|  |
| --- |
| **PM592: Regression Analysis for Data Science**  **Exam 2 – Fall 2023** |

**Instructions**

* Answer questions directly on the exam sheet and show all work.
* You may use your class notes, R software, and a calculator.
* You may **not** consult with any resources that are not a part of this class, including obtaining outside help through websites or talking to others about this exam.
* You may not discuss this exam with classmates until after the final due date.
* Unless otherwise stated, use α = .05 when testing statistical hypotheses.
* You have 180 minutes to submit the exam after accessing it. Plan ahead as the submission process may take longer than expected. If you encounter difficulties uploading the exam, e-mail a copy to [tpickeri@usc.edu](mailto:tpickeri@usc.edu).
* **If you submit the exam late, you will be penalized 4 points for each minute (or fraction thereof) past the due time.**

**Statement of Academic Integrity**

For this exam, I affirm the following:

* This exam reflects only my own work. I did not receive assistance from any other individual, nor did I provide assistance to any other student taking this exam.
* While I may use my own notes, I did not refer to any online source during the exam.
* I understand that acts of academic dishonesty may be penalized in accordance with Section 13 of the University of Southern California Community Standards, including possible “F” in the course, notation on transcript, and/or dismissal from academic programs (<https://sjacs.usc.edu/students/academic-integrity/>).

I affirm by typing my name below.

Flemming Wu\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_12/4/2023\_\_\_\_\_\_\_\_

Name Date

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **A** |  |  |  |  | [25 points] |  |

Gupta et al. (2020) examined the role of delays in treatment for traumatic brain injury on subsequent mortality. Specifically, they looked at the “third delay” – the delay between arriving at the hospital and actually receiving treatment. They studied 6,278 individuals that were seen across four different large trauma centers in India from 2013-2015. “Delay” was classified as minimal (<10 minutes), moderate (10-60 minutes), or extended (>60 minutes). Mortality was defined as whether the patient died within 24 hours from being admitted to the hospital.

Y = 24-hour Mortality (1=Dead, 0=Alive).

They fit the following model:

With corresponding parameter estimates:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| -1.92 | 0.43 | 1.21 | 1.34 |

|  |  |
| --- | --- |
|  | A1. In notation, what is the null and alternative hypothesis that would test if duration of delay is associated with 24-hour mortality? |

|  |  |
| --- | --- |
|  | A2. According to this model, what is the odds ratio of mortality for a patient who was delayed 30 minutes vs. 5 minutes, adjusting for ~~age~~ sex? |

The equation of the best fit line is given by:

The odds of mortality for a patient who was delayed 30 minutes is 3.35 times the odds of mortality for a patient who was delayed 5 minutes, adjusting for sex.

|  |  |
| --- | --- |
|  | A3. According to this model, what is the odds ratio of mortality for a ~~22-year-old~~ patient with delay of 65 minutes vs. a patient of the same ~~age~~ sex with a delay of 8 minutes? |

The odds of mortality for a patient who was delayed 65 minutes is 3.82 times the odds of mortality for a patient who was delayed 8 minutes, adjusting for sex.

|  |  |
| --- | --- |
|  | A4. According to this model, what is the odds ratio of morality for a female patient compared to a male patient, adjusting for delay? |

The odds of mortality for a female patient is 1.54 times that of a male patient, adjusting for delay.

|  |  |
| --- | --- |
|  | A5. Can the intercept term be meaningfully interpreted? If yes, provide that interpretation. If no, state why not. |

In order for the intercept term to be meaningfully interpreted, it must make sense for all coefficients in the equation to have a value for zero. While this makes sense for the duration variable, where if and were both zero it would represent someone who had a delay of less than 10 minutes, it would not make sense to assign a value of zero for the value of . The gender can only either be one or two, representing male or female, so the intercept term cannot be meaningfully interpreted.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **B** |  |  |  |  | [25 points] |  |

Brouwer et al. (2019) studied the relationship between receiving the HPV vaccination and variables related to sexual behavior in college-aged men and women. They asked the participants about whether they had received the HPV vaccine and retrospectively ascertained three behavioral outcomes. Each of these three distinct outcomes was analyzed using the following methods:

* Occurrence of sexual debut (whether or not the participant had their first sexual experience) – logistic regression
* Number of sexual partners – Poisson regression
* Age at sexual debut – Cox proportional hazards regression

The following table presents results from their paper. Note that their use of “incidence ratio” could alternately have been written as a “rate ratio.”

A black and white document with numbers and text

Description automatically generated

|  |  |
| --- | --- |
|  | B1. For the Poisson regression model, the authors state the outcome is “number of sexual partners (per year)”. If they used number of sexual partners since sexual debut as the count outcome, what must the offset term have been? |

If the number of sexual partners since sexual debut was the count outcome for the Poisson regression, the offset term must have been “years sexually active” to model the outcome as the rate: “number of sexual partners (per year)”.

|  |  |
| --- | --- |
|  | B2. In the “number of sexual partners (per year)” model, can the intercept term be meaningfully interpreted? If yes, provide that interpretation. If no, state why not. |

Yes, the intercept term can be interpreted in the Poisson regression model. The intercept term would represent the expected number of sexual partners per year for someone who is 18 years old, male, not White, not a current or binge drinker, not an ever user of marijuana, and not vaccinated for HPV. The expected mean number of sexual partners per year for someone who meets the above criteria is 1.47.

|  |  |
| --- | --- |
|  | B3. Explain why there is no intercept term for the “age at sexual debut” model. |

The “age at sexual debut” model is a Cox proportional hazards model, which models the effect of covariates on some baseline hazard function. The baseline hazard function is the function that models the baseline probability of hazard at a given time without the effect of covariates and thereby already represents the function when all covariates are equal to zero. Therefore, it would not make sense to include an intercept term, which would be interpreted as the hazard for when all covariates equal zero, since this would be included in the baseline hazard function.

|  |  |
| --- | --- |
|  | B4. Briefly explain how being vaccinated for HPV affects each of the three outcomes, with an indication of whether each effect is statistically significant. |

According to the logistic regression model, being vaccinated for HPV is associated with 0.80 times the odds of a person already having their sexual debut compared to someone who is not vaccinated for HPV, adjusting for all other covariates, but this association is not statistically significant . According to the Poisson regression model, being vaccinated for HPV is associated with an expected number of sexual partners per year that is 0.81 the expected number of partners for someone who isn’t vaccinated for HPV, adjusting for all other covariates, and this association is not statistically significant . According to the Cox proportional hazards model, being vaccinated for HPV is associated with 1.27 times the hazard of sexual debut compared to not being vaccinated for HPV, adjusting for all other covariates, and this association is not statistically significant .

|  |  |
| --- | --- |
|  | B5. The authors did not indicate that they checked the fit of the Poisson regression model. If there was a violation of the dispersion assumption of Poisson regression, are the p-values for this model too low or too high? Why? |

If there was a violation of the dispersion assumption of Poisson regression, the p-values for the Poisson model would be too low. The Poisson regression model assumes that the mean is equal to the variance for all X values. If the true variances were higher (meaning overdispersion), then the standard error calculation would be artificially low, i.e. the confidence intervals of the coefficients would be narrower than they should be, leading to artificially smaller p-values of coefficient estimates.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **C** |  |  |  |  | [20 points] |  |

Dr. Hellman was examining periprosthetic joint infection in his patients. The current laboratory standard indicates joint infection when white blood cell (WBC) count is >3000. Blood samples were taken and analyzed both at Keck and through an external lab. Dr. Heckmann was interested in whether the two labs (“Keck” and “External”) concurred in their assessment of WBC count.

For the purposes of this analysis, the Keck measure was treated as the “predicted” and the External measure was treated as the “reference.”

> Conf(

+ dat15192\_test$Keck,

+ dat15192\_test$External

+ )

Confusion Matrix and Statistics

Reference

Prediction 1 0

1 60 5

0 0 128

Total n : 193

Accuracy : 0.9741

95% CI : (0.9408, 0.9889)

No Information Rate : 0.6891

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.9409

Mcnemar's Test P-Value : 0.0736

Sensitivity : 1.0000

Specificity : 0.9624

Pos Pred Value : 0.9231

Neg Pred Value : 1.0000

Prevalence : 0.3109

Detection Rate : 0.3368

Detection Prevalence : 0.3109

Balanced Accuracy : 0.9812

F-val Accuracy : 0.9600

Matthews Cor.-Coef : 0.9425

'Positive' Class : 1

|  |  |
| --- | --- |
|  | C1. What percent of the time did Keck and the External lab agree on their assessment of joint infection (WBC > 3000)? |

Keck and the External lab agreed on their assessment 97.41% of the time. The statistic that represents their overall agreement is the accuracy.

|  |  |
| --- | --- |
|  | C2. When Keck predicted infection, what percent of the time did the External lab agree? |

When Keck predicted infection (65 times), the External lab agreed on 60 of the predictions. The percentage of agreement is . This statistic is given by the positive predictive value.

|  |  |
| --- | --- |
|  | C3. When Keck predicted that there was NOT an infection, what percent of the time did the External lab agree? |

When Keck predicted no infection (128 times), the External lab agreed 100% of the time. The statistic that represents this is the negative predictive value.

|  |  |
| --- | --- |
|  | C4. Which method (Keck or External) was more likely to diagnose a patient with infection? |

Keck was more likely to diagnose a patient with infection, as it classified some patients without infection as having infection. On the other hand, the External lab did not misclassify any patients without infection as having infection.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **D** |  |  |  |  | [30 points] |  |

Dr. Frank Rivolous wanted to test the efficacy of a new supplement: the Homeopathic-Allopathic Prebiotic-Probiotic Yielding (HAPPY) vitamin. It was believed that use of this vitamin could prevent COVID and flu symptoms. He recruited 157 individuals and recorded whether they did not currently take HAPPY vitamins, took the low dose formula, or took the high dose formula. Participants were followed to determine number of days that elapsed until participant was diagnosed with either COVID or flu. Censored participants were those lost to follow up.

Dr. Rivolous realized he didn’t actually know what he was doing with the analysis and decided to seek professional help. He’s asking you to interpret this output to form a cohesive report on what was performed. The main research question is whether use/dosage of the HAPPY vitamin was related to hazard of contracting COVID or flu. Because condition was not randomized, there was concern that participant age might affect the results.

Based only on the output in the appendix, write brief report detailing the methods, results, and conclusions from the available analyses. Your report must be in paragraph format (i.e., no bullet points). Any text that appears after 350 words will be deleted and not graded.

You should comment on:

* The type of analysis performed
* The steps involved in building and selecting the best model
* Which final model(s) you chose to address the research question
* An interpretation of the parameters of interest in final model, including relevant coefficients and p-values
* Information about how well the final model fits (if provided)
* Any missing or next steps that may be appropriate

Please state your word count here: \_\_\_\_\_\_348\_\_\_\_\_\_\_

|  |  |
| --- | --- |
|  | *The following Cox proportional hazards model addresses whether use/dose of HAPPY vitamin was related to hazard of getting COVID or flu (status), and whether age affects this relationship. First, the analyst examined counts and distributions of all the variables of interest. Then, Kaplan-Meier curves were fitted for dose and age quartile. The Log-Rank test shows that status did not differ for different doses of HAPPY vitamin (p=0.81), but did differ for age quartile (p=0.001). Next, the functional form for age was assessed using fractional polynomials and the Martingale residuals. The fractional polynomials approach suggested a cubic term transformation for age offered the best fit, and the Martingale residuals plot seems to confirm this, showing the straightest line for the cubic transformation of age. Next, two Cox proportional hazard models were fit, first an unadjusted model just using dose, and then one adjusting for age cubed. Since the addition of age appears to change the coefficients for dose by a considerable amount, age cubed was included in the final model. For the final model, the assumption of proportional hazards was checked by the Schoenfeld test, which revealed that the assumptions are met for both covariates. According to the final model, a low dose of vitamin is associated with 0.89 times the hazard of getting COVID or flu compared to control (not taking the vitamin), and this association is not significant (p=0.64). A high dose of vitamin is associated with 1.1 times the hazard of getting COVID or flu compared to not taking the vitamin, and this association is also not significant (p=0.64). Finally, the model fit was assessed by examining the Cox-Snell and deviance residuals, and Dbetas. The Cox-Snell residuals appear to follow a 45 degree line, supporting the earlier findings that the proportional hazards assumption is met without stratification. However, the deviance residuals and Dbetas reveal that there are many individual points that were not fit well by the model. This is likely due to the fact that the associations found in the final model were not significant and indicates no relationship between HAPPY vitamin usage and hazard.* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Appendix** |  |  |  |  |  |  |

**time: time from study enrollment to flu/Covid diagnosis (in days)**

**age: age of participant (in years)**

**dose: dosage of HAPPY vitamin given (or control group)**

**status: 1=diagnosed with flu/Covid; 0=censored**

**> dat14675 %>% count(status)**

status n

1 0 55

2 1 102

**> dat14675 %>% count(dose)**

dose n

1 Control 53

2 Low Dose 53

3 High Dose 51

**> dat14675 %>% skimr::skim(age)**

── Variable type: numeric ───────────────────────────────────────────

skim\_variable n\_missing complete\_rate mean sd p0 p25 p50 p75 p100 hist

1 age 0 1 41.7 10.7 12 36 44 49 64 ▁▂▅▇▂

**> dat14675 %>% group\_by(dose) %>% summarise(mean\_diagnosed = mean(status))**

# A tibble: 3 × 2

dose mean\_diagnosed

<fct> <dbl>

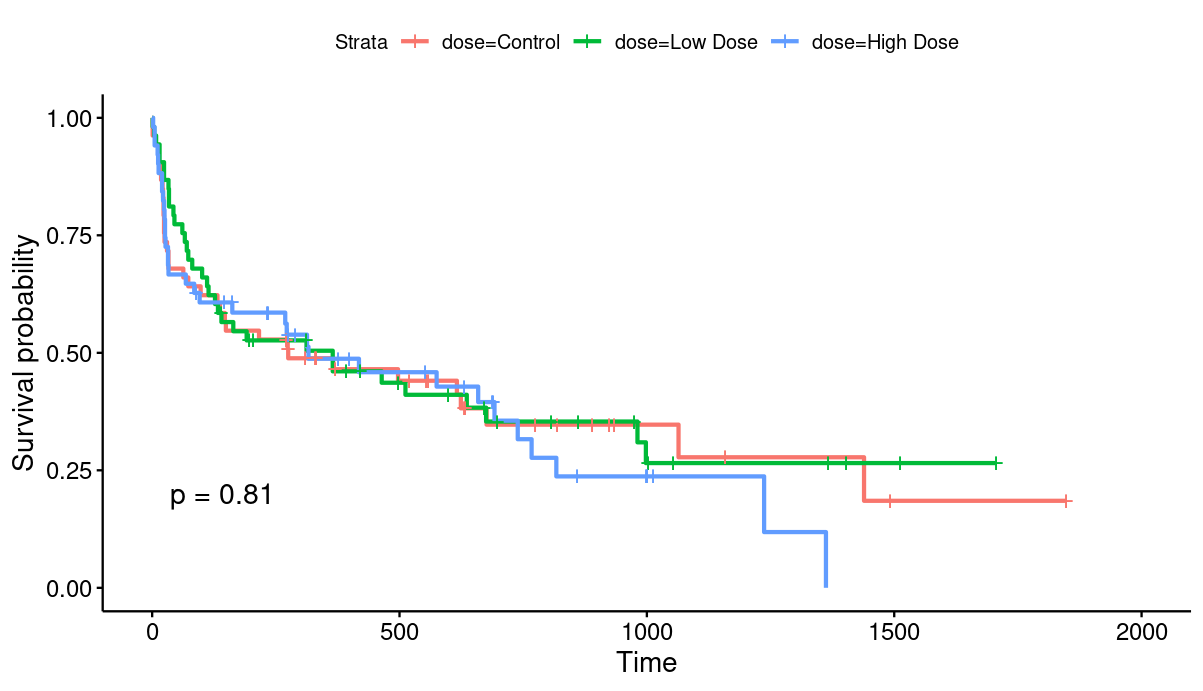
1 Control 0.642

2 Low Dose 0.642

3 High Dose 0.667

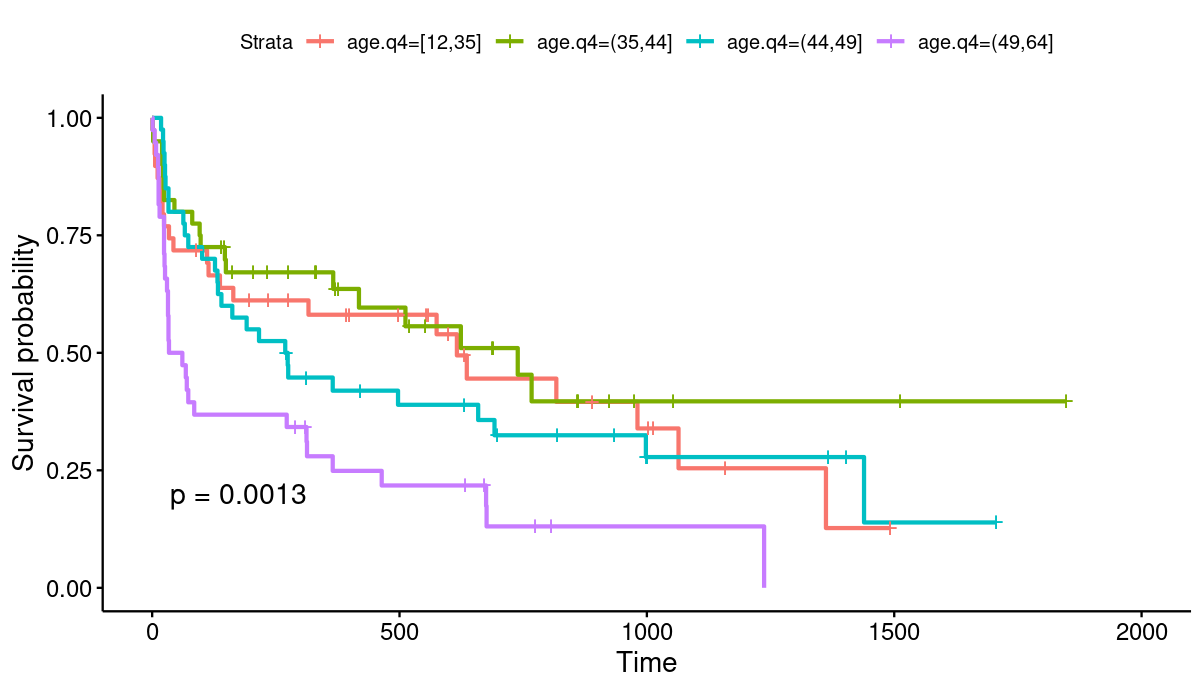
**> surv\_dose.km <- survfit(surv\_object ~ dose, data=dat14675)**

**> ggsurvplot(surv\_dose.km, data=dat14675, pval=T)**

****

**> surv\_age.km <- survfit(surv\_object ~ age.q4, data=dat14675)**

**> ggsurvplot(surv\_age.km, data=dat14675, pval=T)**

****

**> mfp(Surv(time, status) ~ fp(age) + dose, family=cox, data=dat14675)**

Call:

mfp(formula = Surv(time, status) ~ fp(age) + dose, data = dat14675,

family = cox)

Deviance table:

Resid. Dev

Null model 902.3751

Linear model 894.0579

Final model 888.8685

Fractional polynomials:

df.initial select alpha df.final power1 power2

age 4 1 0.05 2 3 .

doseLow Dose 1 1 0.05 1 1 .

doseHigh Dose 1 1 0.05 1 1 .

Transformations of covariates:

formula

age I((age/100)^3)

dose dose

coef exp(coef) se(coef) z p

age.1 7.851e-06 1.0000 2.152e-06 3.6482 0.000264

doseLow Dose.1 -1.117e-01 0.8943 2.432e-01 -0.4595 0.646000

doseHigh Dose.1 1.157e-01 1.1227 2.438e-01 0.4747 0.635000

Likelihood ratio test=13.51 on 3 df, p=0.00366 n= 157

Deviance Residuals:

Min 1Q Median 3Q Max

-4.765 -2.346 -1.381 1.433 26.675

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.25555 1.47410 2.887 0.004312 \*\*

eras -2.65809 0.72620 -3.660 0.000321 \*\*\*

white.fNon-White -0.98422 0.79621 -1.236 0.217837

age 0.02096 0.02159 0.971 0.332818

eras.fERAS:white.fNon-White 1.83254 1.13999 1.608 0.109498

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 16.17932)

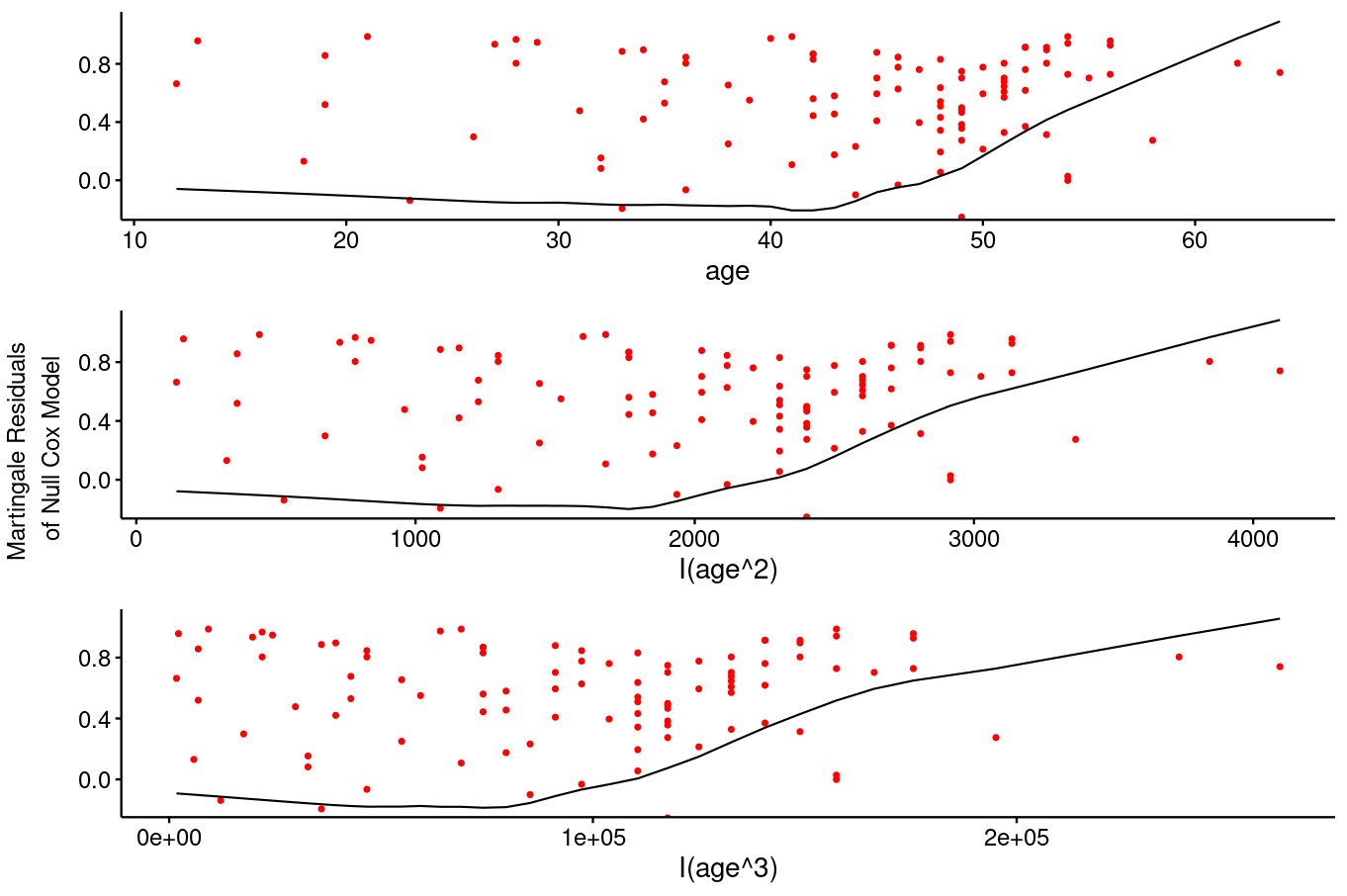
Null deviance: 3530.7 on 207 degrees of freedom

Residual deviance: 3284.4 on 203 degrees of freedom

AIC: 1176.2

Number of Fisher Scoring iterations: 2

**> ggcoxfunctional(surv\_object ~ age + I(age^2) + I(age^3), data = dat14675)**

****

**> model1 <- coxph(surv\_object ~ dose, data=dat14675)**

**> summary(model1)**

Call:

coxph(formula = surv\_object ~ dose, data = dat14675)

n= 157, number of events= 102

coef exp(coef) se(coef) z Pr(>|z|)

doseLow Dose -0.06805 0.93421 0.24287 -0.280 0.779

doseHigh Dose 0.08910 1.09319 0.24356 0.366 0.714

exp(coef) exp(-coef) lower .95 upper .95

doseLow Dose 0.9342 1.0704 0.5804 1.504

doseHigh Dose 1.0932 0.9148 0.6782 1.762

Concordance= 0.513 (se = 0.029 )

Likelihood ratio test= 0.42 on 2 df, p=0.8

Wald test = 0.42 on 2 df, p=0.8

Score (logrank) test = 0.42 on 2 df, p=0.8

**> model2 <- coxph(surv\_object ~ I(age^3) + dose, data=dat14675)**

**> summary(model2)**

Call:

coxph(formula = surv\_object ~ I(age^3) + dose, data = dat14675)

n= 157, number of events= 102

coef exp(coef) se(coef) z Pr(>|z|)

I(age^3) 7.836e-06 1.000e+00 2.154e-06 3.637 0.000275 \*\*\*

doseLow Dose -1.121e-01 8.939e-01 2.431e-01 -0.461 0.644624

doseHigh Dose 1.124e-01 1.119e+00 2.438e-01 0.461 0.644856

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

exp(coef) exp(-coef) lower .95 upper .95

I(age^3) 1.0000 1.0000 1.0000 1.000

doseLow Dose 0.8939 1.1187 0.5550 1.440

doseHigh Dose 1.1189 0.8937 0.6939 1.804

Concordance= 0.596 (se = 0.034 )

Likelihood ratio test= 13.42 on 3 df, p=0.004

Wald test = 13.6 on 3 df, p=0.004

Score (logrank) test = 13.61 on 3 df, p=0.003

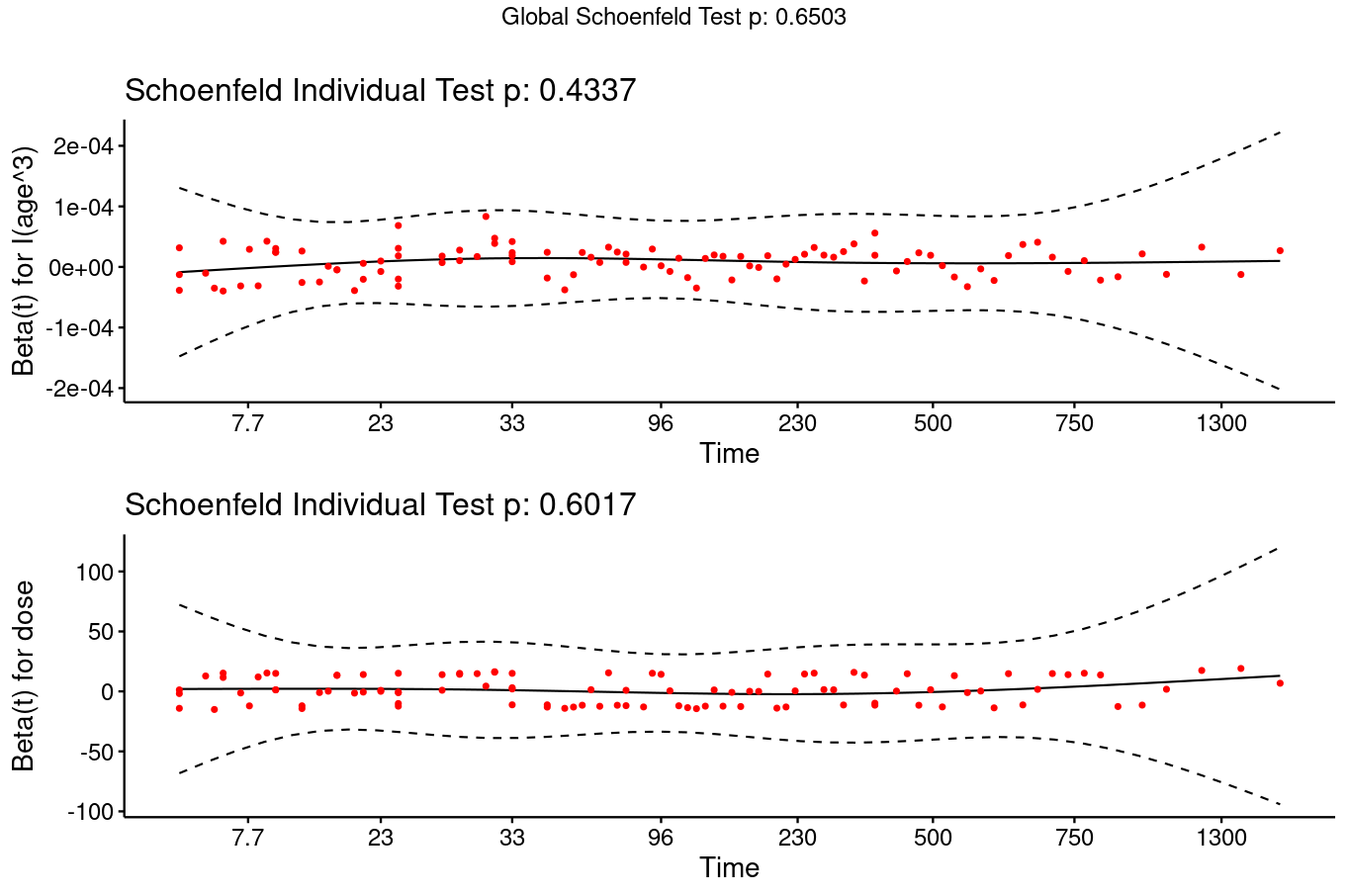
**> cox.zph(model2)**

chisq df p

I(age^3) 0.613 1 0.43

dose 1.016 2 0.60

GLOBAL 1.640 3 0.65

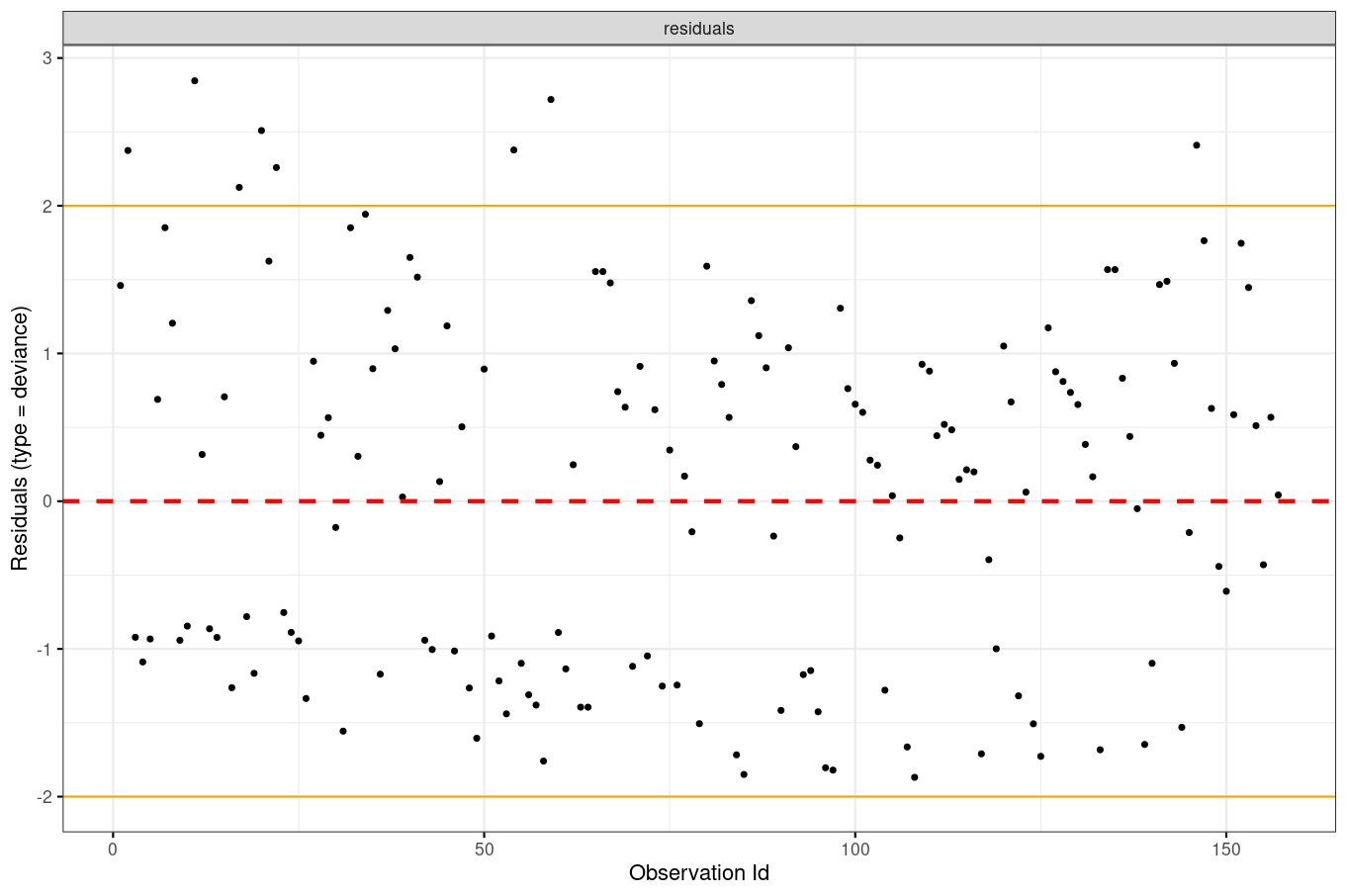


**> ggcoxdiagnostics(model2, type = "deviance", sline = F,**

**+ ox.scale = "observation.id") +**

**+ geom\_hline(yintercept = 2, color = "orange") +**

**+ geom\_hline(yintercept = -2, color = "orange")**



**> ggcoxdiagnostics(model1, type = "dfbeta", sline = F) +**

**+ geom\_hline(data = bind\_rows(**

**+ tibble(val = abs(coefficients(model1)\*.1),**

**+ covariate = names(coefficients(model1))),**

**+ tibble(val = -abs(coefficients(model1)\*.1),**

**+ covariate = names(coefficients(model1)))),**

**+ aes(yintercept = val), color = "orange"**

**A screenshot of a graph

Description automatically generated**

**> coxph(**

**+ Surv(dat14675$status - residuals(model2, type = "martingale"), status) ~ 1,**

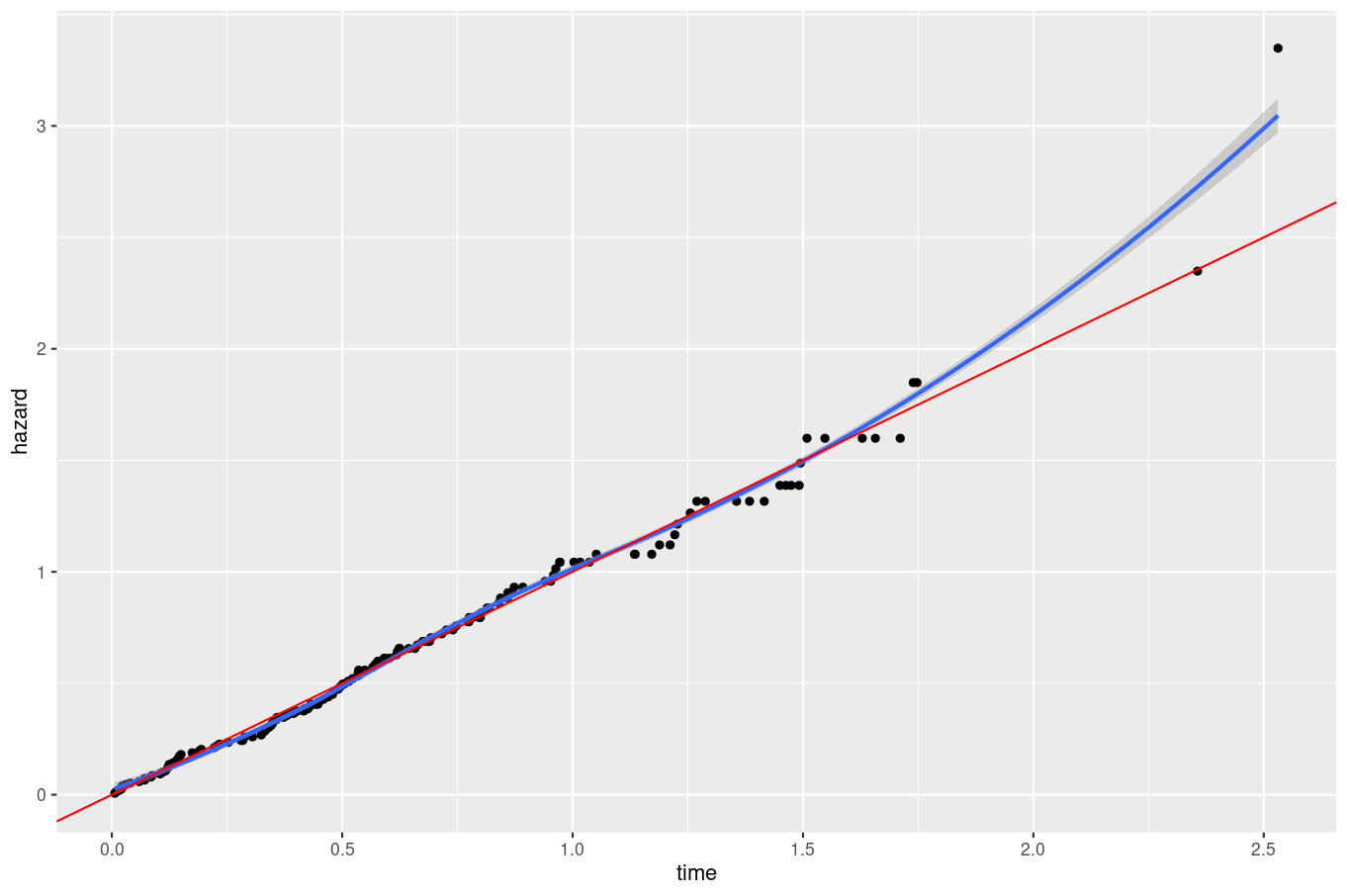
**+ data = dat14675) %>%**

**+ basehaz() %>%**

**+ ggplot(aes(x = time, y = hazard)) +**

**+ geom\_point() +**

**+ geom\_smooth() +**

****