, thereby violating the proportional hazards assumption.

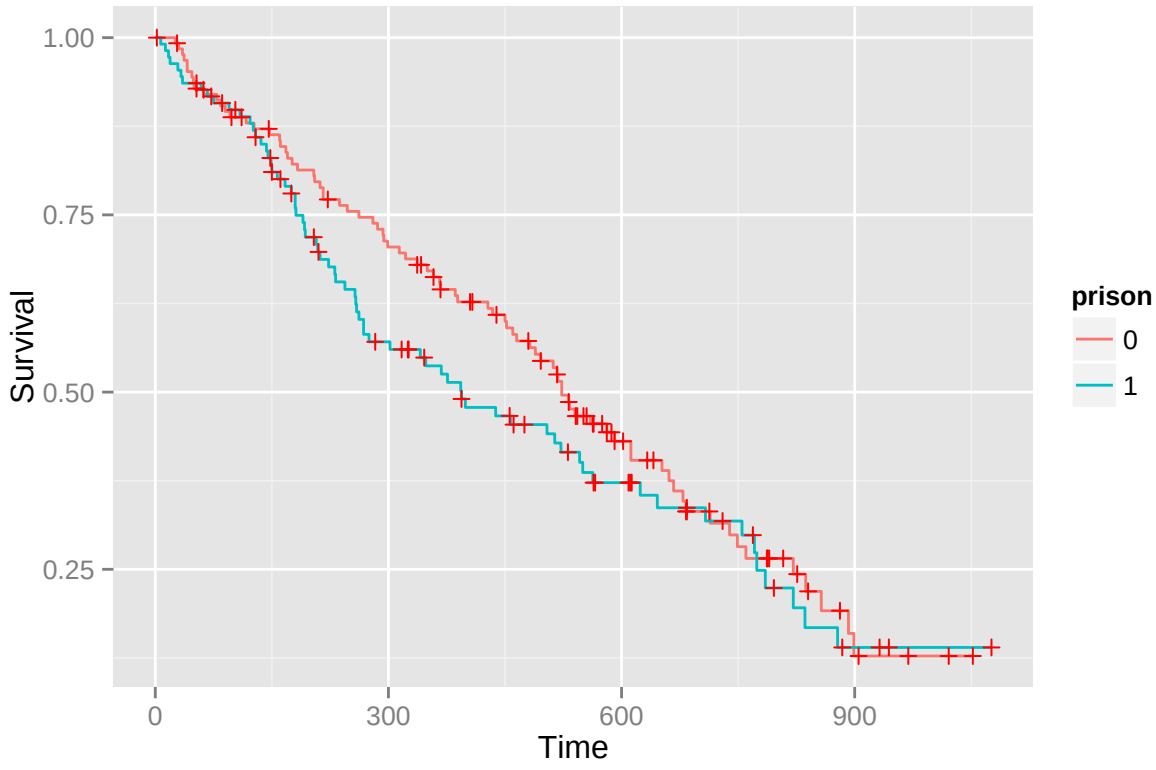
```
library(GGally)

ggsurv(survfit(formula = Surv(time = survt, event = status) ~ clinic,
            data = dat))
```

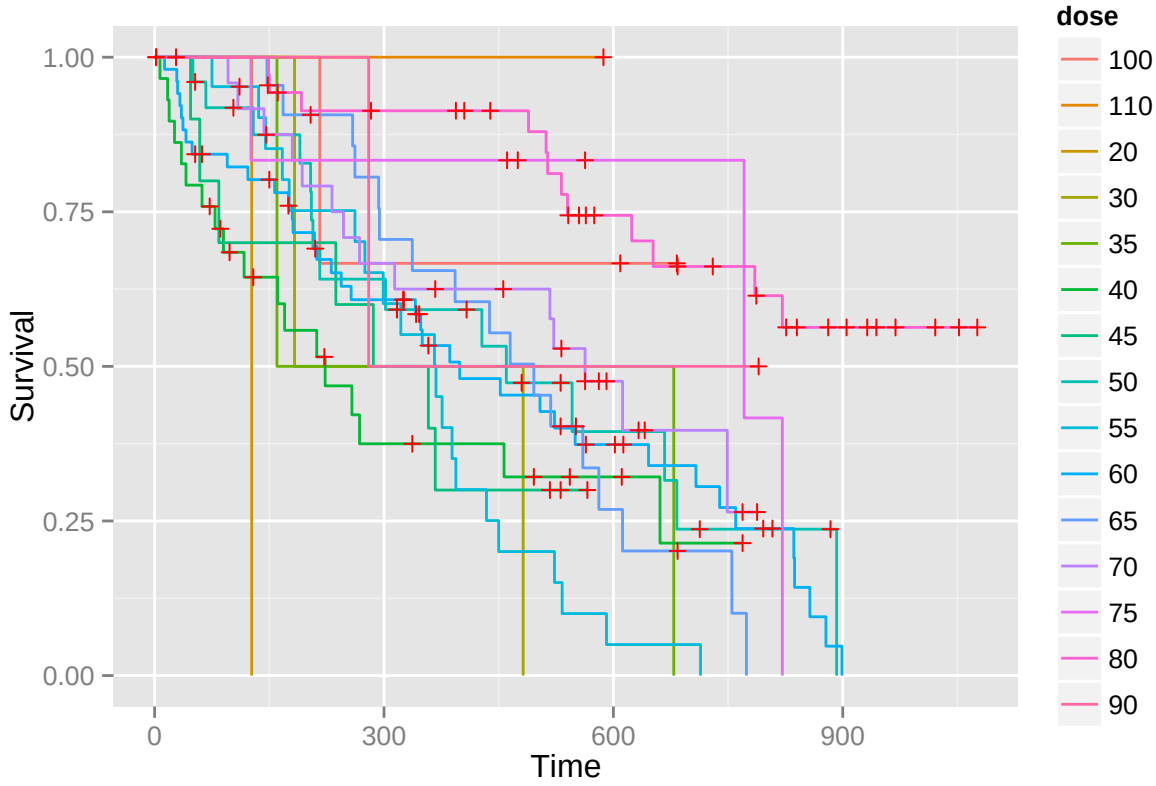
```
## Loading required package: scales
```




```
ggsurv(survfit(formula = Surv(time = survt, event = status) ~ prison,
            data = dat))
```




```
ggsurv(survfit(formula = Surv(time = survt, event = status) ~ dose,
               data = dat))
```



The Kaplan-Meier curve method of testing the proportional hazards assumption is not well suited to small data data sets where, consequently, curves may overlap without the proportional hazards assumption having been violated.

We can also see that, where there an explanatory variable has a lot of categories or is continuous (such as with *survt*), the graph is particularly difficult to read.

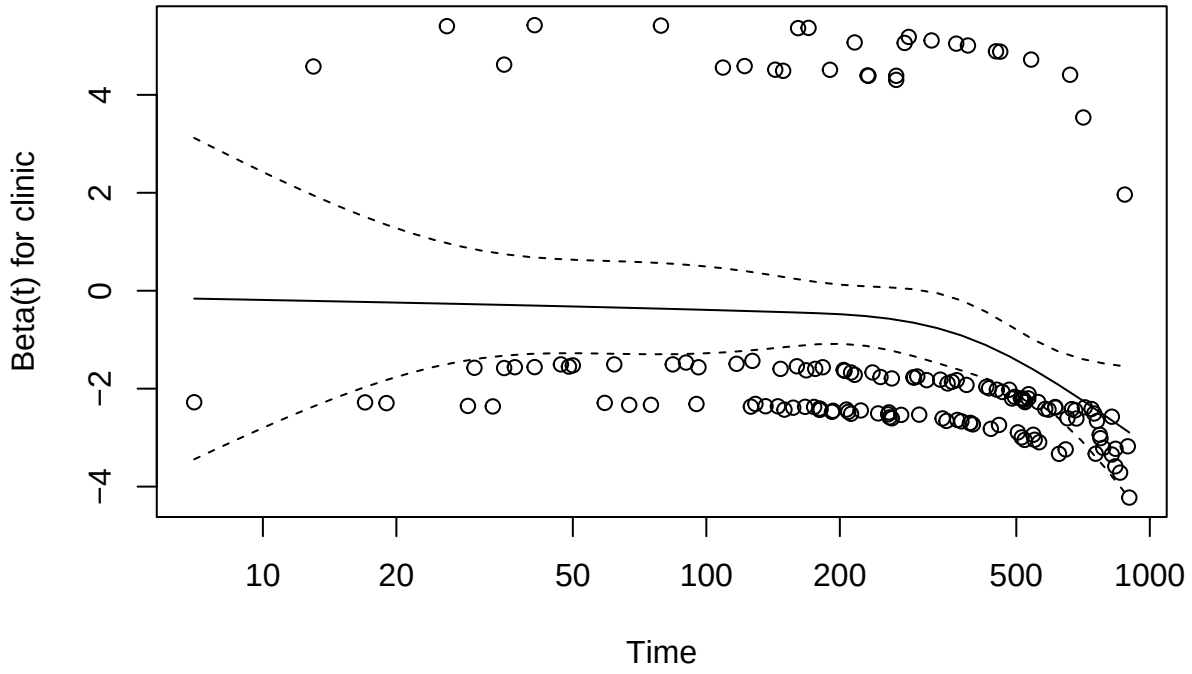
Visualising the Schoenfeld residues

As we've shown, the Kaplan-Meier graphical method for evaluating whether our model conforms to the proportional hazard assumption is not perfect and in cases where our explanatory variable is continuous, is impossible to interpret.

The `cox.zph` function used in **Part A** of this question allows us to plot the Schoenfeld residues for each of the covariates. This graphs show Beta(t) against log(time). Ideally we want to see a straight line. A slope is evidence against proportionality.

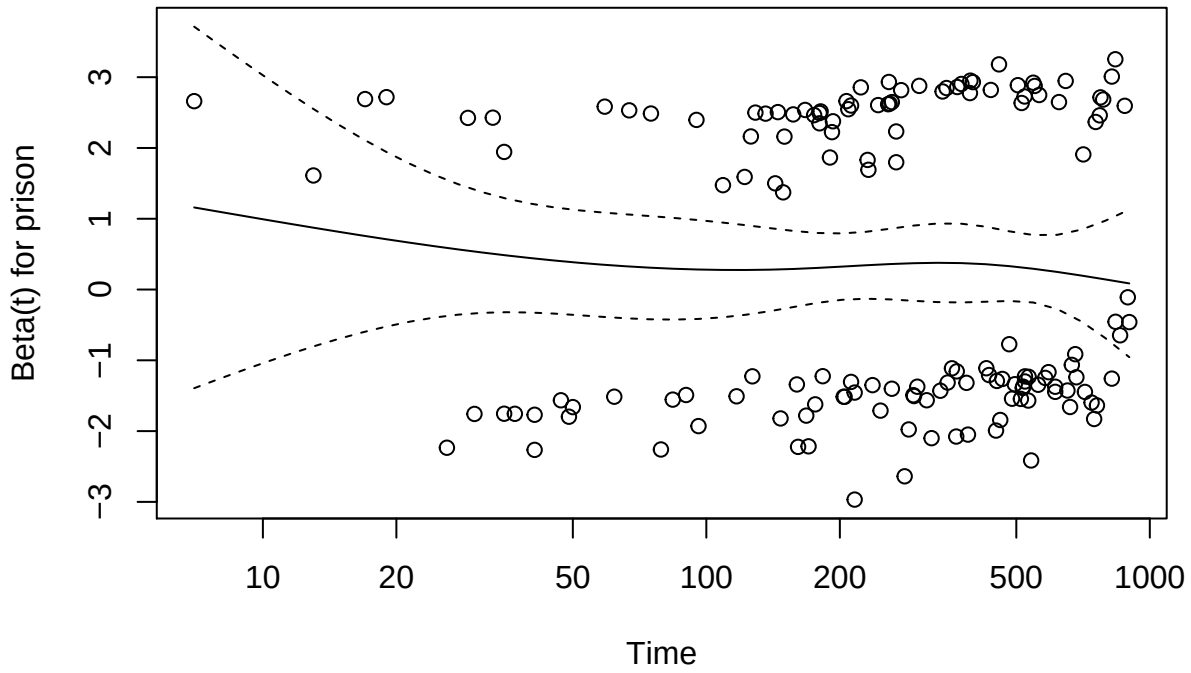
We will use our unimproved model, M1, to show that we can come to a similar conclusion to that of Part A.

```
plot(proportionality.test[1])
```



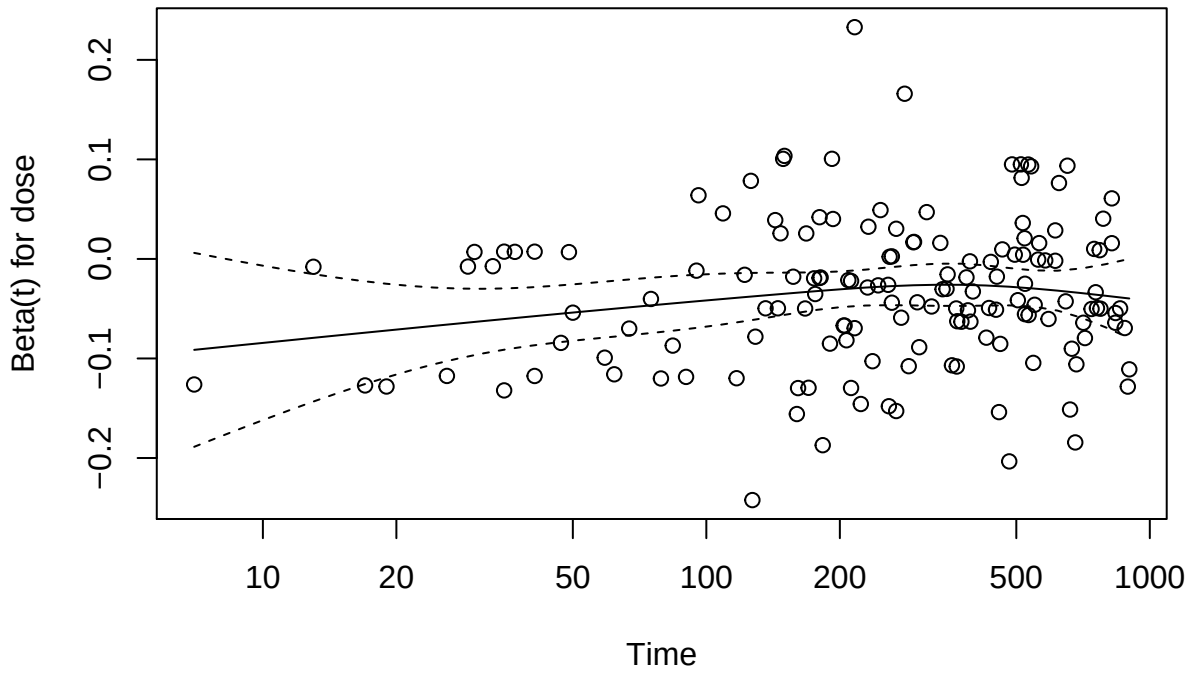
Plotting clinic's Schoenfeld residual shows a Beta(t) that is only straight for the first 200 days. After this Beta(t) quickly slopes downwards. Like in Part A, we could use this to justify removing clinic from the model.

```
plot(proportionality.test[2])
```



Beta(t) over time decreases with respect to time for the prison covariate.

```
plot(proportionality.test[3])
```



Finally, although the graph shows what seems to be a sharper curve for Beta(t) over time for dose than for prison, we can see that the scale of Beta(t) shows that it fluctuates between -0.1 and 0.0. This is reflected in our results table when we used the `cox.zph` function in Part A, showing that dose had the highest p-value with the p-value for clinic being far lower.

Part D

Discuss the difference between 3 Part A and 2.

Looking at the coefficients in both in both the logistic and the Cox proportional hazards models, we can see that, with regard to explanatory variables which seem to confound the model, they both disagree. The logistic regression summary points to the fact that the p-value of prison is very high at a value of 0.8871 meaning that it is highly likely that the magnitude of the correlation coefficient is down to chance. If we were to use the logistic regression to model a person's likelihood of dropping out of a treatment clinic, we would come to the conclusion that the prison explanatory factor should be removed before creating our model.

With Cox's proportional hazards model, however, we would come to the conclusion that, due to our analysis of Schoenfeld residuals that we should dispense with the clinic explanatory factor from its model due to the fact that it would make our model no longer conform to the proportional hazards assumption.

Question 4

Define a Cox Proportional Hazard Model (M2) for the covariate `dose` and stratify on `prison`. Provide a discussion of the results.

```
M2 <- coxph(formula = Surv(time = survt, event = status) ~ dose * strata(prison))
summary(M2)
```

```
## Call:
## coxph(formula = Surv(time = survt, event = status) ~ dose * strata(prison))
##
##   n= 238, number of events= 150
##
##                                 coef exp(coef)  se(coef)      z Pr(>|z|)
## dose                         -0.044556  0.956422  0.008037 -5.544 2.95e-08
## dose:strata(prison)prison=1  0.019241  1.019428  0.012296  1.565    0.118
##
## dose                         ***
## dose:strata(prison)prison=1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                              exp(coef) exp(-coef) lower .95 upper .95
## dose                            0.9564     1.0456    0.9415    0.9716
## dose:strata(prison)prison=1     1.0194     0.9809    0.9952    1.0443
##
## Concordance= 0.656  (se = 0.037 )
## Rsquare= 0.153   (max possible= 0.994 )
## Likelihood ratio test= 39.48  on 2 df,   p=2.668e-09
## Wald test            = 38.14  on 2 df,   p=5.228e-09
## Score (logrank) test = 38.54  on 2 df,   p=4.267e-09
```

Observing the model stratified by prison, we can see that, for those patients that have not been to prison (prison=0), the coefficient calculated indicates that an increase in dose lower the risk of a patient dropping out over time. We also see, from the p-value of 2.95e-08 that this result. This means that, from this sample, we could conclude that it is highly unlikely that this coefficient occurs by chance.

Looking at the case where a patient *has* been to prison, we notice that the coefficient value suggests the antithesis to the above. It implies that an increased dose for those who have been to jail have a marginal increased risk of dropping out of a clinic over time if they have an increased dose of methadone. However, when we observe the p value for this case, the p-value for this case is greater than the 5 percent threshold which means we it is likely that this is a random occurrence.

As such, M2 suggests that, when we stratify for whether a patient has been to prison on dose, it is only useful to consider the amount of methadone prescribed to a patient as a contributor to their risk factor of dropping out of a clinic if they have never been to prison.

Question 5

`Dose` is a continuous covariate. In order to stratify on `dose` one needs to categorize the variables first.

Part A

Decide if 2 or 3 categories are preferred and how to define the corresponding dose-intervals based on your interpretation of the data.

In statistical analysis, categorisation of data can be problematic in that it can lead to a loss of power. When categorisation is required, as discussed in [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2972292/], dichotomisation of data is discouraged (in other words, splitting data into two groups). If we categorise data into two groups, it makes it impossible for in our analysis to identify non-linearity in the exposure-outcome relationship. For this reason we will use the minimal 3 categorisation as suggested.

Using the `quantile` function in R, we can divide the data into three categories by taking the tercile values of the dosage amounts

```
probs <- seq(from = 0, to = 1, by = 1/3)
tercile <- quantile(x = dose, probs = probs, na.rm=T)
tercile
```

```
##        0% 33.33333% 66.66667%      100%
##        20        55        65       110
```

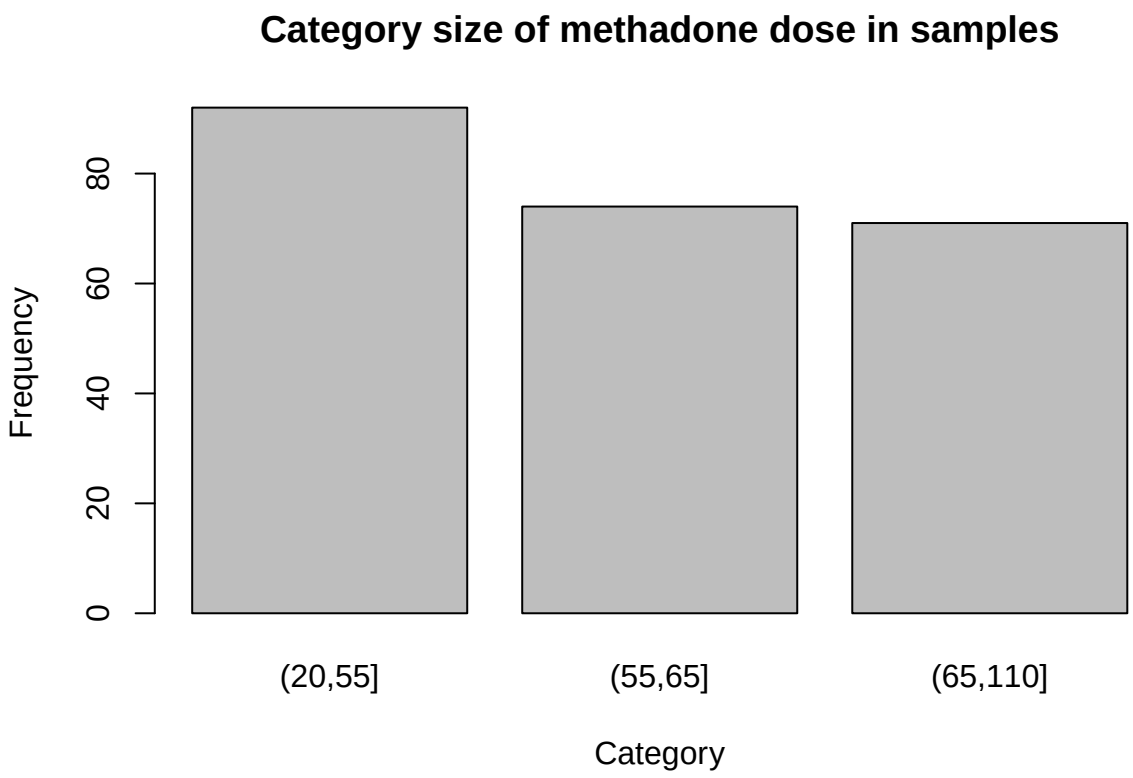
and take the tercile values as the boundaries to our categories. We can use the `cut` and `unique` functions in R to aid the selection of the categories from this result.

```
dose.category <- cut(x = dose, breaks = tercile)
unique(dose.category)
```

```
## [1] (20,55]  (55,65]  (65,110] <NA>
## Levels: (20,55] (55,65] (65,110]
```

We can also observe how many of the samples are in each category:

```r
plot(dose.category, main = "Category size of methadone dose in samples",
     xlab = "Category", ylab = "Frequency",
     col = "grey")
```




```r
table(dose.category)
```

```
## dose.category
##  (20,55]  (55,65] (65,110]
##       92       74       71
```

Part B

Define a Cox Proportional Hazard Model (M3) for the covariate `prison` and stratify on `dose`. Provide a discussion of the results

---

```r
M3 <- coxph(formula = Surv(time = survt, event = status) ~ prison * strata(dose.category))
summary(M3)
```

```
## Call:
## coxph(formula = Surv(time = survt, event = status) ~ prison *
##     strata(dose.category))
##
##   n= 237, number of events= 149
##    (1 observation deleted due to missingness)
```

```
##
##                                       coef exp(coef) se(coef)      z
## prison                                -0.01456   0.98555  0.26254 -0.055
## prison:strata(dose.category)(55,65]    0.32182   1.37964  0.38293  0.840
## prison:strata(dose.category)(65,110]   0.67107   1.95633  0.44704  1.501
##                                       Pr(>|z|)
## prison                                   0.956
## prison:strata(dose.category)(55,65]      0.401
## prison:strata(dose.category)(65,110]     0.133
##
##                                       exp(coef) exp(-coef) lower .95
## prison                                   0.9855     1.0147    0.5891
## prison:strata(dose.category)(55,65]      1.3796     0.7248    0.6513
## prison:strata(dose.category)(65,110]     1.9563     0.5112    0.8146
##                                       upper .95
## prison                                    1.649
## prison:strata(dose.category)(55,65]       2.922
## prison:strata(dose.category)(65,110]      4.699
##
## Concordance= 0.52  (se = 0.039 )
## Rsquare= 0.019   (max possible= 0.988 )
## Likelihood ratio test= 4.53  on 3 df,   p=0.2098
## Wald test            = 4.51  on 3 df,   p=0.2114
## Score (logrank) test = 4.64  on 3 df,   p=0.2004
```

For the case where a patient has been to prison and had the lowest category of dose `(20,55]`, we find that the Cox proportional model suggests from the coefficient value that going to prison lowers one's risk of dropping out of a clinic. However, the p-value of 0.956, above the 0.05 threshold, indicates very strongly that we cannot refute the null hypothesis that the model is consistent with the null distribution.

Knowing a patient has been taking a dose in the range of `(55,65]`, the Cox model summary indicates that if the patient has gone to prison they have an elevated risk of dropping out of the clinic. Unfortunately, like with the first dose category, this model cannot refute the null hypothesis either with a p-value of 0.401 which is higher than the necessary 5% threshold.

Finally, the category in which the dose is `(65,110]`, we might deduce from the coefficient that there is a higher correlation with having gone to prison and dropping out of a clinic given they have this dose of methadone. The null hypothesis cannot be refuted in this case either meaning that, in all categories, we have no statistical plausibility for our model to dictate whether having gone to prison is a good indicator of whether a patient will drop out of prison given any of the dosage categories suggested above.

Running the code in this document

This document was compiled using the `knitr` extension to R. Knitr is an implementation of the *Literate Programming* paradigm conceived of by Donald Knuth. This allows an author to *weave* code and text together into one document.

To run the code from this document, follow these instructions.

1. Open the zipped directory containing this pdf
2. (Alternatively, download the whole project from http://github.com/hiraethus/cox-proportional-assignment)
3. Open RStudio and load this project by choosing `File > Open Project` and navigating to the file `cox_proportional_assignment.Rproj`.

4. From the file browser in RStudio, choose `jonesm_assignment2.Rmd`
5. Use the `chunks` menu located on the text editor opened in R to run each of the chunks in the file
6. Run all the code chunks in RStudio by using typing `Ctrl + Alt + R`.