PM592: Regression Analysis for Health Data Science

Lab 12 – Regression for Count Outcomes

Data Needed: *hmohiv.dta*

This lab is devoted entirely to the exercise.

Lab 12 Exercises

Objective(s):	Implement the tools for survival analysis, including Kaplan-Meier curves and Cox proportional hazards regression. Assess the fit of these models and interpret their output.
Datasets Required:	hmohiv

An HMO was evaluating the survival times of members that had been diagnosed with HIV.

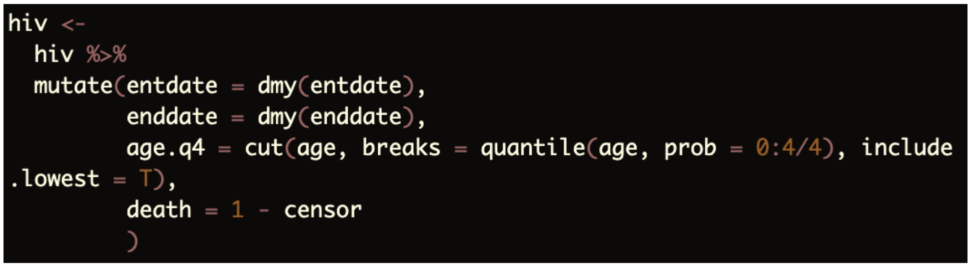
Does history of IV drug use increase risk of mortality?

Variable	Description
ID	Subject ID
entdate	Date of entry into the study
enddate	Date at which participant was censored
time	Time (months) of follow-up
age	Age (years) of participant at beginning of study
drug	History of IV drug use
censor	1=censored, 0=death

Prepare the data for analysis.

a) Convert the date variables to a date-type object.

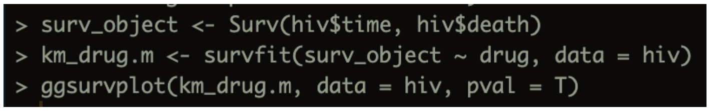
b) Create a variable that indicates whether the participant died.



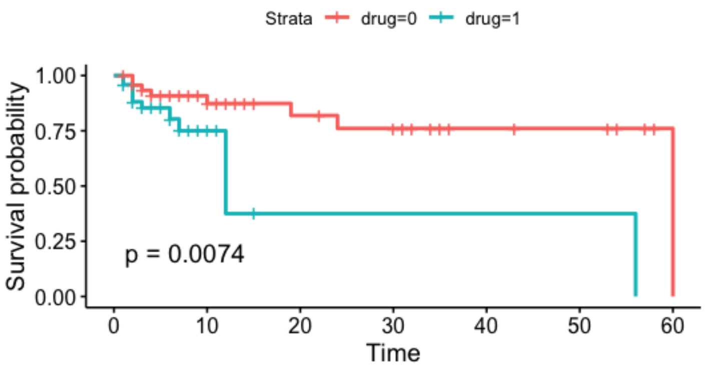
```
hiv <-
  hiv %>%
  mutate(entdate = dmy(entdate),
         enddate = dmy(enddate),
         age.q4 = cut(age, breaks = quantile(age, prob = 0:4/4), include
.lowest = T),
         death = 1 - censor
         )
```

1) Perform exploratory data analysis by examining the Kaplan-Meier curves.

c) Does IV drug use appear related to survival?

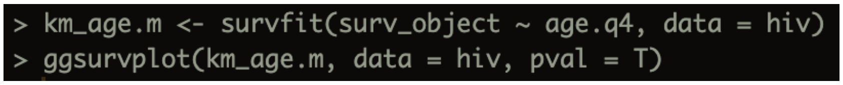


```
> surv_object <- Surv(hiv$time, hiv$death)
> km_drug.m <- survfit(surv_object ~ drug, data = hiv)
> ggsurvplot(km_drug.m, data = hiv, pval = T)
```

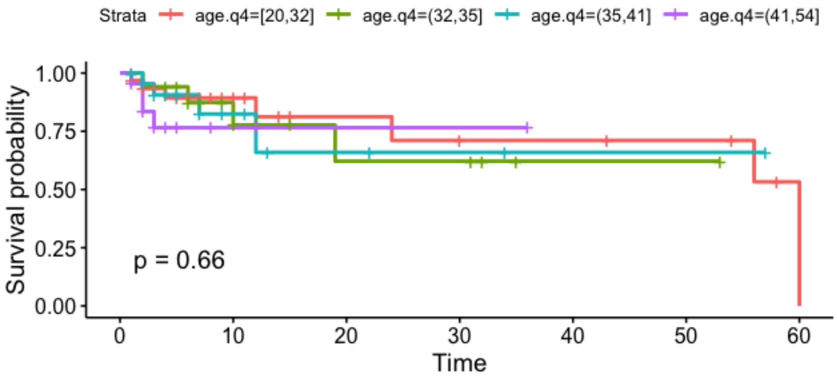


Yes, IV drug use does appear to be related to survival (p=0.0074). Those with history of IV drug use appear to die earlier than those with no history of IV drug use.

d) Does age appear related to survival? (Create a quartile variable.)



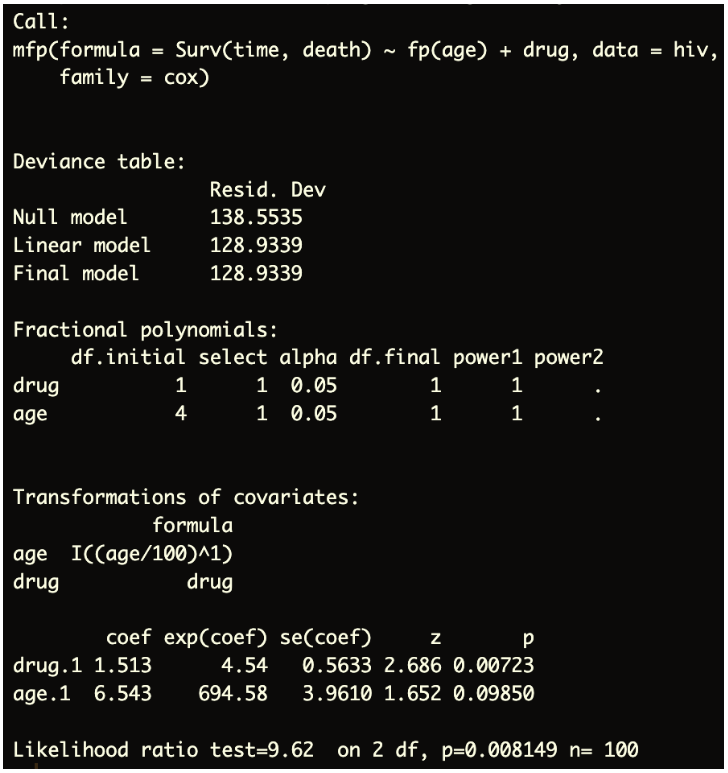
```
> km_age.m <- survfit(surv_object ~ age.q4, data = hiv)
> ggsurvplot(km_age.m, data = hiv, pval = T)
```



Age does not appear to be related to survival (p=0.66), the curves look similar across the four quartiles.

2) Examine the functional form for age.

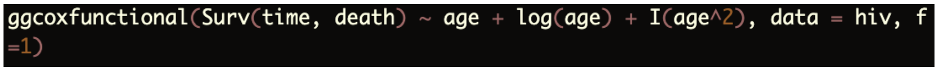
a) Is baseline age related to hazard, based on the MFP procedure?



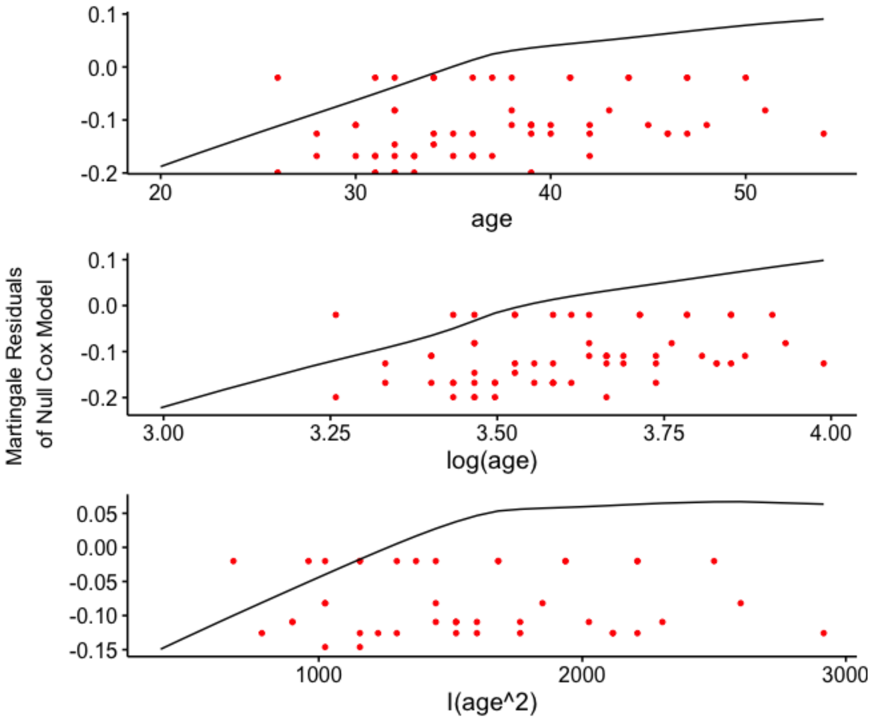
```
Call:
mfp(formula = Surv(time, death) ~ fp(age) + drug, data = hiv,
    family = cox)


Deviance table:
                 Resid. Dev
Null model       138.5535
Linear model     128.9339
Final model      128.9339

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


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

        coef exp(coef) se(coef)     z       p
drug.1 1.513      4.54   0.5633 2.686 0.00723
age.1  6.543    694.58   3.9610 1.652 0.09850

Likelihood ratio test=9.62  on 2 df, p=0.008149 n= 100
```

According to fractional polynomials, it appears that baseline age is linearly related to hazard.

b) Examine the Martingale residuals vs. age. Is the relationship roughly linear?



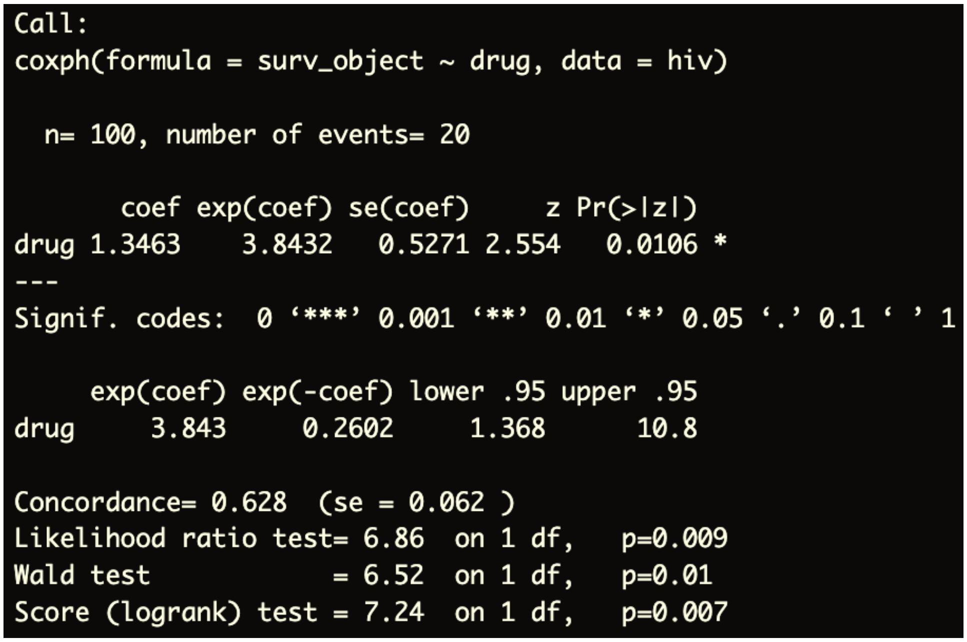
```
ggcoxfunctional(Surv(time, death) ~ age + log(age) + I(age^2), data = hiv, f
=1)
```



It appears roughly linear in the middle, less so in at the tails.

3) Construct the Cox PH models.

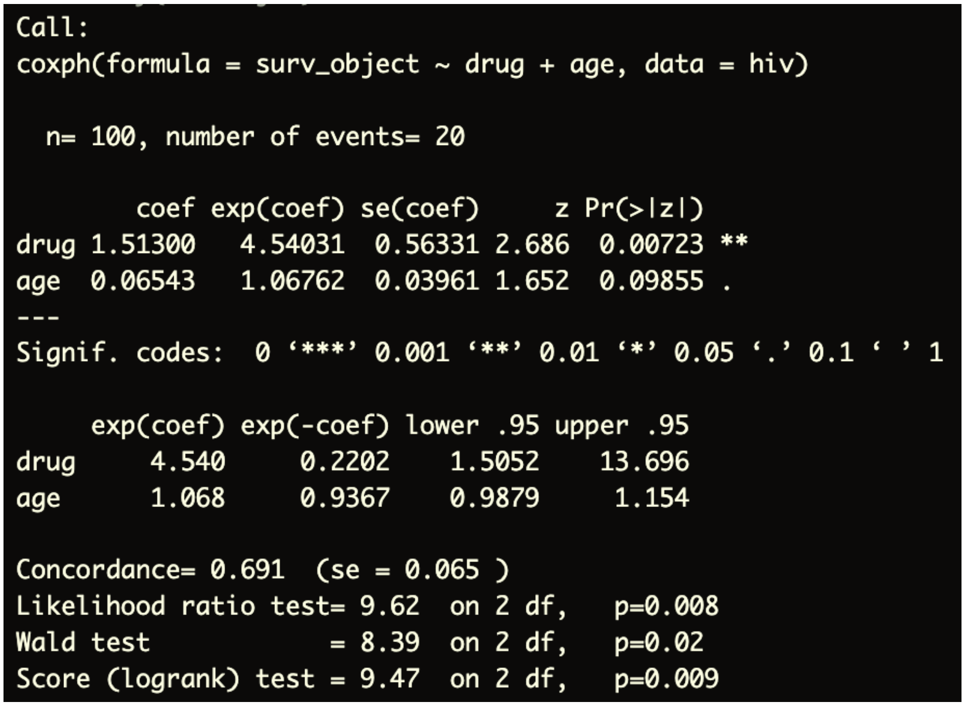
a) Construct a model with drug use predicting death. What is the parameter estimate, and what is the interpretation of this estimate?



```
Call:
coxph(formula = surv_object ~ drug, data = hiv)

  n= 100, number of events= 20

       coef exp(coef) se(coef)     z Pr(>|z|)
drug 1.3463    3.8432   0.5271 2.554   0.0106 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
drug     3.843     0.2602     1.368      10.8

Concordance= 0.628  (se = 0.062 )
Likelihood ratio test= 6.86  on 1 df,   p=0.009
Wald test            = 6.52  on 1 df,   p=0.01
Score (logrank) test = 7.24  on 1 df,   p=0.007
```

Compared to those who do not have a history of IV drug use, those who do have a history of IV drug use have 3.84 times the hazard of death (p=0.011, concordance = 0.628).

b) Add age to the model. Does it change the parameter estimate for drug use? By how much?

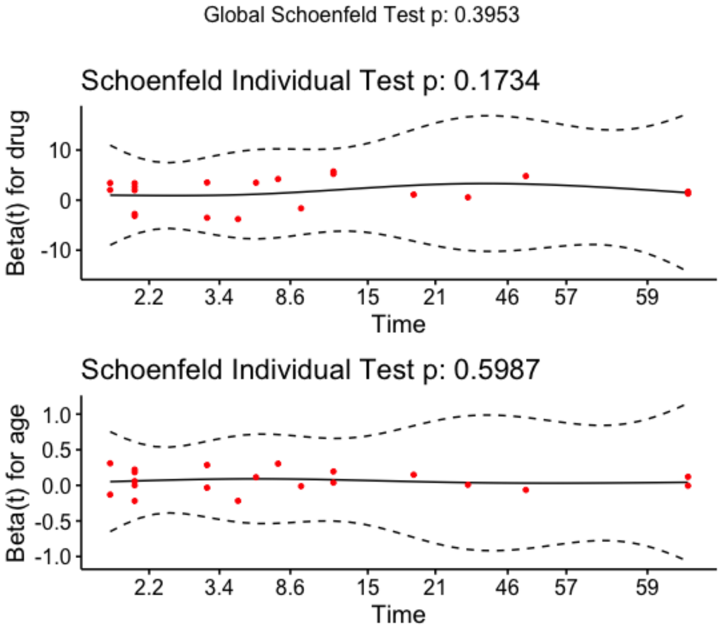


```
Call:
coxph(formula = surv_object ~ drug + age, data = hiv)

  n= 100, number of events= 20

        coef exp(coef) se(coef)     z Pr(>|z|)
drug 1.51300   4.54031  0.56331 2.686  0.00723 **
age  0.06543   1.06762  0.03961 1.652  0.09855 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
drug     4.540     0.2202    1.5052    13.696
age      1.068     0.9367    0.9879     1.154

Concordance= 0.691  (se = 0.065 )
Likelihood ratio test= 9.62  on 2 df,   p=0.008
Wald test            = 8.39  on 2 df,   p=0.02
Score (logrank) test = 9.47  on 2 df,   p=0.009
```

Yes, age does change the parameter estimate for drug from 1.3463 to 1.513, which is a change of about 11%. Age should be included in the model, those that are older are more likely to have history of drug use and are more likely to die as well.

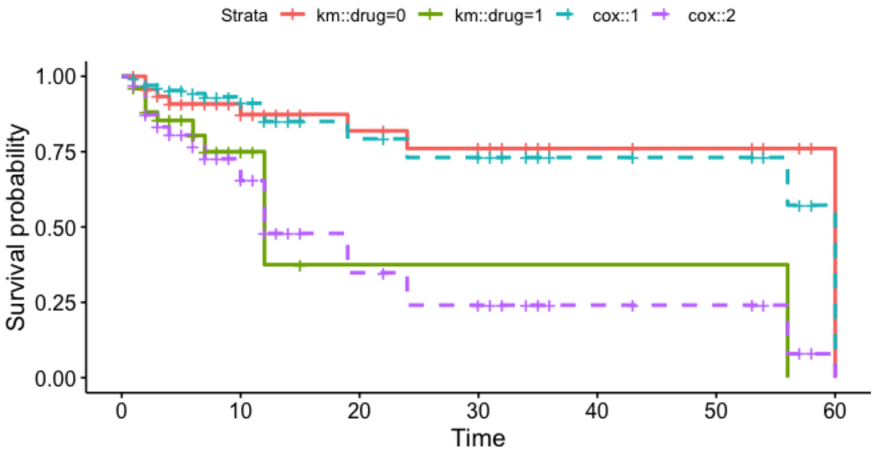
4) Assess the proportional hazards assumption.

a) Do the Schoenfeld residuals indicate that the proportional hazards assumption is violated for any variable?



No, the Schoenfeld residuals are relatively flat, indicating that the effect of age and drug use is constant across time. The proportional hazards assumption is not violated for age (p=0.60) and not violated for drug (p=0.17).

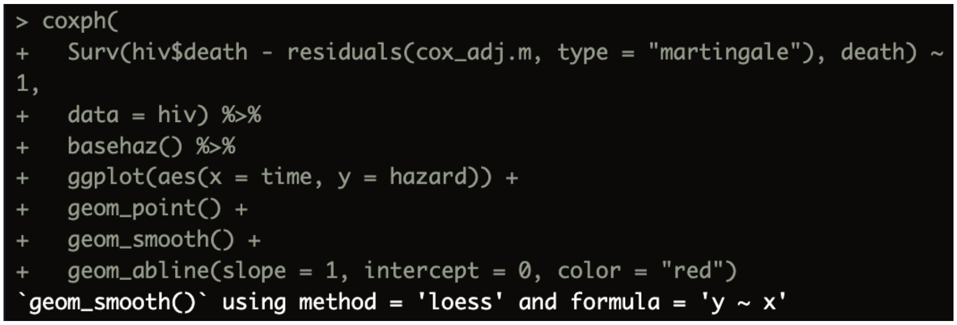
b) Use the code in the lab file to produce a plot combining the Kaplan-Meier and Cox-estimated survival curves. Do the Cox curves appear to approximate the KM curves?



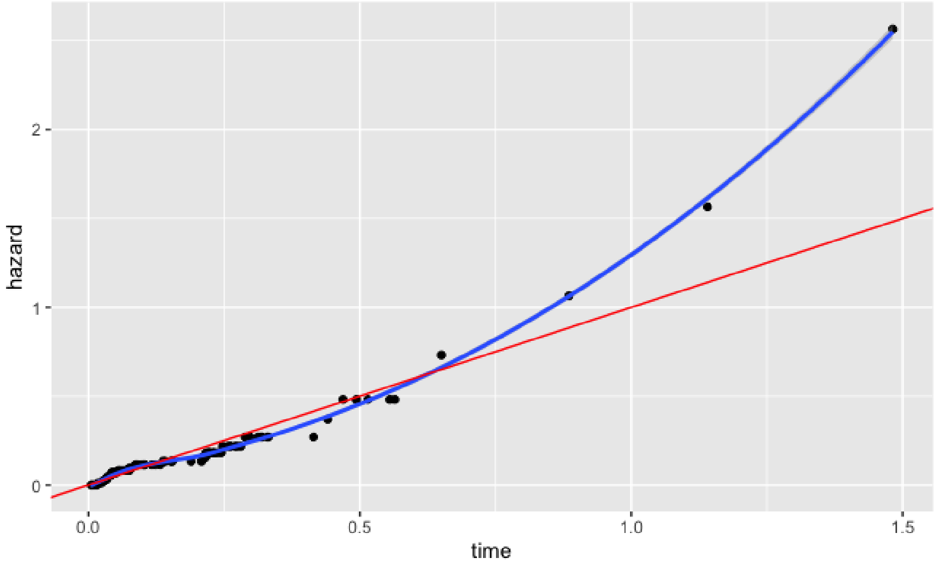
Yes the cox curves appear to approximate the KM curves. The solid lines are the KM curves, and the dotted lines are the Cox curves.

5) Assess model goodness of fit.

a) When the Cox-Snell residuals are used as the time variable, is the estimated hazard rate approximately equal to 1?

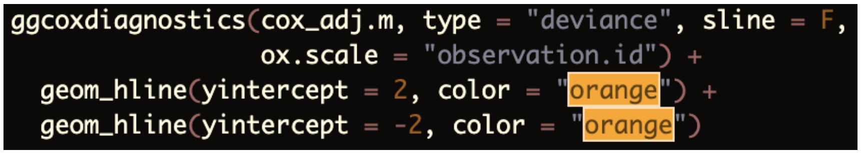


```
> coxph(
+   Surv(hiv$death - residuals(cox_adj.m, type = "martingale"), death) ~
1,
+   data = hiv) %>%
+   basehaz() %>%
+   ggplot(aes(x = time, y = hazard)) +
+   geom_point() +
+   geom_smooth() +
+   geom_abline(slope = 1, intercept = 0, color = "red")
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

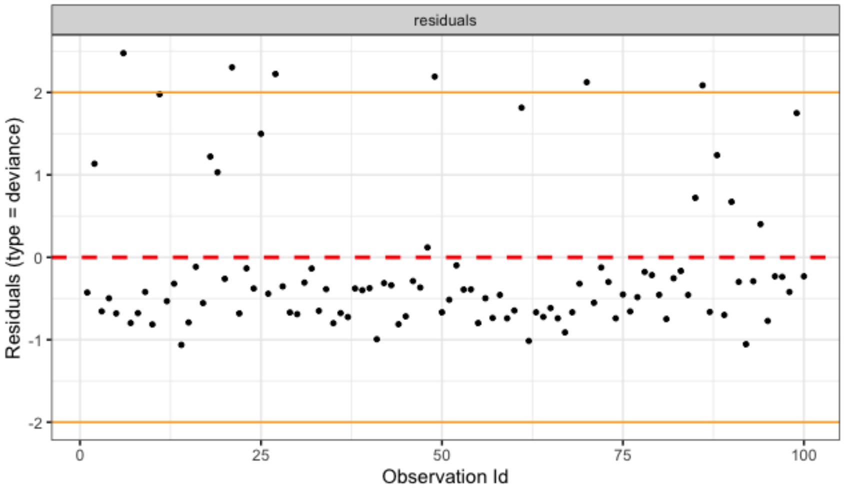


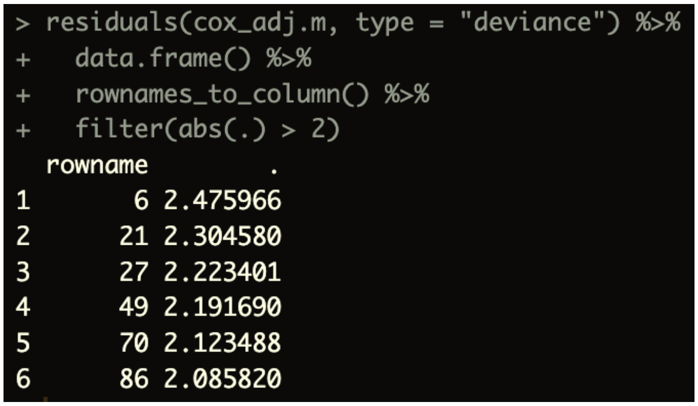
For the majority of data points, they do follow a 45 degree line (slope of 1).

b) Do the deviance residuals indicate any influential observations? What is the value of the deviance residual with the highest magnitude? Why did the observation have such a high deviance residual value?



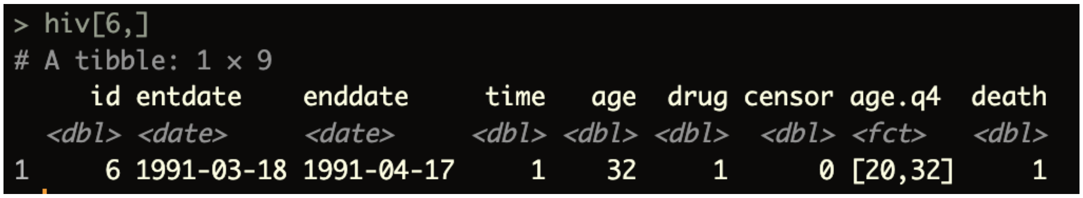
```
ggcoxdiagnostics(cox_adj.m, type = "deviance", sline = F,
                 ox.scale = "observation.id") +
  geom_hline(yintercept = 2, color = "orange") +
  geom_hline(yintercept = -2, color = "orange")
```





```
> residuals(cox_adj.m, type = "deviance") %>%
+   data.frame() %>%
+   rownames_to_column() %>%
+   filter(abs(.) > 2)
  rowname        .
1       6 2.475966
2      21 2.304580
3      27 2.223401
4      49 2.191690
5      70 2.123488
6      86 2.085820
```

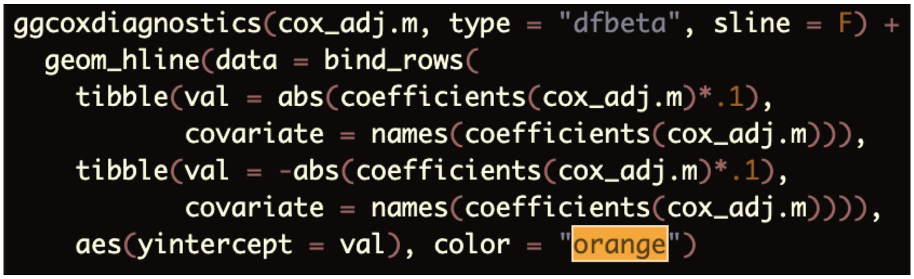
It appears that observation 6 is the most influential with the highest residual.



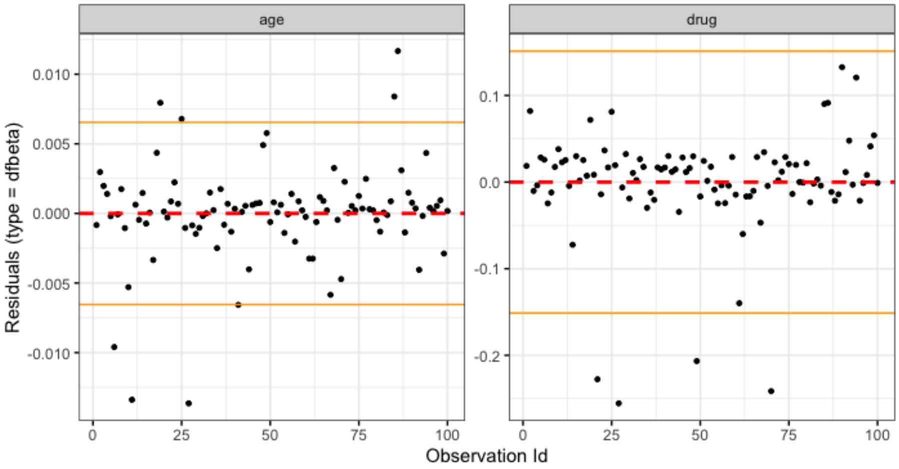
```
> hiv[6,]
# A tibble: 1 × 9
     id entdate    enddate     time   age  drug censor age.q4  death
  <dbl> <date>     <date>     <dbl> <dbl> <dbl>  <dbl> <fct>   <dbl>
1     6 1991-03-18 1991-04-17     1    32     1      0 [20,32]     1
```

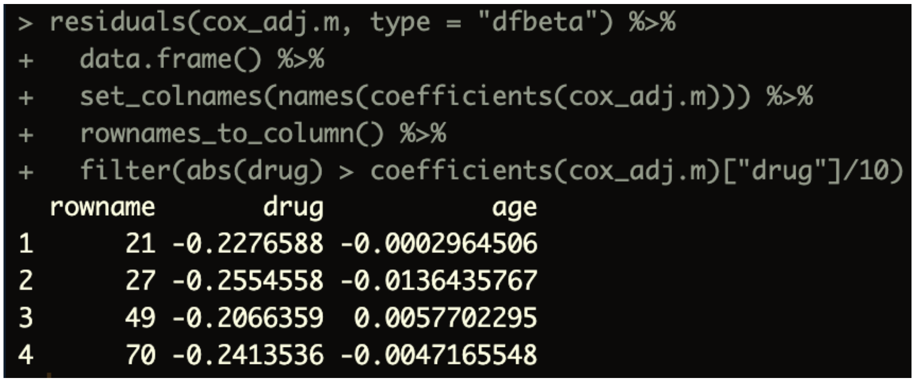
Aged 32, had history of IV drug use and died within a month.

c) Do the dfbetas indicate any influential observation with regard to age or drug? Which observations? Why did these observations have such a high dfbeta value?

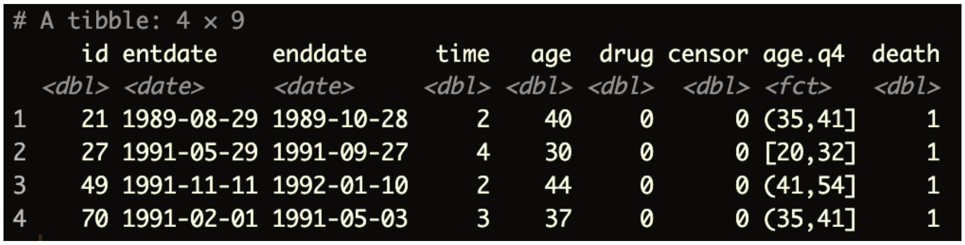


```
ggcoxdiagnostics(cox_adj.m, type = "dfbeta", sline = F) +
  geom_hline(data = bind_rows(
    tibble(val = abs(coefficients(cox_adj.m)*.1),
           covariate = names(coefficients(cox_adj.m))),
    tibble(val = -abs(coefficients(cox_adj.m)*.1),
           covariate = names(coefficients(cox_adj.m)))),
    aes(yintercept = val), color = "orange")
```





```
> residuals(cox_adj.m, type = "dfbeta") %>%
+   data.frame() %>%
+   set_colnames(names(coefficients(cox_adj.m))) %>%
+   rownames_to_column() %>%
+   filter(abs(drug) > coefficients(cox_adj.m)["drug"]/10)
  rowname       drug           age
1      21 -0.2276588 -0.0002964506
2      27 -0.2554558 -0.0136435767
3      49 -0.2066359  0.0057702295
4      70 -0.2413536 -0.0047165548
```



```
# A tibble: 4 × 9
     id entdate    enddate     time   age  drug censor age.q4  death
  <dbl> <date>     <date>     <dbl> <dbl> <dbl>  <dbl> <fct>   <dbl>
1    21 1989-08-29 1989-10-28     2    40     0      0 (35,41]     1
2    27 1991-05-29 1991-09-27     4    30     0      0 [20,32]     1
3    49 1991-11-11 1992-01-10     2    44     0      0 (41,54]     1
4    70 1991-02-01 1991-05-03     3    37     0      0 (35,41]     1
```

Observations 21, 27, 49, and 70 were influential observations with regard to drug. All had no history of IV drug use but had event early on (within 2 – 4 months).

6) Write a conclusion paragraph explaining the analysis you performed and results you found, specifically addressing the research question.

A survival analysis was performed to analyze whether history of IV drug use is associated with higher mortality in HIV patients. First Kaplan-Meier curves were constructed to look at differences in survival between drug vs no drug and between age quartile groups. IV drug use was found to be related to survival (p=0.0074) but age was not (p=0.66). Next, the functional form for age was determined through the fractional polynomials approach and looking at age vs the Martingale residuals. Both methods suggested that age as a linear variable was related to hazard. Then, an unadjusted Cox Proportional Hazards model was constructed with drug use as the covariate of interest, which found that compared to those who do not have a history of IV drug use, those who do have a history of IV drug use have 3.84 times the hazard of death (p=0.011, concordance = 0.628). Then, another Cox Proportional Hazards model was constructed adjusting for age, which did change the parameter estimate for drug from 1.3463 to 1.513, about 11%. The hazard ratio increased to 4.54 for those with history of drug use. For the adjusted model, the Schoenfeld residuals were analyzed for the proportional hazards assumption, which appears to be met. The Cox-Snell residuals also show conformity to a 45 degree line, indicating the model is fit well. Lastly, the deviance residuals and dfbeta values were assessed for influential points. Some observations were found to influence parameter estimates because of a low time to event when there was no history of drug use, but overall the dataset as a whole was fit well by the Cox PH model.