Introduction to TTE modeling: Workbook 4

Cox regression

2023-07-11

Contents

Preliminaries for R examples

```
library(tidyverse)
library(stringr)
library(survival)
library(survminer)
library(texreg)
library(mgcv)
library(flexsurv)
library(muhaz)
library(Hmisc)

theme_set(theme_bw())
```

```
load('../data/aedat.RDS')

aedat <-
  aedat %>%
  mutate(AETOXGR = factor(aedat$AETOXGR, 0:3, labels=c("None","Mild","Moderate","Severe")),
         ae_any = AETOXGR != 'None') %>%
  group_by(USUBJID) %>%
  # End of study for patients without a severe event
  mutate(TTE_SEVERE = case_when(
    STUDYID=="PROTA" ~ 2,
    STUDYID=="PROTB" ~ 6
  ),
  # Time of severe event for those that had one
  TTE_SEVERE = ifelse(AETOXGR=="Severe", TTE, TTE_SEVERE)
  )

# Both for EDA and for model-checking, it's generally helpful to have quartiles of exposure:
dat_use <-
  aedat %>% arrange(USUBJID, TTE_SEVERE) %>% slice(1) %>%
  group_by(PBO) %>%
  mutate(Quartile = ifelse(PBO == "PBO", "PBO",
```

```
                            paste0("Q", ntile(CAVGSS, n = 4)))) %>%
  ungroup() %>%
  mutate(rowid = 1:n())
```

Workbook 4: Cox regression

We'll start by fitting a model to time to a severe event as a function of CAVGSS

```
cox_model01 <- coxph(Surv(TTE_SEVERE, AE01) ~ CAVGSS, data=dat_use)
```

```
print(cox_model01)
```

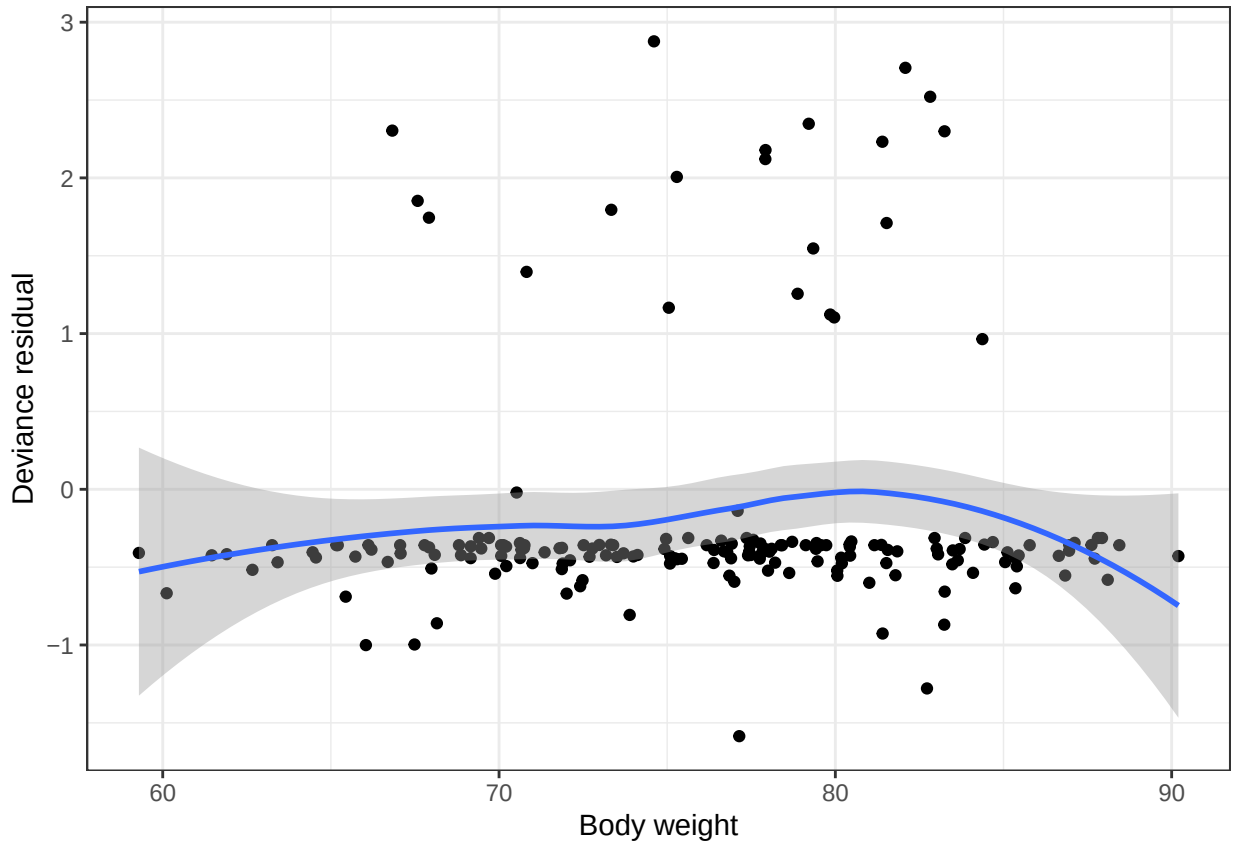
```
. Call:
. coxph(formula = Surv(TTE_SEVERE, AE01) ~ CAVGSS, data = dat_use)
.
.           coef exp(coef) se(coef)     z        p
. CAVGSS 0.51681   1.67667  0.09061 5.704 1.17e-08
.
. Likelihood ratio test=24.45  on 1 df, p=7.624e-07
. n= 180, number of events= 24
```

Let's look at the deviance residuals

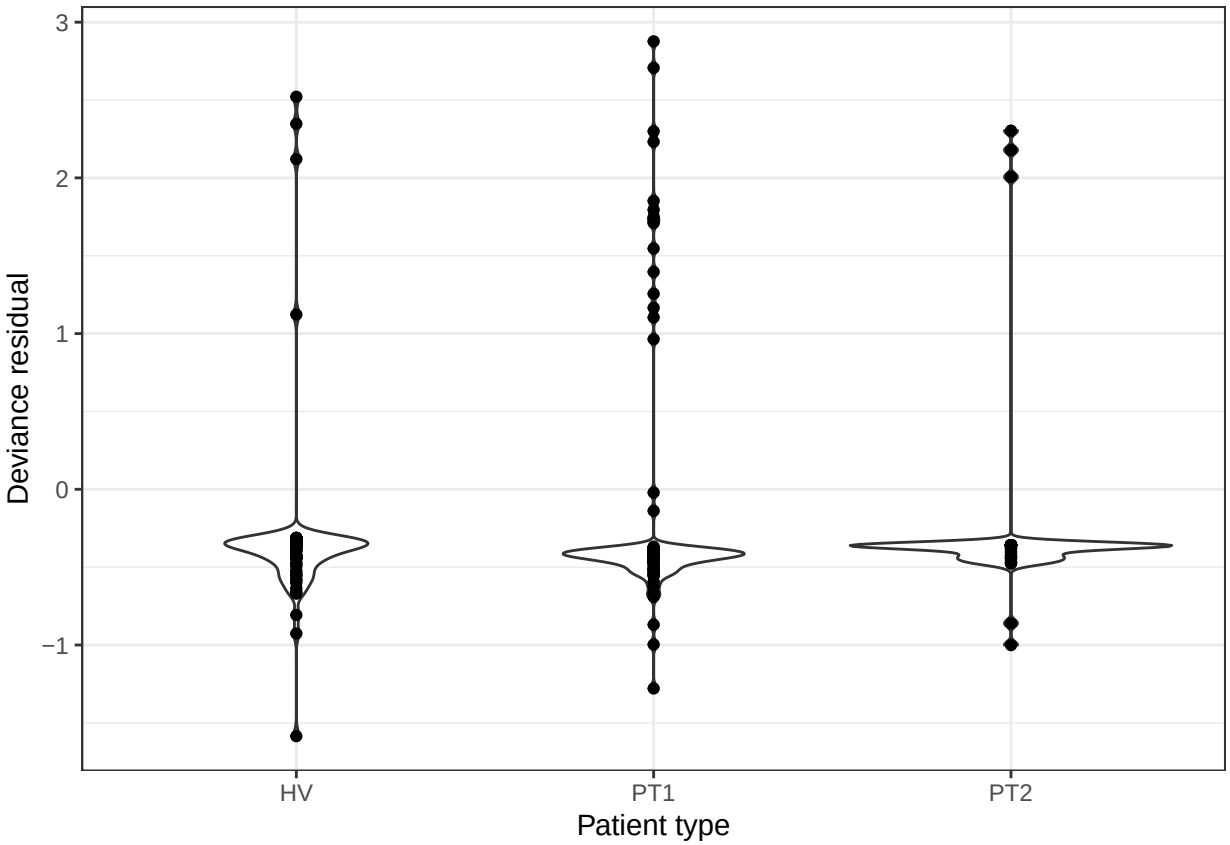
- Versus body weight and patient type, to see if there is a relationship with covariates not in our model
- Versus CAVGSS to see if we've modeled the relationship reasonably well

```
dat_use <- dat_use %>%
  mutate(resid01 = residuals(cox_model01, type='deviance'))
```

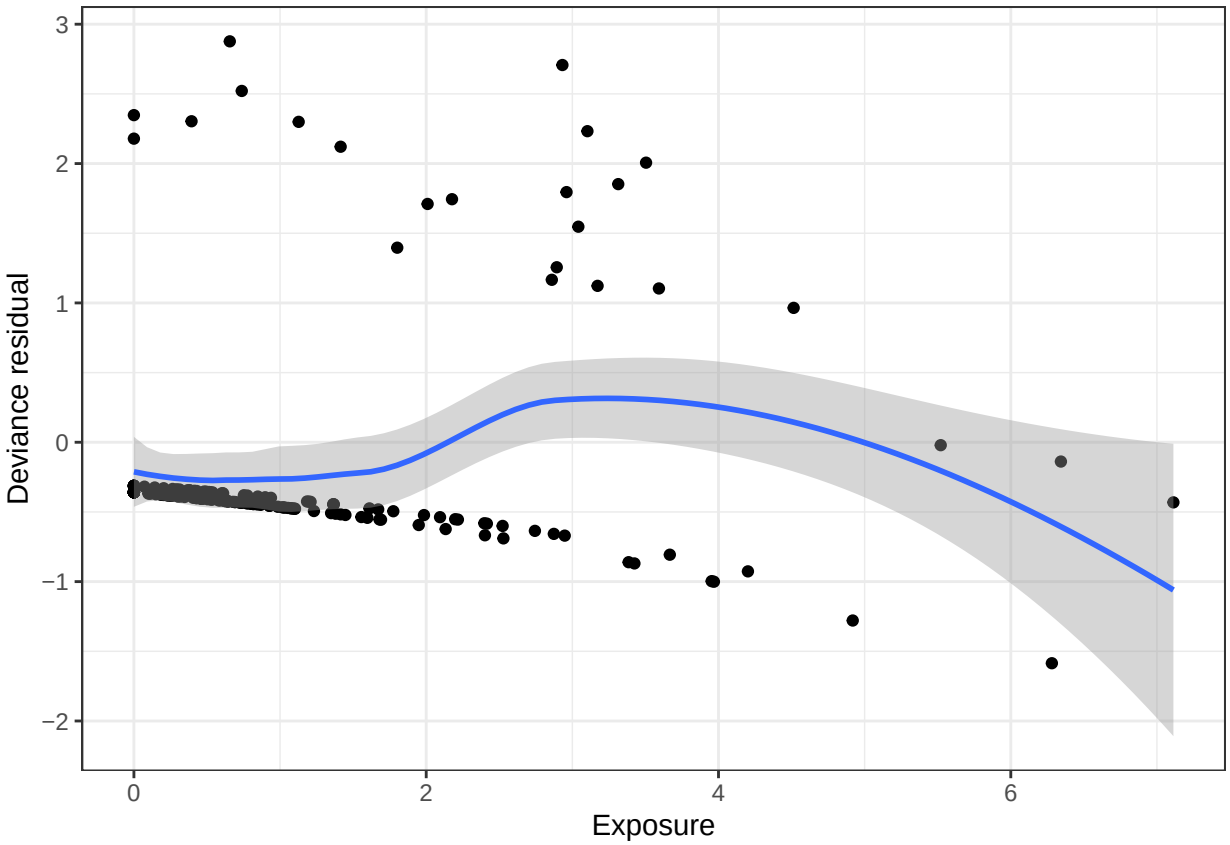
```
dat_use %>%
  ggplot(aes(x=BWT, y=resid01)) +
  geom_point() +
  geom_smooth() +
  labs(x='Body weight', y='Deviance residual')
```




```
dat_use %>%
  ggplot(aes(x=PTTYPE, y=resid01)) +
  geom_violin() +
  geom_point() +
  labs(x='Patient type', y='Deviance residual')
```




```
dat_use %>%
  ggplot(aes(x=CAVGSS, y=resid01)) +
  geom_point() +
  geom_smooth() +
  labs(x='Exposure', y='Deviance residual')
```



Do you see any evidence that we should modify our model based on these plots? If so, what leads you to that decision and how might you change the model?

Answer:

The residual plots versus body weight and patient type look to be reasonably centered around 0. However, the plot versus average steady-state exposure shows a trend indicating that the exposure-response relationship may not be linear on the log scale.

Let's make some predictions and overlay the model fits

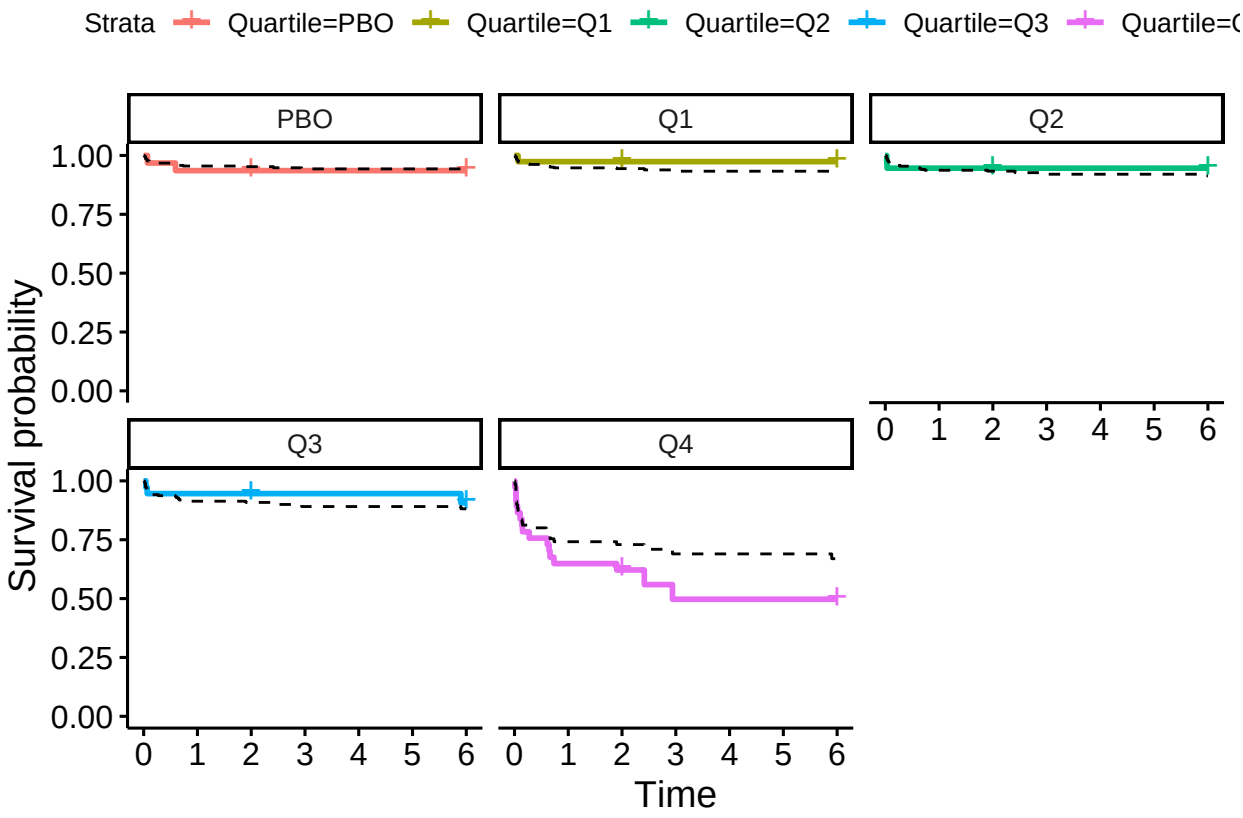
```
preds_mod01 <- survfit(cox_model01, newdata = dat_use) %>% survfit0()
```

```
# Make a tall version of the data
survival_preds_fit1 <- preds_mod01$surv %>%
  as.data.frame() %>%
  mutate(time = preds_mod01$time) %>%
  pivot_longer(cols=-time) %>%
  # Add a row id column for merging with the covariate data
  mutate(rowid = as.numeric(name))

# Merge predictions and covariates for plotting
survival_preds_exposure <- survival_preds_fit1 %>%
  left_join(dat_use) %>%
  # Because CAVGSS is continuous, we need to plot using categories
```

```
  # So, we'll use exposure quartiles
  group_by(time, Quartile) %>%
  # Calculate expected value for S(t) across subjects in each group
  summarise(est=mean(value))

# Plot the observed data (the Kaplan-Meier estimate)
ggsurvplot(survfit(Surv(TTE_SEVERE, AE01) ~ Quartile, data=dat_use),
           data=dat_use)$plot +
  # Overlay the predictions
  geom_step(data=survival_preds_exposure,
            aes(x=time, y=est), linetype='dashed') +
  facet_wrap(~Quartile
             )
```



Does this plot align with your expectation based on the residual plots?

Answer:

Yes. The residual plot generally showed a negative average residual through exposures of 2 (corresponding to the first three quartiles) and positive residuals in the fourth quartile. This is reflected in the predicted survival as slight over prediction of S(t) in Q1-Q3 and under prediction in Q4.

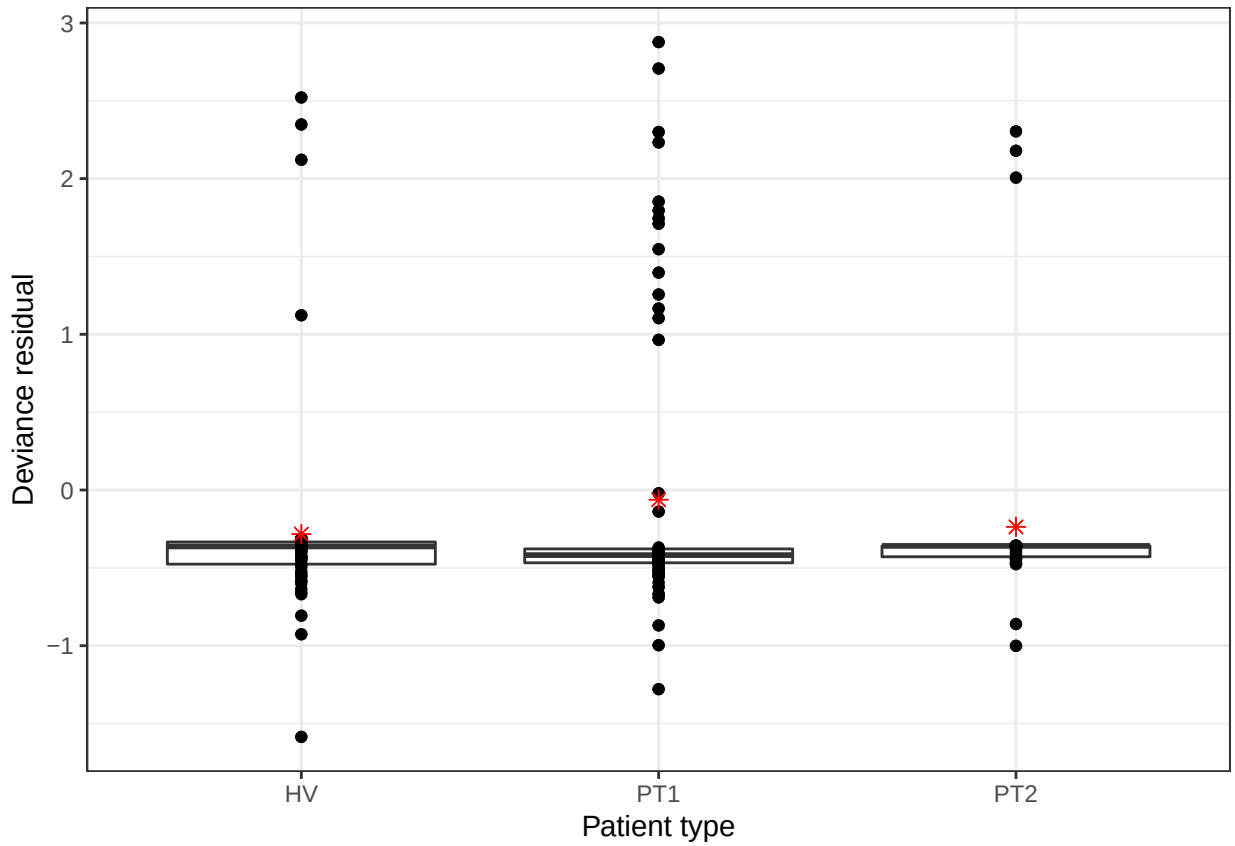
Exercises

- Make similar plots for patient type and body weight. Do these align with your expectation based on the residual plots?
  - Hint: you'll need to make a categorical variable from body weight. The `ntile()` function is a handy tool for this.

Answer:

First, patient type. Recall, the deviance residual plot didn't show much of a trend:

```
dat_use %>%
  ggplot(aes(x=PTTYPE, y=resid01)) +
  geom_boxplot() +
  geom_point() +
  stat_summary(fun=mean, geom='point',col='red', shape='asterisk', size=2) +
  labs(x='Patient type', y='Deviance residual')
```

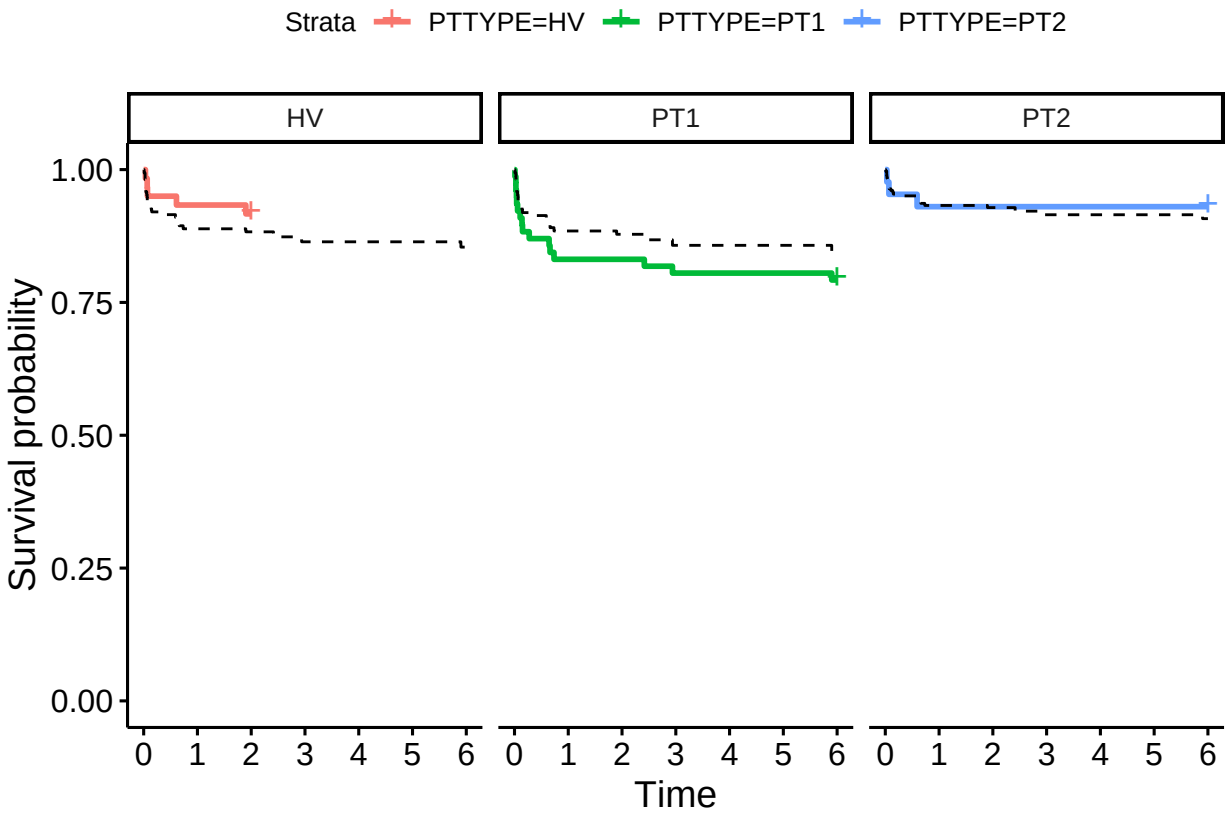


However, the plot based on the predicted survival curves shows a different pattern, with a hit of underprediction of survival in the PT1 group.

```
# Merge predictions and covariates for plotting
survival_preds_pttype <- survival_preds_fit1 %>%
  left_join(dat_use) %>%
  # Group by PTTYPE
  group_by(time, PTTYPE) %>%
  # Calculate expected value for S(t) across subjects in each group
```

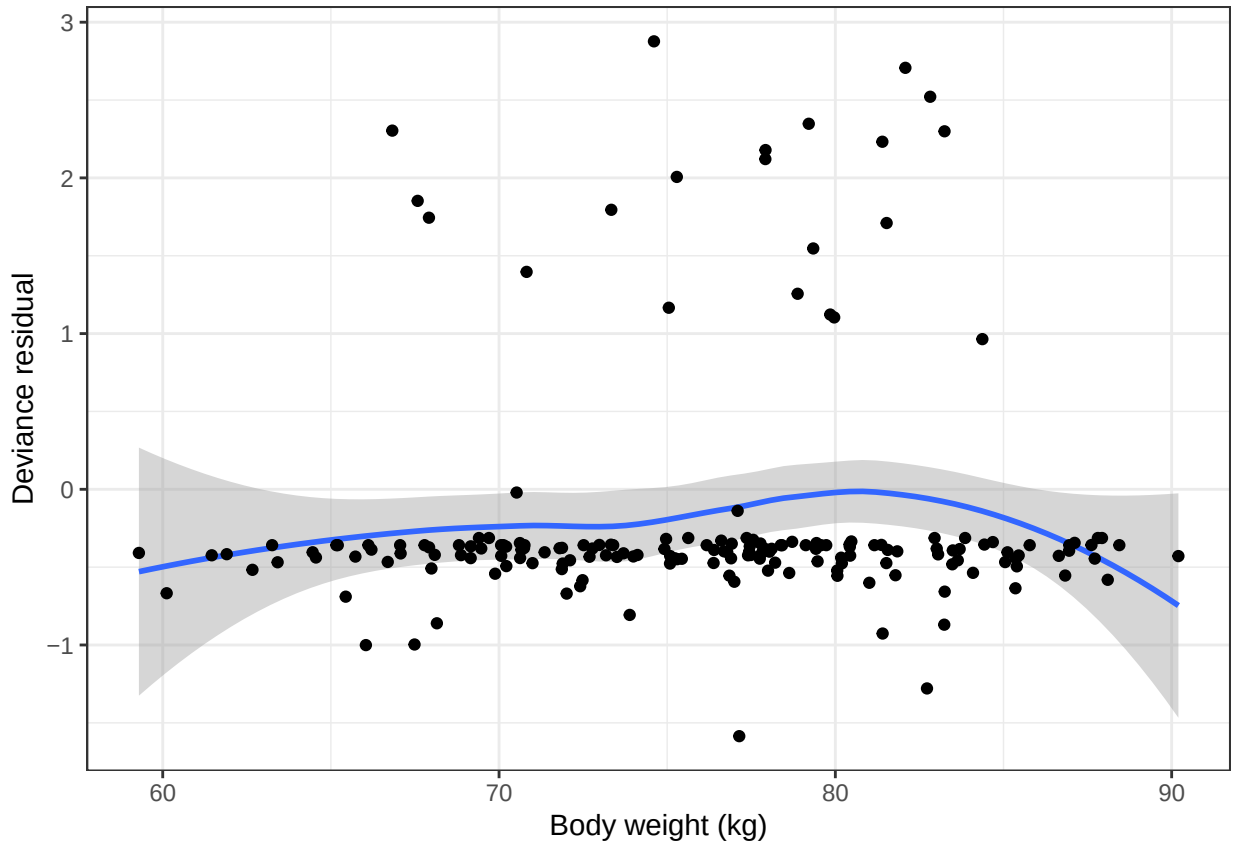
```
  summarise(est=mean(value))

# Plot the observed data (the Kaplan-Meier estimate)
ggsurvplot(survfit(Surv(TTE_SEVERE, AE01) ~ PTTYPE, data=dat_use),
           data=dat_use)$plot +
  # Overlay the predictions
  geom_step(data=survival_preds_pttype,
            aes(x=time, y=est), linetype='dashed') +
  facet_wrap(~PTTYPE
             )
```



Second, body weight. Recall, the deviance residual plot didn't show much of a trend:

```
dat_use %>%
  ggplot(aes(x=BWT, y=resid01)) +
  geom_smooth() +
  geom_point() +
  labs(x='Body weight (kg)', y='Deviance residual')
```

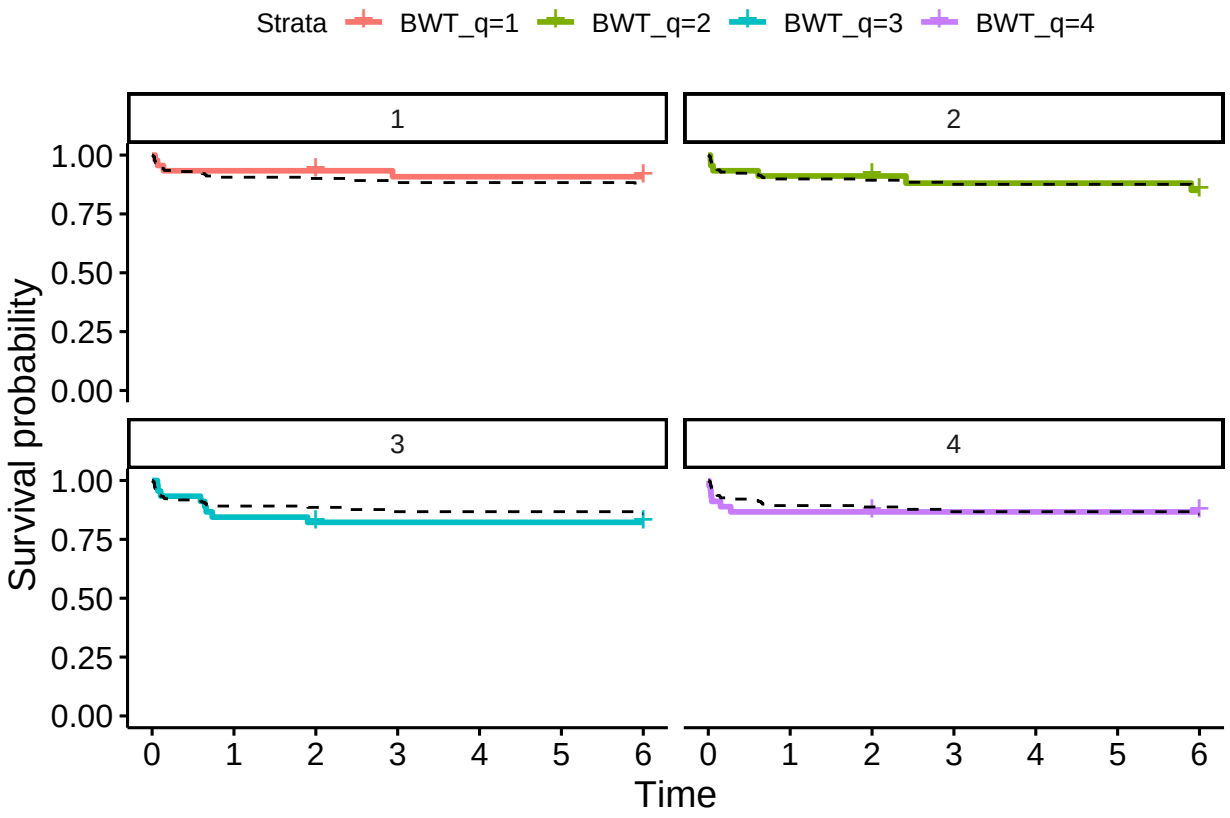


For body weight, the predictions align closely to the observed.

```
# Make body weight quartile variable
dat_use <- dat_use %>% mutate(BWT_q = ntile(BWT,4))

# Merge predictions and covariates for plotting
survival_preds_bwt <- survival_preds_fit1 %>%
  left_join(dat_use) %>%
  # Group by BWT quartile
  group_by(time, BWT_q) %>%
  # Calculate expected value for S(t) across subjects in each group
  summarise(est=mean(value))

# Plot the observed data (the Kaplan-Meier estimate)
ggsurvplot(survfit(Surv(TTE_SEVERE, AE01) ~ BWT_q, data=dat_use),
           data=dat_use)$plot +
  # Overlay the predictions
  geom_step(data=survival_preds_bwt,
            aes(x=time, y=est), linetype='dashed') +
  facet_wrap(~BWT_q
             )
```


Strata
BWT_q=1
BWT_q=2
BWT_q=3
BWT_q=4
1
2
3
4
Survival probability
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
5
6
Time

Now, let's fit three more models for comparison

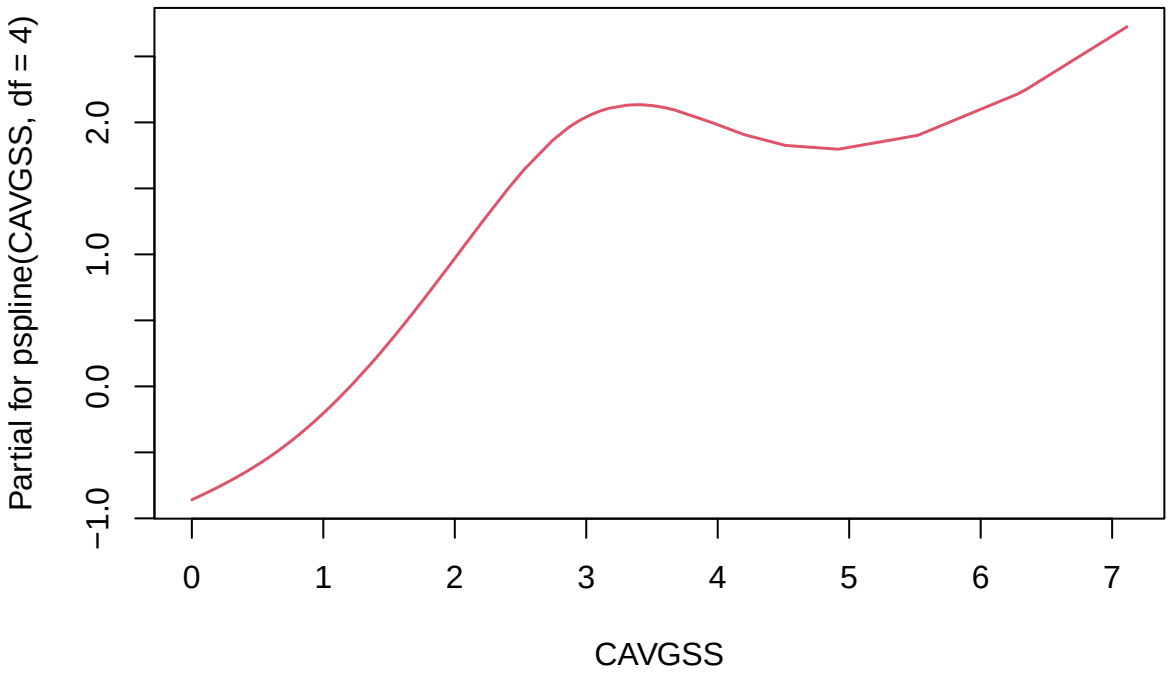
- Surv(TTE_SEVERE,AE01) ~ CAVGSS + PTTYPE
- Surv(TTE_SEVERE, AE01) ~ Quartile
- Surv(TTE_SEVERE,AE01) ~ pspline(CAVGSS, df=4)

The last model will fit a non-parametric (smooth) function of CAVGSS using penalized splines with 4 degrees of freedom.

```
cox_model02 <- update(cox_model01, . ~ . + PTTYPE)
cox_model03 <- coxph(Surv(TTE_SEVERE, AE01) ~ Quartile, data=dat_use)
cox_model04 <- coxph(Surv(TTE_SEVERE, AE01) ~ pspline(CAVGSS, df=4), data=dat_use)
```

To get an idea of how the effects look on the log-hazard scale, we can use the termplot function

```
termplot(cox_model04,terms = 1)
```



We'll compare these models initially using concordance.

```
concordance(cox_model01,cox_model02,cox_model03,cox_model04)
```

```
. Call:
. concordance.coxph(object = cox_model01, cox_model02, cox_model03,
.     cox_model04)
.
. n= 180
.             concordance     se
. cox_model01      0.7729 0.0542
. cox_model02      0.7747 0.0570
```

```
. cox_model03      0.7698 0.0499
. cox_model04      0.7822 0.0555
.
.             concordant discordant tied.x tied.y tied.xy
. cox_model01       2950        846     59      0       0
. cox_model02       2972        854     29      0       0
. cox_model03       2622        542    691      0       0
. cox_model04       2986        810     59      0       0
```

Exercise

1. Based on the c-index, is there one model that is clearly better than the others?

Answer: While the model which uses the spline function of exposure (model 4) has the highest concordence, it is not notably higher than the linear model (model 1).

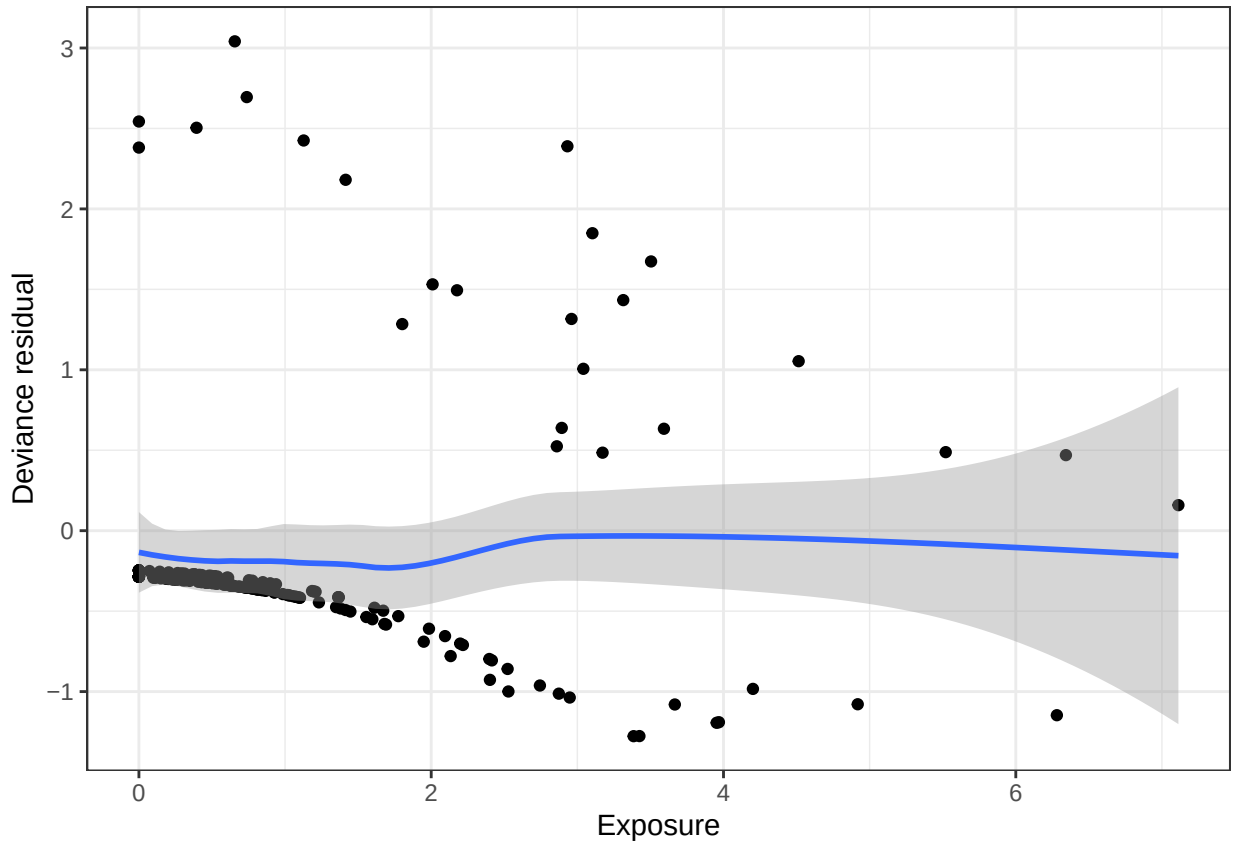
---

2. For comparison,
   a. Plot the deviance residuals from model `cox_mod04` vs CAVGSS. Do they look better than the residuals from model 01?
   b. Make the observed and predicted plot for model `cox_mod04`. How do these compare to `cox_mod01`?

Answer: The residual plot looks much better, as does the comparison of predicted and observed survival functions.

```
dat_use <- dat_use %>%
  mutate(resid04 = residuals(cox_model04, type='deviance'))
```

```
dat_use %>%
  ggplot(aes(x=CAVGSS, y=resid04)) +
  geom_point() +
  geom_smooth() +
  labs(x='Exposure', y='Deviance residual')
```

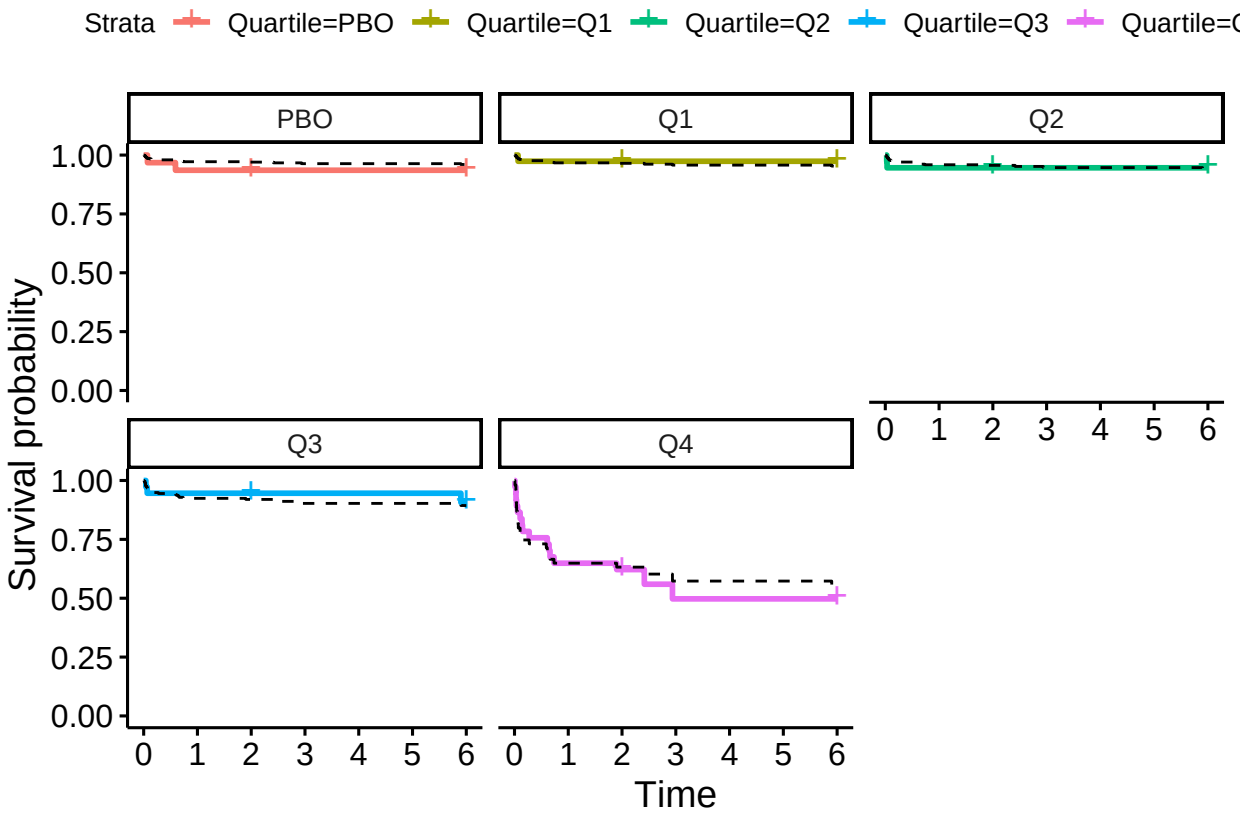



```
preds_mod04 <- survfit(cox_model04, newdata = dat_use) %>% survfit0()

# Make a tall version of the data
survival_preds_fit4 <- preds_mod04$surv %>%
  as.data.frame() %>%
  mutate(time = preds_mod04$time) %>%
  pivot_longer(cols=-time) %>%
  mutate(rowid = as.numeric(name))

# Merge predictions and covariates for plotting
survival_preds_exposure4 <- survival_preds_fit4 %>%
  left_join(dat_use) %>%
  group_by(time, Quartile) %>%
  summarise(est=mean(value))

# Plot the observed data (the Kaplan-Meier estimate)
ggsurvplot(survfit(Surv(TTE_SEVERE, AE01) ~ Quartile, data=dat_use),
           data=dat_use)$plot +
  # Overlay the predictions
  geom_step(data=survival_preds_exposure4,
            aes(x=time, y=est), linetype='dashed') +
  facet_wrap(~Quartile
             )
```



3. Based on all of this information, which of these models would you select going forward? Why?

Answer: Based on this information, some non-linear function of exposure seems to be better than a linear one. If we are going to stick with the Cox model, then I would choose model 4 over model 1. As we move to parametric models, we'll see that we can include parametric non-linear functions of exposure (such as an Emax model).