Lin_ST625_HW9

Frances Lin

3/5/2022

1

Proportional hazard means that the hazard ratio remains constant over time. We can use graphical methods such as Schoenfeld residuals to check this assumption, and we want the residuals to exhibit a random pattern.

2a

Model 1:

$h(t) = h_0(t)exp(\beta_0 + \beta_1 P27 + \beta_2 CYCLINE + \beta_3 NODES + \beta_4 SIZE2 + \beta_5 SIZE3 + \beta_6 Age + \beta_7 Year)$

Model 2:

$h(t) = h_0(t)exp(\beta_0 + \beta_1 P27 + \beta_2 NODES + \beta_3 SIZE2 + \beta_4 SIZE3 + \beta_5 Age + \beta_6 Year + \beta_7 NODES * P27 + \beta_8 NODES * CYCLINE)$

2b

It is estimated that the risk of death is $exp(-0.259) = 0.771823$ times lower for the other group (`NODES` = 1: if cancer has spread to lymph nodes and `CYCLINE` = 1: if protein CYCLINE is abnormal), as compared to the reference group (`NODES` = 0: otherwise and `CYCLINE` = 0: if normal), while keeping all else constant. (i.e. Hazard ratio is higher if cancer has spread to lymph nodes and if protein CYCLINE is abnormal.)

2c

To test $H_0 : \beta_7 = \beta_8 = 0$, we can use either LR (commonly used to compare nested models) or score test. df (degree of freedom) would be $p - k = 9 - 7 = 2$ in both cases since under the H_0, both test statistic $\sim \chi^2(p - k)$.

From **2e**, we know that there is no evidence that the full model does better (LRT: 0.313, p-val = 0.1448). As result, we want to keep the reduced model. This is the same as saying that there is no evidence that the interaction terms are significant.

Individually, we can also use Wald test to test $H_0 : \beta_7 = 0$ and $H_0 : \beta_8 = 0$. This will yield p-val = 0.9595 and p-val = 0.5750, which match the output given.

2d

$\hat{S}(t|Z) =$

```
## [1] 0.58995
```

since $S(t|Z) = S_0(t)^{exp(Z^T\beta)}$, where $S_0(t)$ is the baseline survival for individuals with $Z = 0$. Here, we are given that $\hat{S}_0(t) = 0.93$.

2e

We can use the LR (likelihood ratio) test to compare nested models (reduced vs. full model).

There is no evidence that the full model does better (LRT: 0.313, p-val = 0.1448). As result, we want to keep the reduced model.

```
LR = 2 * (Log-likelihood(full) - (Log-likelihood(reduced))) =
```

```
## [1] 0.313
```

```
df(full) - df(reduced) = 9 - 7 = 2
```

```
p-val = pchisq(LR, 2, lower.tail = FALSE) =
```

```
## [1] 0.1448685
```

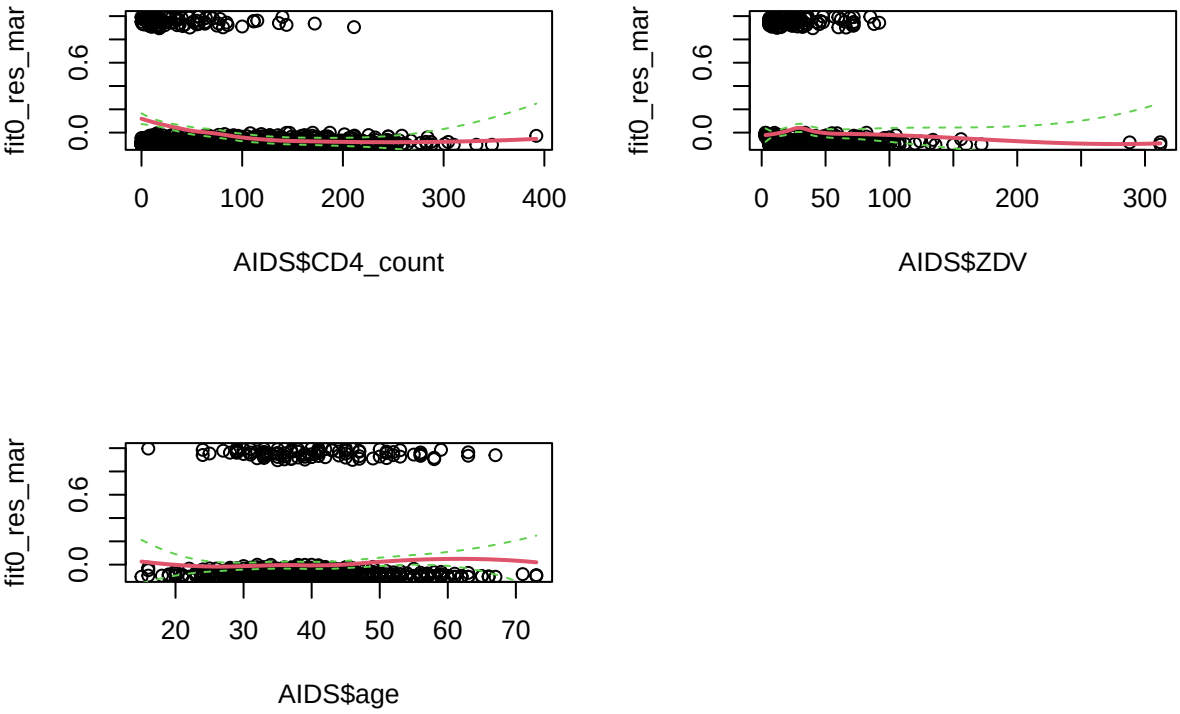
2f

We can use graphical methods such as Schoenfeld residuals to access the proportional hazard assumption for P27.

3a

They look linear to me, but perhaps consider taking quadratic forms of `CD4_count` and `ZDV` and cubic form of `age`.

Note. For full-sized plot, remove `par(mfrow=c(2,2))` and replot.



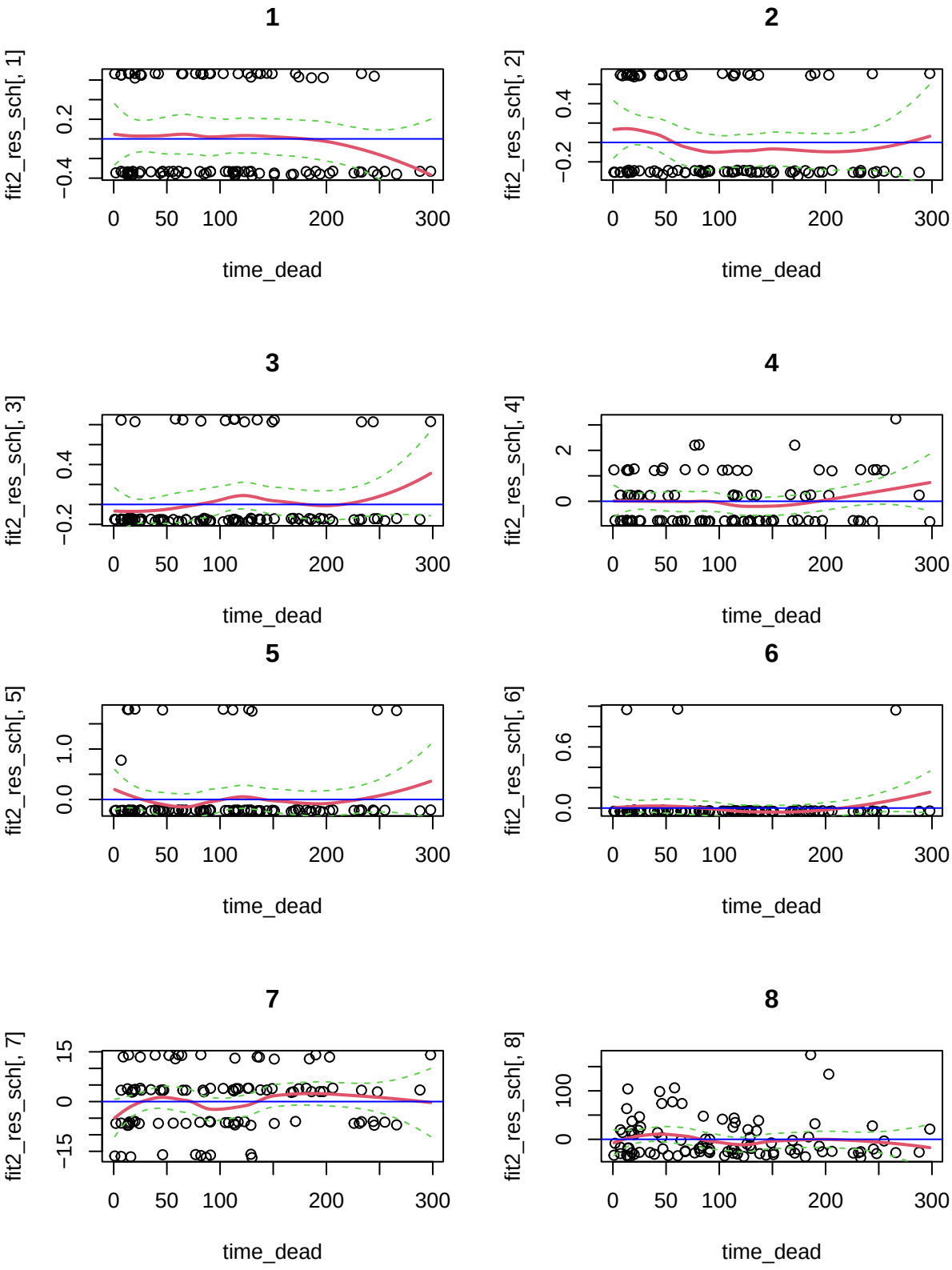
3b

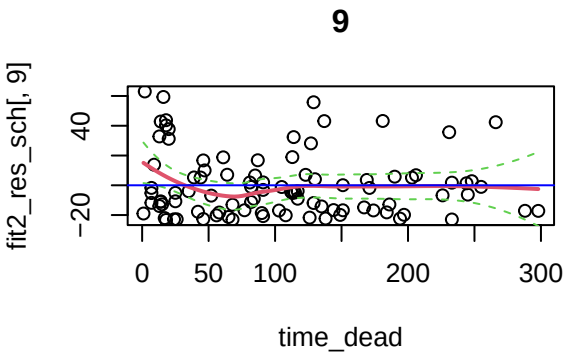
`ZDV2` (p-val = 0.079153) is significant at the $\alpha = 0.1$ level, so we'll keep this term in the Cox model.

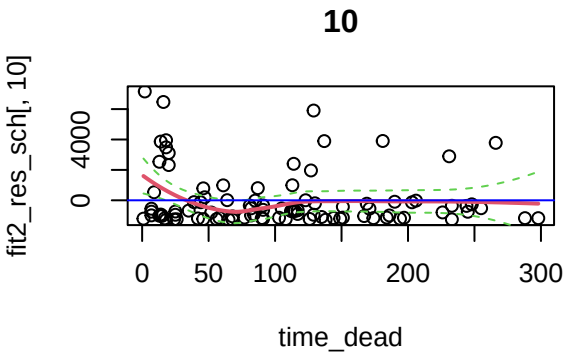
Flat line means that PH assumption is satisfied. Evidence of non-proportionality are shown in all plots except plot 6 (`hamophiliac`), 8 (`CD4_count`) and 11 (`age`).

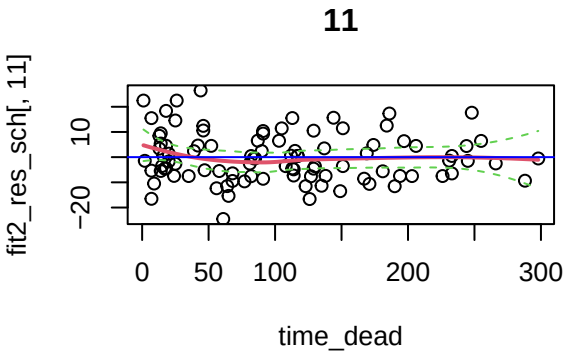
Note.

1. Plot 1-11 are for `trt`, `CD4_stratum`, `sex`, `race`, `IV`, `hamophiliac`, `KPS`, `CD4_count`, `ZDV`, `ZDV2`, `age`, respectively.
2. For full-sized plot, remove `par(mfrow=c(2,2))` and replot.


1
fit2_res_sch[, 1]
−0.4
0.2
0
50
100
200
300
time_dead
2
fit2_res_sch[, 2]
−0.2
0.4
0
50
100
200
300
time_dead
3
fit2_res_sch[, 3]
−0.2
0.4
0
50
100
200
300
time_dead
4
fit2_res_sch[, 4]
0
2
0
50
100
200
300
time_dead
5
fit2_res_sch[, 5]
0.0
1.0
0
50
100
200
300
time_dead
6
fit2_res_sch[, 6]
0.0
0.6
0
50
100
200
300
time_dead
7
fit2_res_sch[, 7]
−15
0
15
0
50
100
200
300
time_dead
8
fit2_res_sch[, 8]
0
100
0
50
100
200
300
time_dead







3c

There is no evidence that the full model does better (LRT: 9.0352, p-val = 0.4712). As result, we want to keep the reduced model that contains the significant covariates (`factor(trt)`, `factor(KPS)`, `CD4_count`, and `age`).

```
LR = 2 * (fit_full$loglik[2] - fit_red$loglik[2]) =
```

```
## [1] 9.035223
```

```
p-val = pchisq(LR, 10, lower.tail = TRUE) =
```

```
## [1] 0.5287641
```

Note. The full model contains all the variables with 17 parameters. The reduced model contains `factor(trt)`, `factor(KPS)`, `CD4_count`, and `age` with 7 parameters.

The results are confirmed using the package `lmtest`:

```
## Likelihood ratio test
##
## Model 1: Surv(time, censor) ~ factor(trt) + factor(KPS) + CD4_count +
##     age
## Model 2: Surv(time, censor) ~ factor(trt) + factor(CD4_stratum) + factor(sex) +
##     factor(race) + factor(IV) + factor(hamophiliac) + factor(KPS) +
##     CD4_count + ZDV + age
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -608.74
## 2  16 -604.23 10 9.0352     0.5288
```

3d

Results of the reduced model:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
factor(trt)1	-0.6575	0.5181	0.2155	-3.051	0.002279
factor(KPS)80	-0.4042	0.6675	0.3654	-1.106	0.2686
factor(KPS)90	-1.061	0.346	0.3626	-2.928	0.003416
factor(KPS)100	-1.499	0.2233	0.4067	-3.687	0.0002273
CD4_count	-0.01451	0.9856	0.002518	-5.762	8.318e-09
age	0.0218	1.022	0.01137	1.917	0.05529

All except `factor(KPS)80` and `age` have significant effects on survival at the $\alpha = 0.05$ level. (There is weak evidence of `age` on survival.)

It is estimated that the risk of death is $exp(-0.6575) = 0.5181$ times lower for the other group (IDV: `trt` $= 1$), as compared to the reference group (control: `trt` $= 0$).

It is estimated that the risk of death is $exp(-1.061) = 0.346$ times lower for the other group (minor symptoms: `KPS` $= 90$), as compared to the reference group (active work impossible: `KPS` $= 70$).

It is estimated that the risk of death is $exp(-1.499) = 0.2233$ times lower for the other group (no evidence of disease: `KPS` $= 100$), as compared to the reference group (active work impossible: `KPS` $= 70$).

It is estimated that the risk of death decreases by $(1 - exp(-0.01451)) * 100 = 1.4405\%$ with 1-unit increase in the CD4 count in cells.

Appreciate your grading for all these assignments!