PM592: Regression Analysis for Data Science **HW11** *Survival Analysis*	Name: Flemmin g Wu

Instructions

- Answer questions directly within this document.
- Upload to Blackboard by the due date & time.
- Clearly indicate your answers to all questions.
- If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
- For the purpose of this assignment, statistical evidence refers to a test statistic and associated p-value.
- If a question requires a conclusion, it must be phrased professionally and coherently.
- There are 2 questions and 30 points possible.

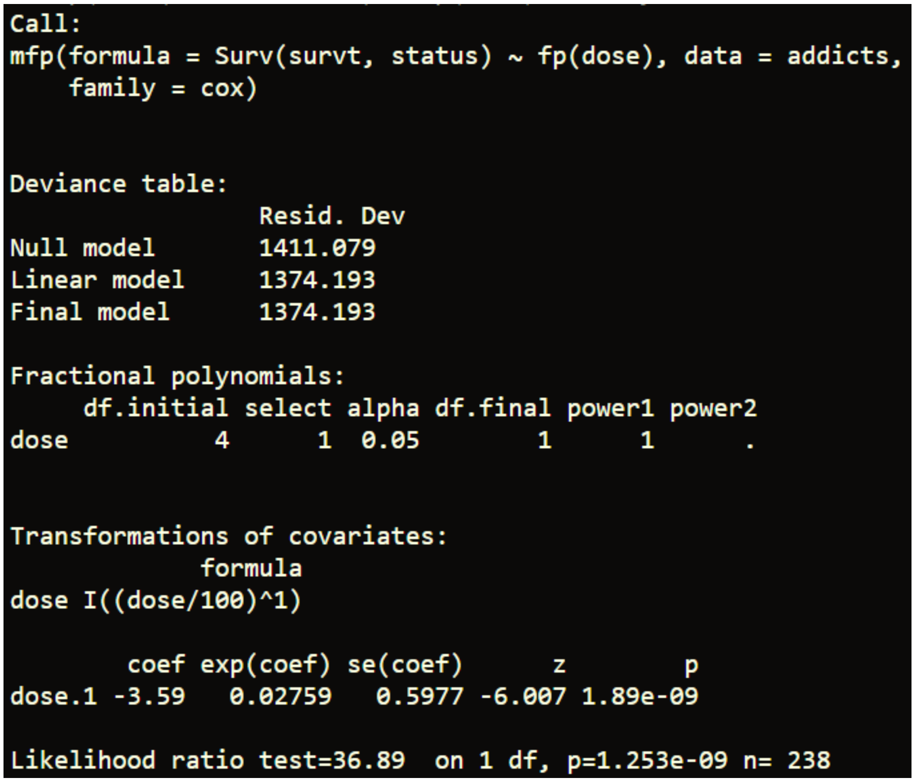
Question 1 [20 points]

In an Australian study on drug addiction, recovering heroin addict patients were examined at two methadone treatment clinics. Patients were examined at these clinics until they either dropped out of the clinic (failure) or were censored. The two clinics differed in terms of their live-in policies for patients. The data is presented in **addicts.dta**.

Does the patient's maximum methadone dose and having a prison record affect their chance of dropping out of the treatment, adjusting for clinic? Answer these questions by performing a Cox proportional hazards regression model.

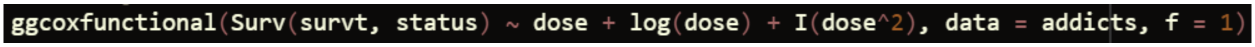
Variable	Description
Id	Patient ID
survt	Time (days) until either drop out (failure) or censored
status	1 = dropped out, 0 = censored
clinic	1 = patient attended Clinic 1, 2 = patient attended Clinic 2
prison	1 = patient had a prison record, 0 = patient did not have a prison record
dose	Patient's maximum methadone dose (mg/day)

1a. [3 points] What is the functional form for dose? Comment on the results from the MFP procedure and Martingale residuals.

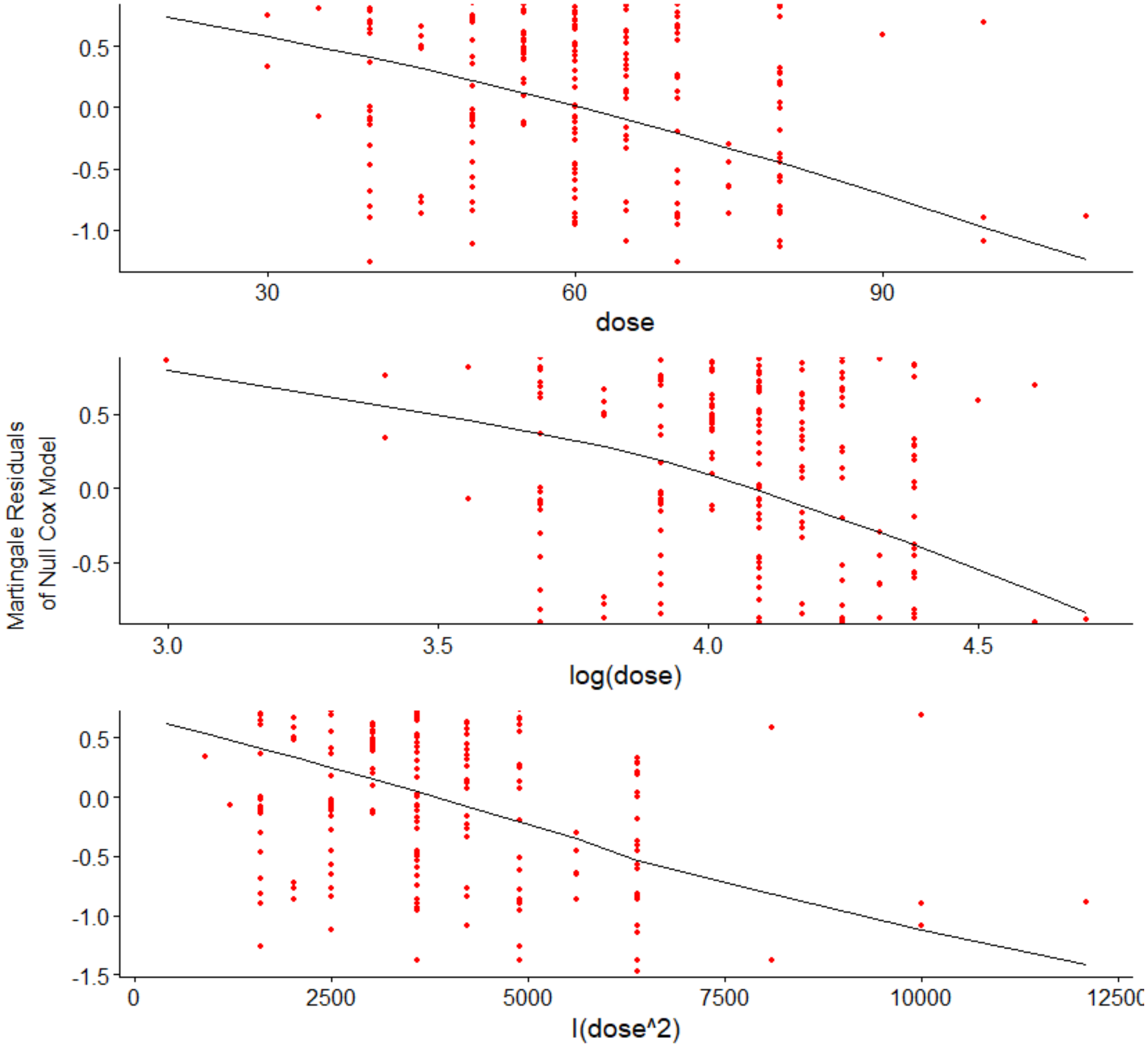


```
Call:
mfp(formula = Surv(survt, status) ~ fp(dose), data = addicts,
    family = cox)


Deviance table:
                 Resid. Dev
Null model       1411.079
Linear model     1374.193
Final model      1374.193

Fractional polynomials:
     df.initial select alpha df.final power1 power2
dose          4      1  0.05        1      1      .


Transformations of covariates:
             formula
dose I((dose/100)^1)

        coef exp(coef) se(coef)      z        p
dose.1 -3.59   0.02759   0.5977 -6.007 1.89e-09

Likelihood ratio test=36.89  on 1 df, p=1.253e-09 n= 238
```

The fractional polynomials approach suggests that the form of dose is linear.

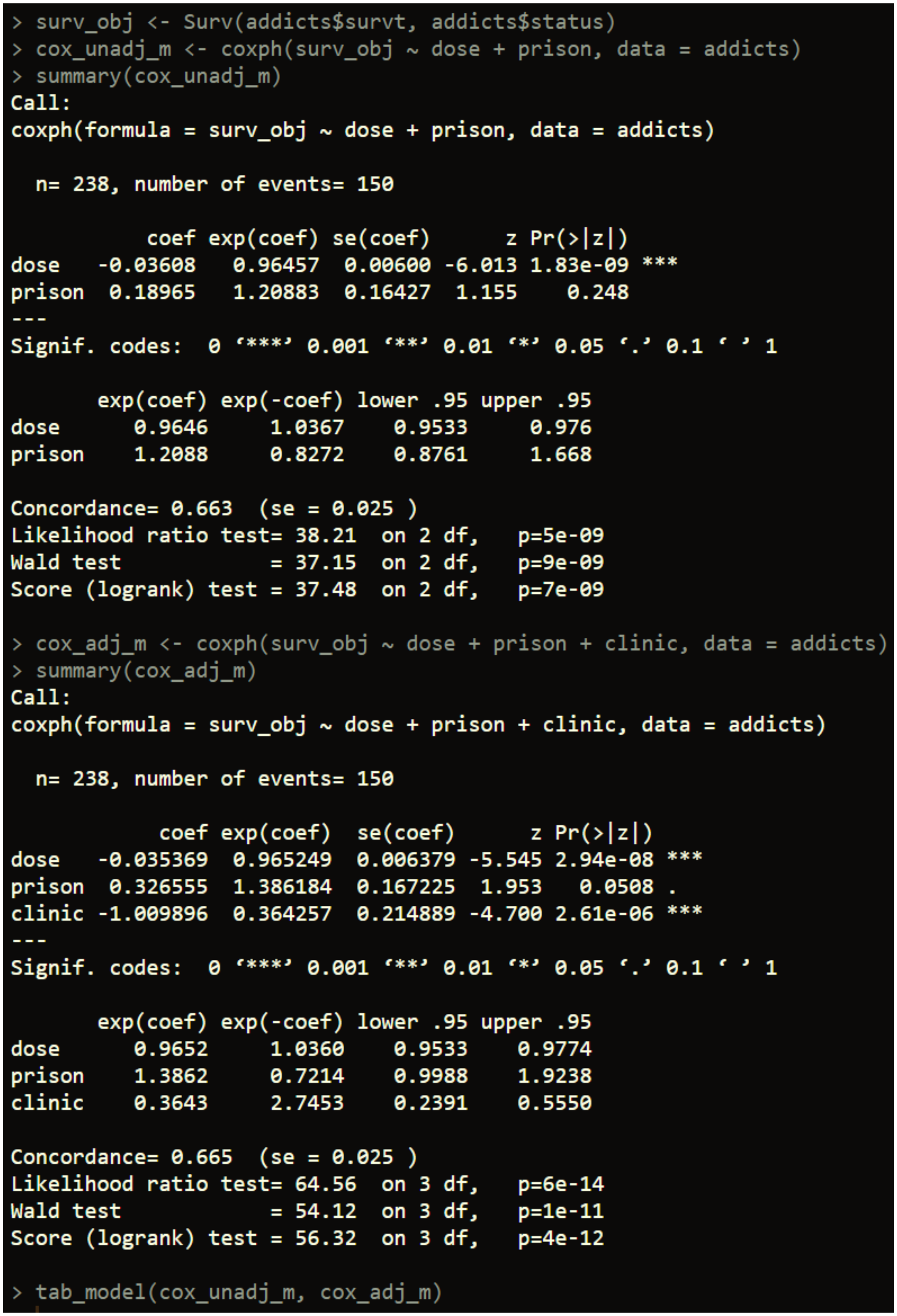


```
ggcoxfunctional(Surv(survt, status) ~ dose + log(dose) + I(dose^2), data = addicts, f = 1)
```

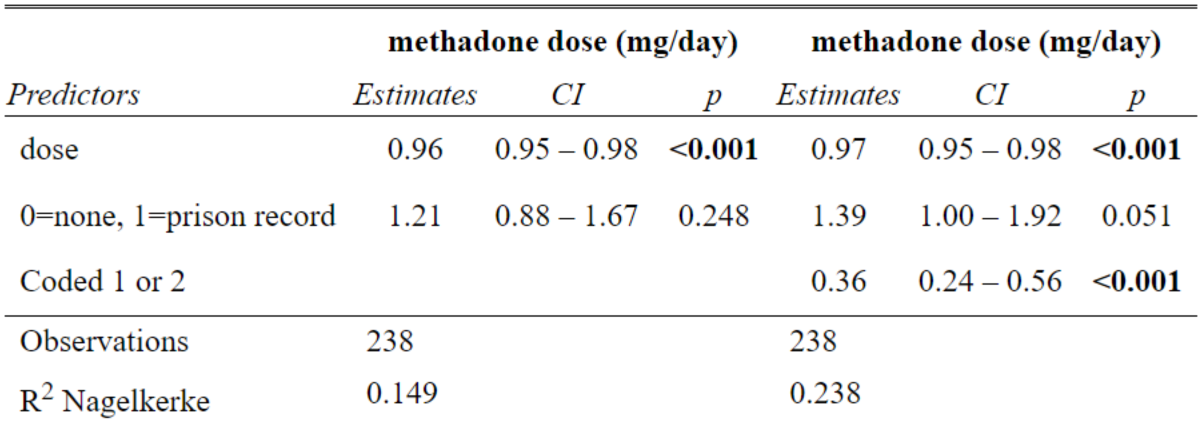


The Martingale residuals also confirm the that the functional form for dose is linear.

1b. [4 points] Construct a model with prison and dose as independent variables. Report the parameter estimates of these variables in a model without clinic, and a model with clinic.



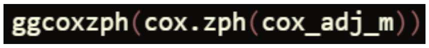
```
> surv_obj <- Surv(addicts$survt, addicts$status)
> cox_unadj_m <- coxph(surv_obj ~ dose + prison, data = addicts)
> summary(cox_unadj_m)
Call:
coxph(formula = surv_obj ~ dose + prison, data = addicts)

  n= 238, number of events= 150 

           coef exp(coef) se(coef)      z Pr(>|z|)    
dose   -0.03608   0.96457  0.00600 -6.013 1.83e-09 ***
prison  0.18965   1.20883  0.16427  1.155    0.248    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
dose      0.9646     1.0367    0.9533     0.976
prison    1.2088     0.8272    0.8761     1.668

Concordance= 0.663  (se = 0.025 )
Likelihood ratio test= 38.21  on 2 df,   p=5e-09
Wald test            = 37.15  on 2 df,   p=9e-09
Score (logrank) test = 37.48  on 2 df,   p=7e-09

> cox_adj_m <- coxph(surv_obj ~ dose + prison + clinic, data = addicts)
> summary(cox_adj_m)
Call:
coxph(formula = surv_obj ~ dose + prison + clinic, data = addicts)

  n= 238, number of events= 150 

            coef exp(coef)  se(coef)      z Pr(>|z|)    
dose   -0.035369  0.965249  0.006379 -5.545 2.94e-08 ***
prison  0.326555  1.386184  0.167225  1.953   0.0508 .  
clinic -1.009896  0.364257  0.214889 -4.700 2.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
dose      0.9652     1.0360    0.9533    0.9774
prison    1.3862     0.7214    0.9988    1.9238
clinic    0.3643     2.7453    0.2391    0.5550

Concordance= 0.665  (se = 0.025 )
Likelihood ratio test= 64.56  on 3 df,   p=6e-14
Wald test            = 54.12  on 3 df,   p=1e-11
Score (logrank) test = 56.32  on 3 df,   p=4e-12

> tab_model(cox_unadj_m, cox_adj_m)
```



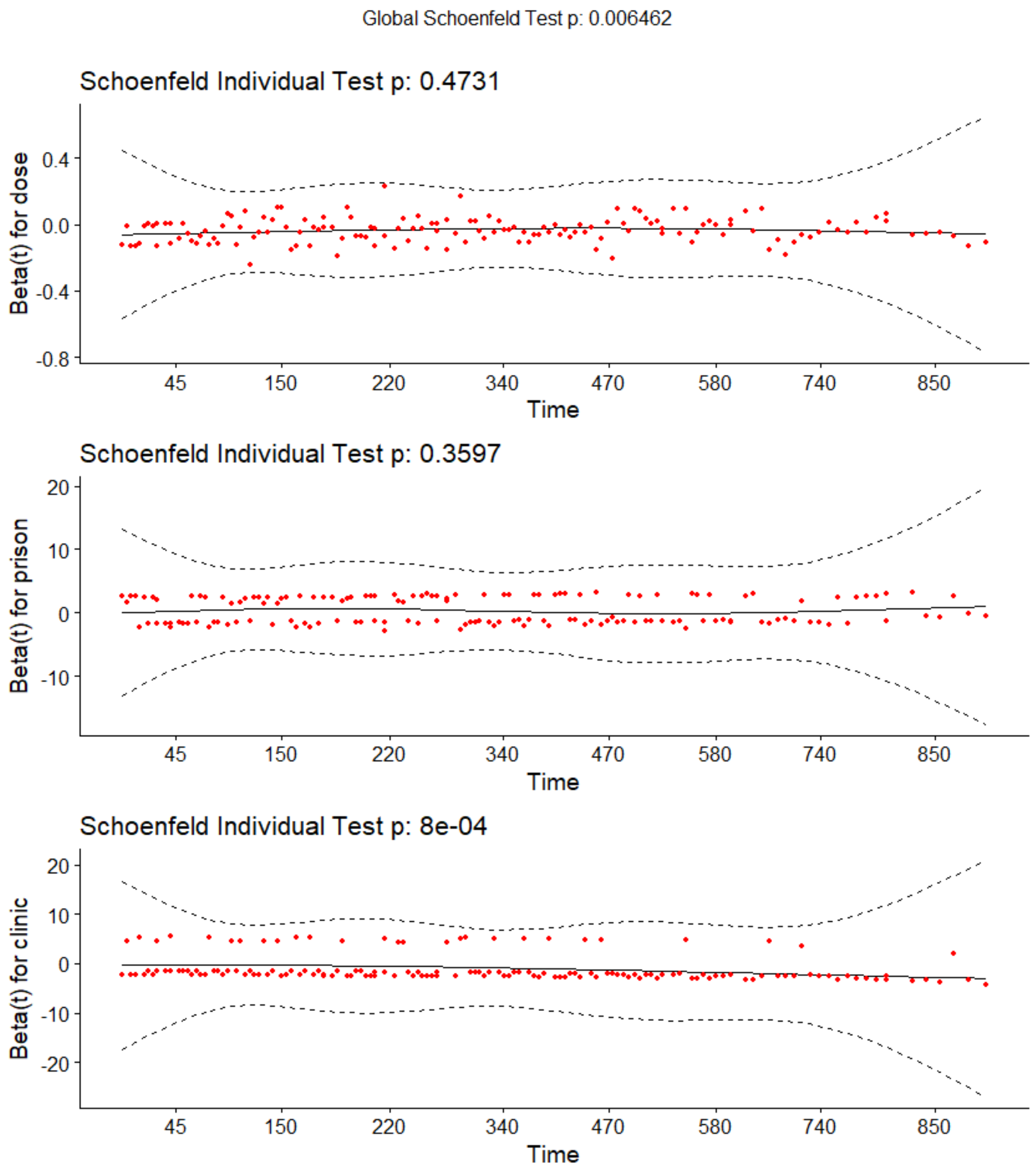
	methadone dose (mg/day)			methadone dose (mg/day)		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
dose	0.96	0.95 – 0.98	**<0.001**	0.97	0.95 – 0.98	**<0.001**
0=none, 1=prison record	1.21	0.88 – 1.67	0.248	1.39	1.00 – 1.92	0.051
Coded 1 or 2				0.36	0.24 – 0.56	**<0.001**
Observations	238			238		
R^2 Nagelkerke	0.149			0.238		

In the unadjusted model, the parameter estimate for dose was -0.036 (CI = (-0.042, -0.030), p<0.001) and 0.189 for prison record (CI = (0.025, 0.354), p=0.248) in the unadjusted model. A one mg per day increase in methadone (drug) is associated with 0.96 times the hazard of death, and compared to those with no prison record, those with a prison record are associated with 1.21 times the hazard of death. After adjusting for clinic, the parameter estimate for dose remained at about -0.036 (CI = (-0.042, -0.0290), p<0.001) and the parameter estimate for prison record was 0.323 (CI = (0.159, 0.494), p=0.051). A one mg per day increase in methadone (drug) is associated with 0.97 times the hazard of death, and compared to those with no prison record, those with a prison record are associated with 1.39 times the hazard of death.

1c. [4 points] In the model adjusting for clinic, assess the proportional hazards assumption. Does it appear to be violated for any variable?

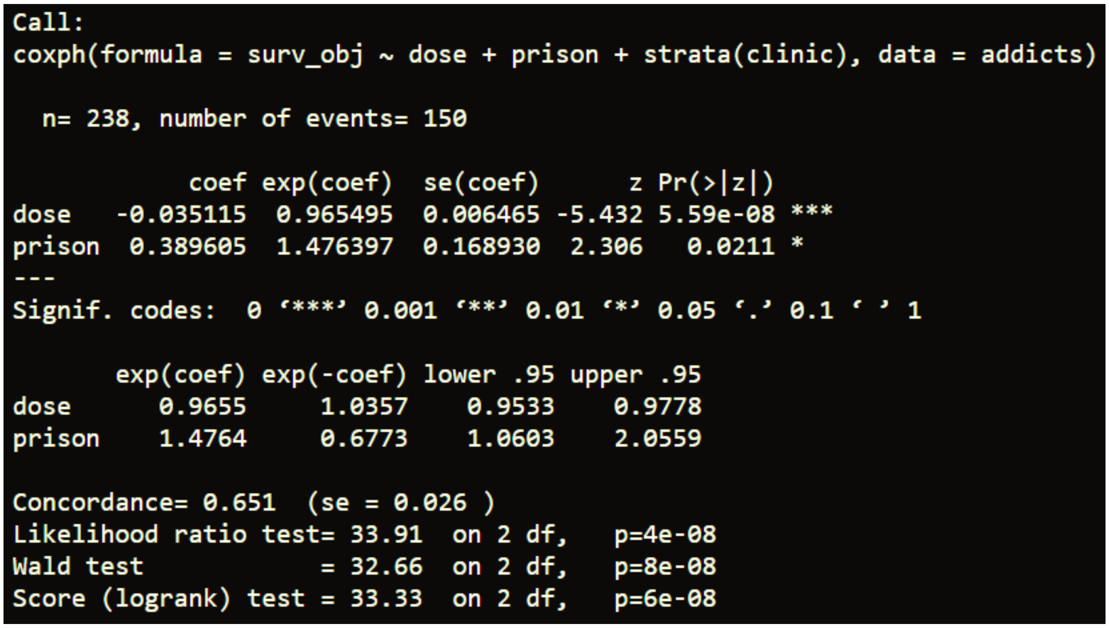


```
ggcoxzph(cox.zph(cox_adj_m))
```

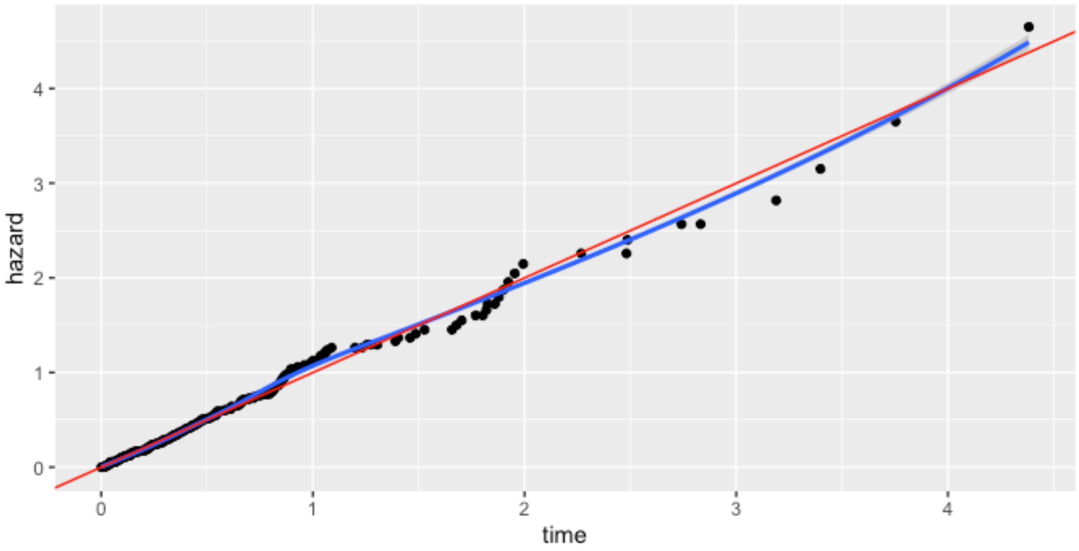


The proportional hazards assumption does not appear to be violated for dose (p=0.473) and for prison (p=0.360). However, the proportional hazards assumption seems to be violated for clinic (p<0.001).

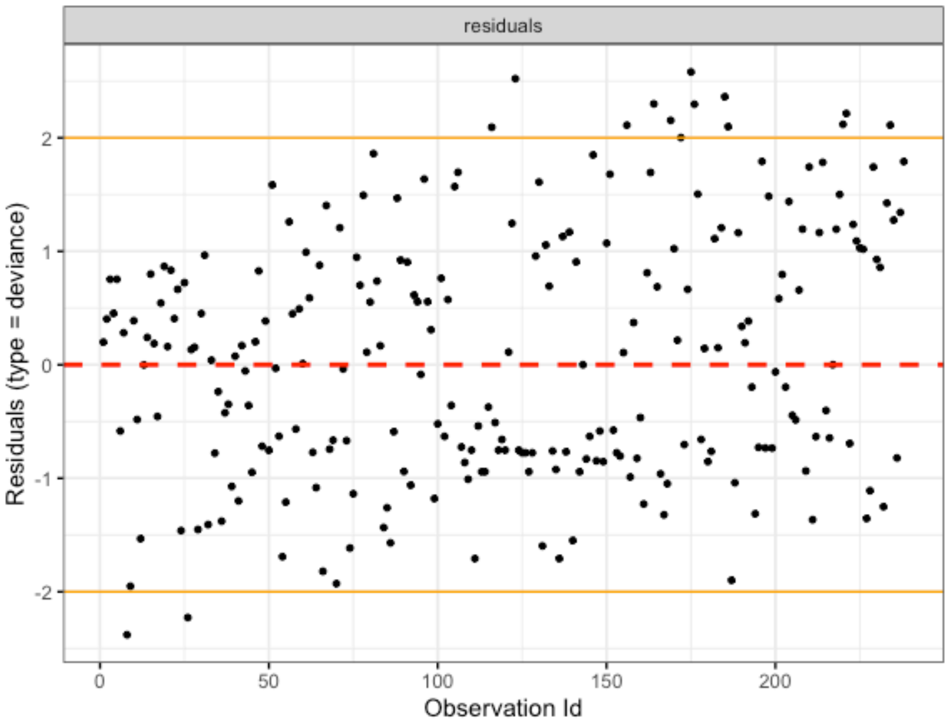
1d. [4 points] Re-fit your model to account for proportional hazards, and then assess your model goodness-of-fit. Comment on the Cox-Snell residuals, deviance residuals, and dfbeta values.

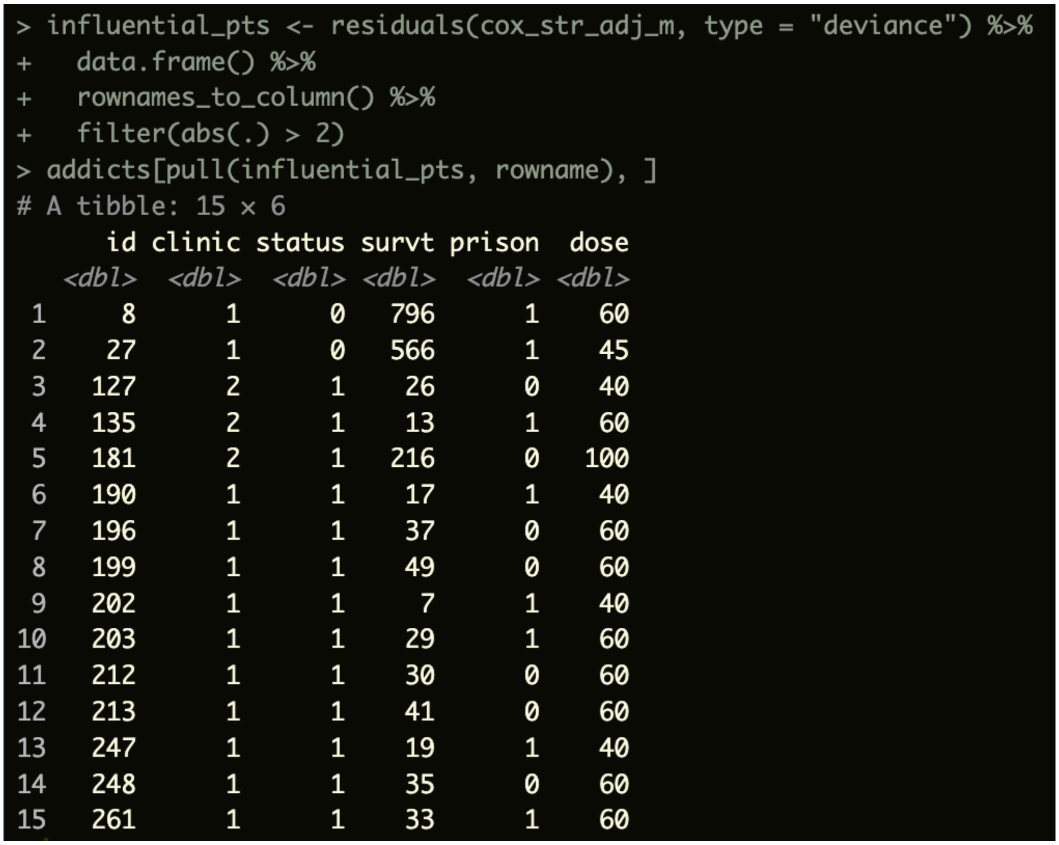


```
Call:
coxph(formula = surv_obj ~ dose + prison + strata(clinic), data = addicts)

  n= 238, number of events= 150 

            coef exp(coef)  se(coef)      z Pr(>|z|)    
dose   -0.035115  0.965495  0.006465 -5.432 5.59e-08 ***
prison  0.389605  1.476397  0.168930  2.306   0.0211 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
dose      0.9655     1.0357    0.9533    0.9778
prison    1.4764     0.6773    1.0603    2.0559

Concordance= 0.651  (se = 0.026 )
Likelihood ratio test= 33.91  on 2 df,   p=4e-08
Wald test            = 32.66  on 2 df,   p=8e-08
Score (logrank) test = 33.33  on 2 df,   p=6e-08
```



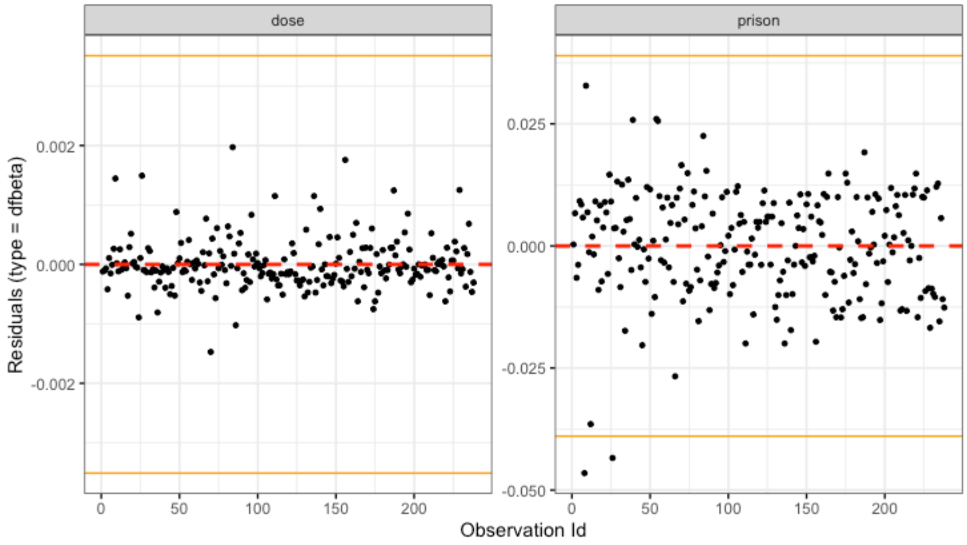
The Cox-Snell residuals follow a 45 degree line, indicating that the stratified Cox PH model is fit well.

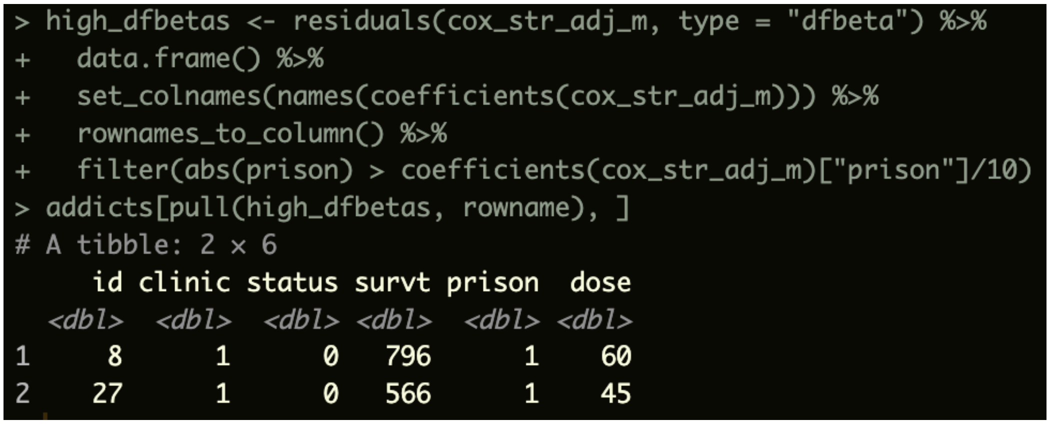




```
> influential_pts <- residuals(cox_str_adj_m, type = "deviance") %>%
+   data.frame() %>%
+   rownames_to_column() %>%
+   filter(abs(.) > 2)
> addicts[pull(influential_pts, rowname), ]
# A tibble: 15 × 6
      id clinic status survt prison  dose
   <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl>
 1     8      1      0   796      1    60
 2    27      1      0   566      1    45
 3   127      2      1    26      0    40
 4   135      2      1    13      1    60
 5   181      2      1   216      0   100
 6   190      1      1    17      1    40
 7   196      1      1    37      0    60
 8   199      1      1    49      0    60
 9   202      1      1     7      1    40
10   203      1      1    29      1    60
11   212      1      1    30      0    60
12   213      1      1    41      0    60
13   247      1      1    19      1    40
14   248      1      1    35      0    60
15   261      1      1    33      1    60
```

Observations 8 and 27 had high negative deviance residuals, likely due to not observing an event for them, even though they had a prison record. The other observations had deviance residuals greater than 2. Most of these individuals had a low survival time, meaning they experienced the event very early into the study, and their drug dosages were about average.





```
> high_dfbetas <- residuals(cox_str_adj_m, type = "dfbeta") %>%
+   data.frame() %>%
+   set_colnames(names(coefficients(cox_str_adj_m))) %>%
+   rownames_to_column() %>%
+   filter(abs(prison) > coefficients(cox_str_adj_m)["prison"]/10)
> addicts[pull(high_dfbetas, rowname), ]
# A tibble: 2 x 6
     id clinic status survt prison  dose
  <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl>
1     8      1      0   796      1    60
2    27      1      0   566      1    45
```

Again, observations 8 and 27 have high dfbeta values for prisonn, like due to not observing an event, even though they had a prison record.

1e. [5 points] Write a conclusion paragraph explaining the analysis you performed and the results you found, specifically addressing the research question.

A survival analysis was performed to determine if prison history and maximum dose of methadone was associated with hazard of death in recovering heroin addict patients. First, since dose is not a dichotomous variable, its functional form was determined by looking at the Martingale residuals as well as through the fractional polynomials method. Both methods confirmed that dose was linearly associated with hazard. Then, a Cox Proportional Hazards model was constructed using drug dose and prison record as independent variables. In the unadjusted model, the parameter estimate for dose was -0.036 (CI = (-0.042, -0.030), p<0.001) and 0.189 for prison record (CI = (0.025, 0.354), p=0.248) in the unadjusted model. A one mg per day increase in methadone (drug) is associated with 0.96 times the hazard of death, and compared to those with no prison record, those with a prison record are associated with 1.21 times the hazard of death. Then, another model was constructed, adjusting for clinic. After the adjustment, the parameter estimate for dose remained at about -0.036 (CI = (-0.042, -0.0290), p<0.001) and the parameter estimate for prison record was 0.323 (CI = (0.159, 0.494), p=0.051). A one mg per day increase in methadone (drug) is associated with 0.97 times the hazard of death, and compared to those with no prison record, those with a prison record are associated with 1.39 times the hazard of death. For the adjusted model, the Schoenfeld residuals were assessed to ensure the proportional hazards assumption was met. It was found that the assumption was violated for clinic, so the model was refit with a stratified Cox Proportional Hazards model, stratified by clinic. The Cox-Snell residuals of the stratified model suggested that the model was well-fit. Lastly, dfbetas and deviance residuals were assessed for influential points. Observations 8 and 27 had high negative dfbeta values due to not observing an event despite having a prison record. Overall, none of the deviance residuals or dfbeta values were extreme outliers, indicating that the model is well-fit.

Question 2 [10 points]

Read the article by Matas et al. (2015) about predictors of dropout in driving simulators.

2a. [5 points] Identify the following in this paper:

- The entry time under which observation began on individuals
- The criteria for determining the event, and the criteria for determining censoring
- Any time-dependent covariates

The entry time under which observation began on individuals was when the practice drive commenced.

The event of interest was defined as dropout or withdrawal during the driving task. Participants who completed all driving tasks (stages 1-5) were considered censored at stage 5.

The authors considered answers to the Simulator Sickness Questionnaire (SSQ) as a time-dependent covariate.

2b. [5 points] What variables did the authors include in their Cox proportional hazards model? What did they do to check the proportional hazards assumption, and was the assumption met? Based on this, do you feel confident about their model estimates?

The variables included in the Cox proportional hazards model were gender, motion sickness, other conditions, SSQ, Age, and MMSE (Mini-Mental State Examination). The authors say that they checked for interactions between the independent variables, and report that the models had met the assumption of proportional hazards for all variables except age. However, the study included older participants only and the authors cite that a similar study in young participants had found that younger people were less likely to drop out. They cite this as strong evidence that older adults are a high-risk group for simulator sickness. Based on this, I am confident in their model estimates.