Assignment 2

Michal Malyska

10/02/2020

```
df1 <- read_rds(path = paste0(here(), "/data/ON_mortality.RDS")) %>%
    mutate(age = as.numeric(if_else(age == "110+", "110", age)))
```

Question 1

$$\lambda(t) = \alpha e^{\beta t}$$

a

$$S(t) = \exp(-\frac{\alpha}{\beta}(e^{\beta t} - 1))$$

Showing $S(t)$ (kind of reverse showing since I show that this $S(t)$ implies our hazard function but still valid.)

$$\begin{aligned}
\lambda(t) &= -\frac{d}{dt} log(S(t)) \\
&= -\frac{d}{dt}\left(-\frac{\alpha}{\beta}(e^{\beta t} - 1)\right) \\
&= \frac{d}{dt}\left(\frac{\alpha}{\beta}(e^{\beta t} - 1)\right) \\
&= \frac{d}{dt}\left(\frac{\alpha}{\beta}e^{\beta t}\right) \\
&= \alpha e^{\beta t}
\end{aligned}$$

$$f(t) = \alpha \exp(\beta t - \frac{\alpha}{\beta}(e^{\beta t} - 1))$$

Showing $f(t)$

$$\lambda(t) = \frac{f(t)}{S(t)} f(t) = \lambda(t) S(t)$$

$$\begin{aligned}
f(t) &= \lambda(t) S(t) \\
&= \alpha e^{\beta t} * \exp(-\frac{\alpha}{\beta}(e^{\beta t} - 1)) \\
&= \alpha \exp(\beta t - \frac{\alpha}{\beta}(e^{\beta t} - 1))
\end{aligned}$$

b

Modal Time of Death (mode of $f(t)$)

$$\frac{d}{dt}f(t) = f(t) * (\beta - \alpha e^{\beta t}) = 0$$
$$\implies (\beta - \alpha e^{\beta t}) = 0 \text{ or } f(t) = 0$$

so the mode is at:

$$t = \frac{log(\frac{\beta}{\alpha})}{\beta}$$

as long as $\alpha < \beta$

otherwise the function is decreasing so:

$$t = 0$$

c

$ h(x) = aeˆ{bx} = eˆ{log(a) + bx}$ So: $log(h(x)) = log(a) + bx$

```
df_c <- df1 %>%
    filter(between(age, 40, 100)) %>%
    mutate(loghx = log(hx))

df_1961 <- df_c %>% filter(year == 1961)
df_2011 <- df_c %>% filter(year == 2011)

lm1961 <- lm(loghx ~ age, data = df_1961)
lm2011 <- lm(loghx ~ age, data = df_2011)

coef1961 <- coef(lm1961)
coef2011 <- coef(lm2011)

a1961 = unname(exp(coef1961[1]))
a2011 = unname(exp(coef2011[1]))

b1961 = unname(coef1961[2])
b2011 = unname(coef2011[2])
```

The values for 1961 are: alpha of 7.1168697×10^{-5} and beta of 0.0892529 compared to the values for 2011 : alpha of 1.4755769×10^{-5} and beta of 0.1006012

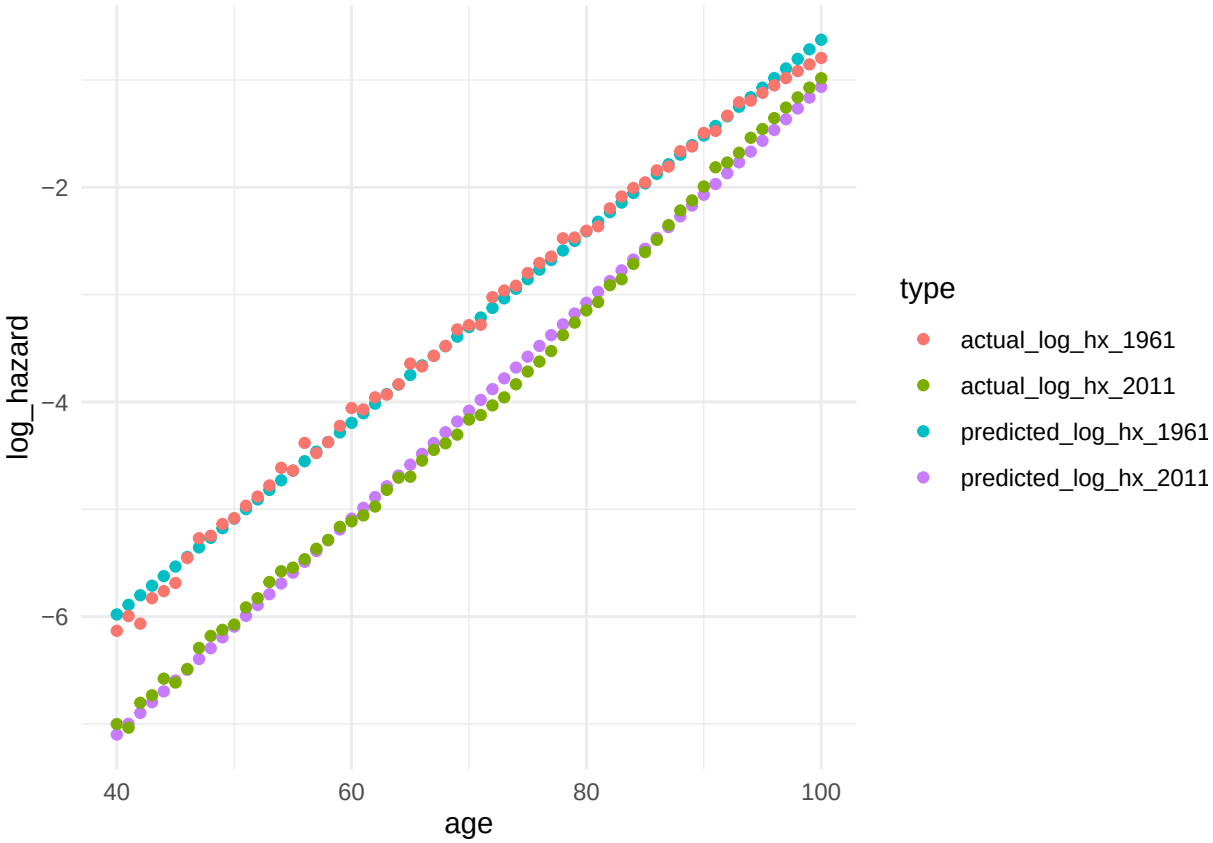
The meaning of alpha is the starting level of mortality (much higher for 1961) and beta gives the increase in mortality over time which surprisingly is higher for 2011. Perhaps lower infant mortality screws with us a tiny bit and makes it seem like people die faster with age in 2011 than they did in 1961 just because so many of them already died before 40 where we start.

d

```
preds_d <- tibble(age = seq(from = 40, to = 100, by = 1))
preds_d$predicted_log_hx_1961 <- predict(lm1961, newdata = preds_d)
preds_d$predicted_log_hx_2011 <- predict(lm2011, newdata = preds_d)
preds_d$actual_log_hx_1961 <- df_1961$loghx
```

```
preds_d$actual_log_hx_2011 <- df_2011$loghx

preds_d_long <- preds_d %>% pivot_longer(cols = c("predicted_log_hx_1961", "predicted_log_hx_2011", "ac
```

```
preds_d_long %>%
    ggplot(aes(x = age, y = log_hazard, color = type)) +
    geom_point() +
    theme_minimal()
```

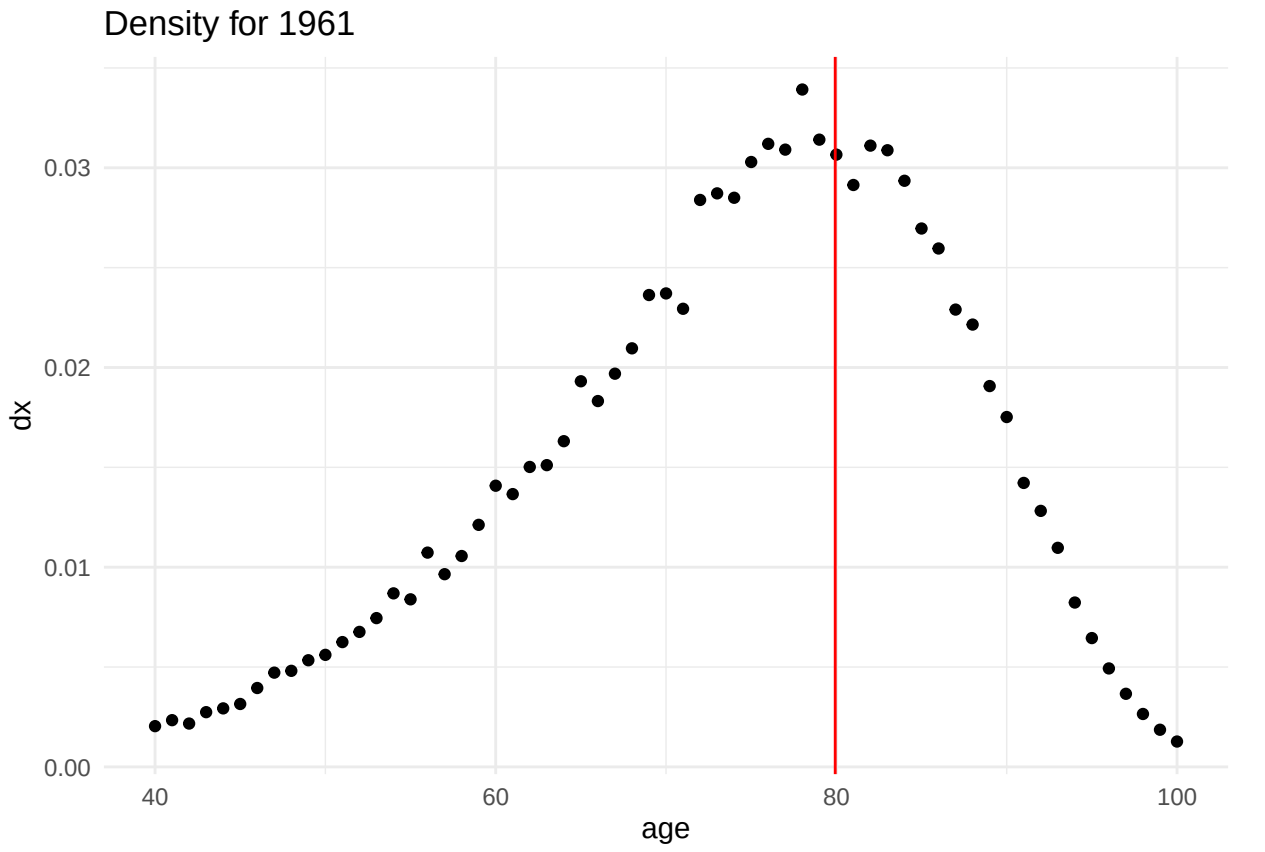


They both seem to fit surprisingly well. There are some minor patterns in the predicted vs actual for 2011 between the ages of 70 and 80 where the actual log hazard seems to be lower, and later on when actual log hazard seems to be higher for those 90+. For 1961 model it seems to be the opposite for the super old - we overestimate the log hazard for those pushing 100. Overall I would say the assumption is quite reasonable.

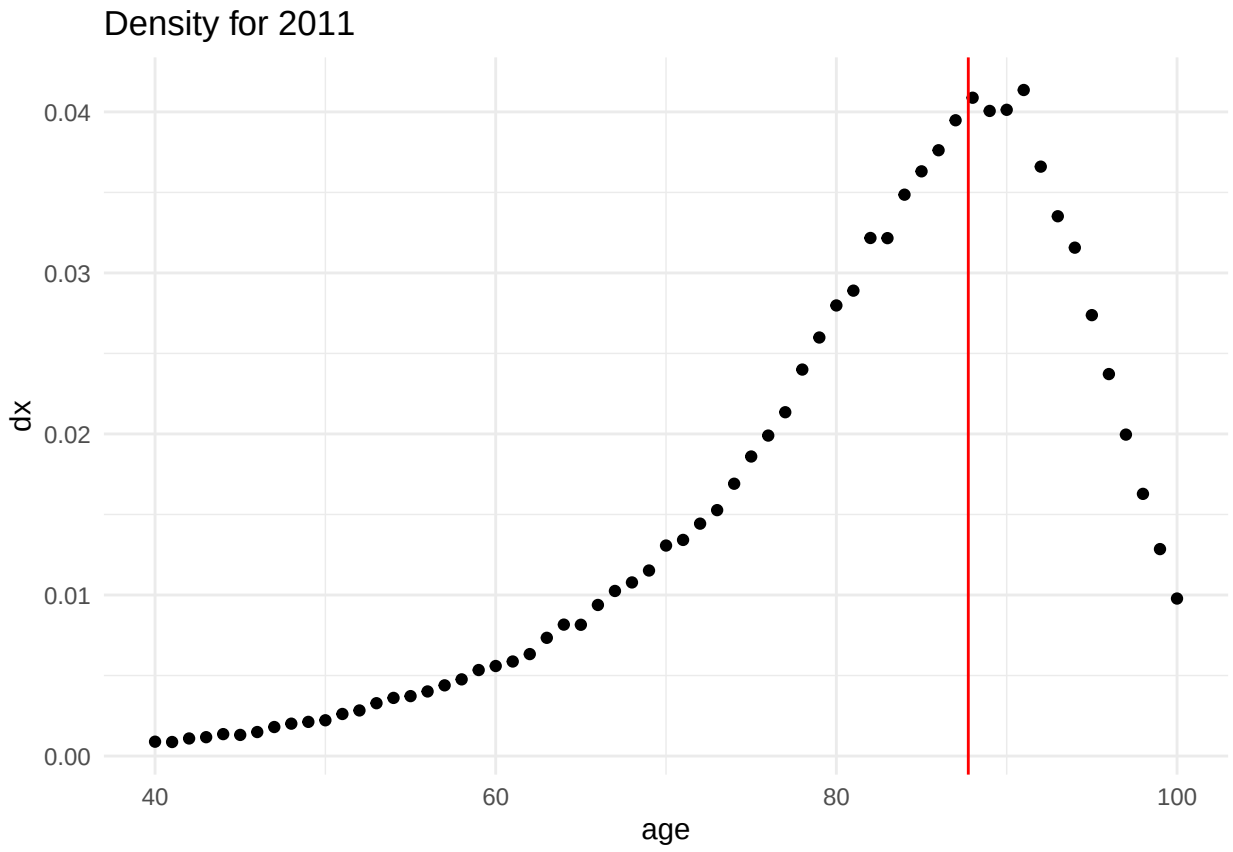
e

```
mode1961 <- log(b1961 / a1961) / b1961
mode2011 <- log(b2011 / a2011) / b2011

df_e <- df_c %>% filter(year %in% c(2011, 1961))

df_e %>% filter(year == 1961) %>%
    ggplot(aes(x = age, y = dx)) +
    geom_point() +
    geom_vline(xintercept = mode1961, color = "red") +
    theme_minimal() +
```

```
    labs(title = "Density for 1961")
```




```
df_e %>% filter(year == 2011) %>%
    ggplot(aes(x = age, y = dx)) +
    geom_point() +
    geom_vline(xintercept = mode2011, color = "red") +
    theme_minimal() +
    labs(title = "Density for 2011")
```

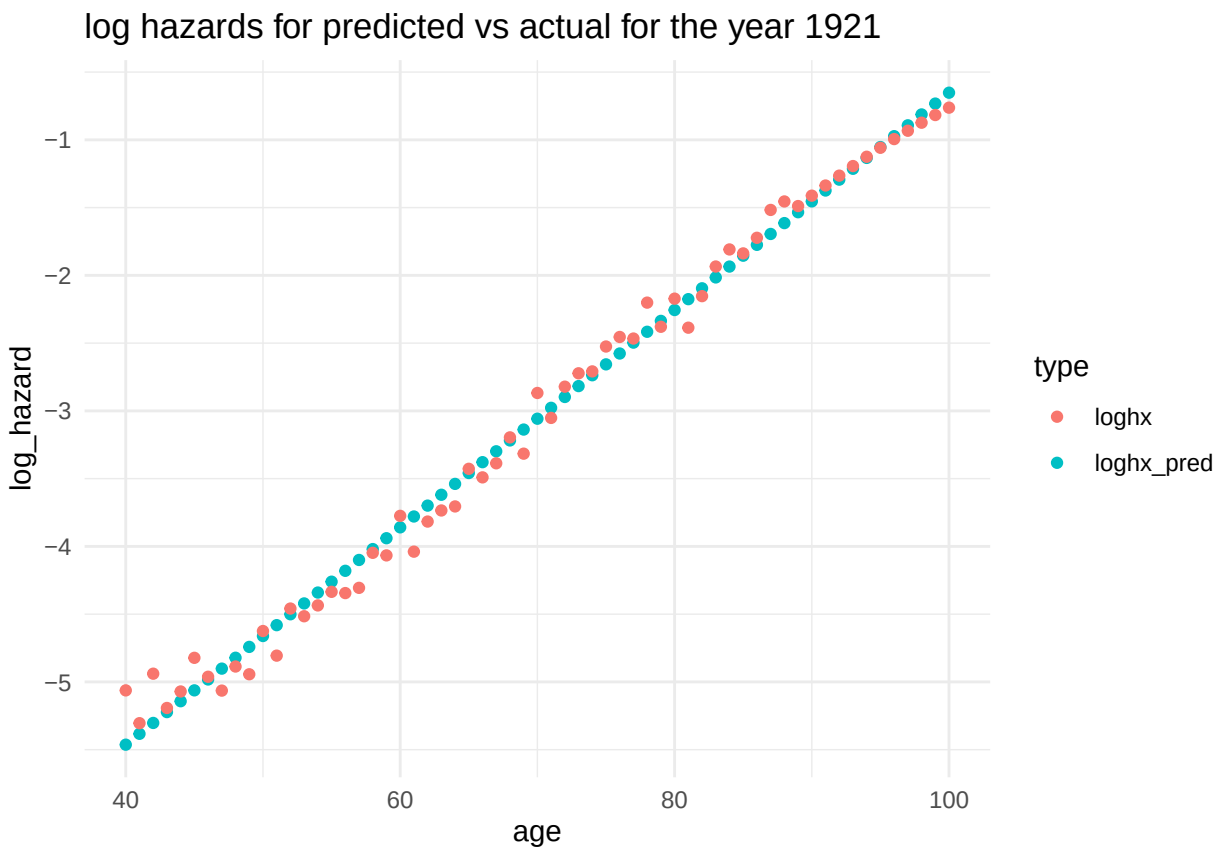


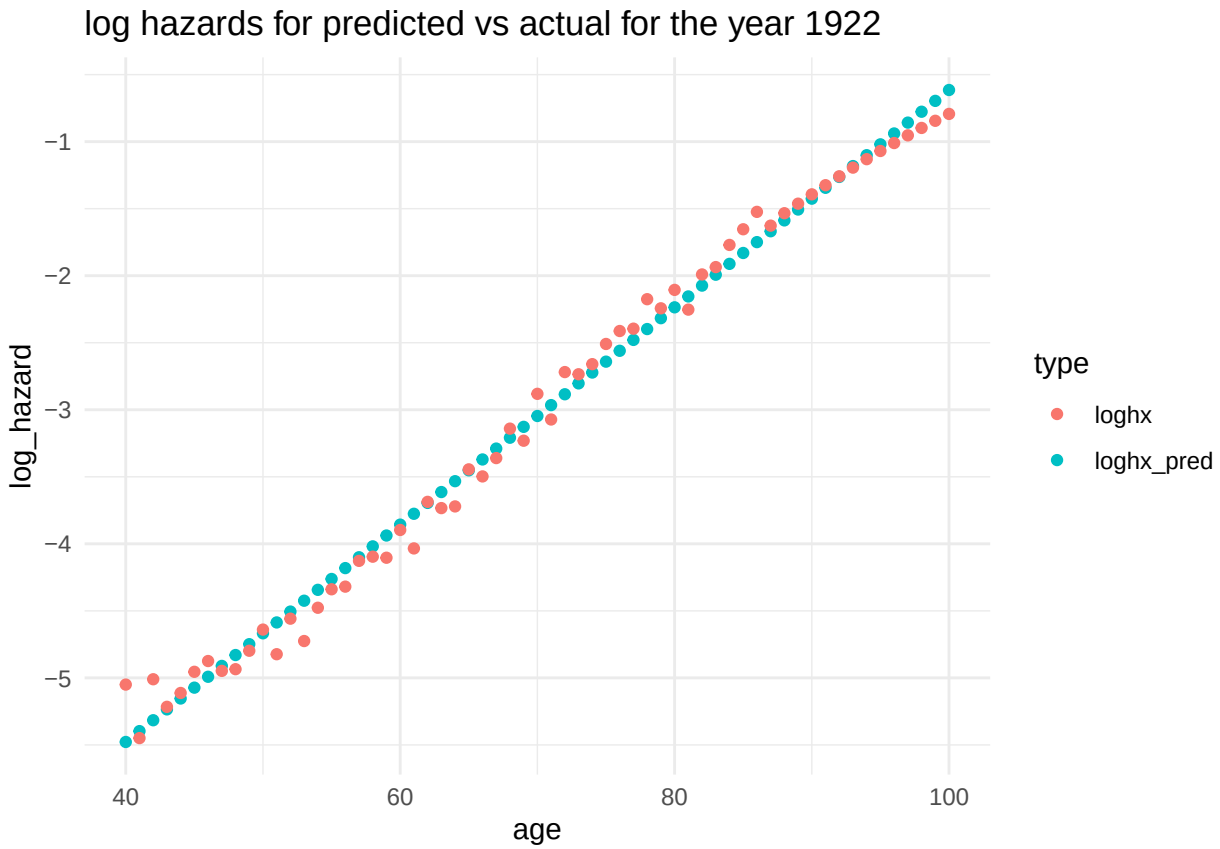
f

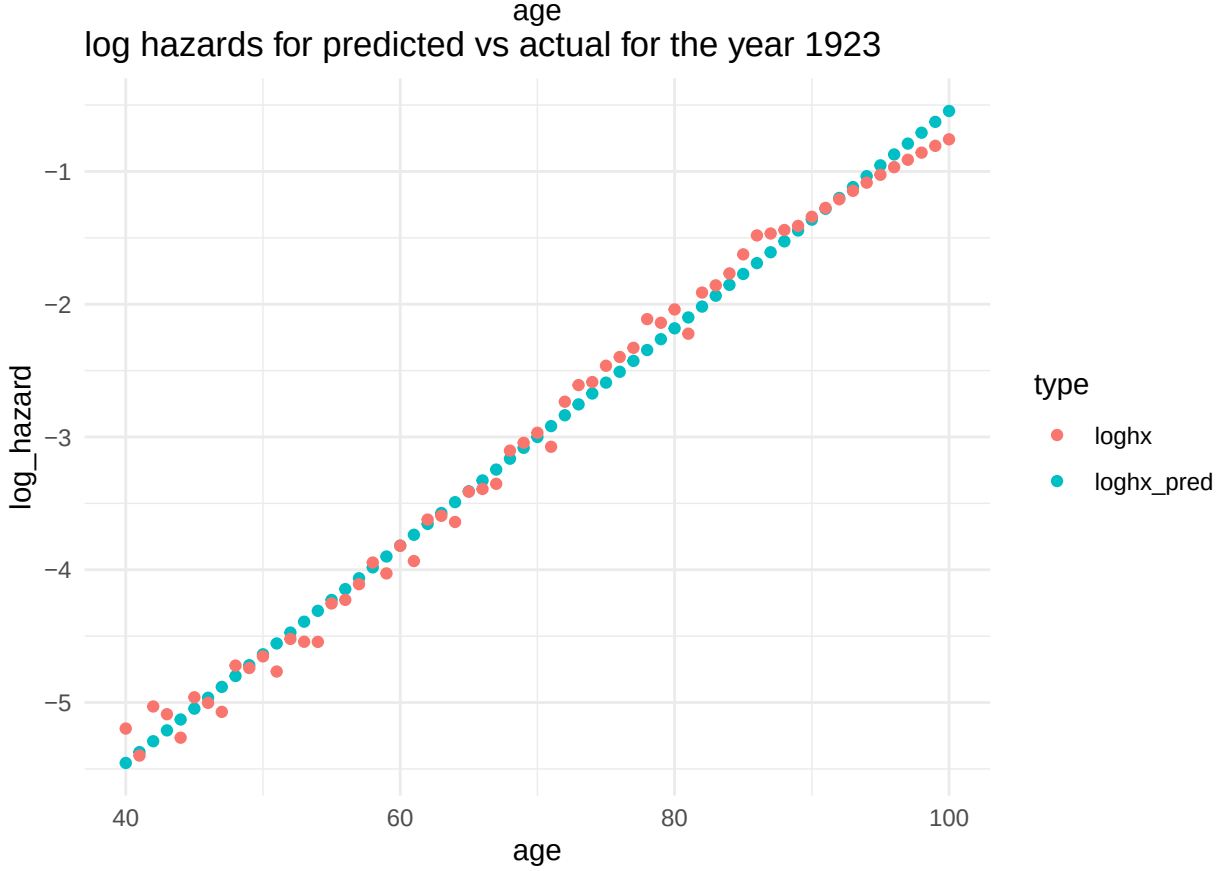
I could probably figure this out with map(lm) but I don't wanna right now.

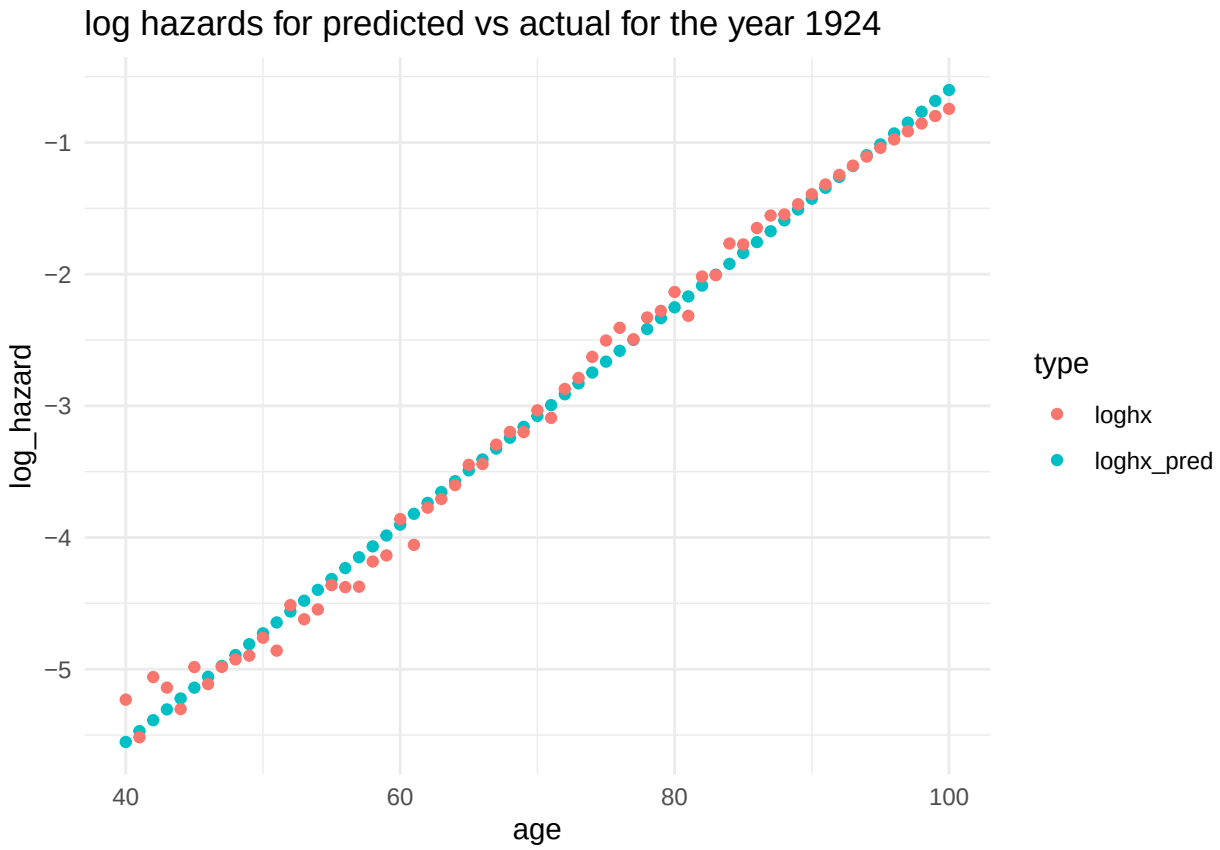
```
years <- unique(df_c$year)
alphas <- rep(NA, length(years))
betas <- rep(NA, length(years))


for (i in 1:length(years)) {
    # print(i)
     # fit model
    df_model <- df_c %>%
        filter(year == years[i])
    lm_loop <- lm(loghx ~ age, data = df_model)
    coef_model <- coef(lm_loop)

    alphas[i] <- unname(exp(coef_model[1]))
    betas[i] <- unname(coef_model[2])

    df_model$loghx_pred <- predict(lm_loop, newdata = df_model)

    df_plot <- df_model %>% pivot_longer(cols = c("loghx_pred", "loghx"), names_to = "type", values_to =
    p <- df_plot %>% ggplot(aes(x = age, y = log_hazard, color = type)) +
    geom_point() +
    theme_minimal() +
    labs(title = paste0("log hazards for predicted vs actual for the year ", years[i]))
```

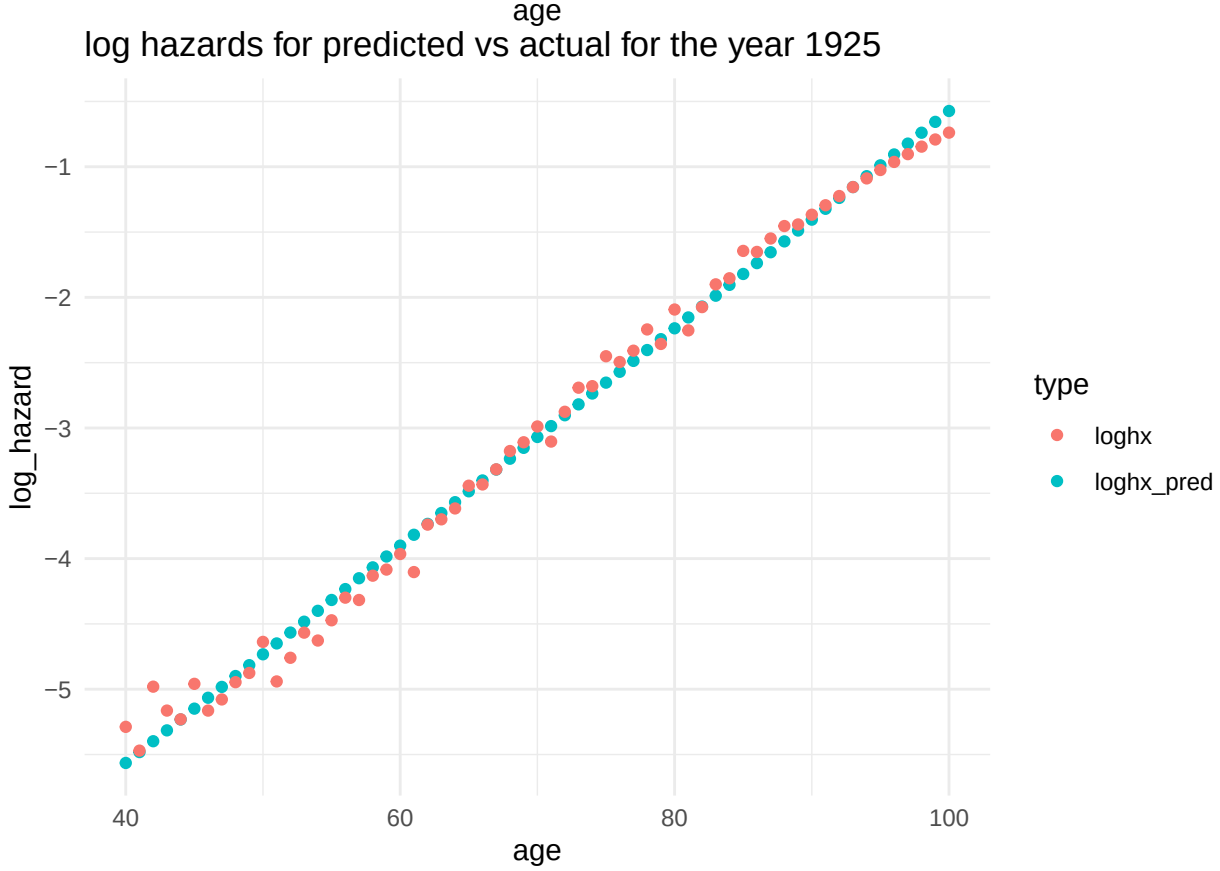
```
    print(p)
}
```

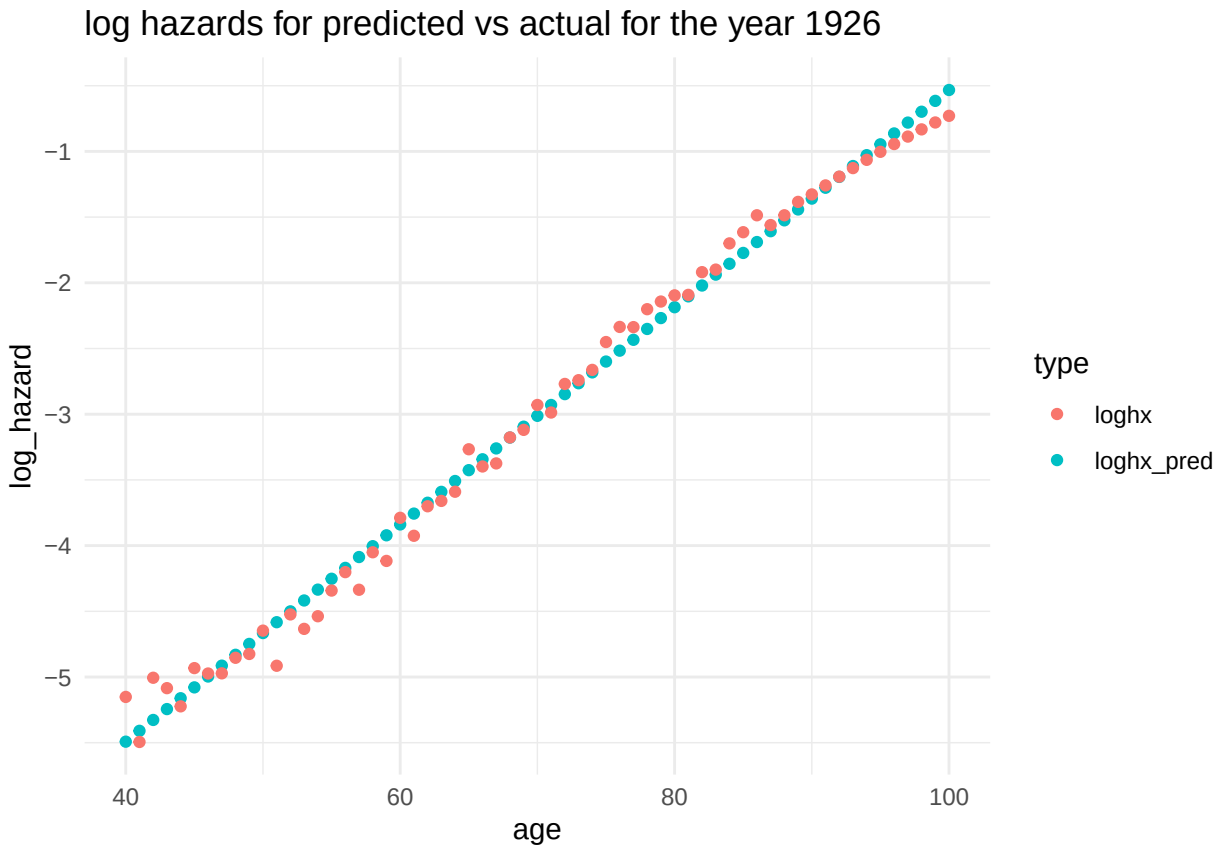

log hazards for predicted vs actual for the year 1921
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

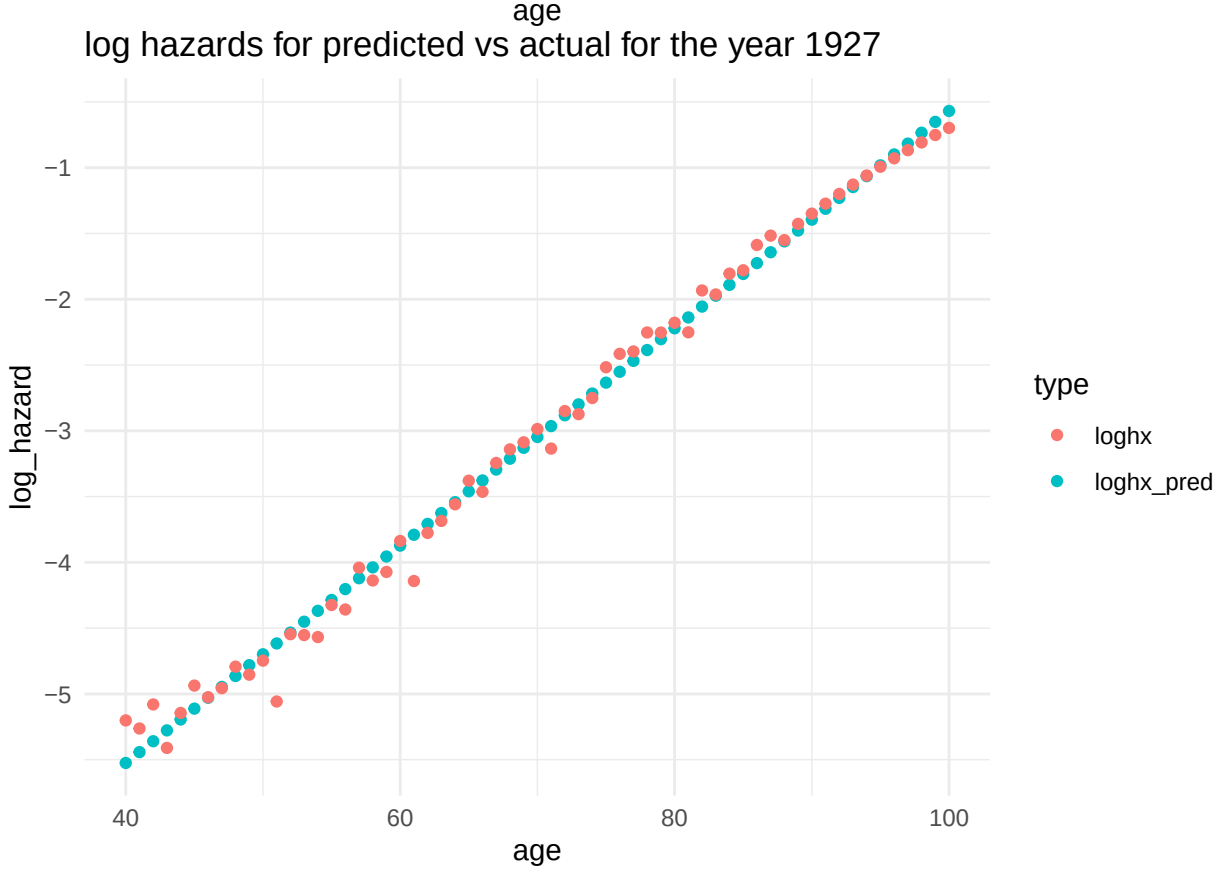

log hazards for predicted vs actual for the year 1922
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

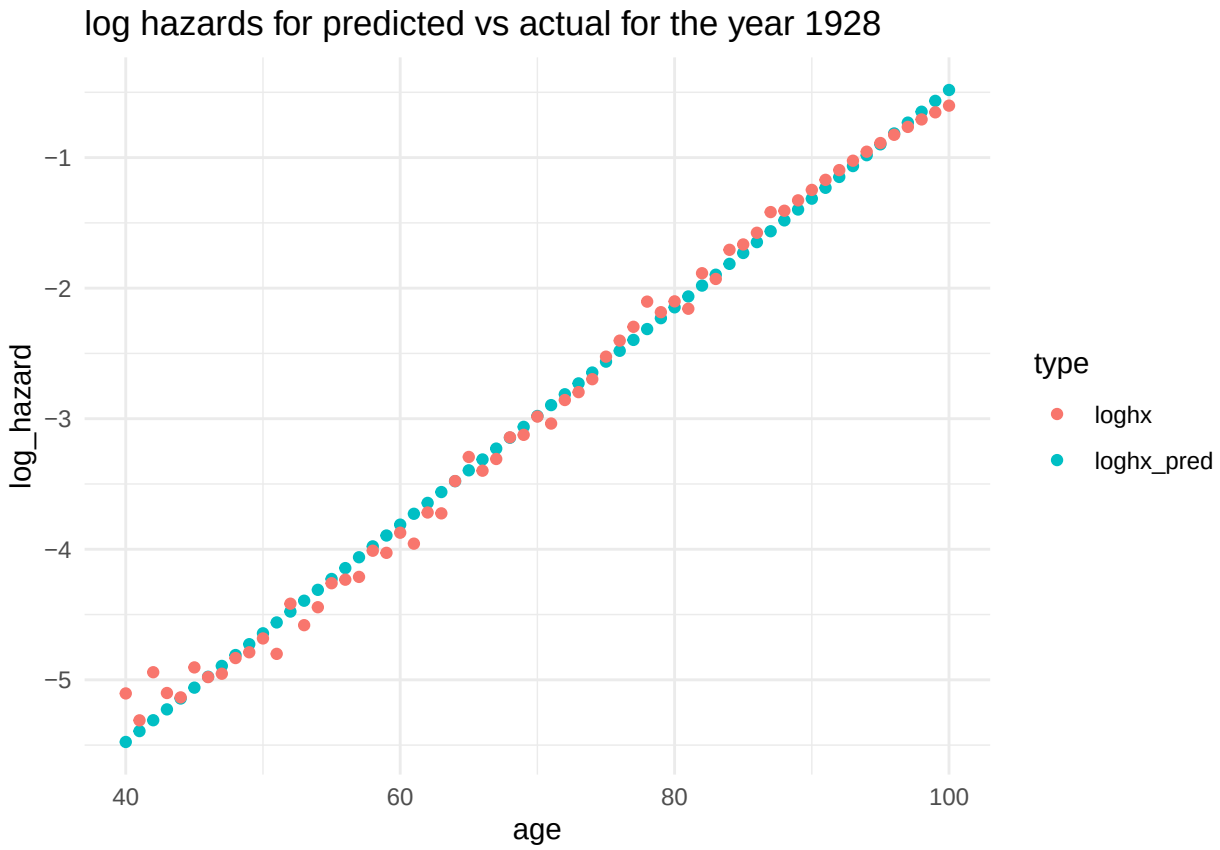

log hazards for predicted vs actual for the year 1923
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

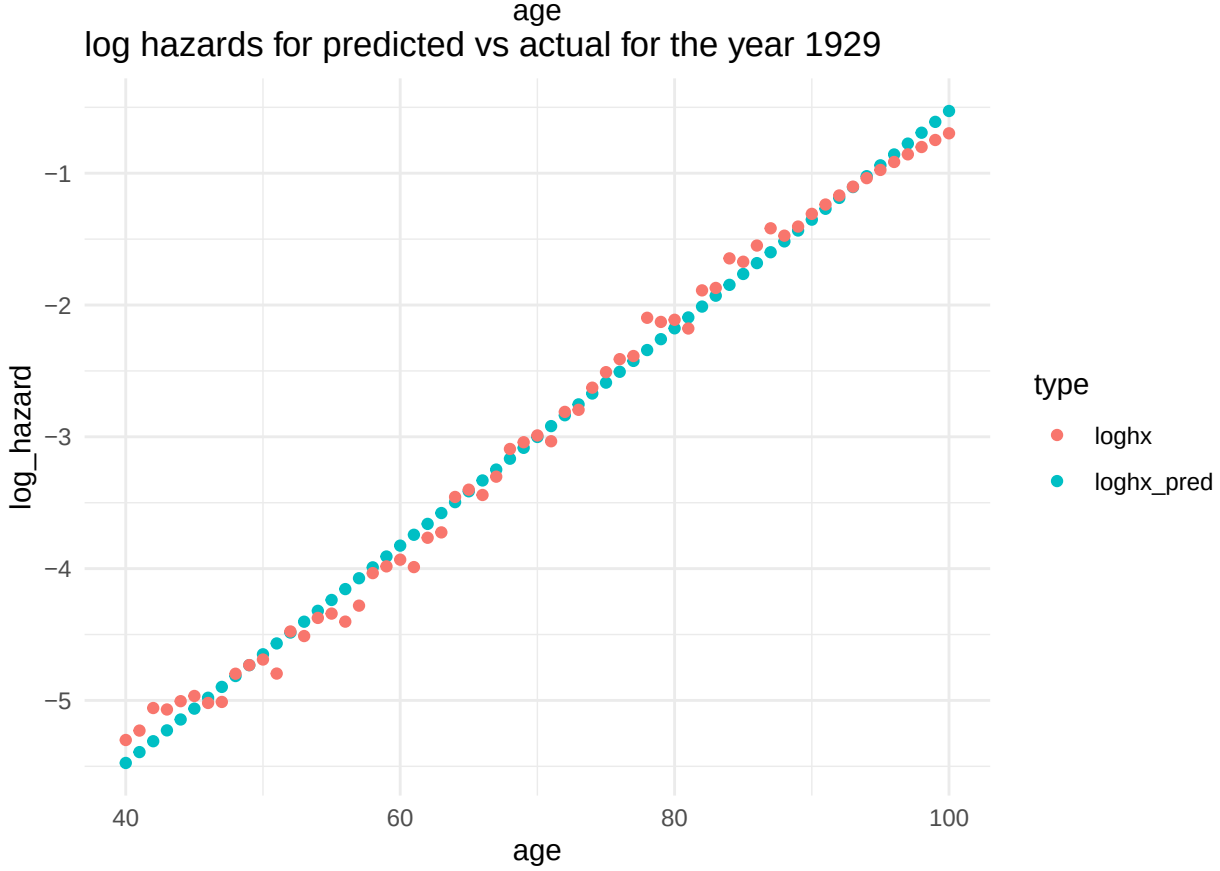

log hazards for predicted vs actual for the year 1924
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

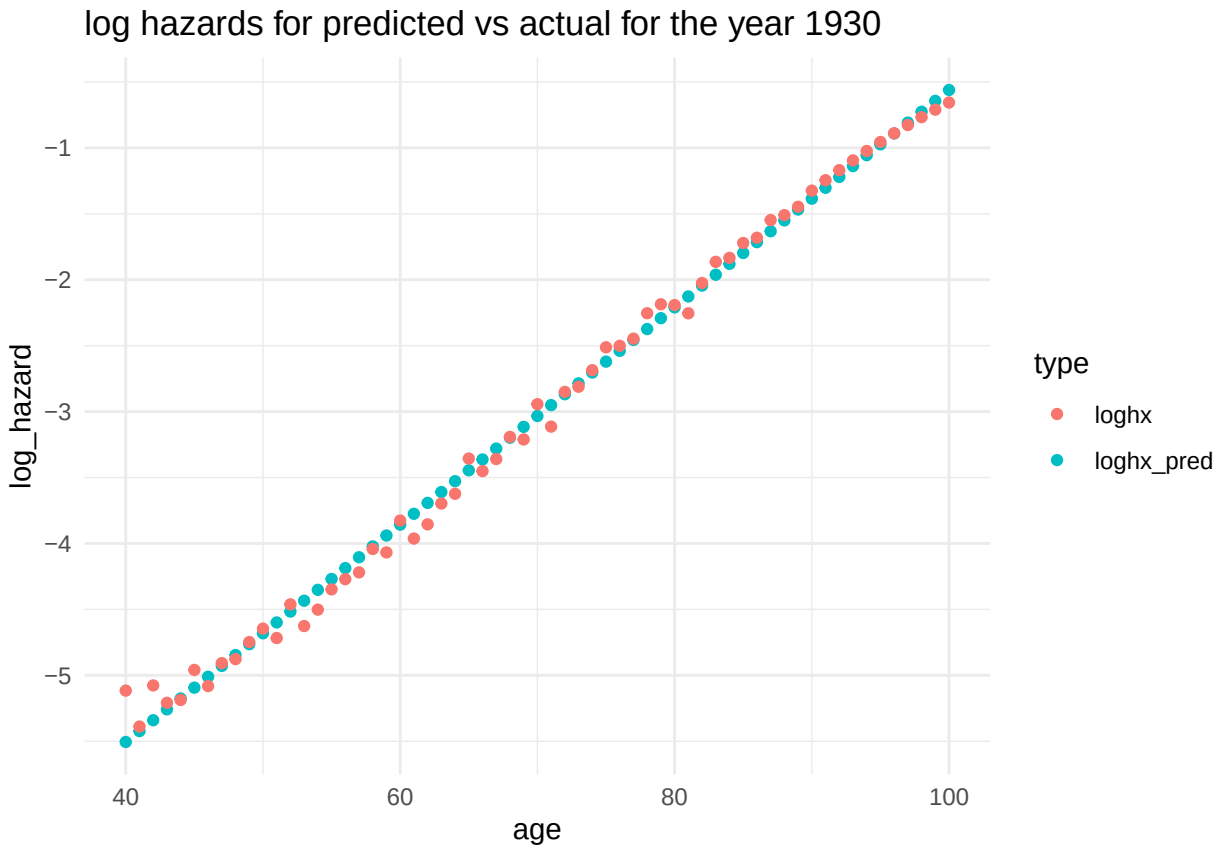

log hazards for predicted vs actual for the year 1925
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

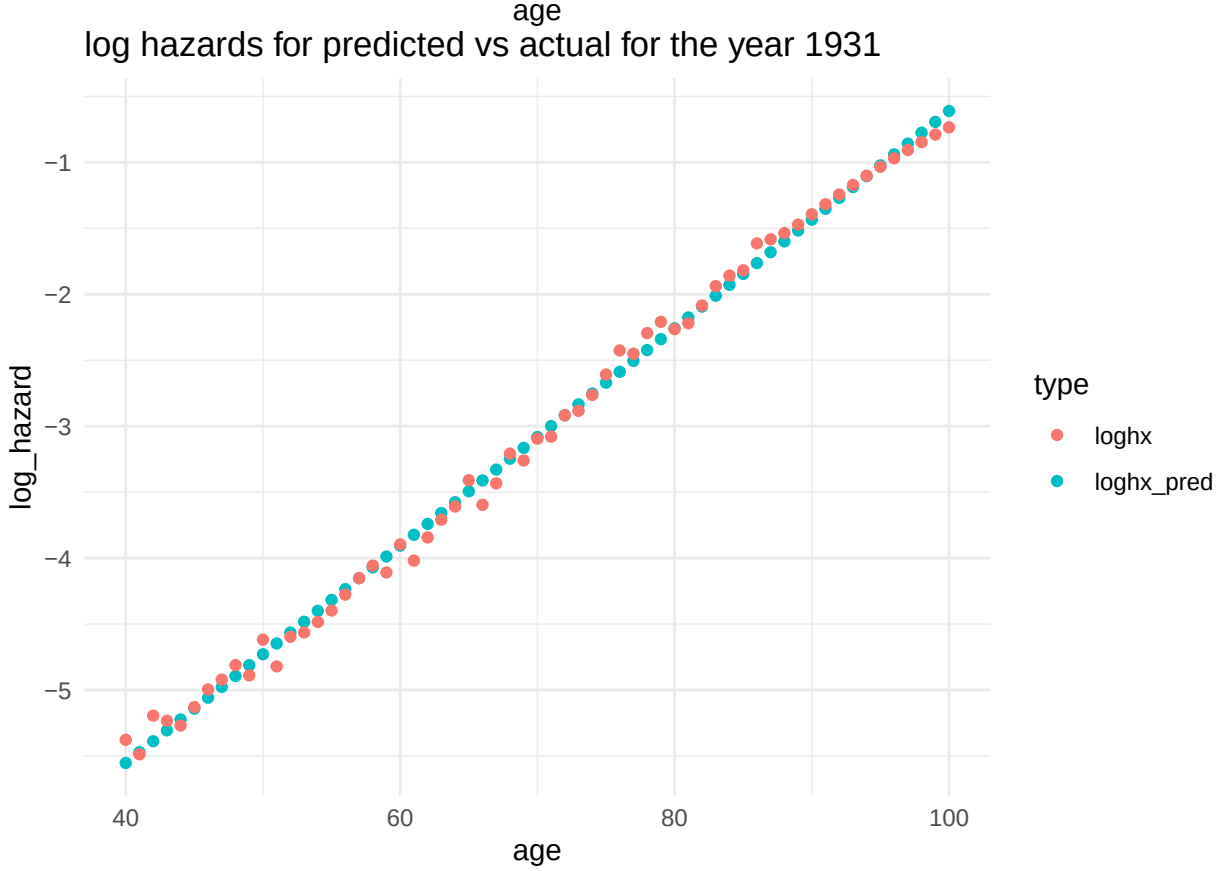

log hazards for predicted vs actual for the year 1926
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

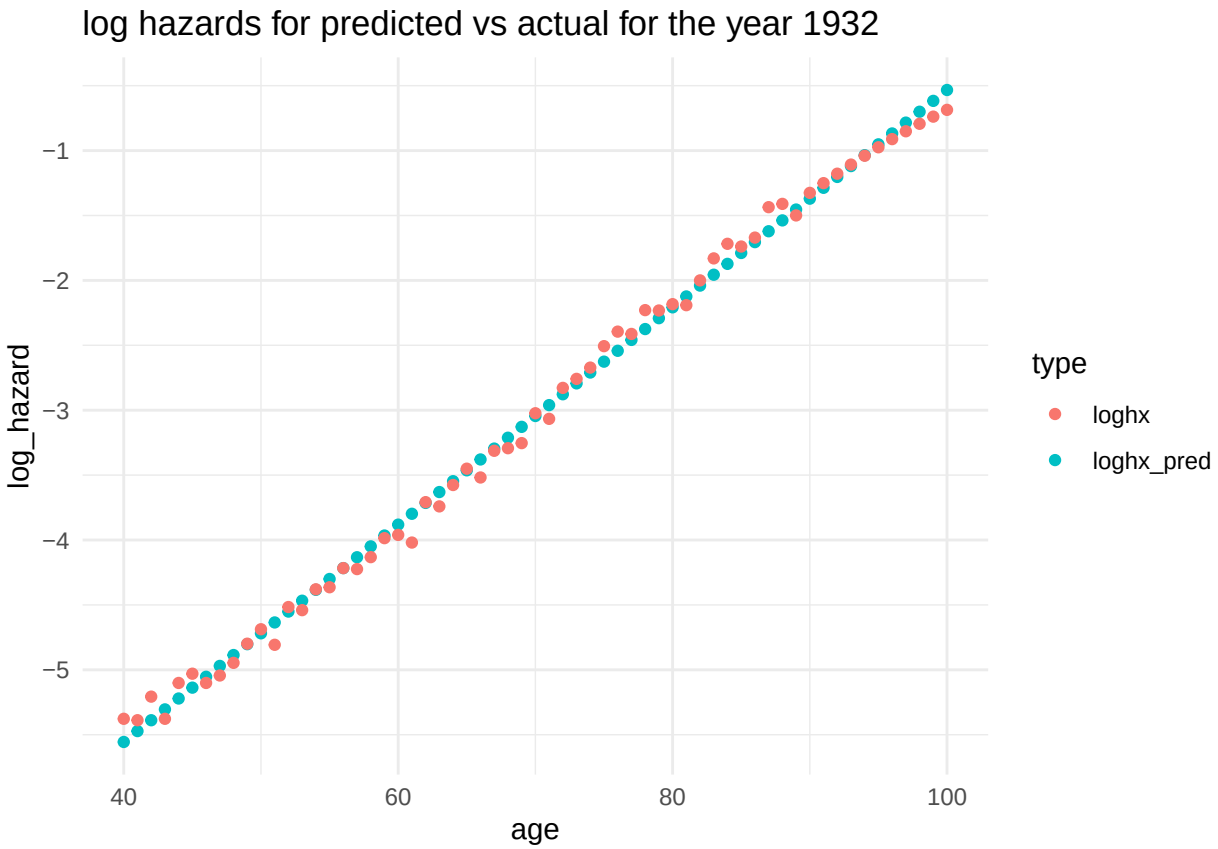

log hazards for predicted vs actual for the year 1927
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

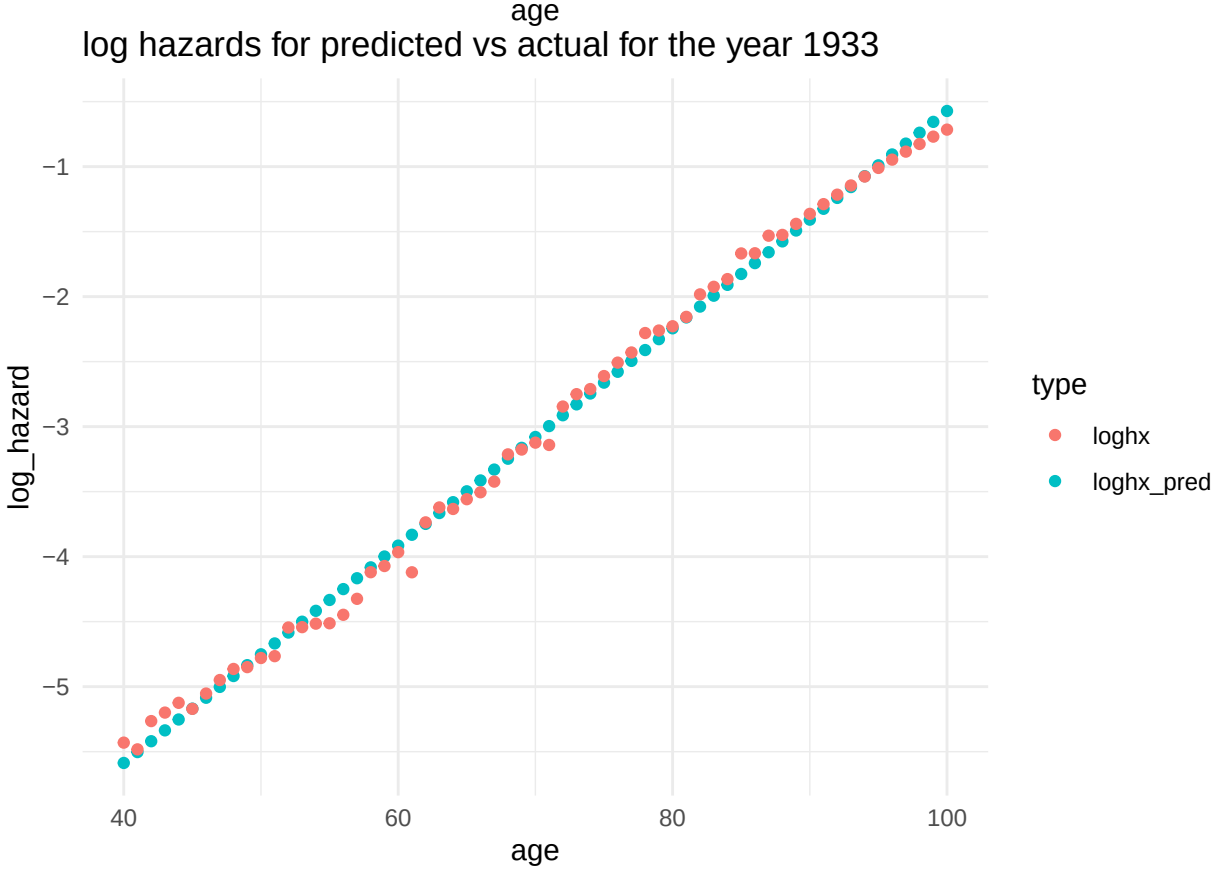

log hazards for predicted vs actual for the year 1928
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

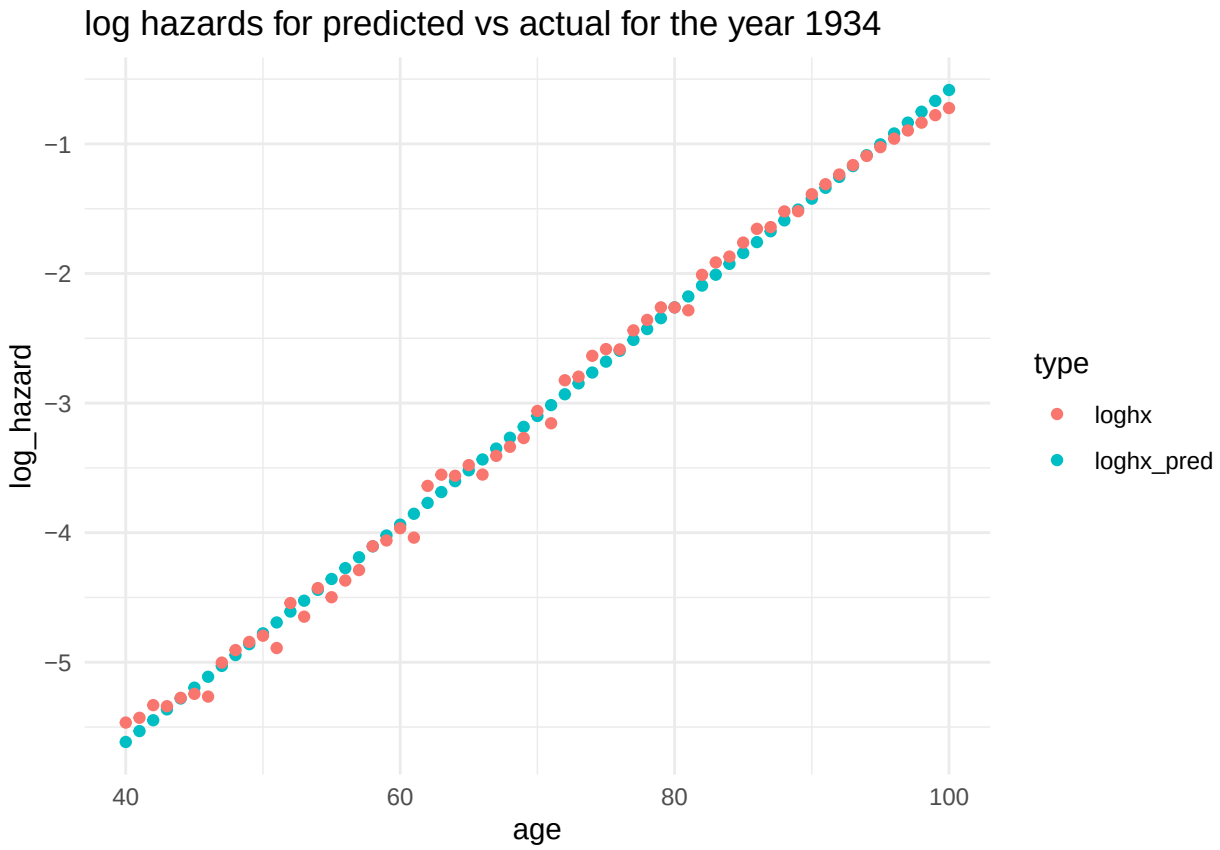

log hazards for predicted vs actual for the year 1929
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

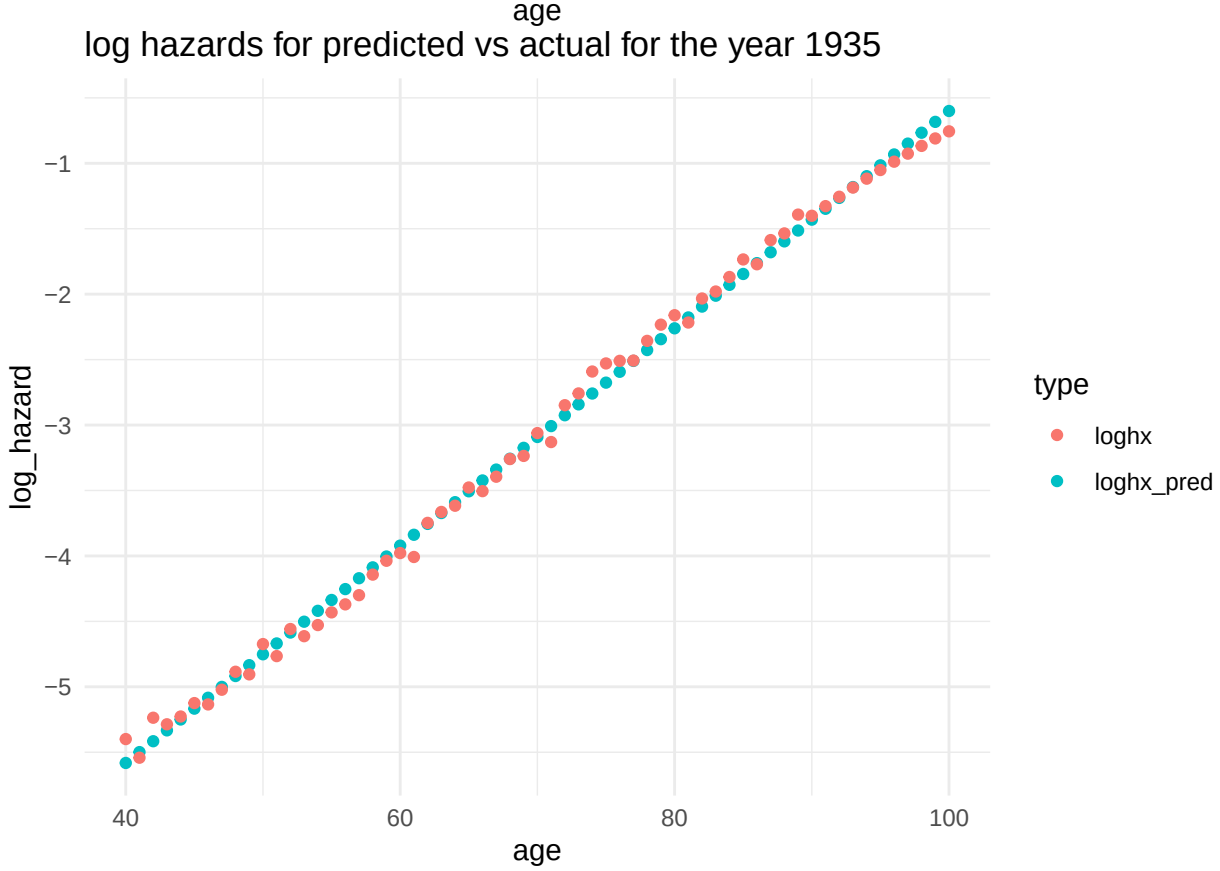

log hazards for predicted vs actual for the year 1930
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

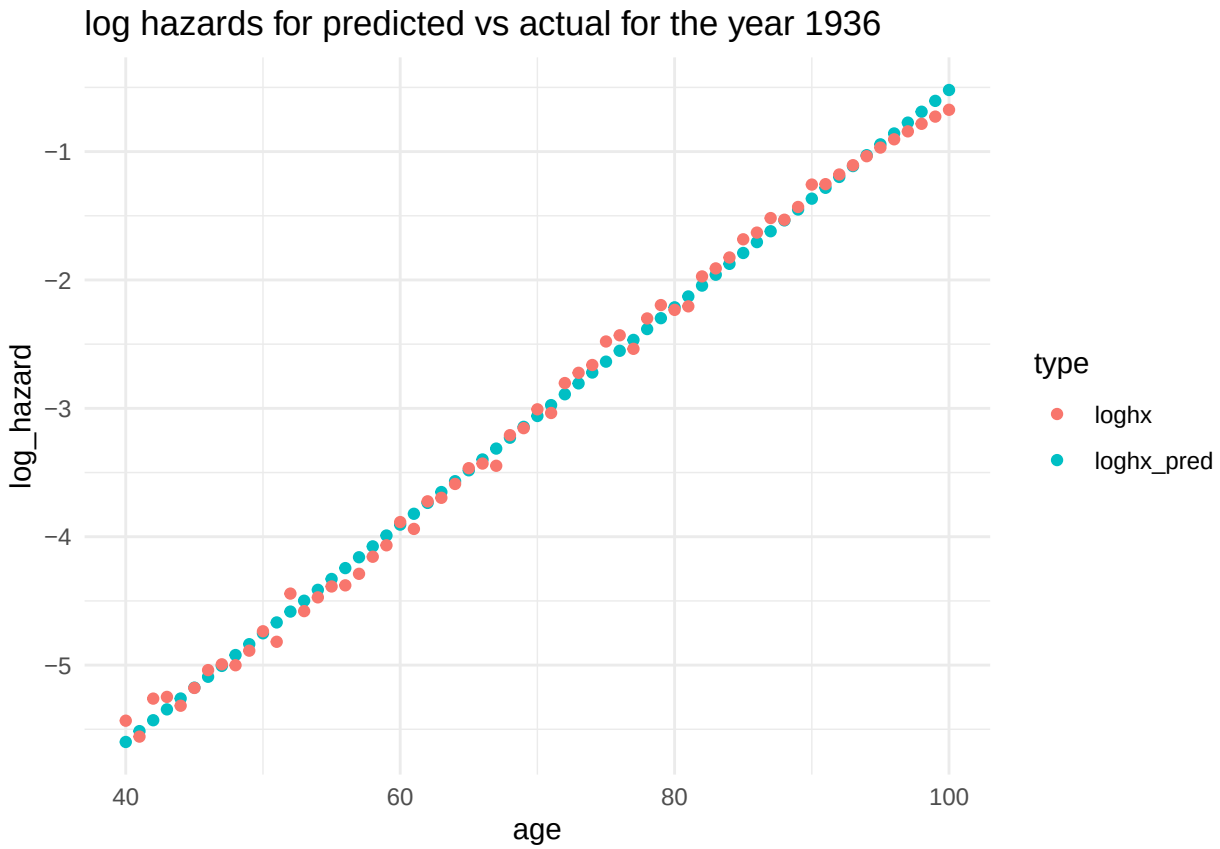

log hazards for predicted vs actual for the year 1931
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

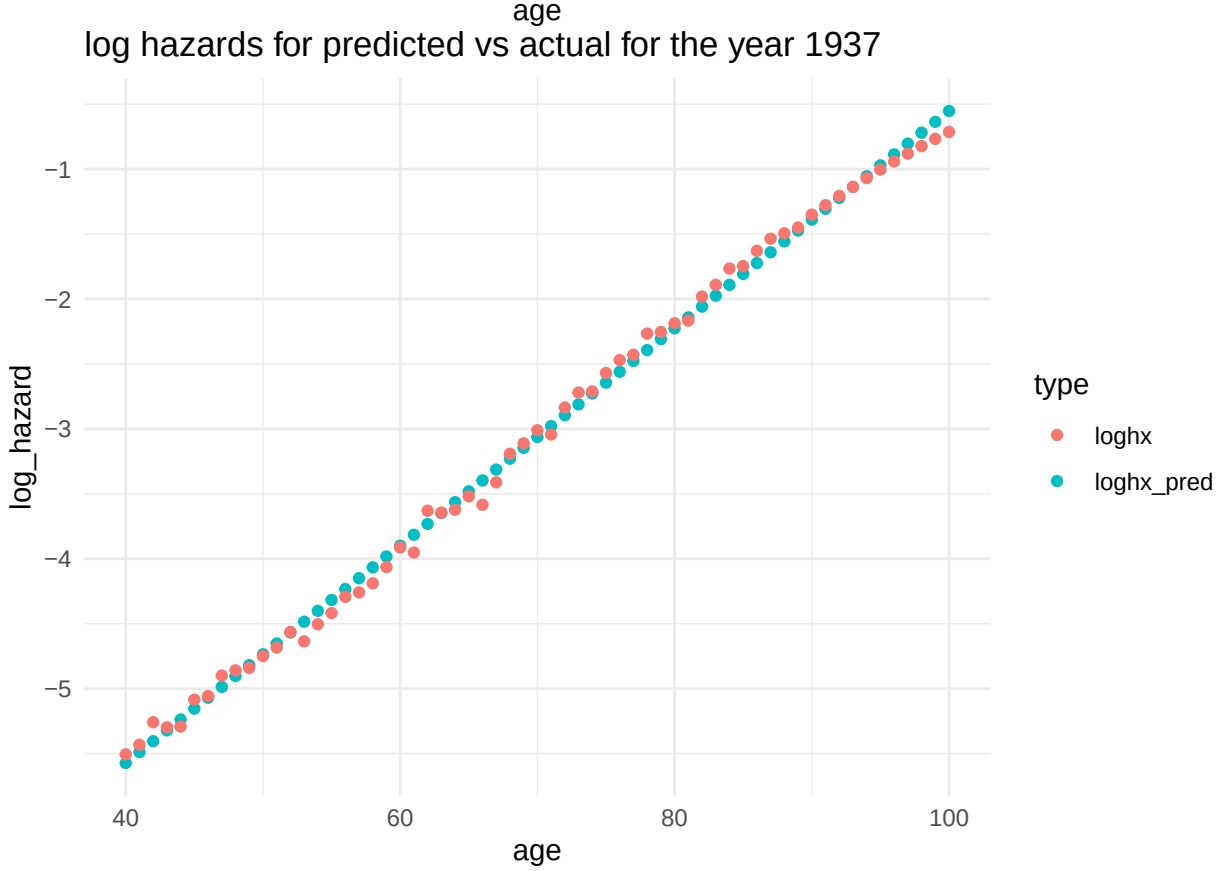

log hazards for predicted vs actual for the year 1932
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

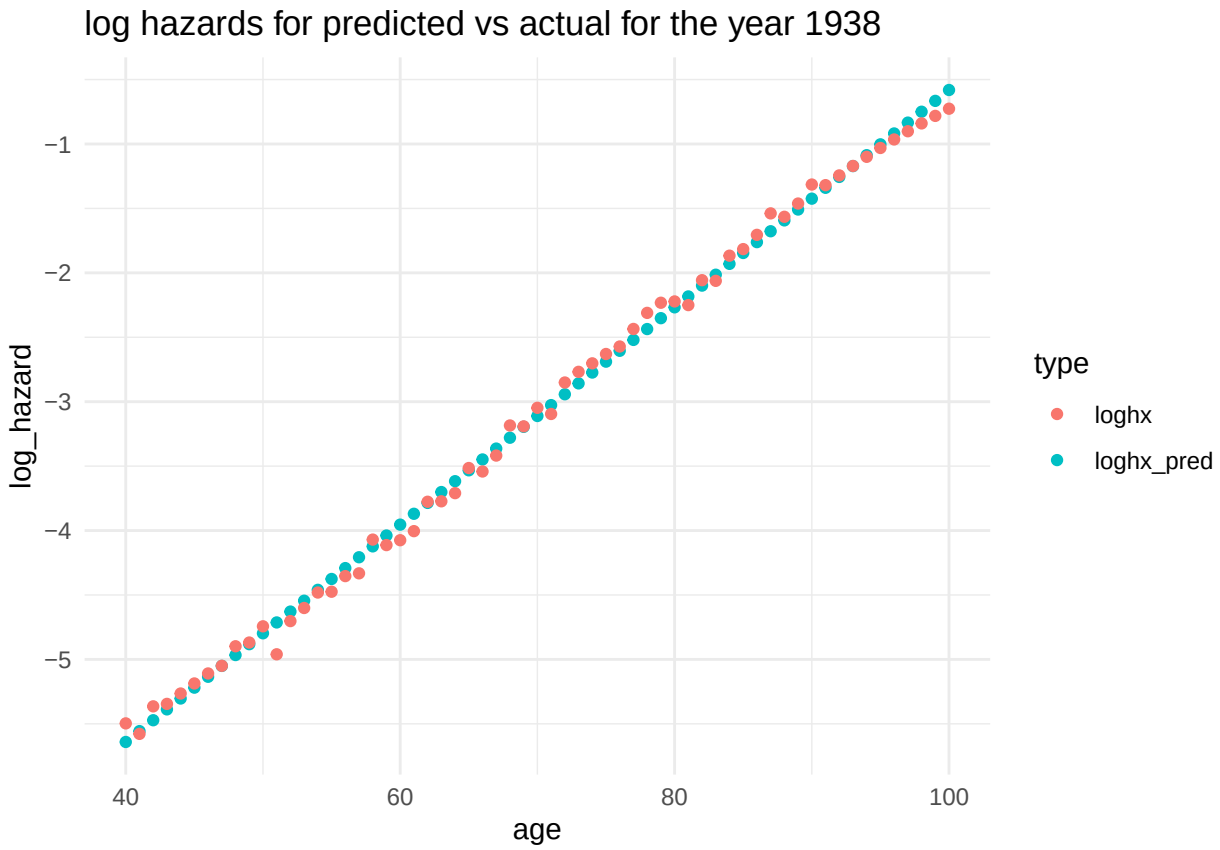

log hazards for predicted vs actual for the year 1933
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

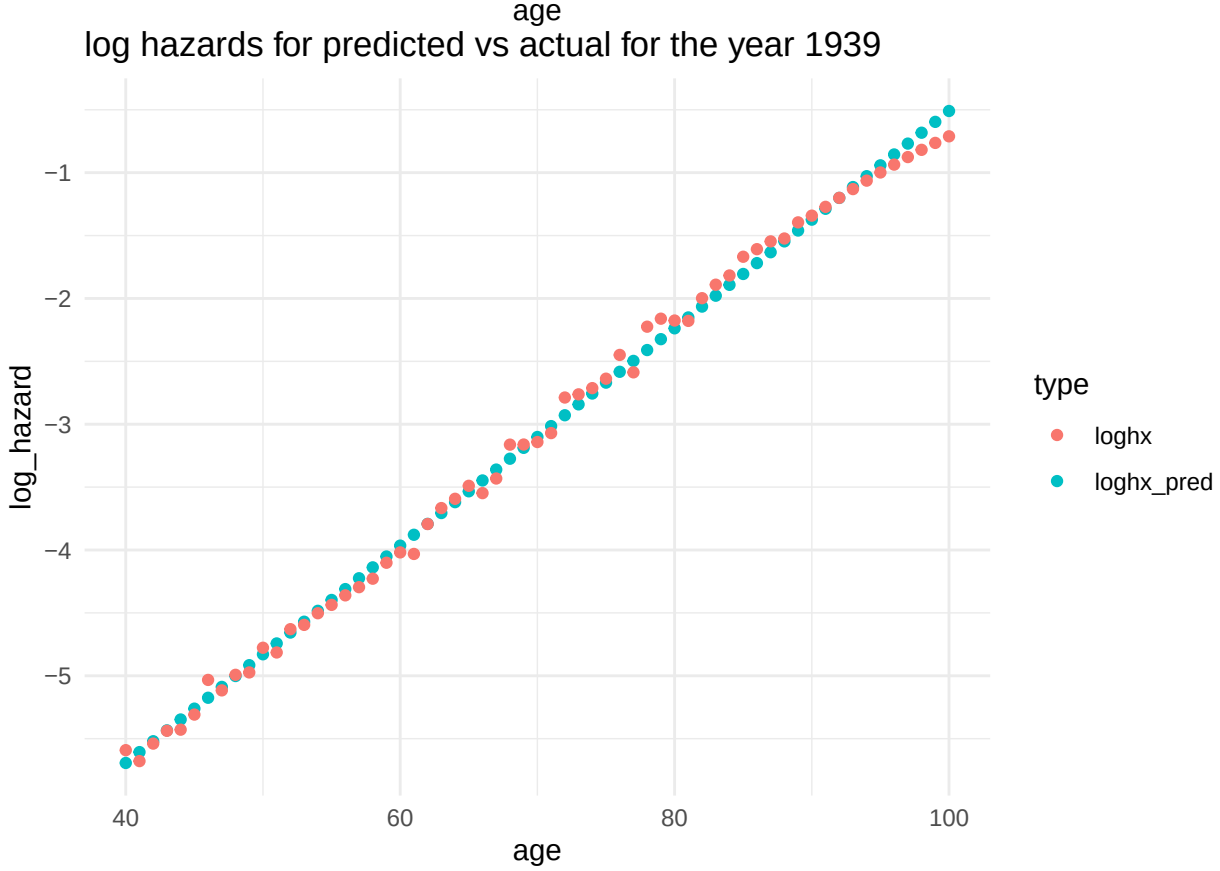

log hazards for predicted vs actual for the year 1934
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

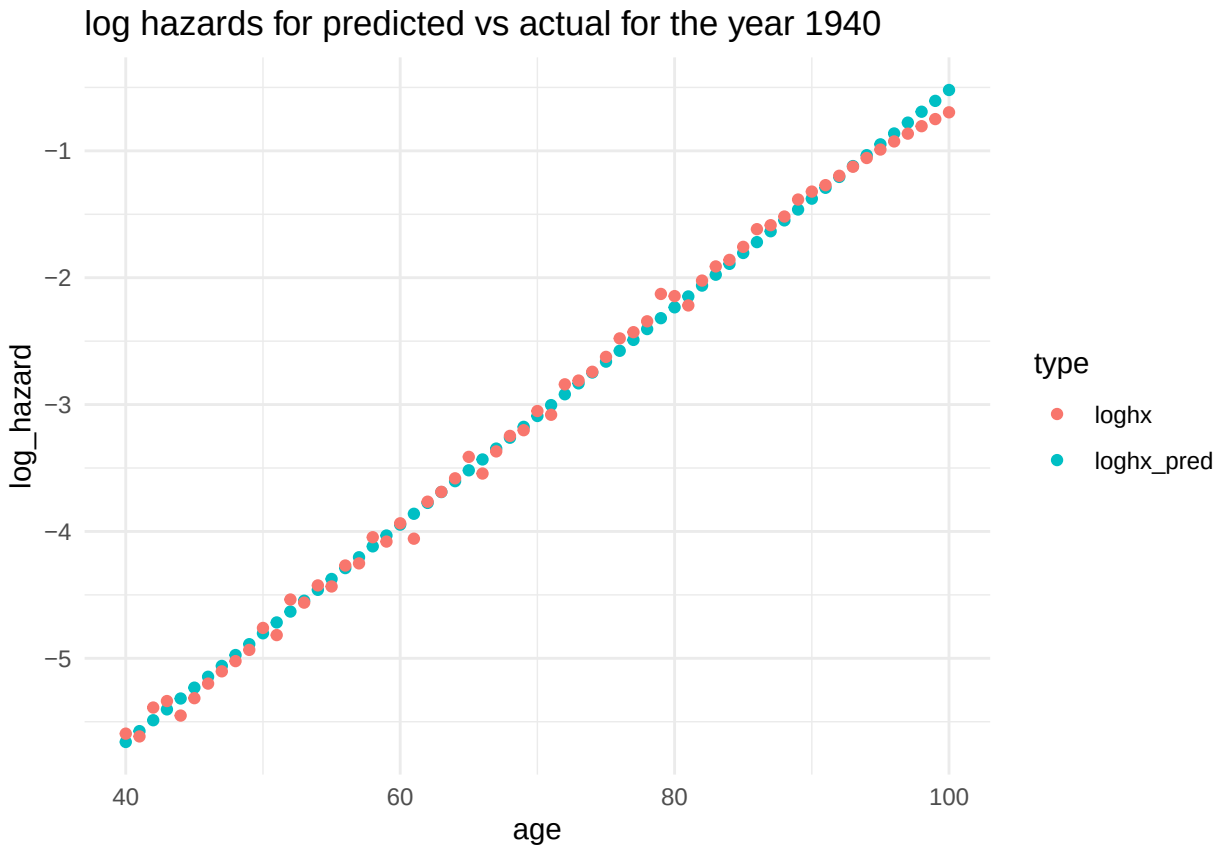

log hazards for predicted vs actual for the year 1935
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

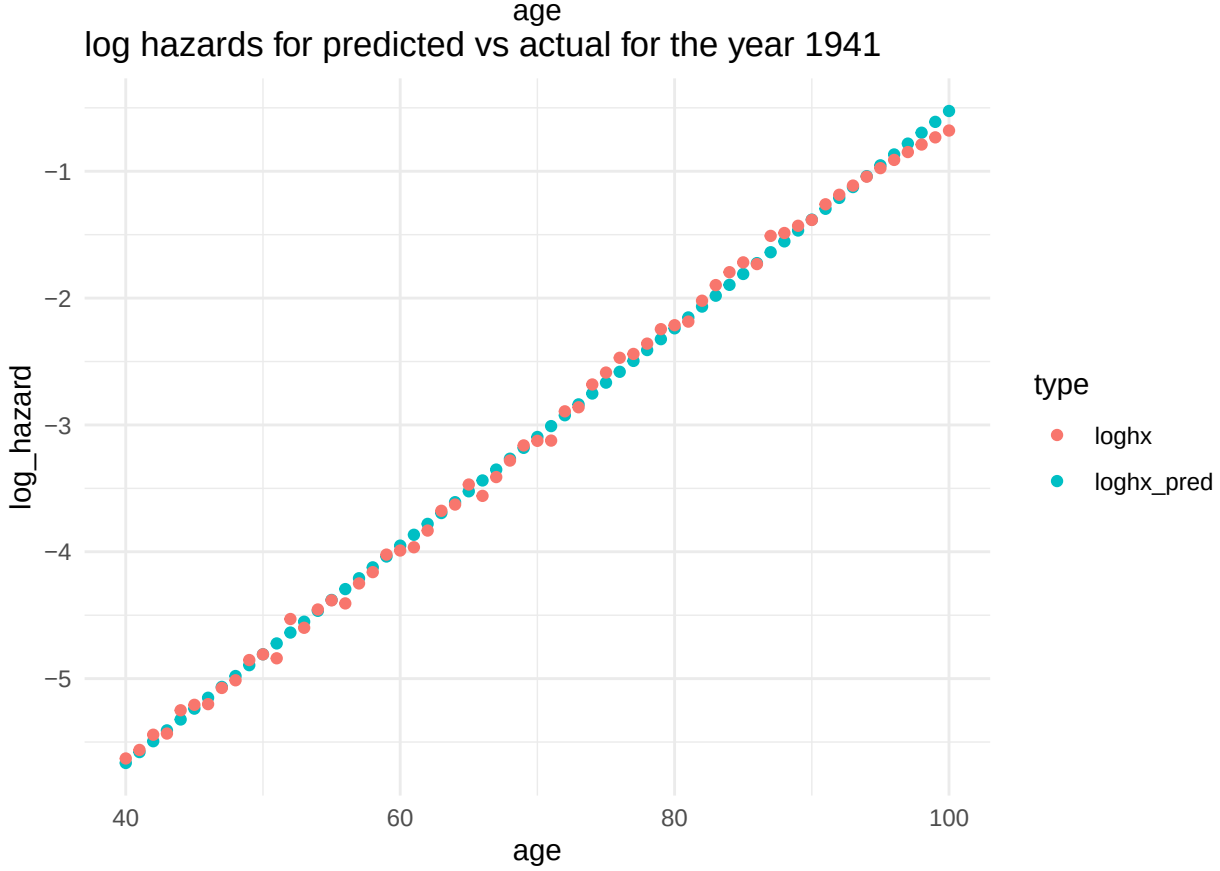

log hazards for predicted vs actual for the year 1936
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

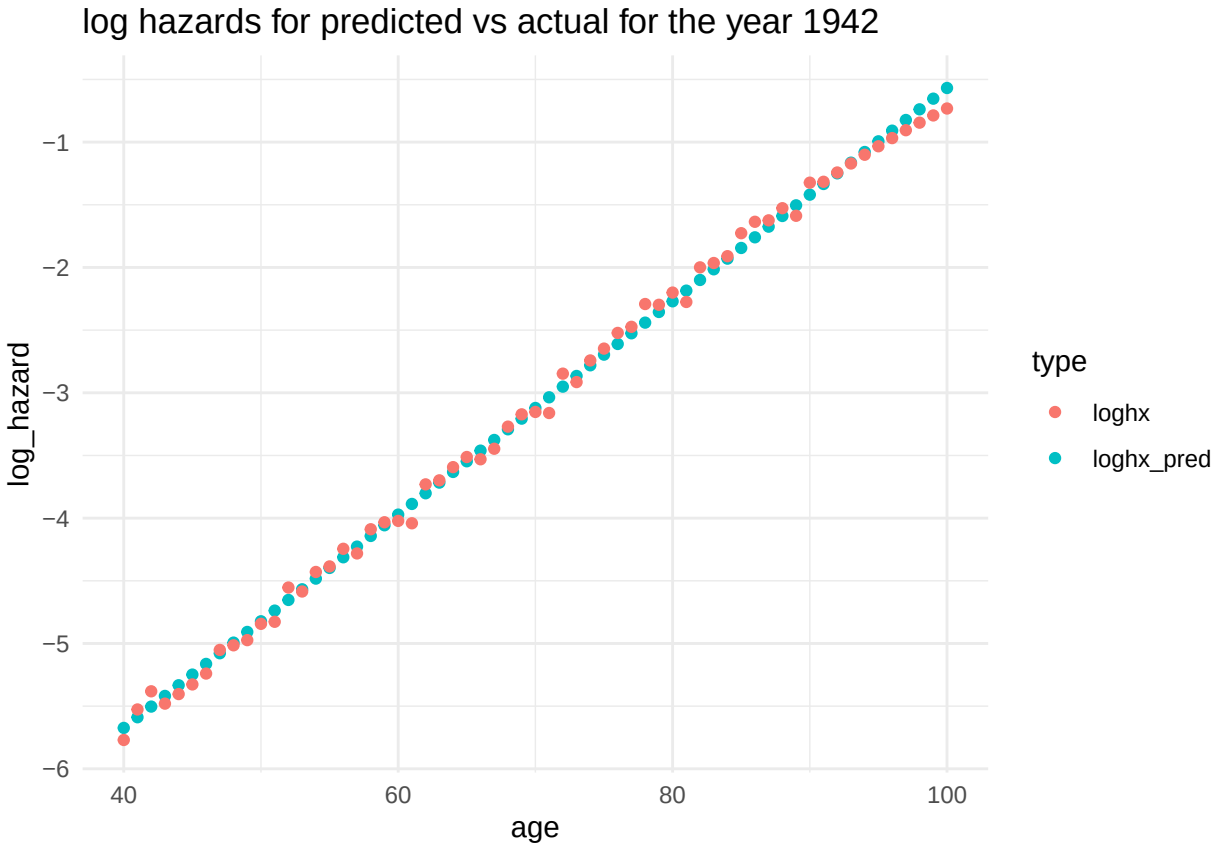

log hazards for predicted vs actual for the year 1937
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

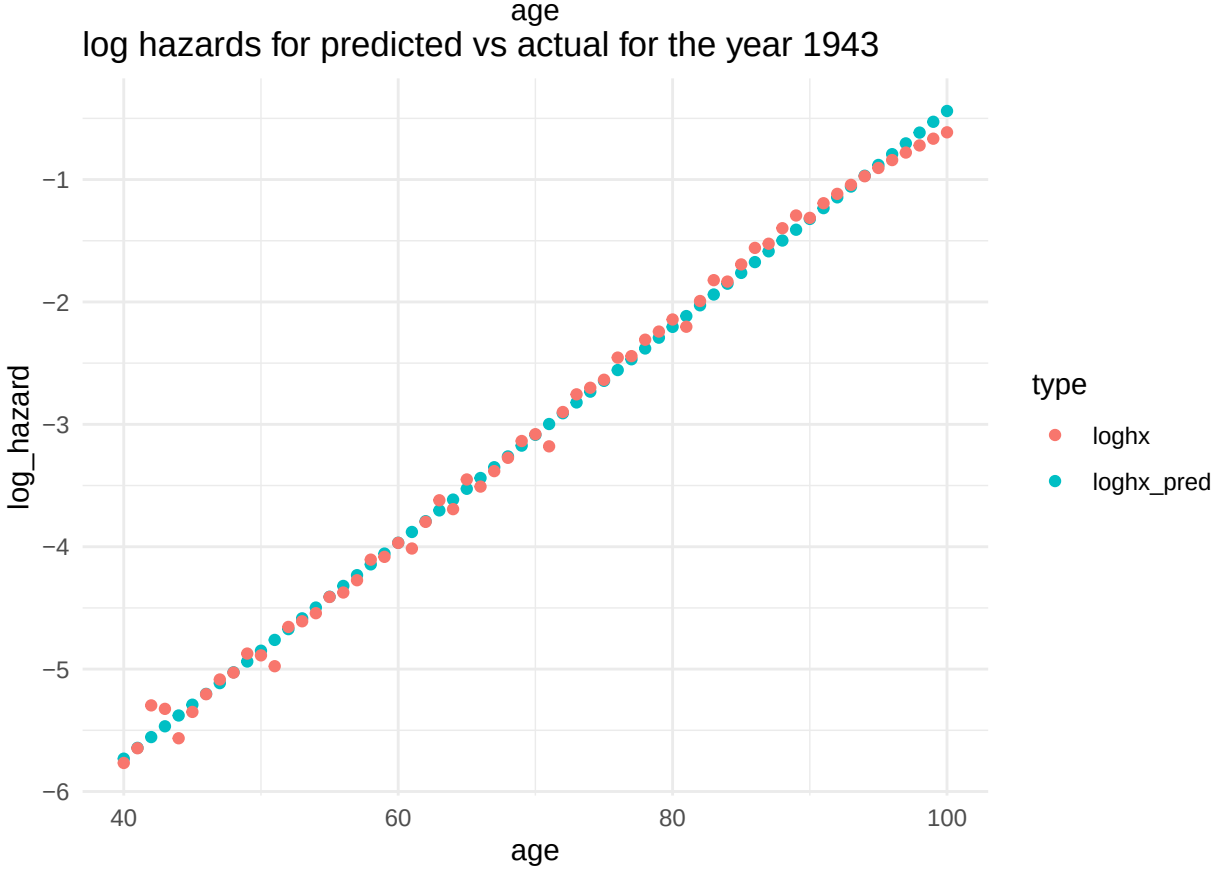

log hazards for predicted vs actual for the year 1938
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

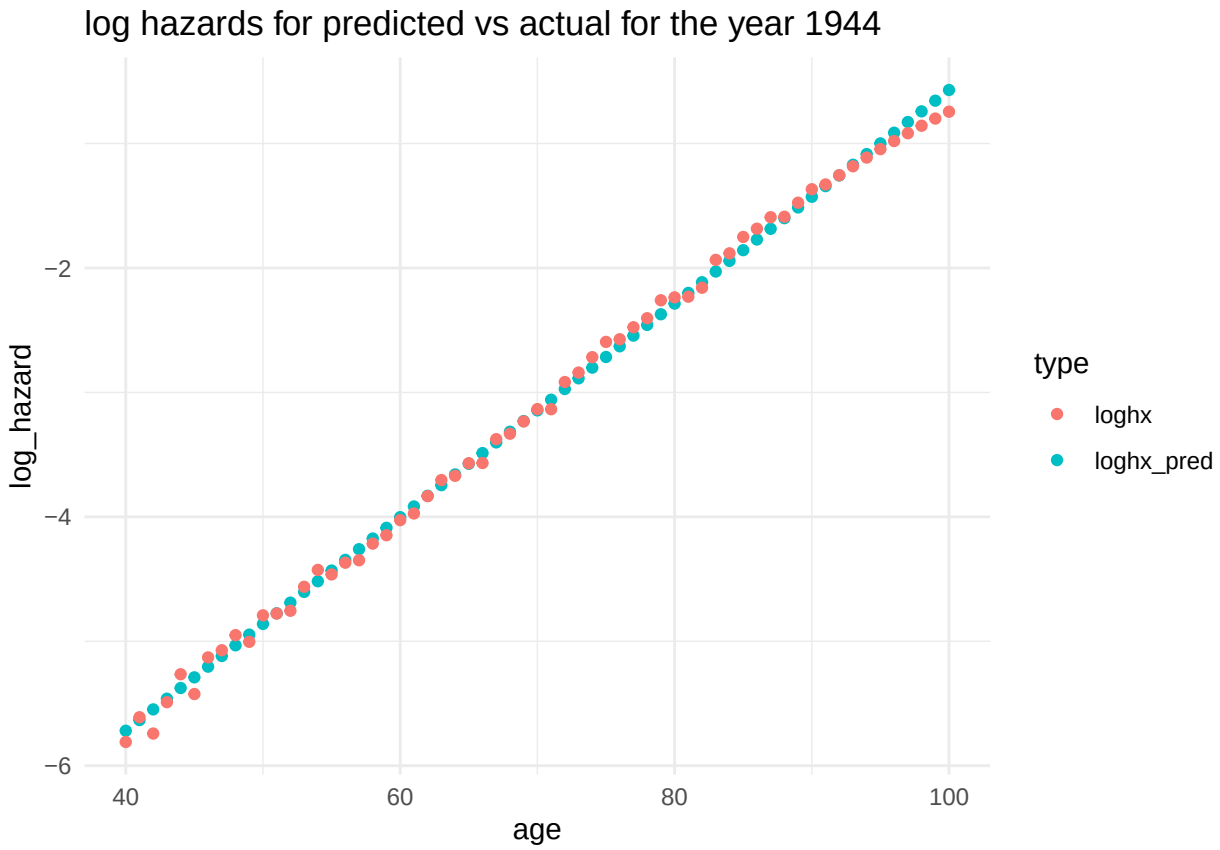

log hazards for predicted vs actual for the year 1939
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

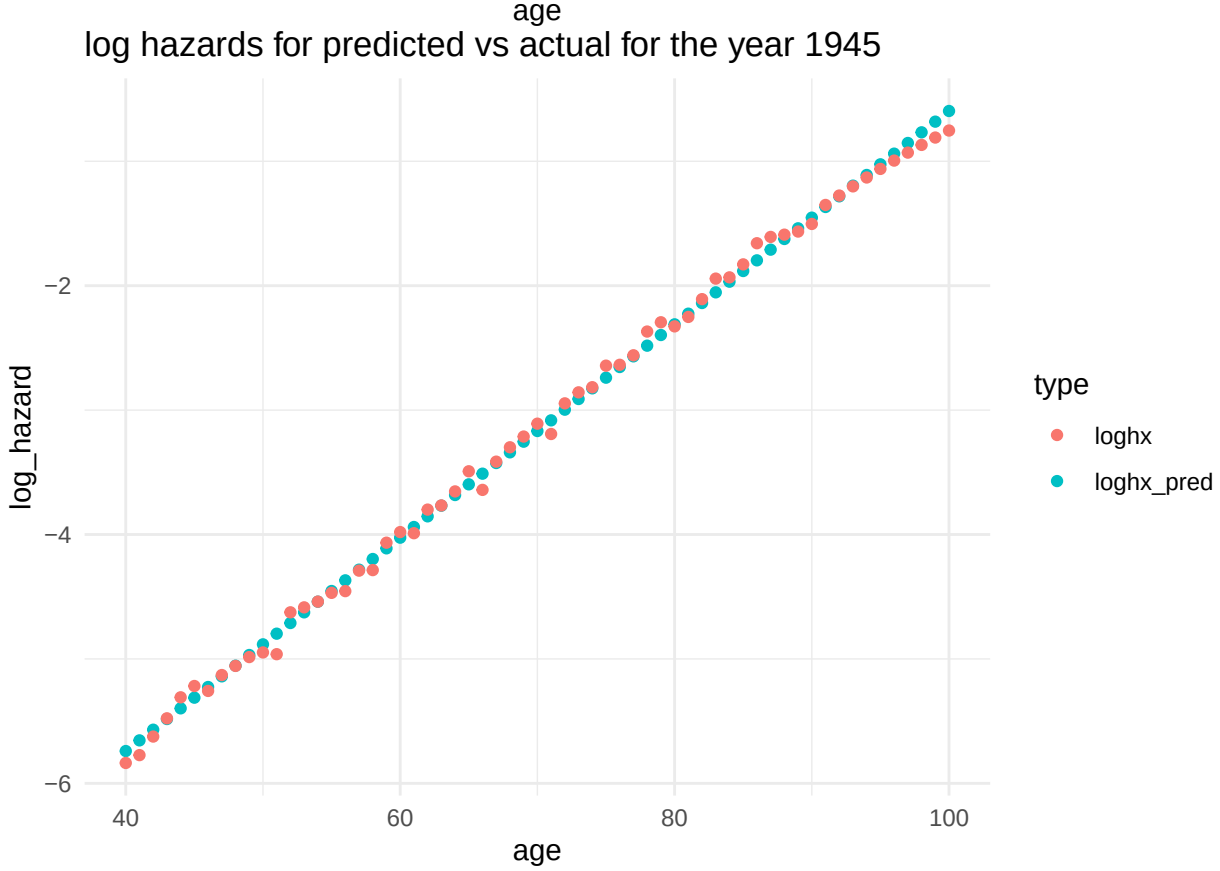

log hazards for predicted vs actual for the year 1940
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

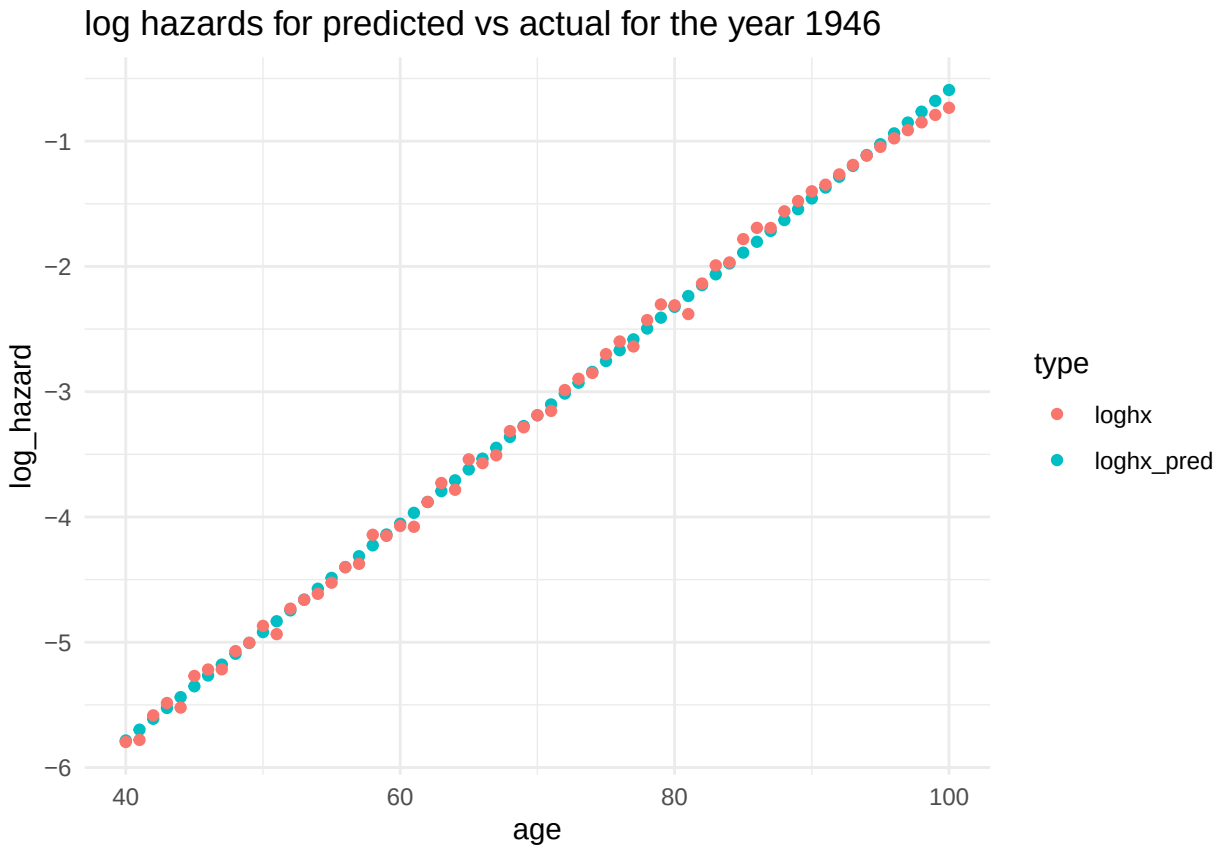

log hazards for predicted vs actual for the year 1941
log_hazard
−1
−2
−3
−4
−5
40
60
80
100
age
type
loghx
loghx_pred

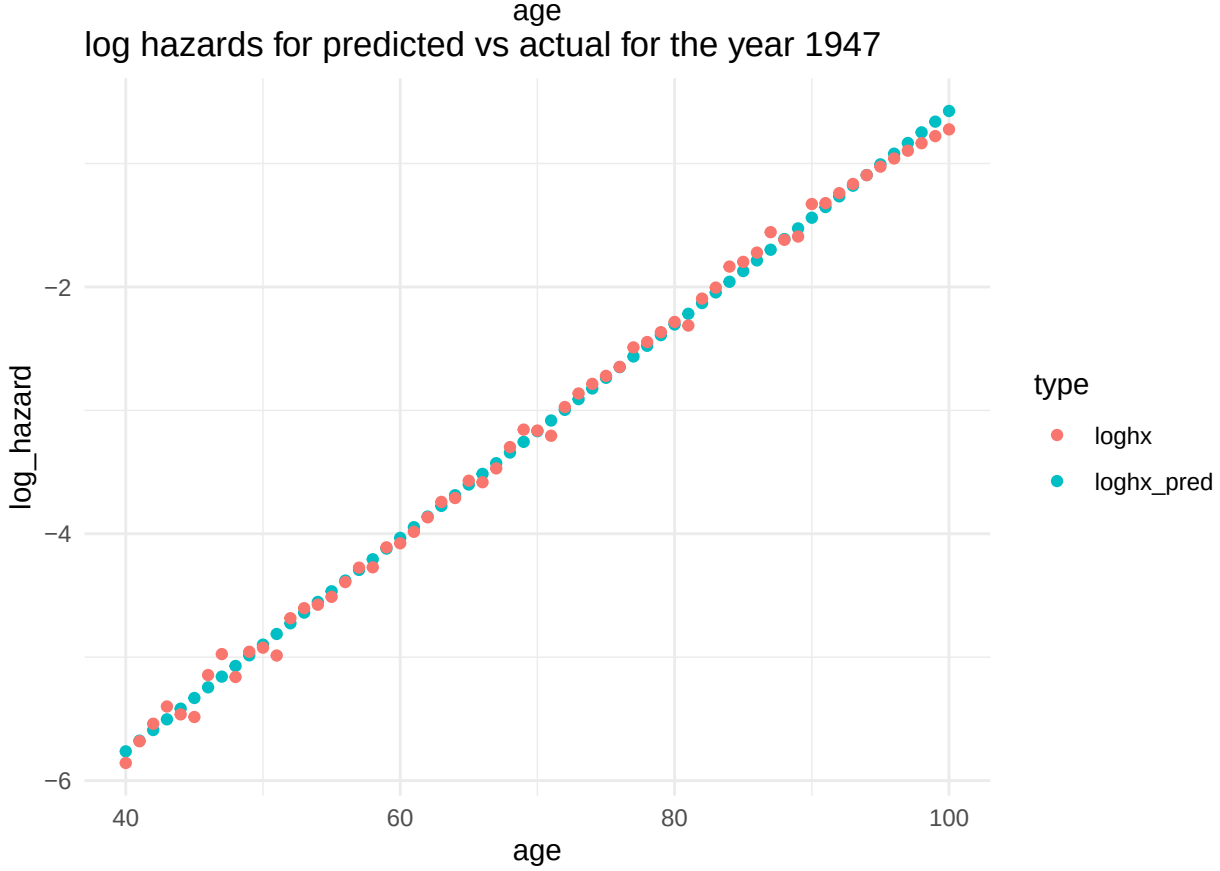

log hazards for predicted vs actual for the year 1942
log_hazard
−1
−2
−3
−4
−5
−6
40
60
80
100
age
type
loghx
loghx_pred

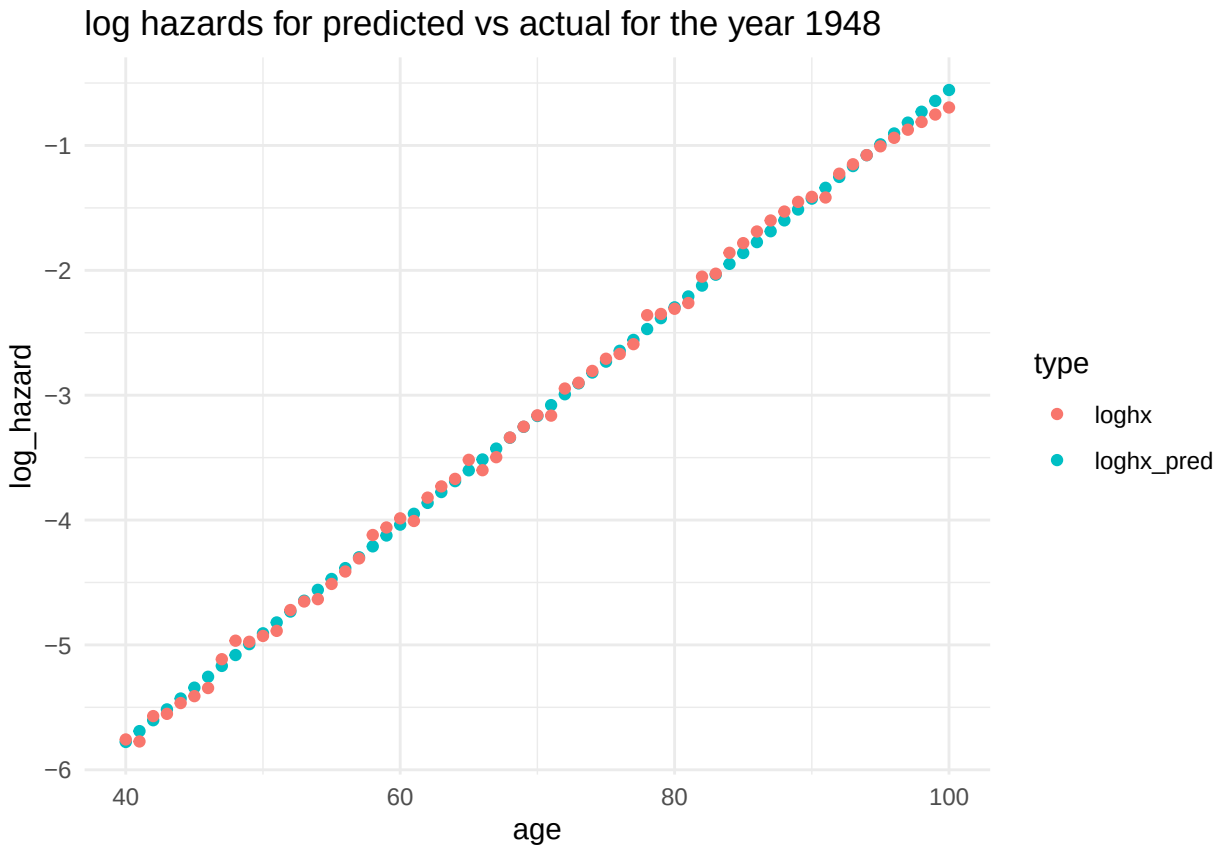

log hazards for predicted vs actual for the year 1943
log_hazard
−1
−2
−3
−4
−5
−6
40
60
80
100
age
type
loghx
loghx_pred

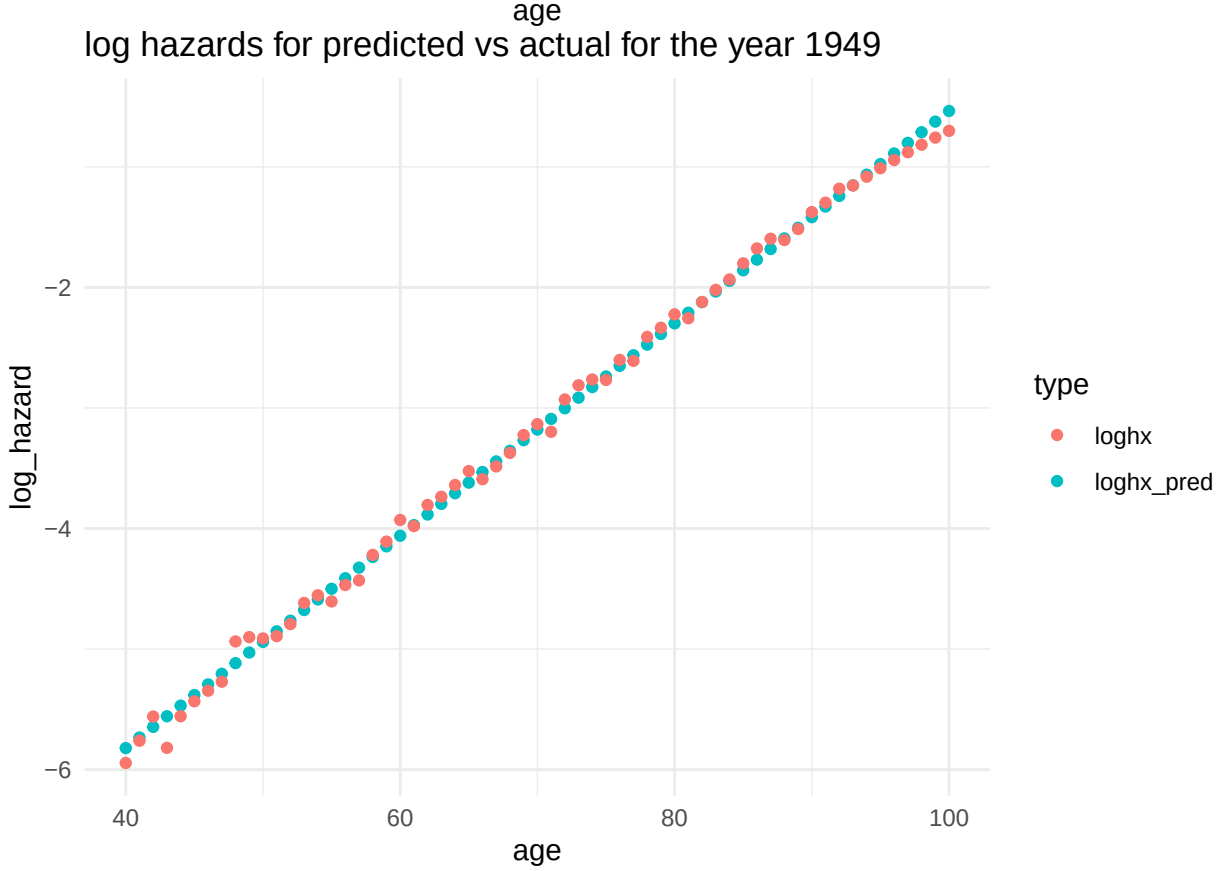

log hazards for predicted vs actual for the year 1944
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

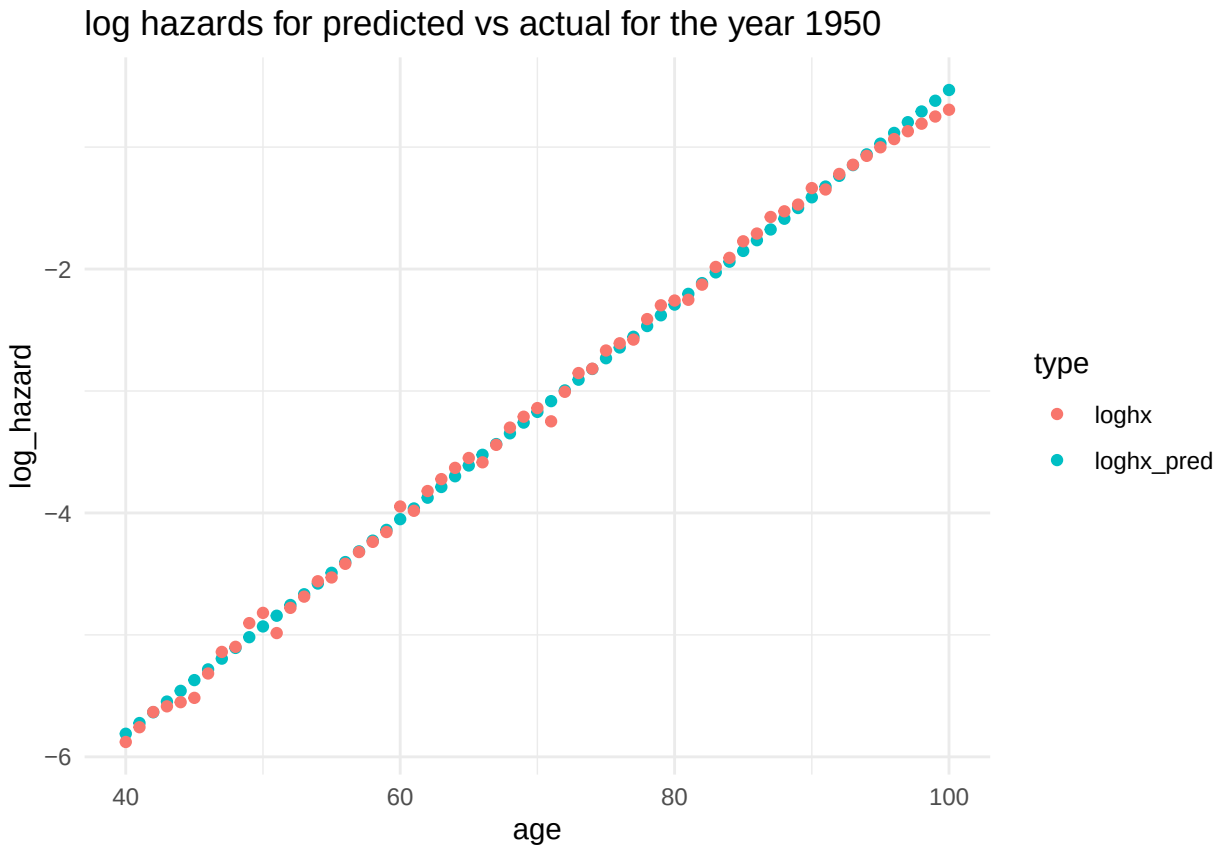

log hazards for predicted vs actual for the year 1945
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

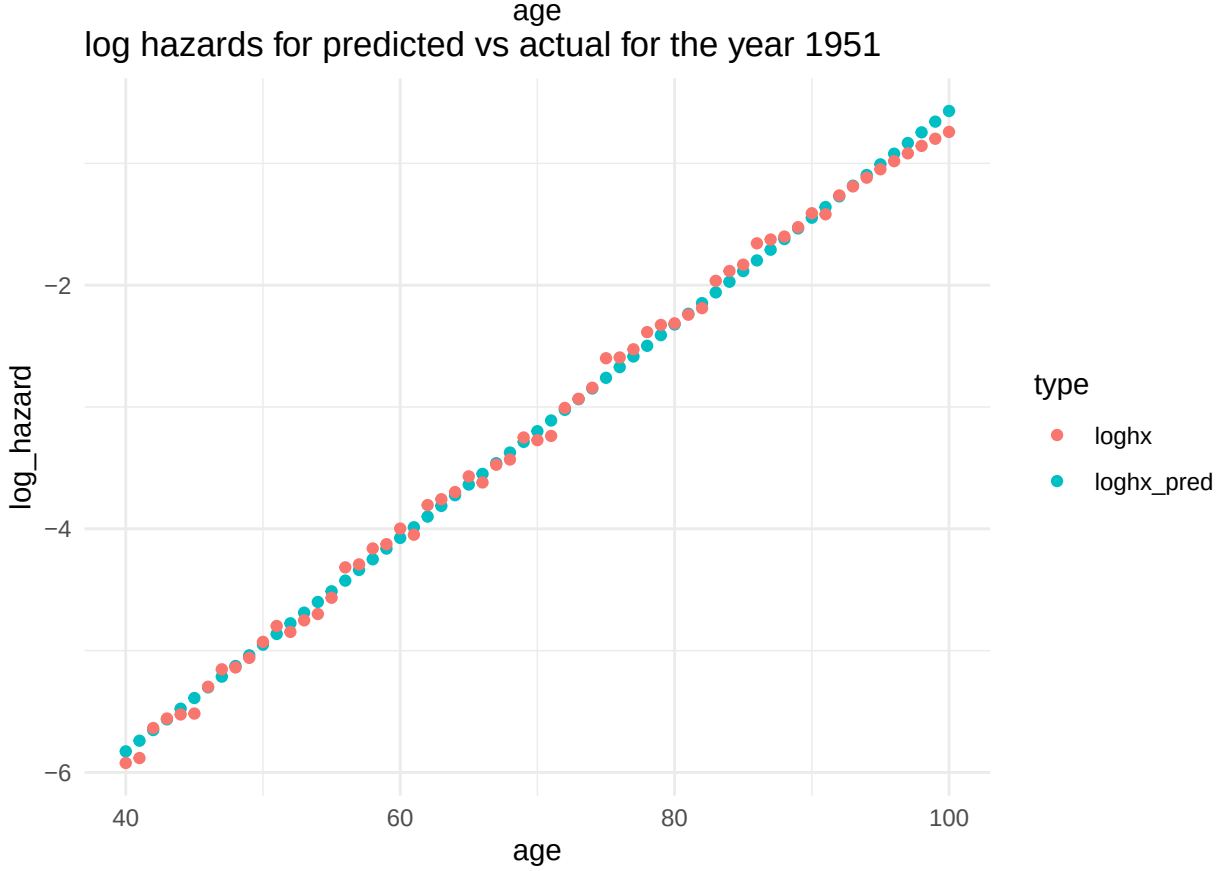

log hazards for predicted vs actual for the year 1946
log_hazard
−1
−2
−3
−4
−5
−6
40
60
80
100
age
type
loghx
loghx_pred

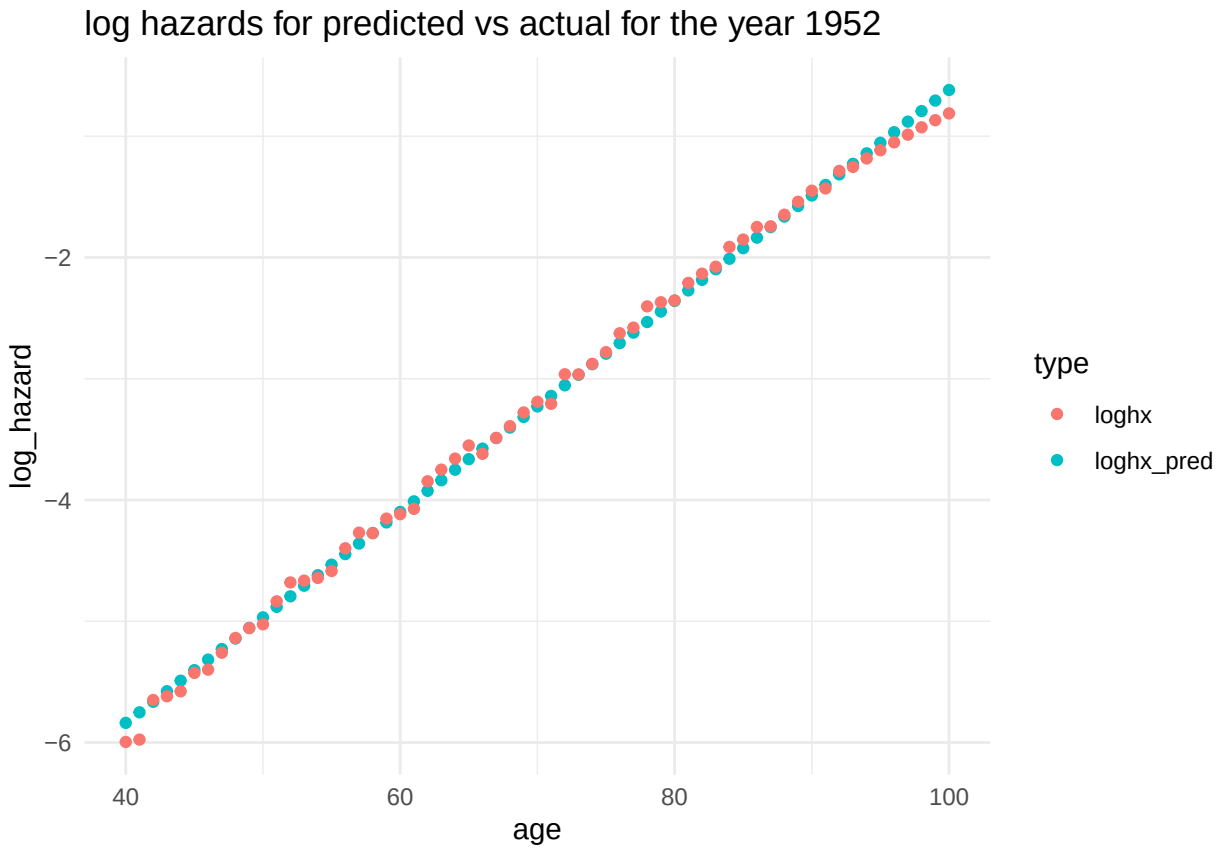

log hazards for predicted vs actual for the year 1947
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

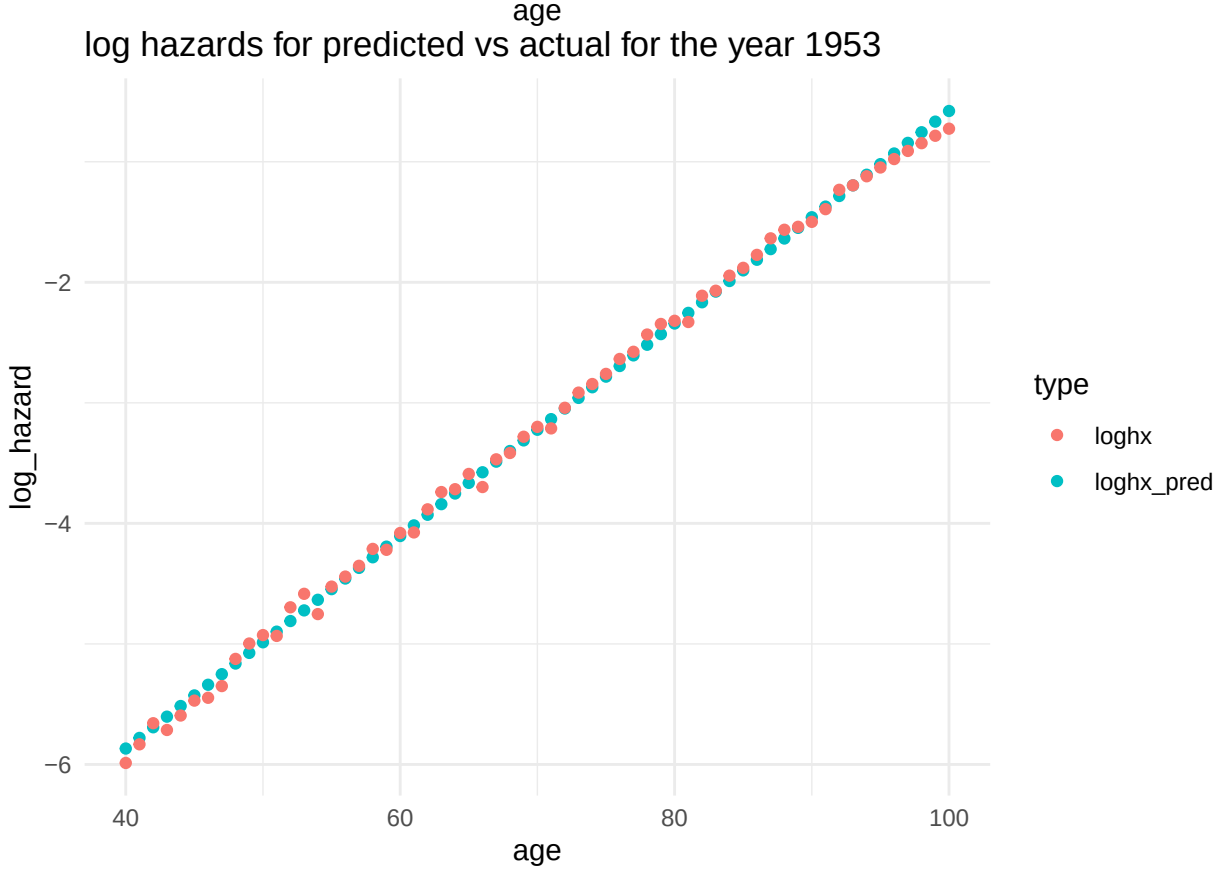

log hazards for predicted vs actual for the year 1948
log_hazard
−1
−2
−3
−4
−5
−6
40
60
80
100
age
type
loghx
loghx_pred

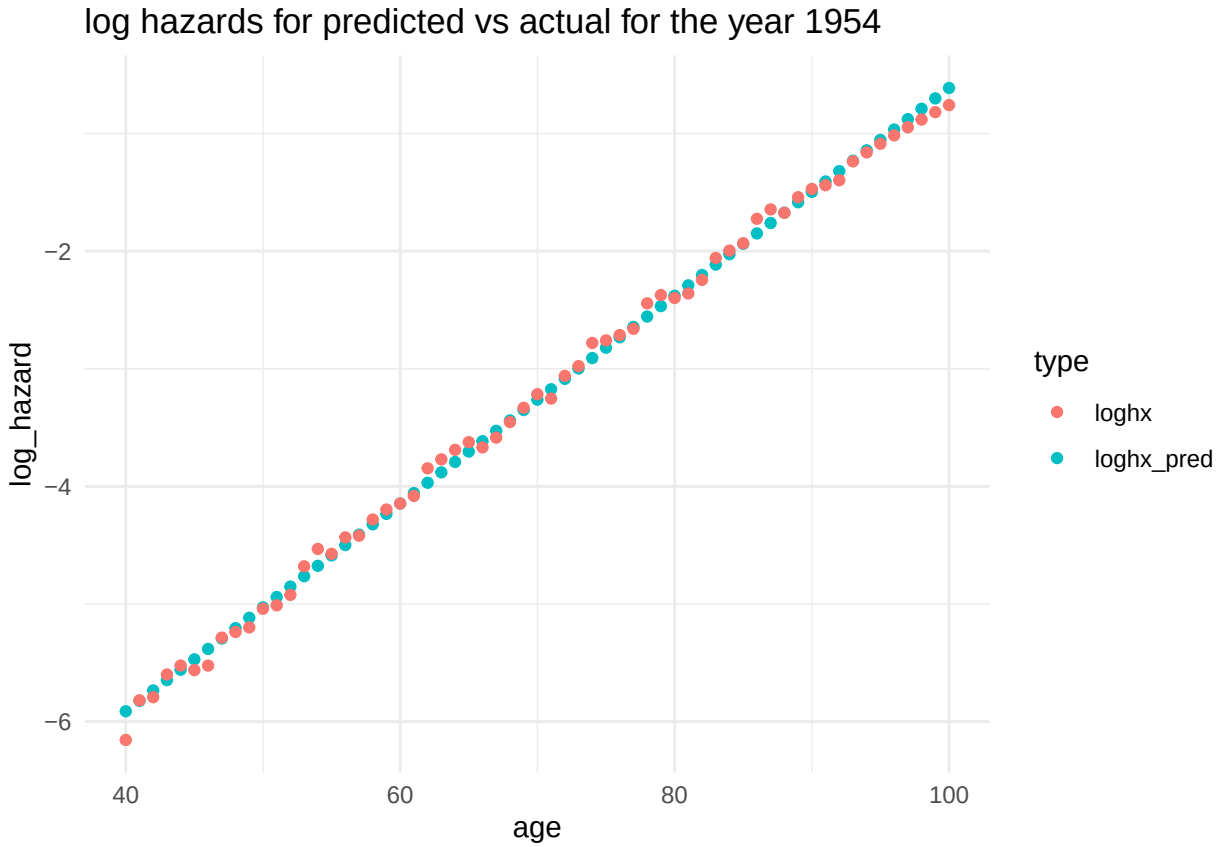

log hazards for predicted vs actual for the year 1949
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

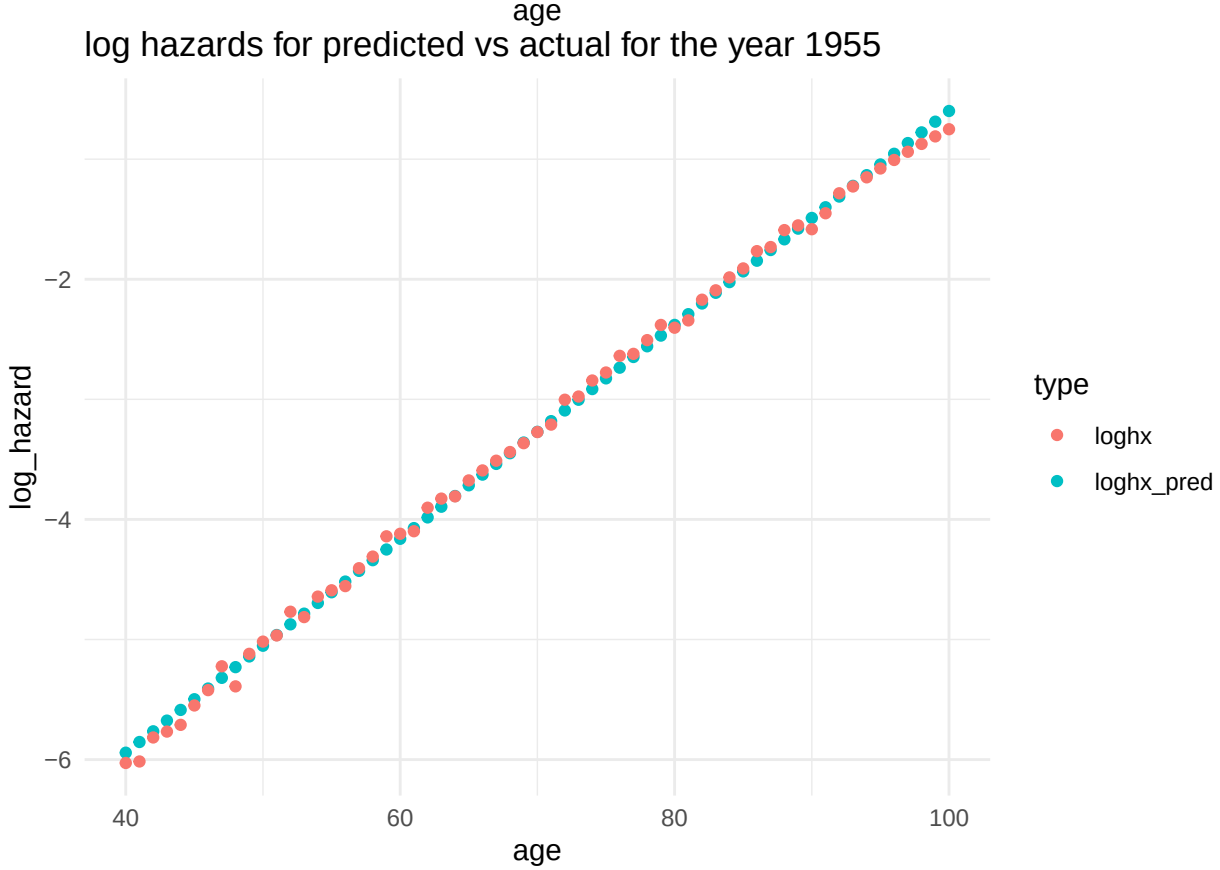

log hazards for predicted vs actual for the year 1950
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

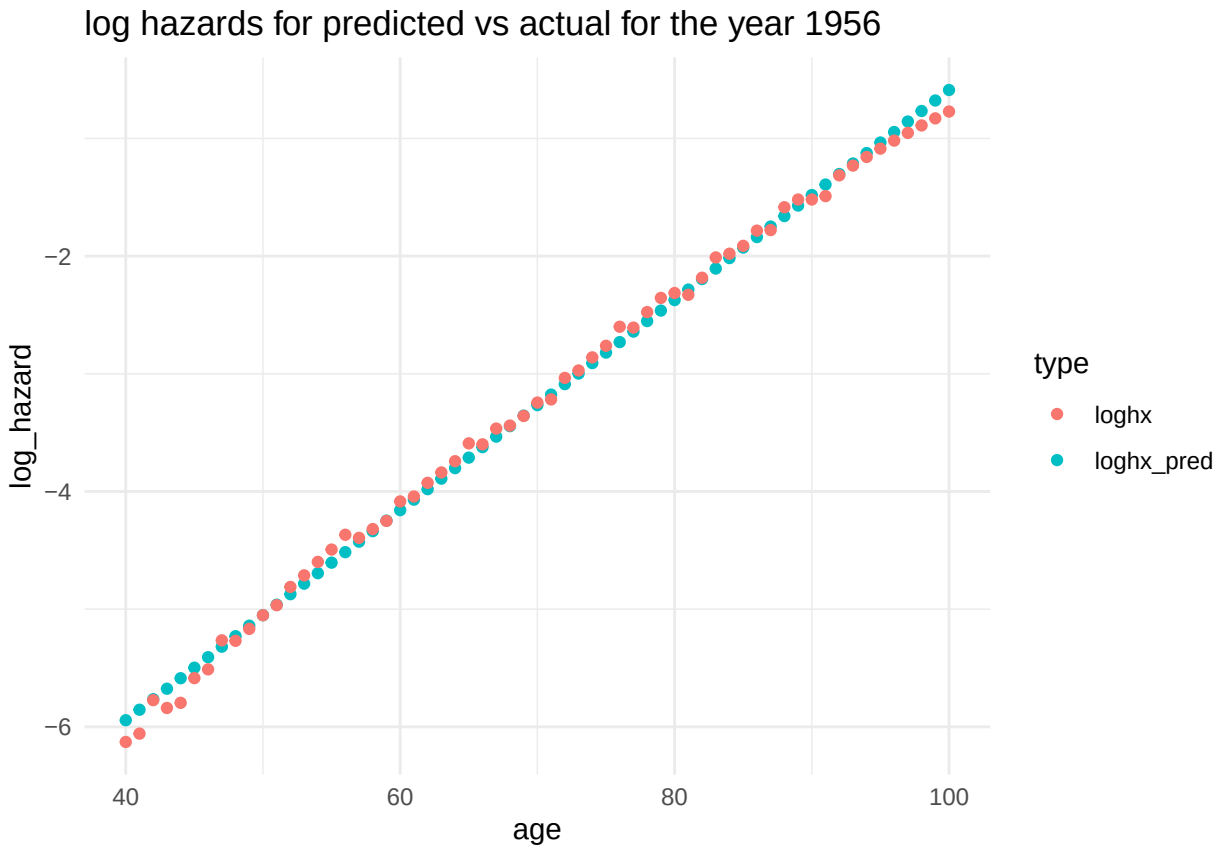

log hazards for predicted vs actual for the year 1951
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

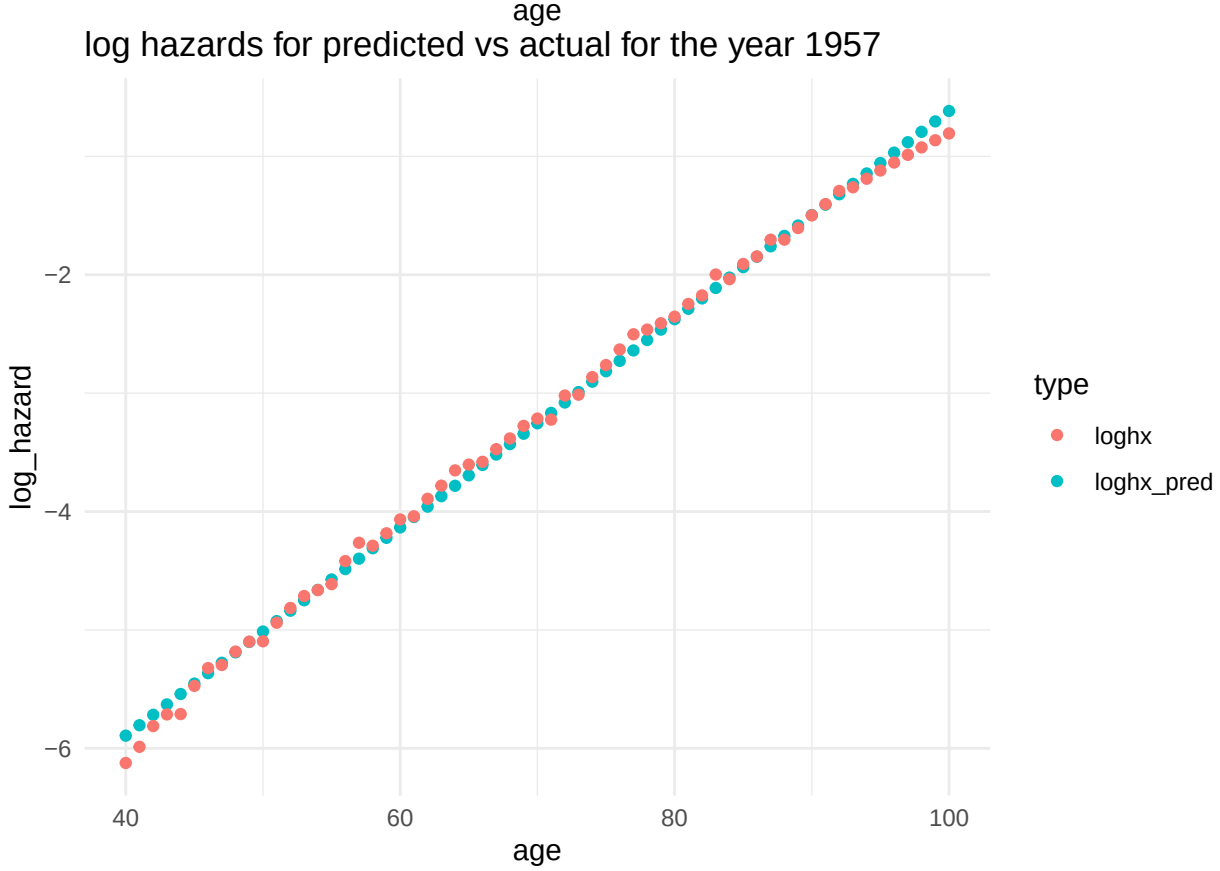

log hazards for predicted vs actual for the year 1952
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

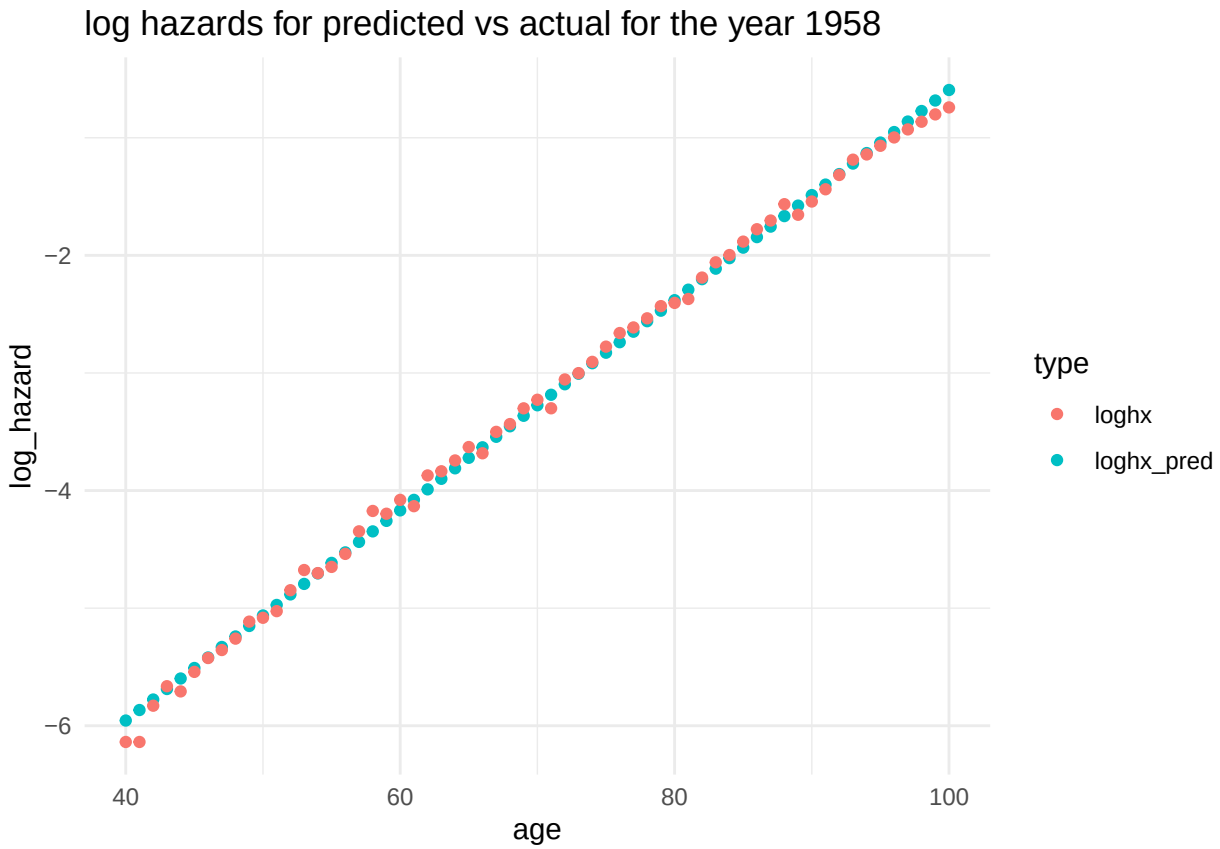

log hazards for predicted vs actual for the year 1953
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

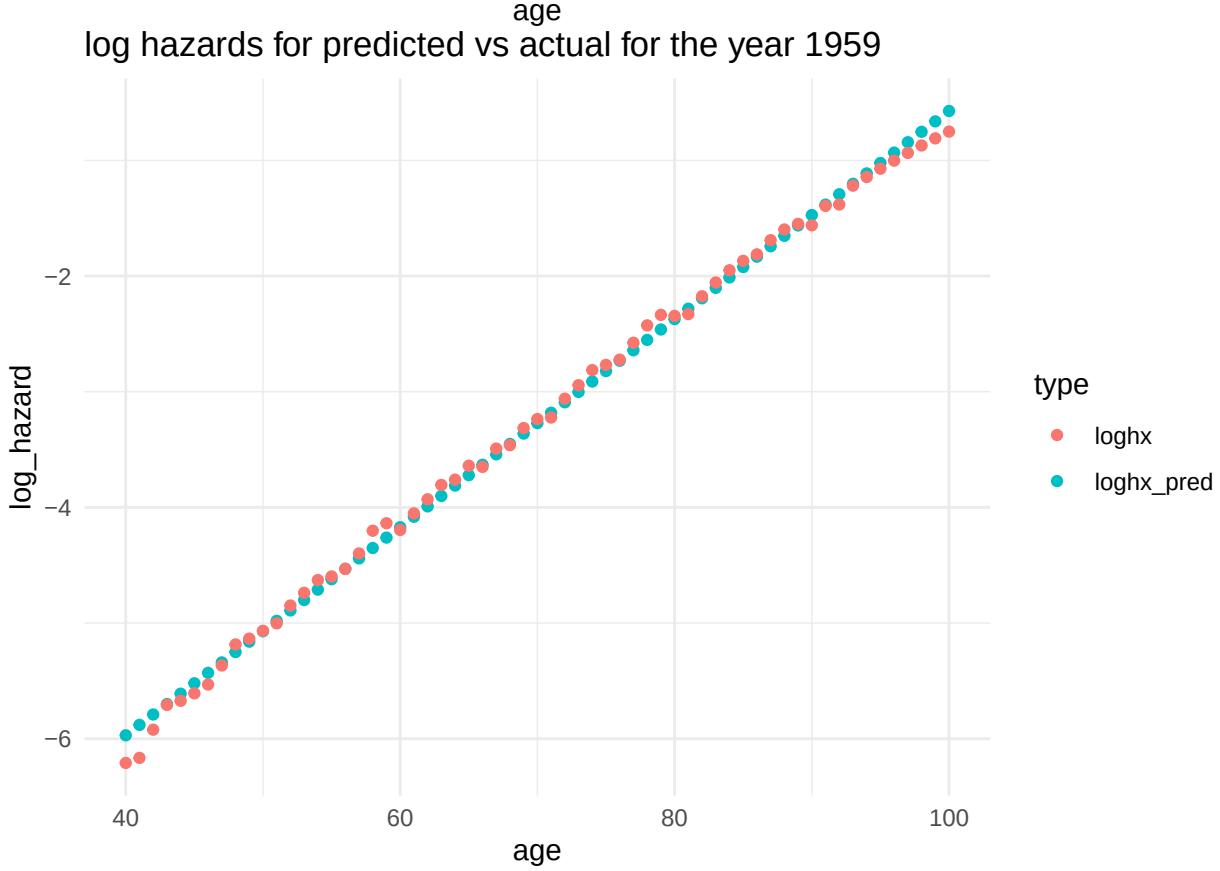

log hazards for predicted vs actual for the year 1954
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

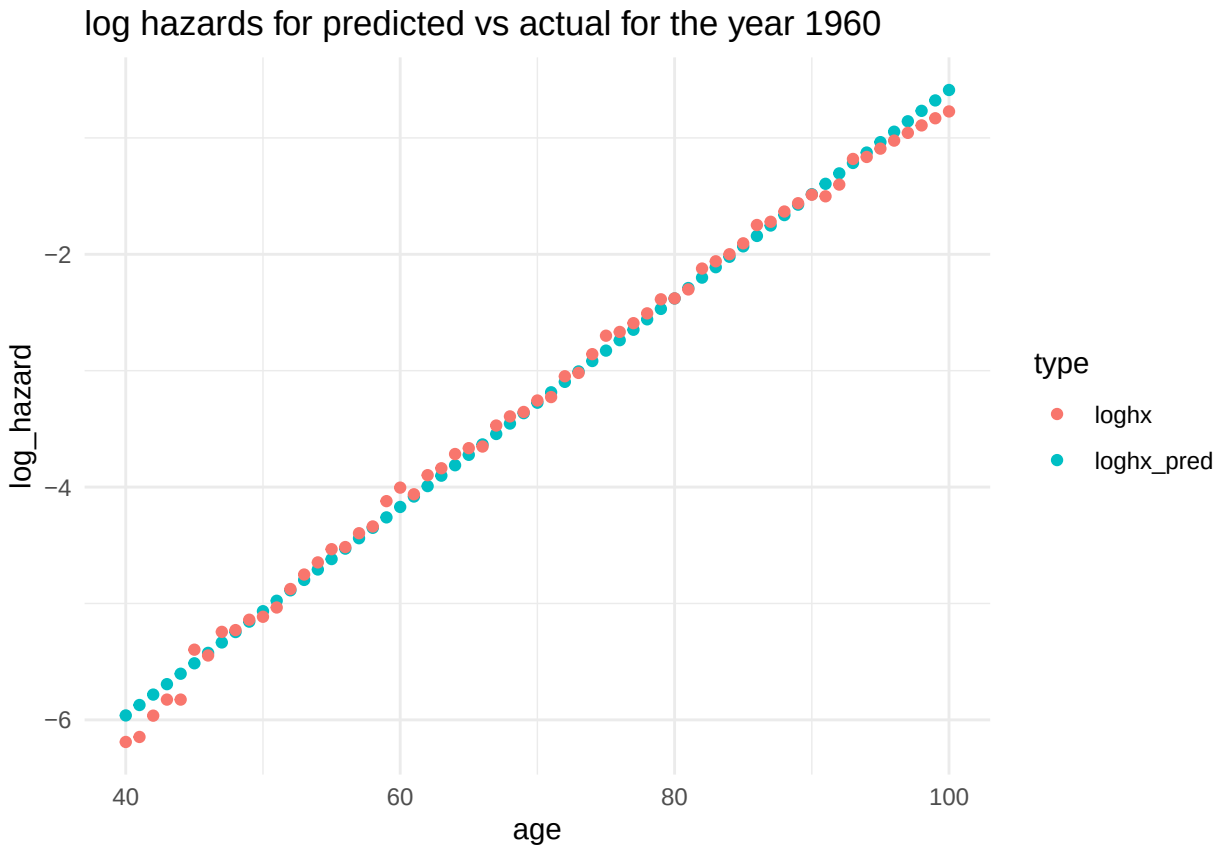

log hazards for predicted vs actual for the year 1955
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

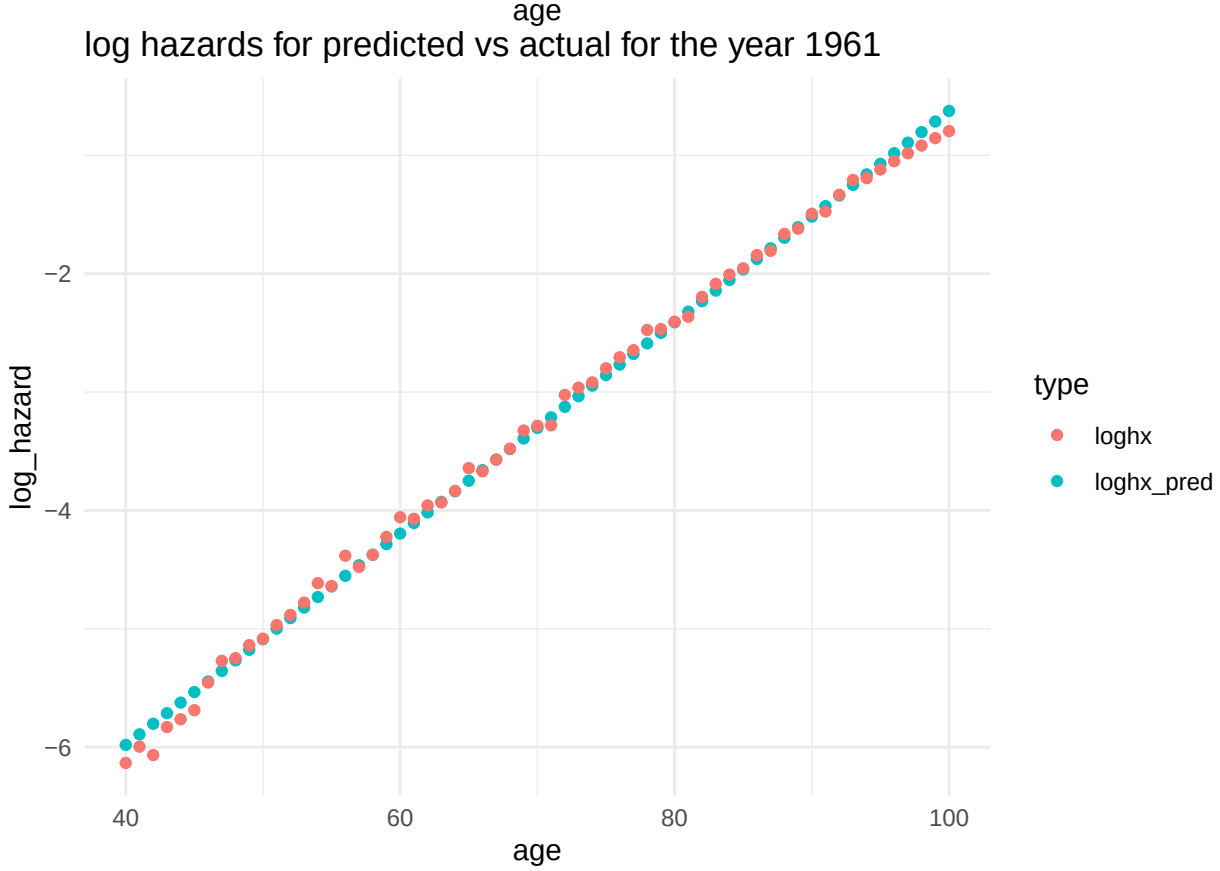

log hazards for predicted vs actual for the year 1956
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

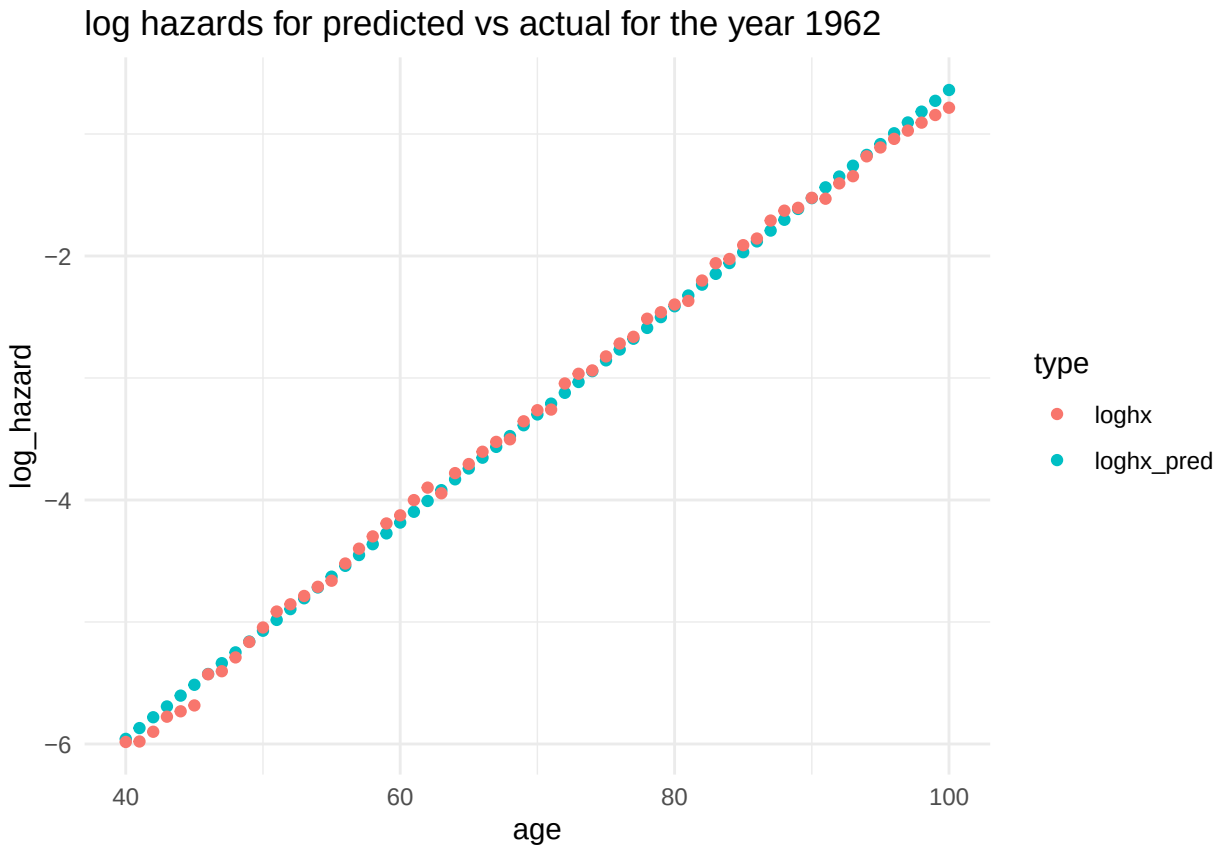

log hazards for predicted vs actual for the year 1957
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

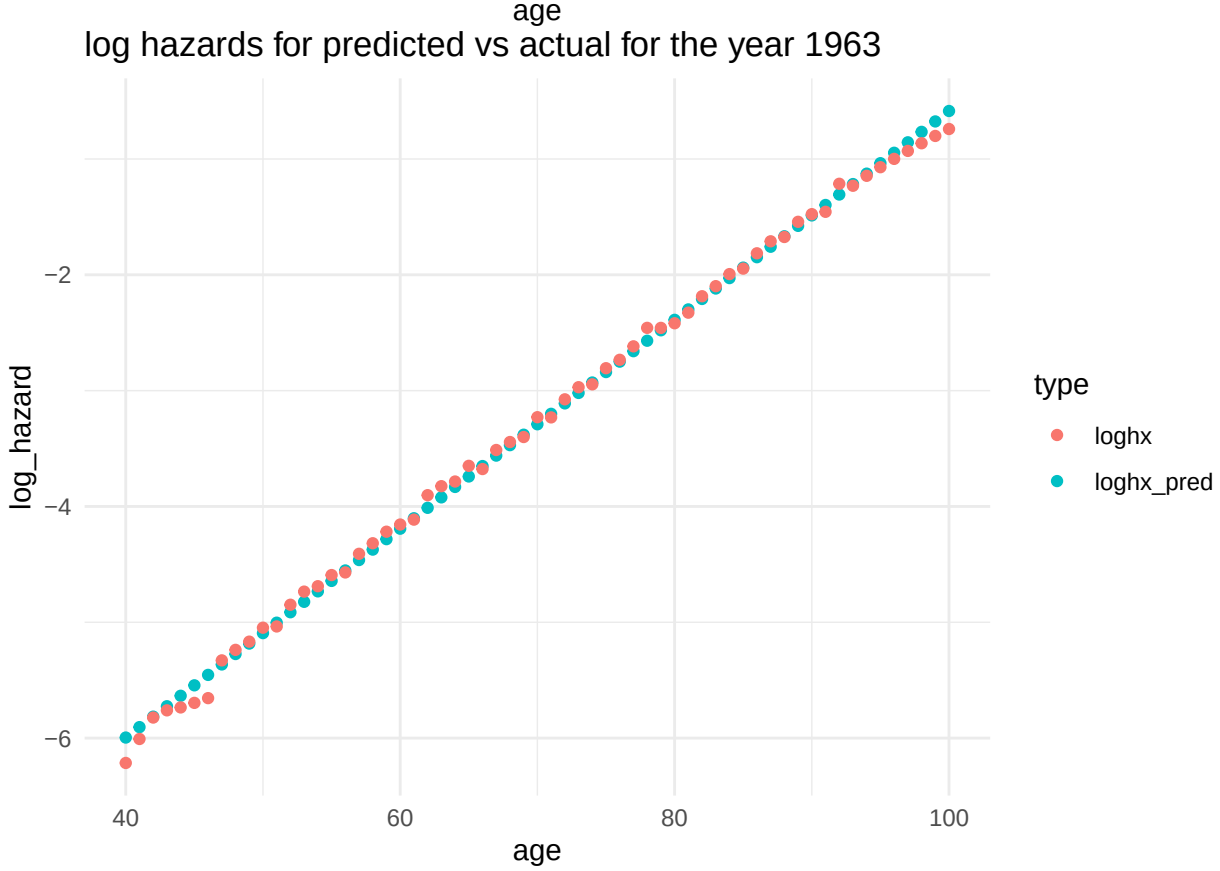

log hazards for predicted vs actual for the year 1958
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

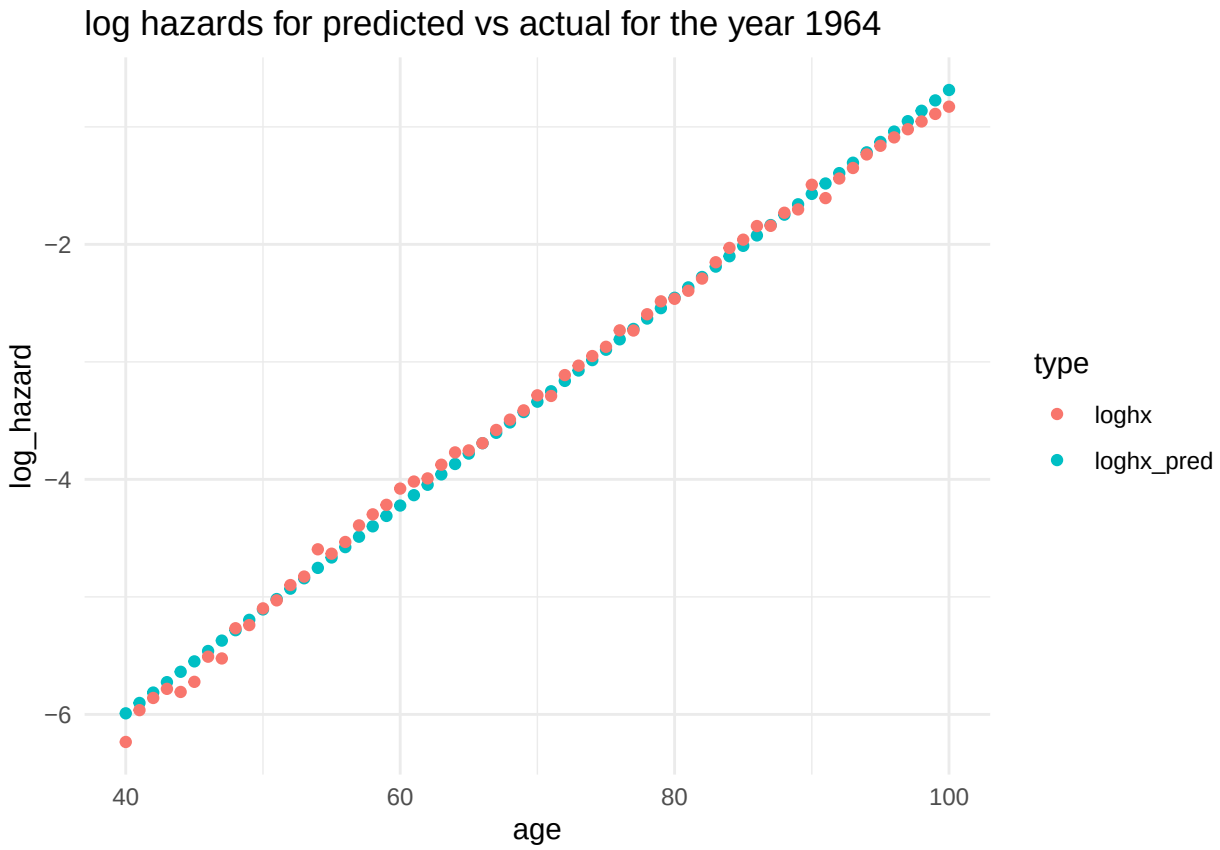

log hazards for predicted vs actual for the year 1959
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

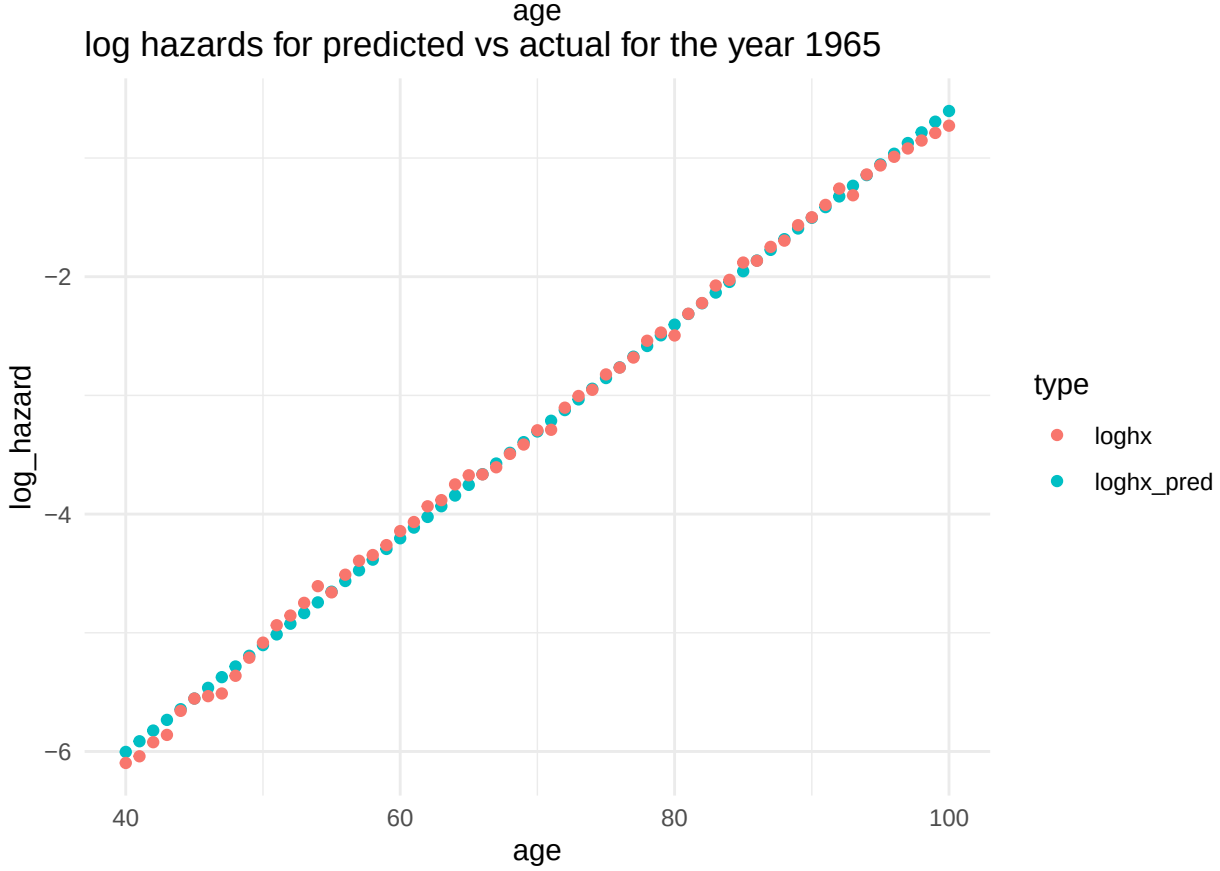

log hazards for predicted vs actual for the year 1960
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

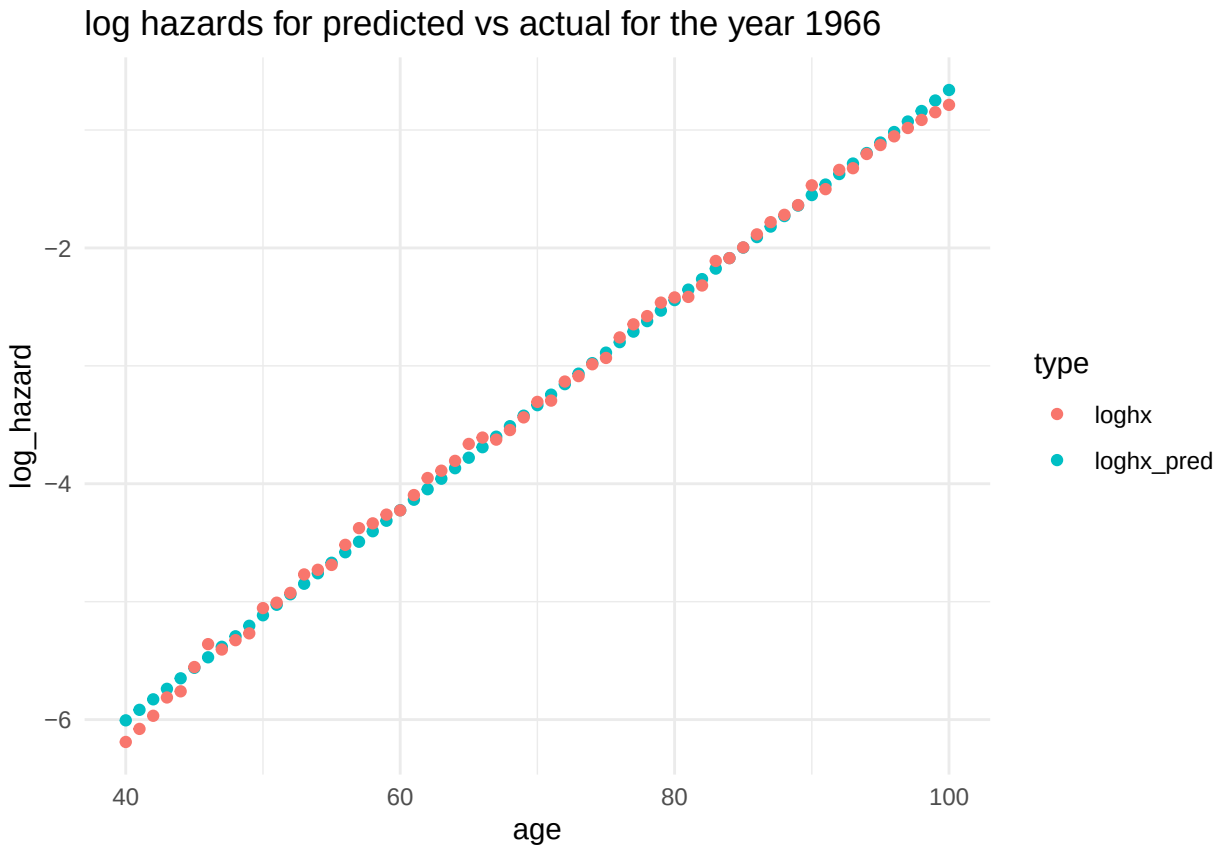

log hazards for predicted vs actual for the year 1961
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

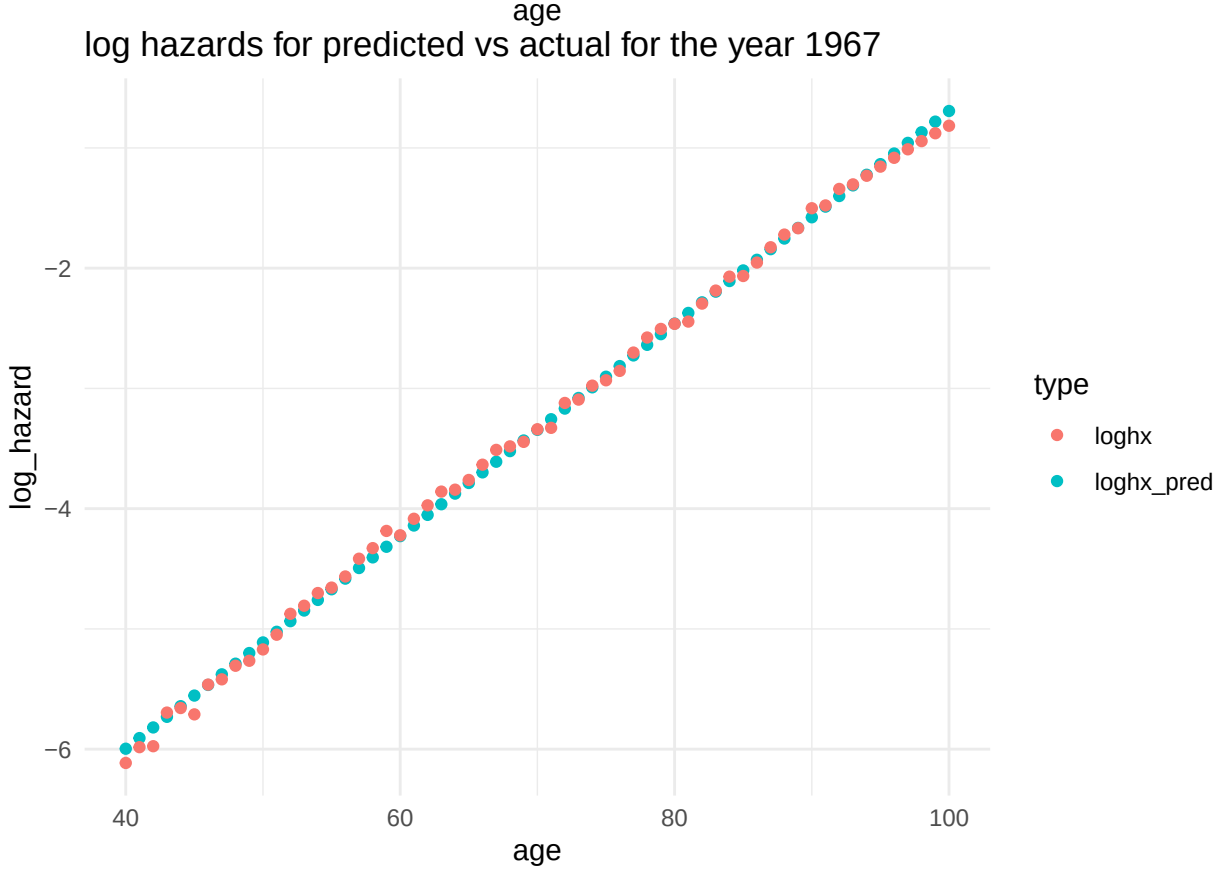

log hazards for predicted vs actual for the year 1962
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

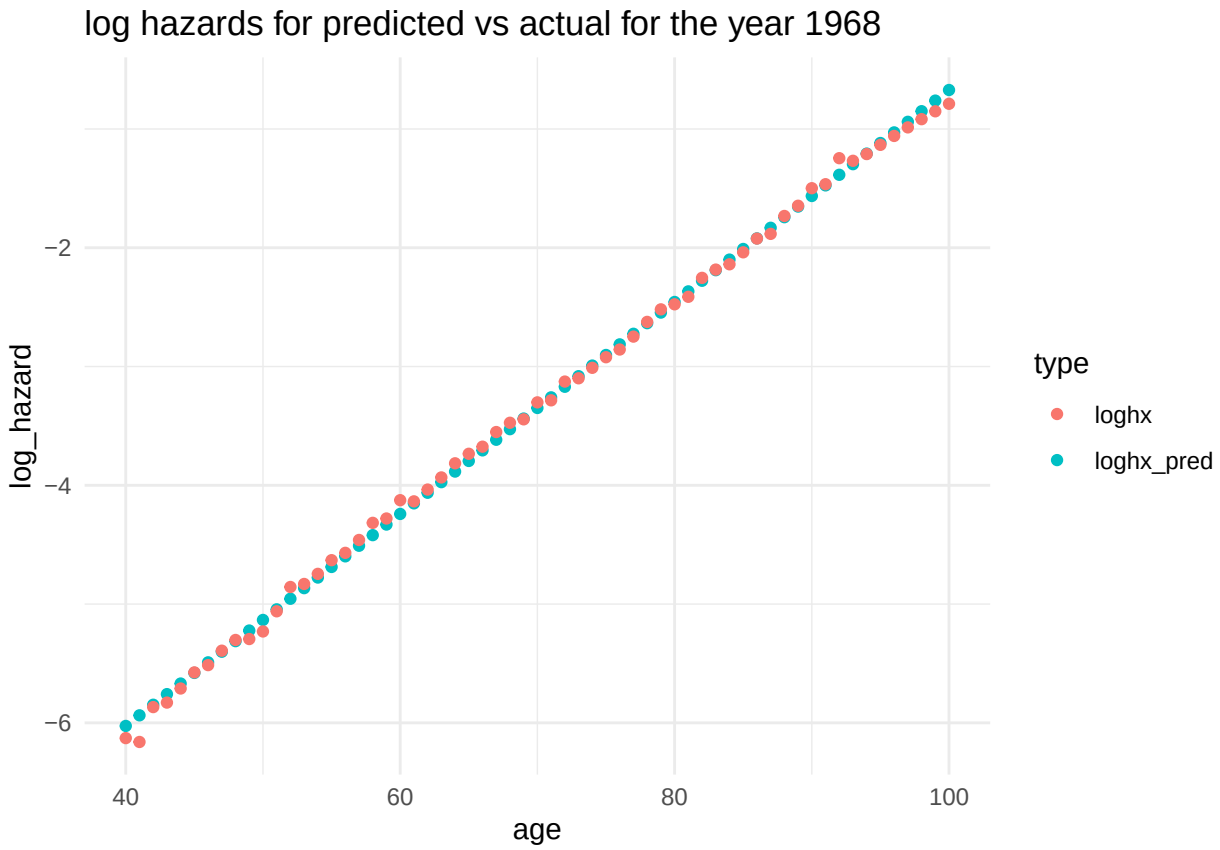

log hazards for predicted vs actual for the year 1963
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

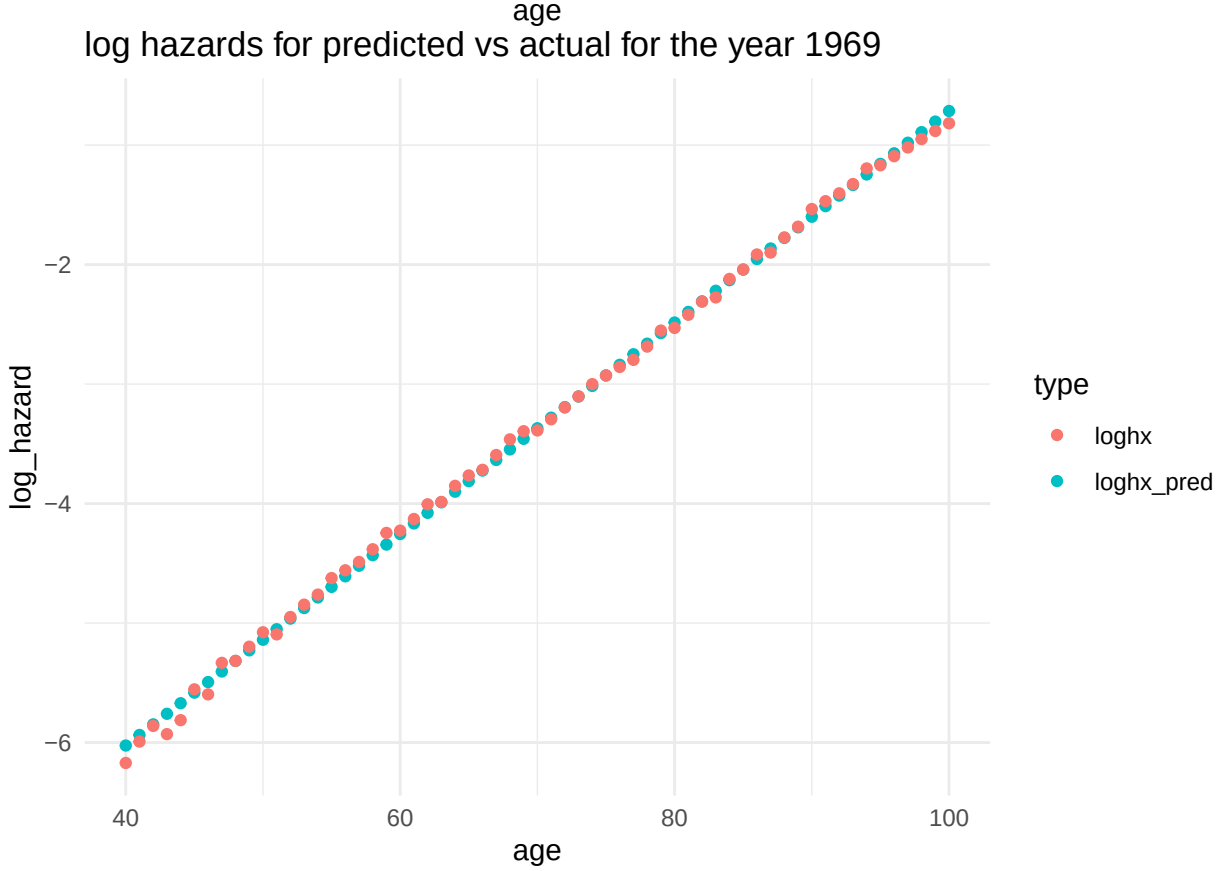

log hazards for predicted vs actual for the year 1964
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

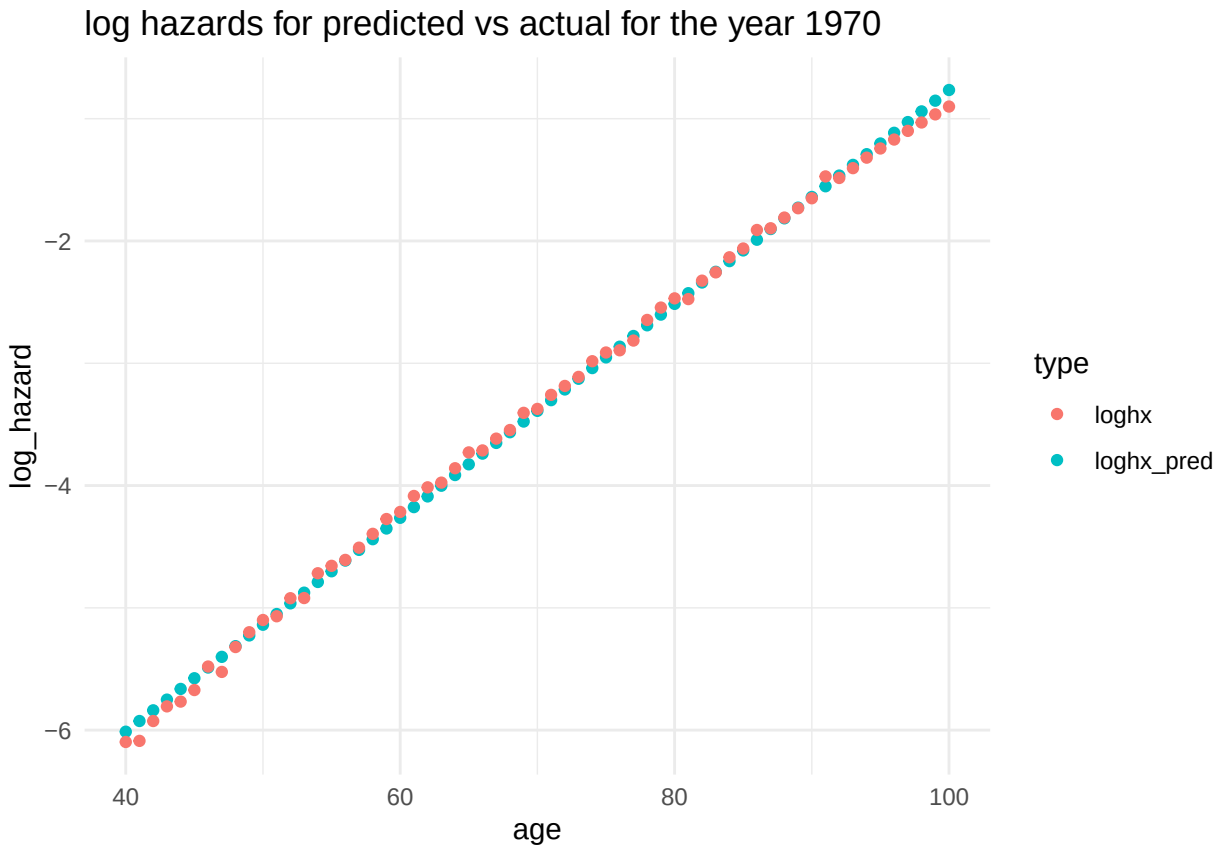

log hazards for predicted vs actual for the year 1965
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

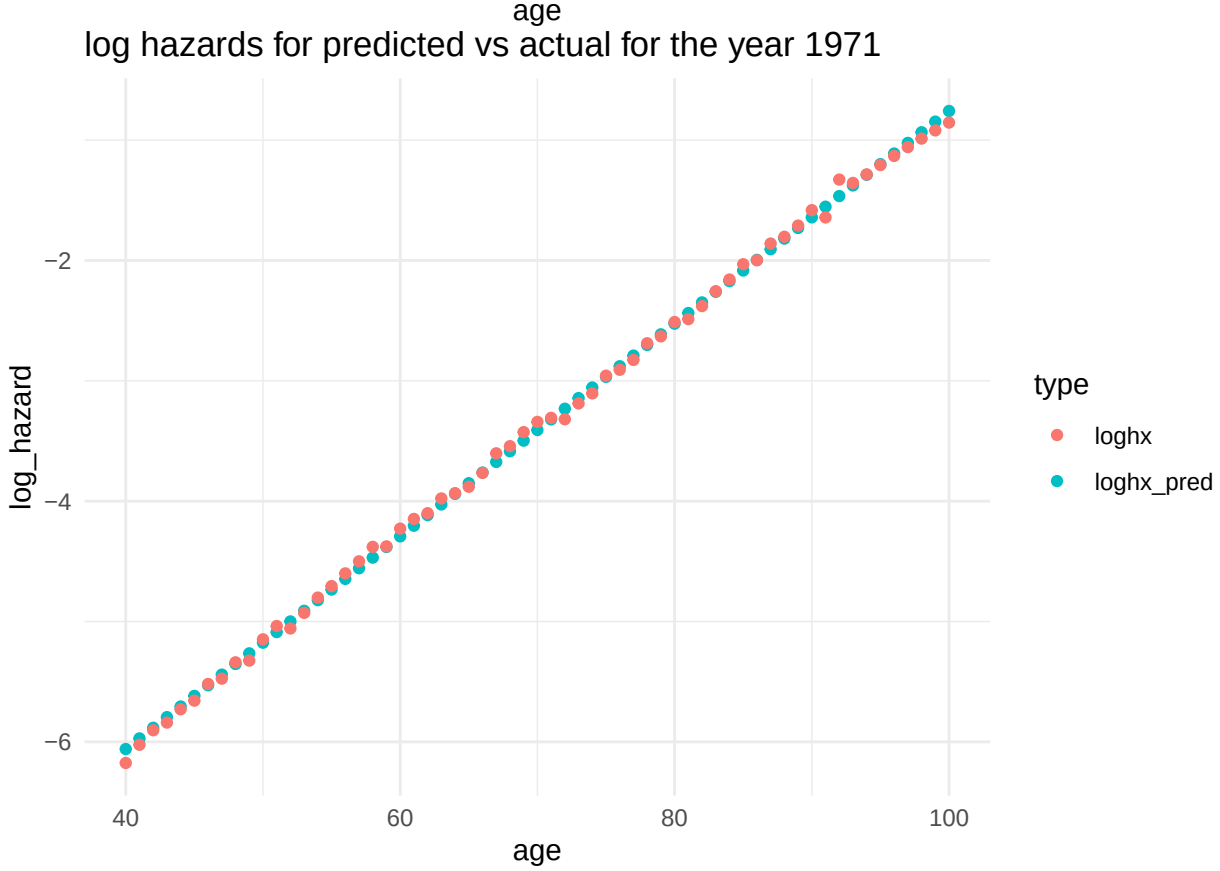

log hazards for predicted vs actual for the year 1966
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

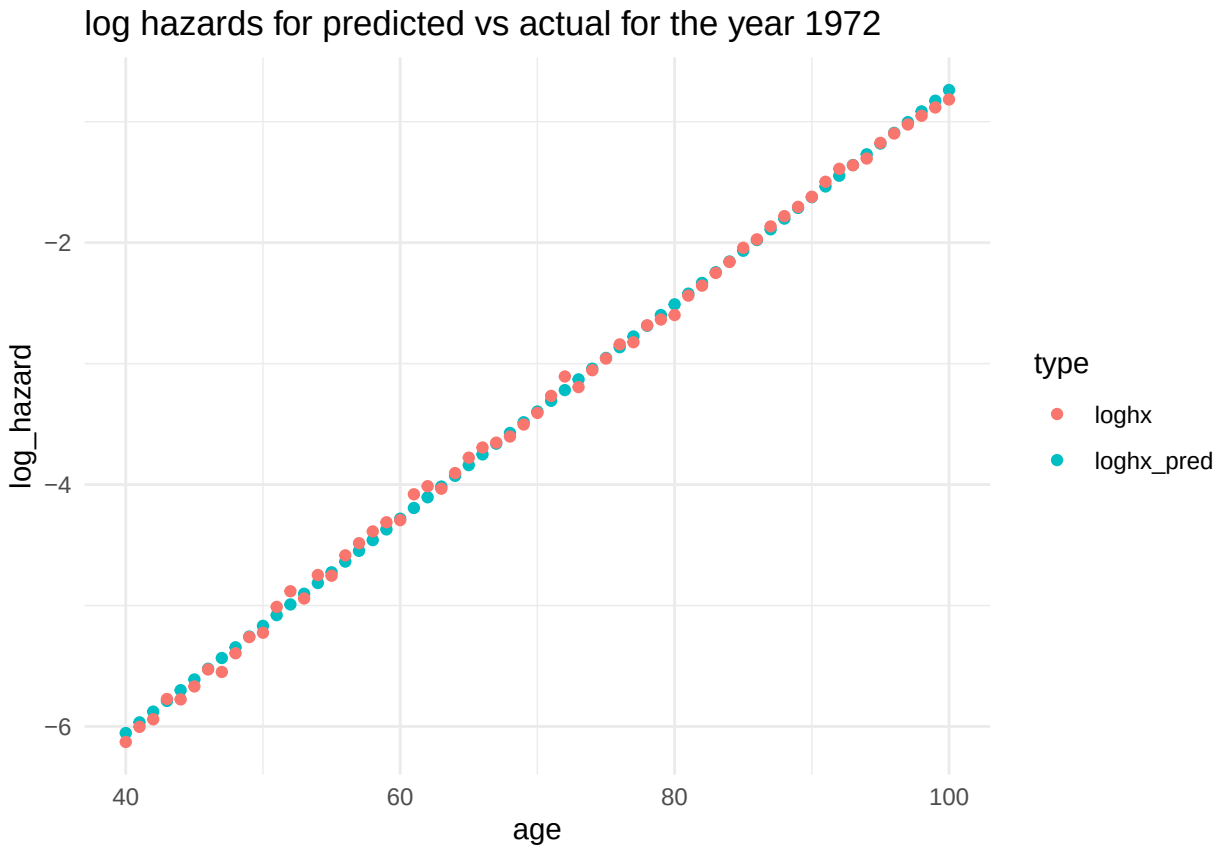

log hazards for predicted vs actual for the year 1967
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

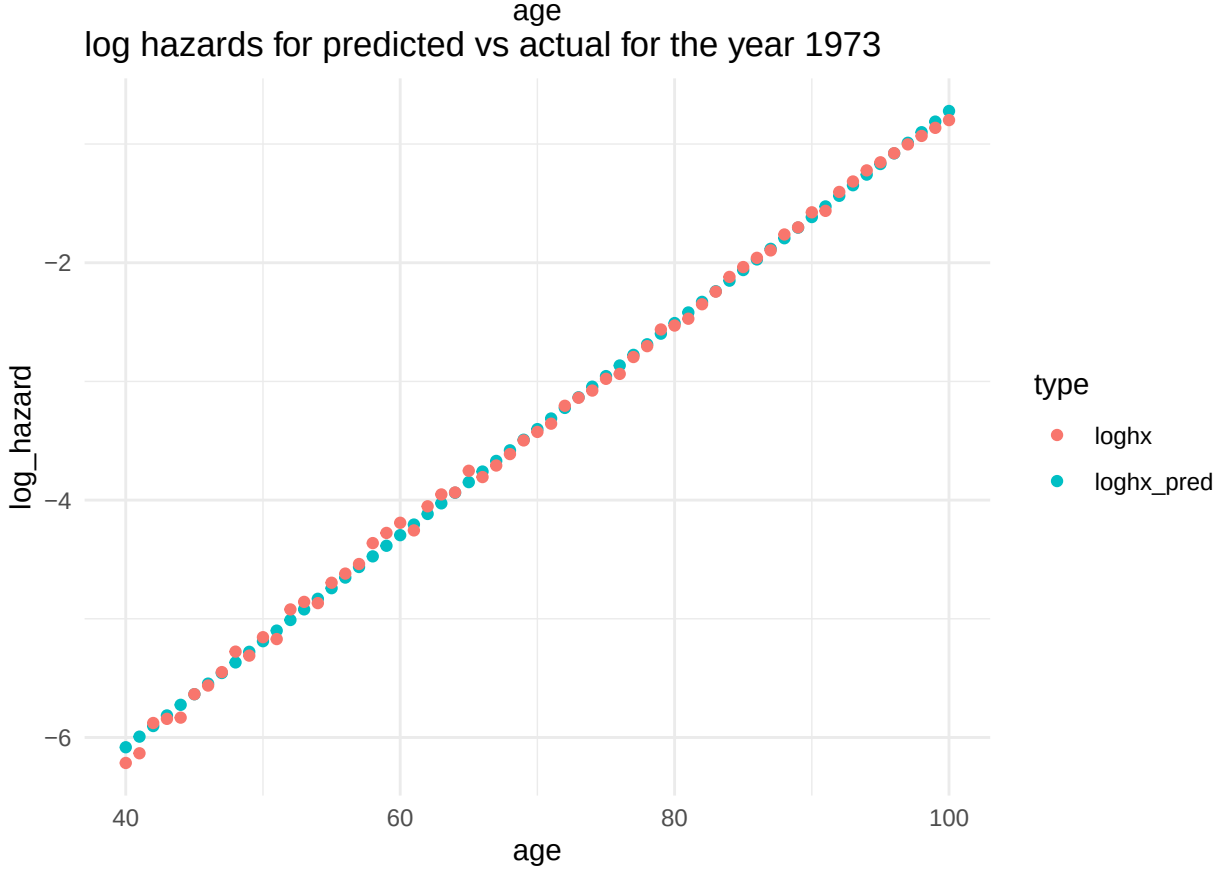

log hazards for predicted vs actual for the year 1968
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

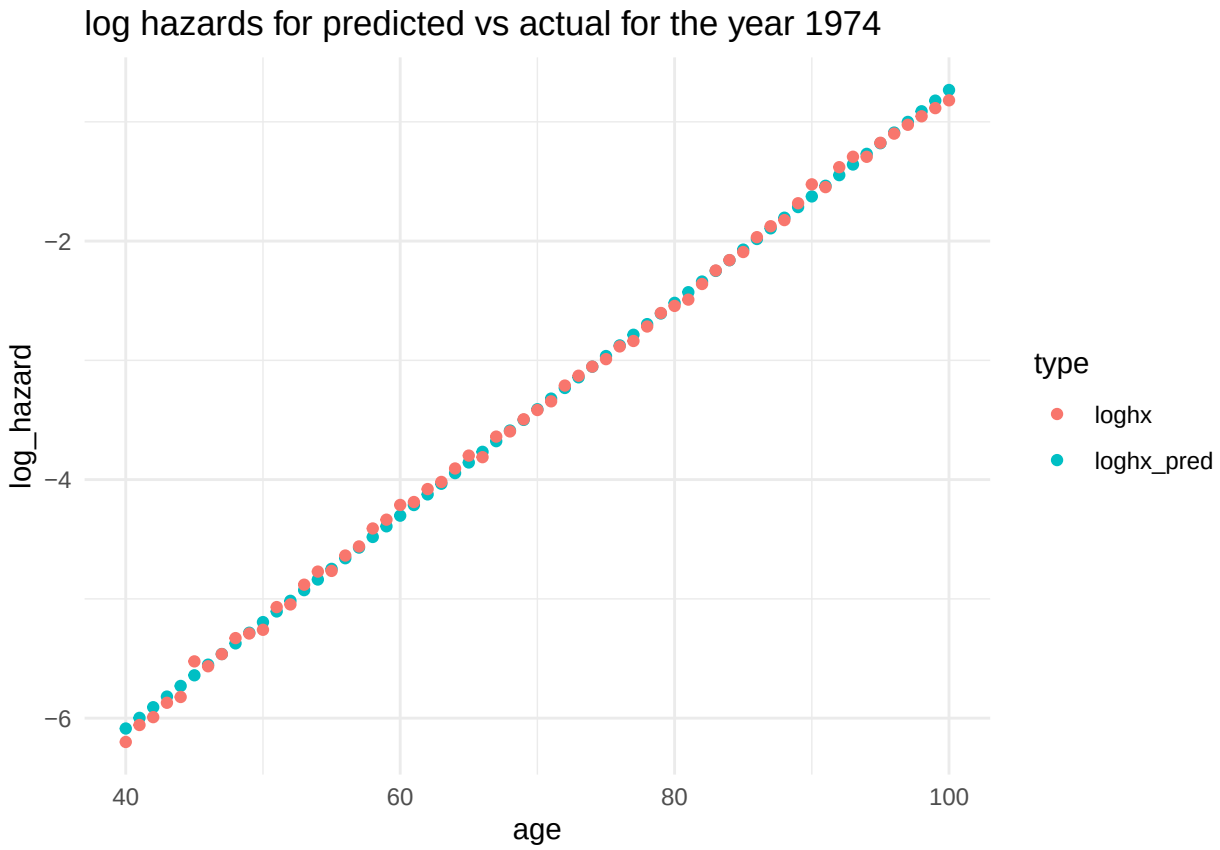

log hazards for predicted vs actual for the year 1969
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

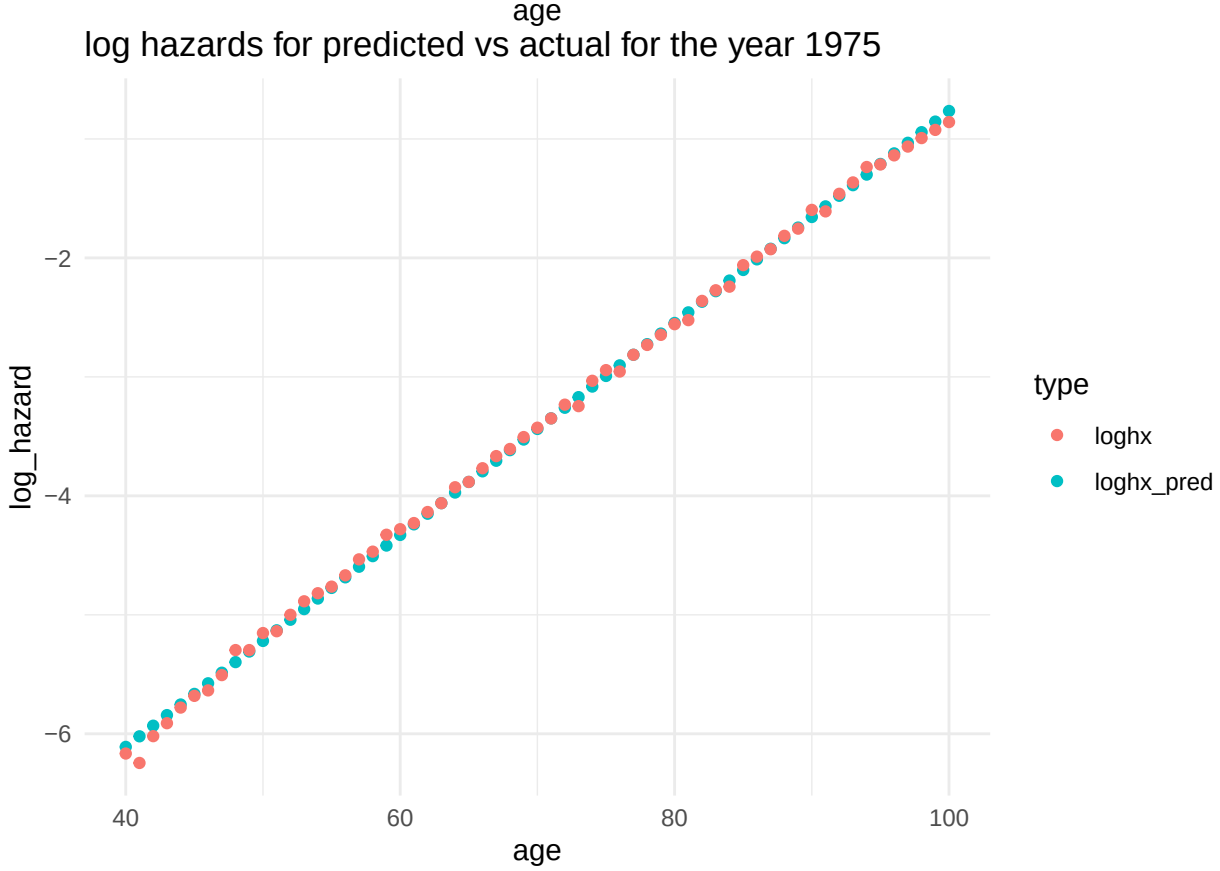

log hazards for predicted vs actual for the year 1970
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

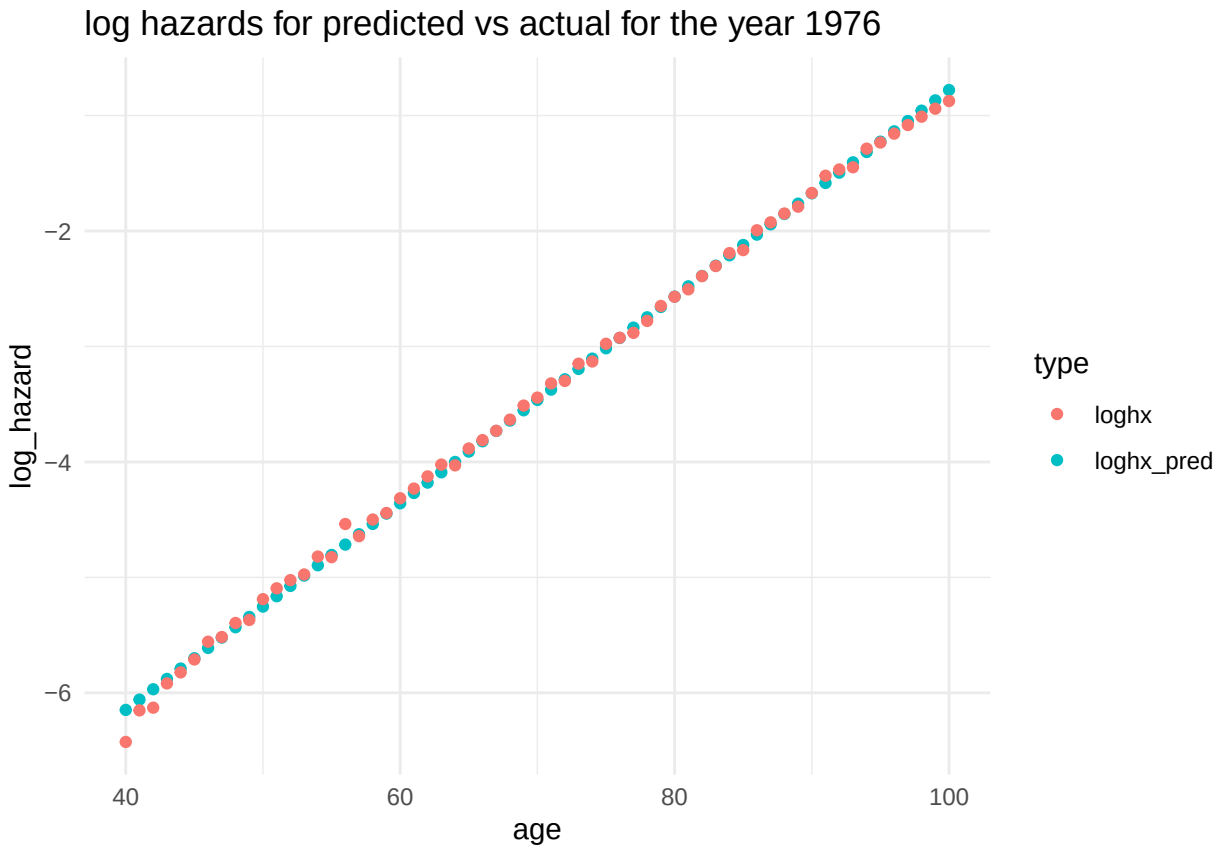

log hazards for predicted vs actual for the year 1971
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

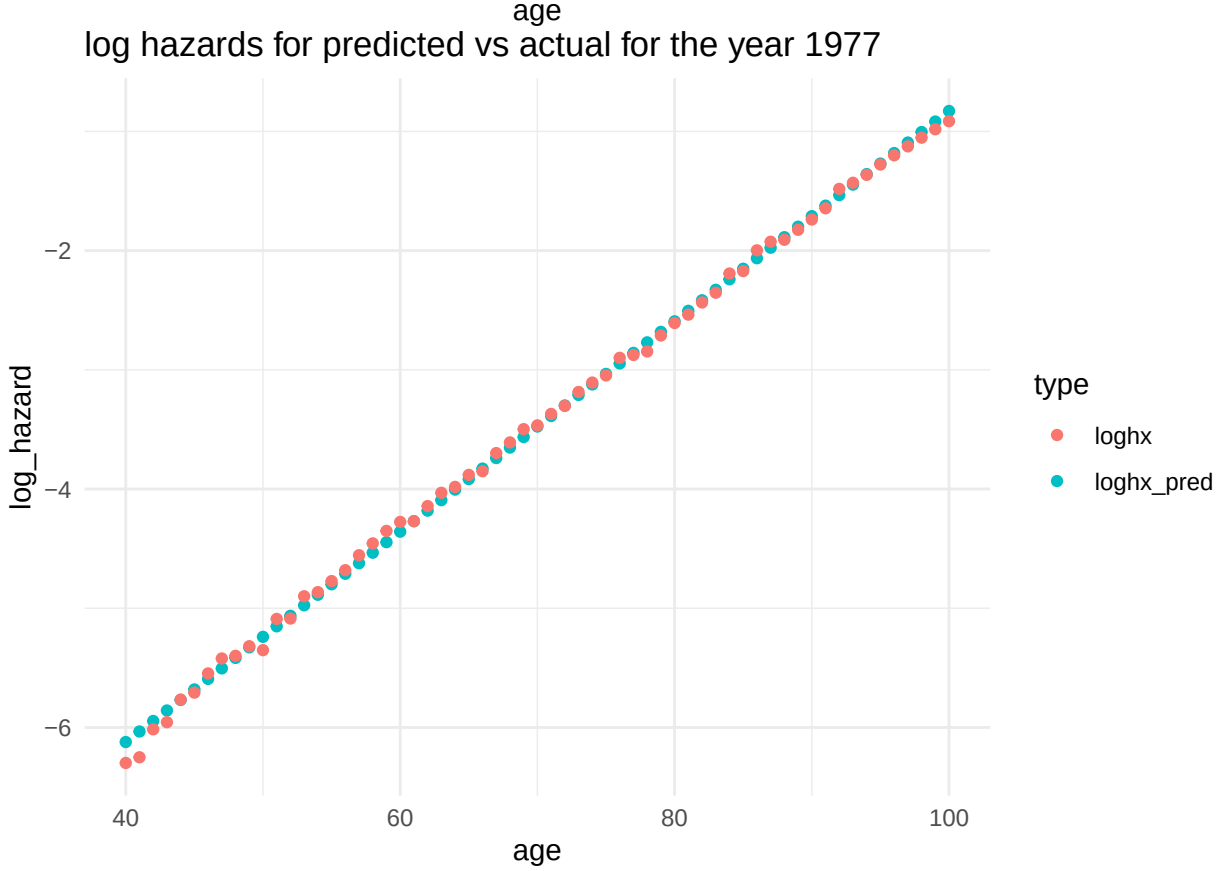

log hazards for predicted vs actual for the year 1972
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

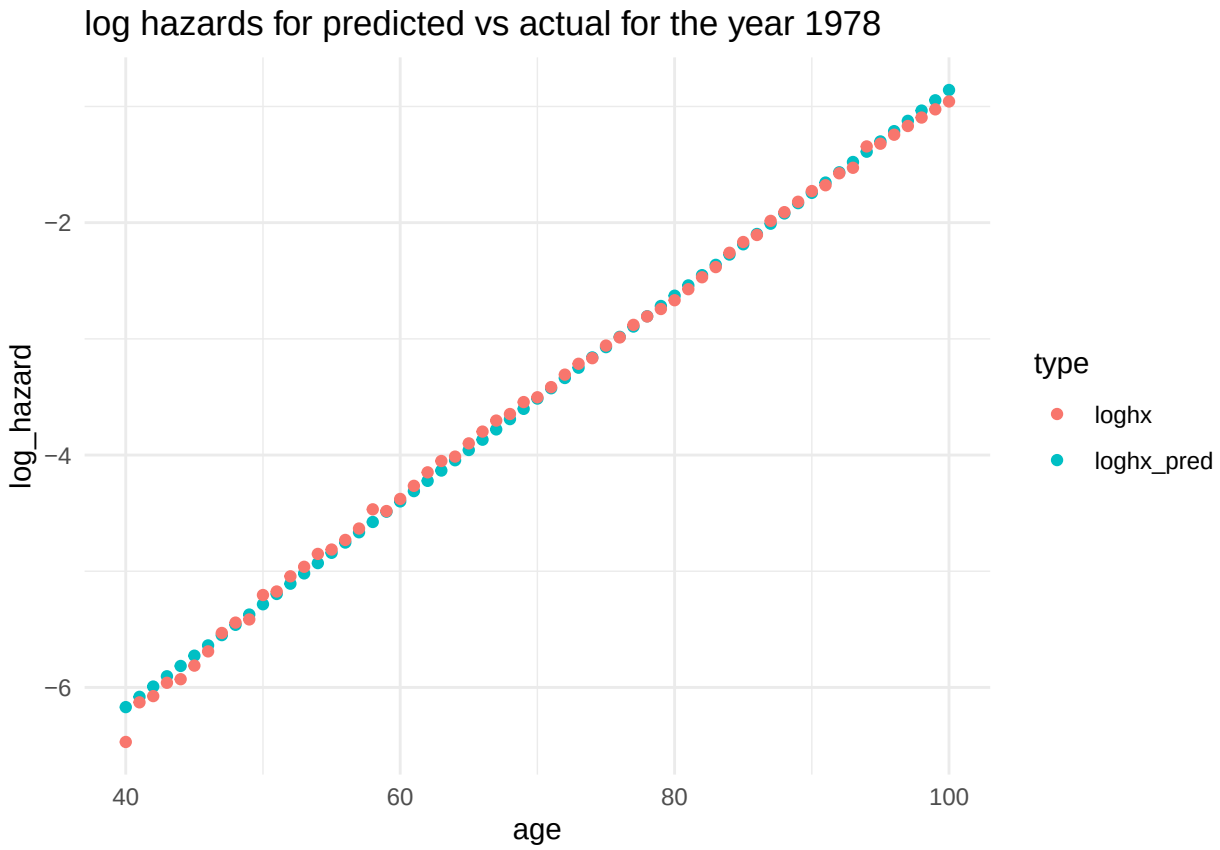

log hazards for predicted vs actual for the year 1973
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

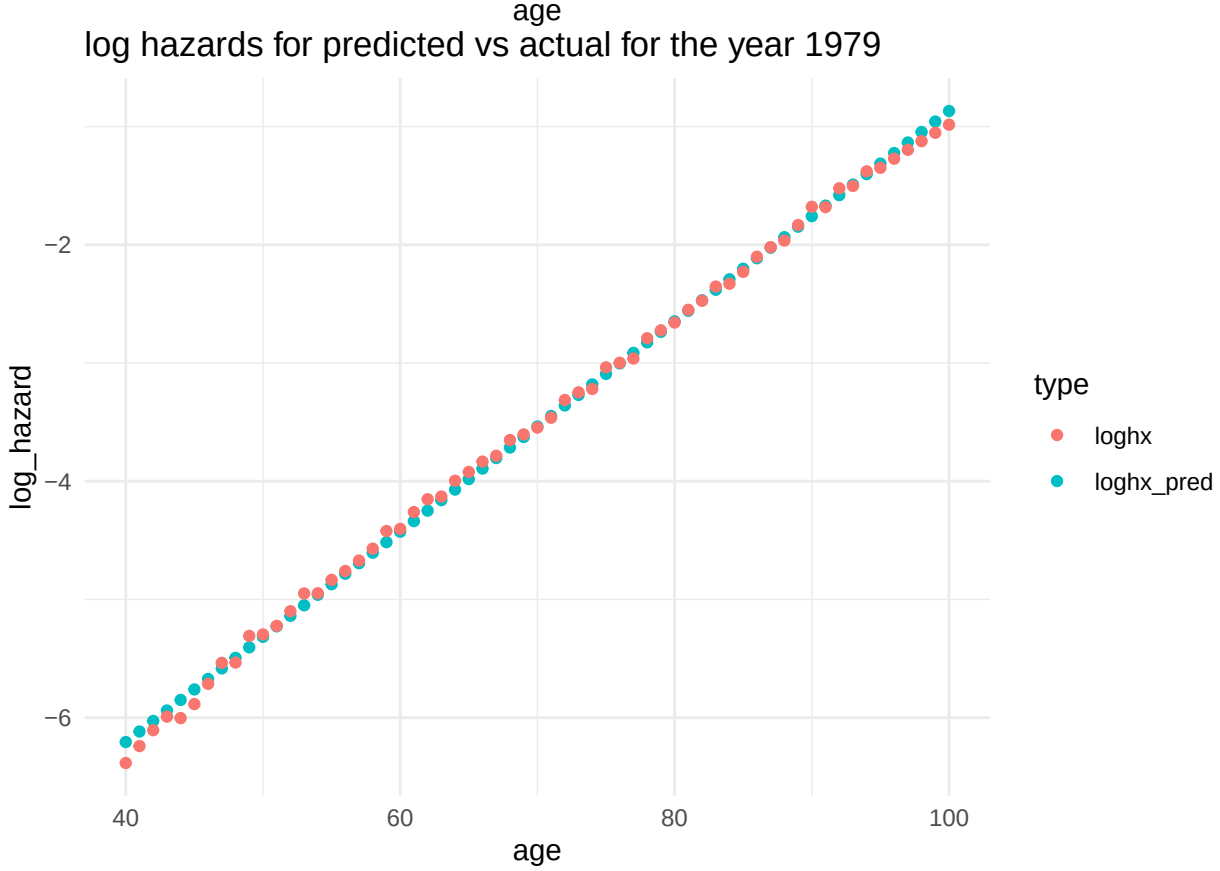

log hazards for predicted vs actual for the year 1974
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

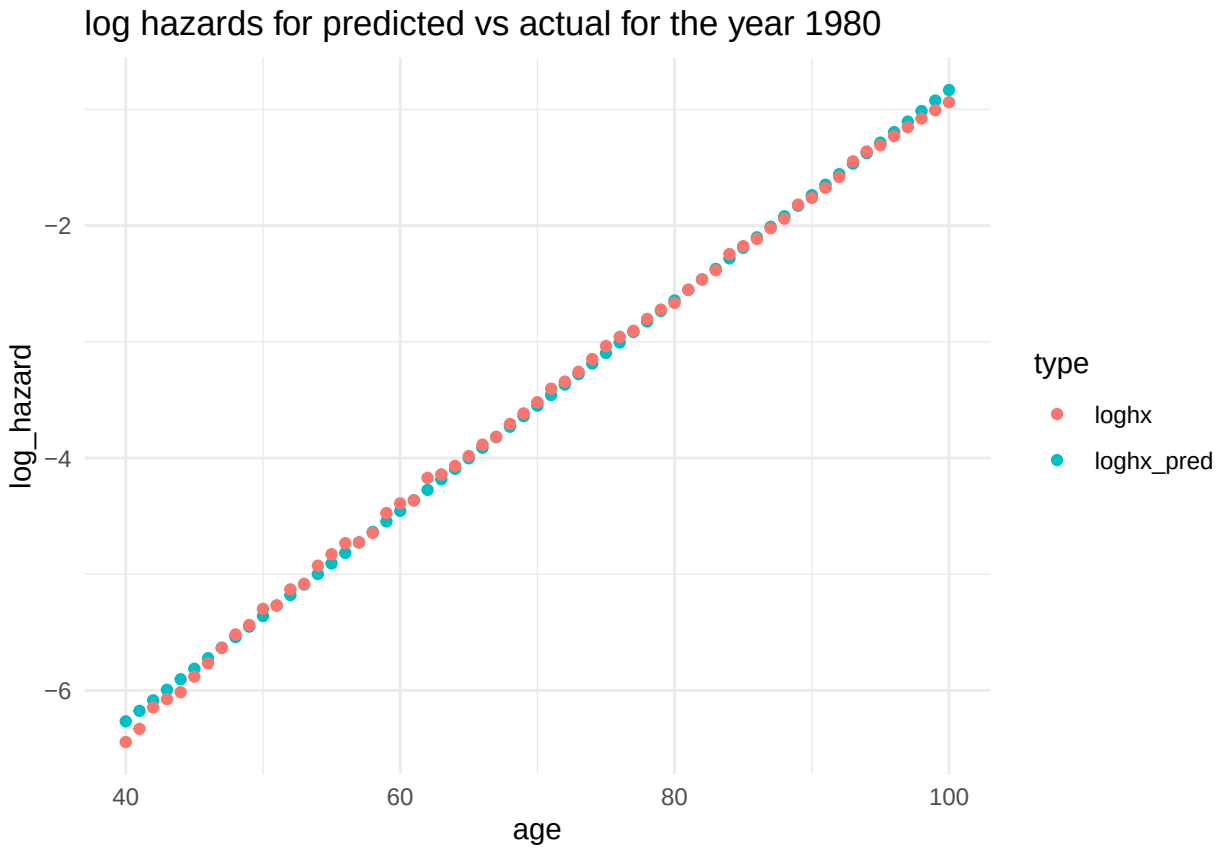

log hazards for predicted vs actual for the year 1975
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

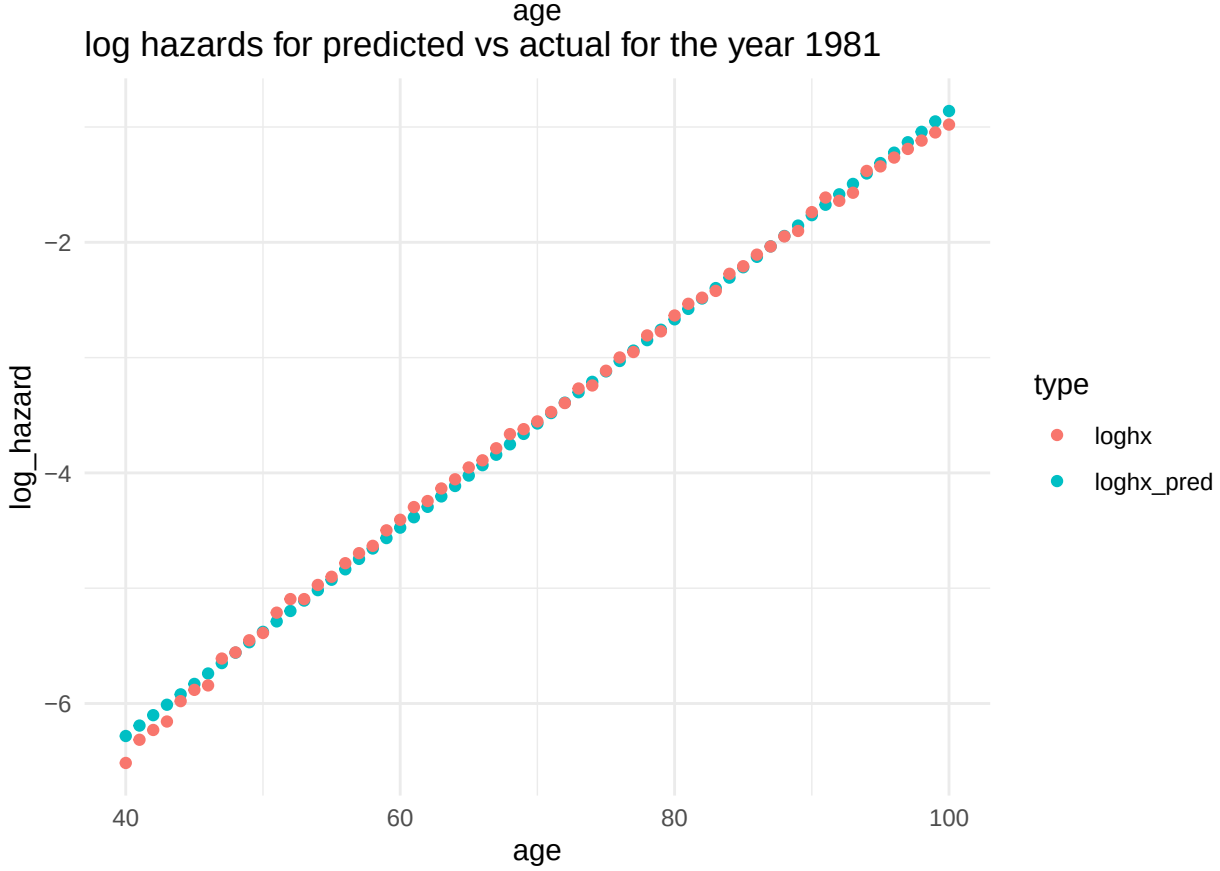

log hazards for predicted vs actual for the year 1976
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

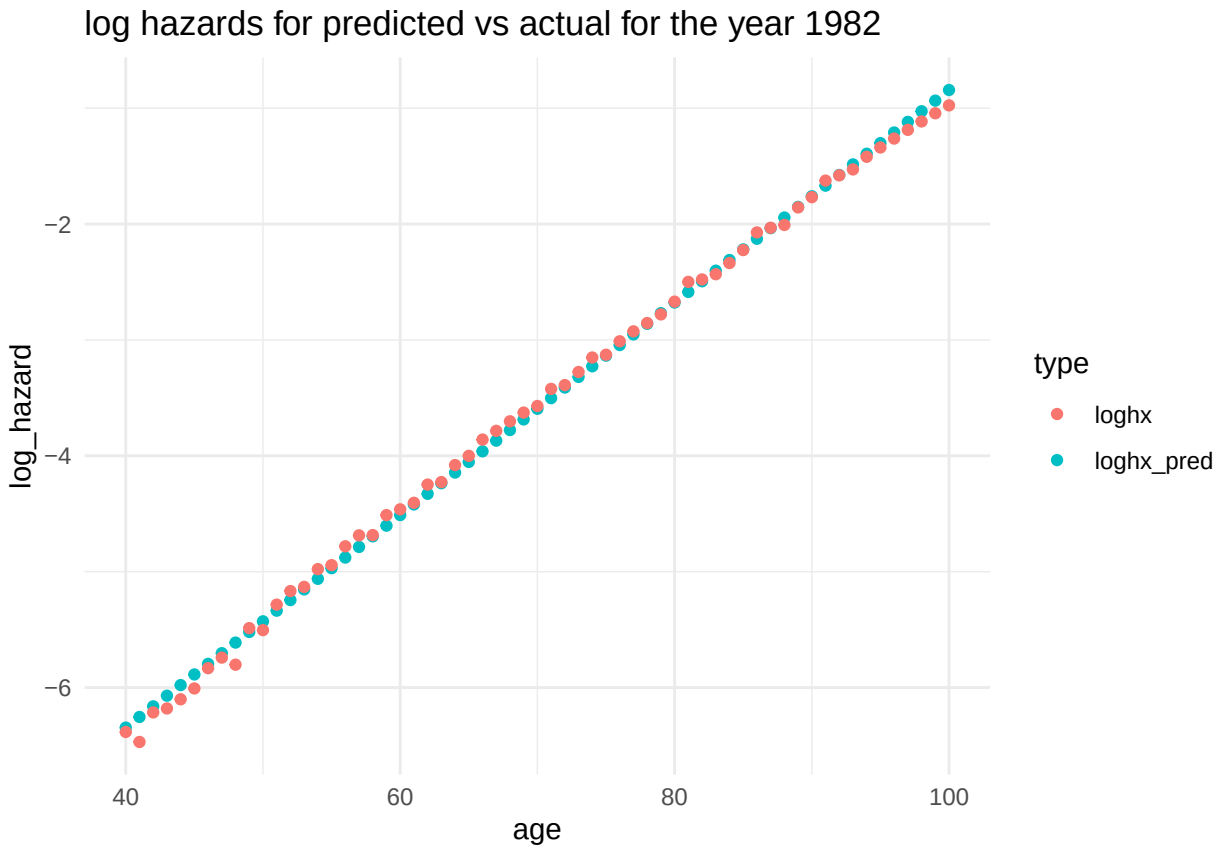

log hazards for predicted vs actual for the year 1977
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

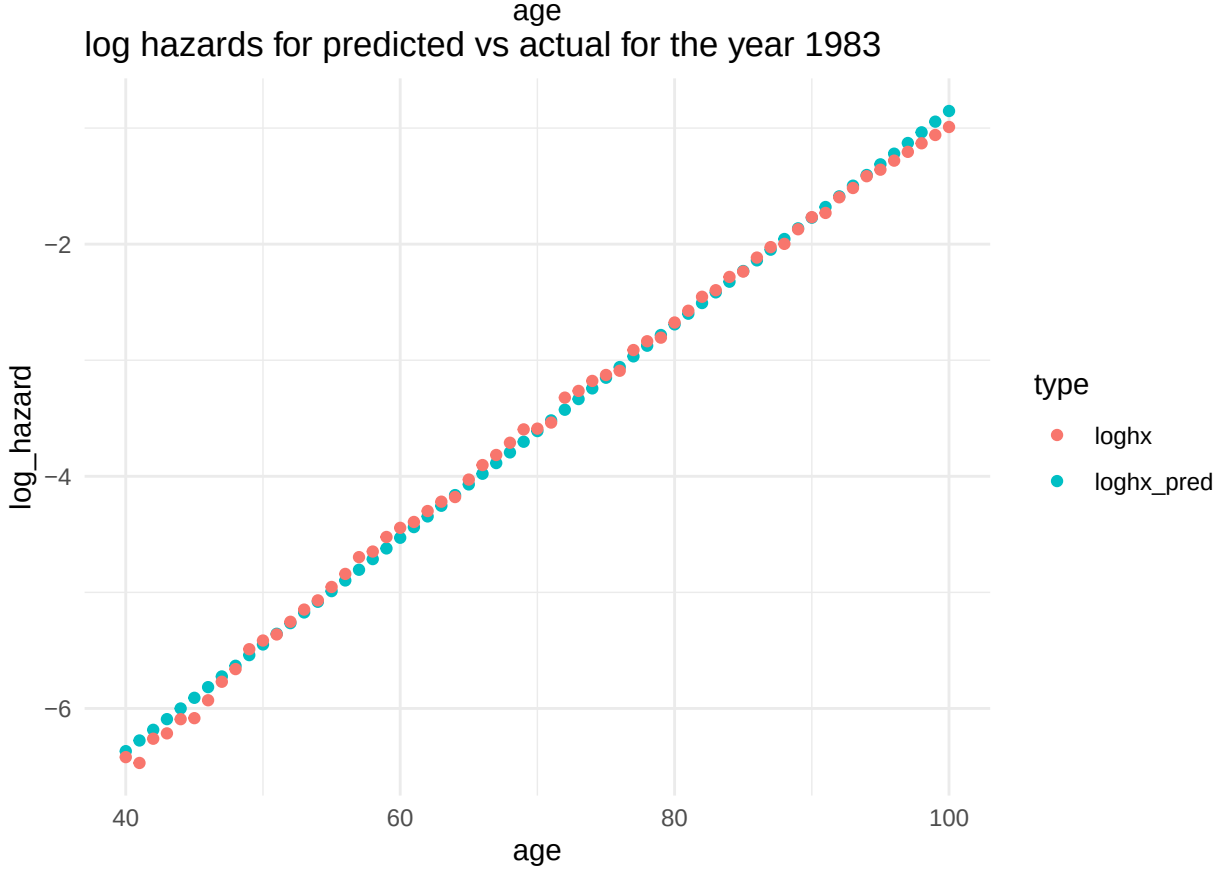

log hazards for predicted vs actual for the year 1978
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

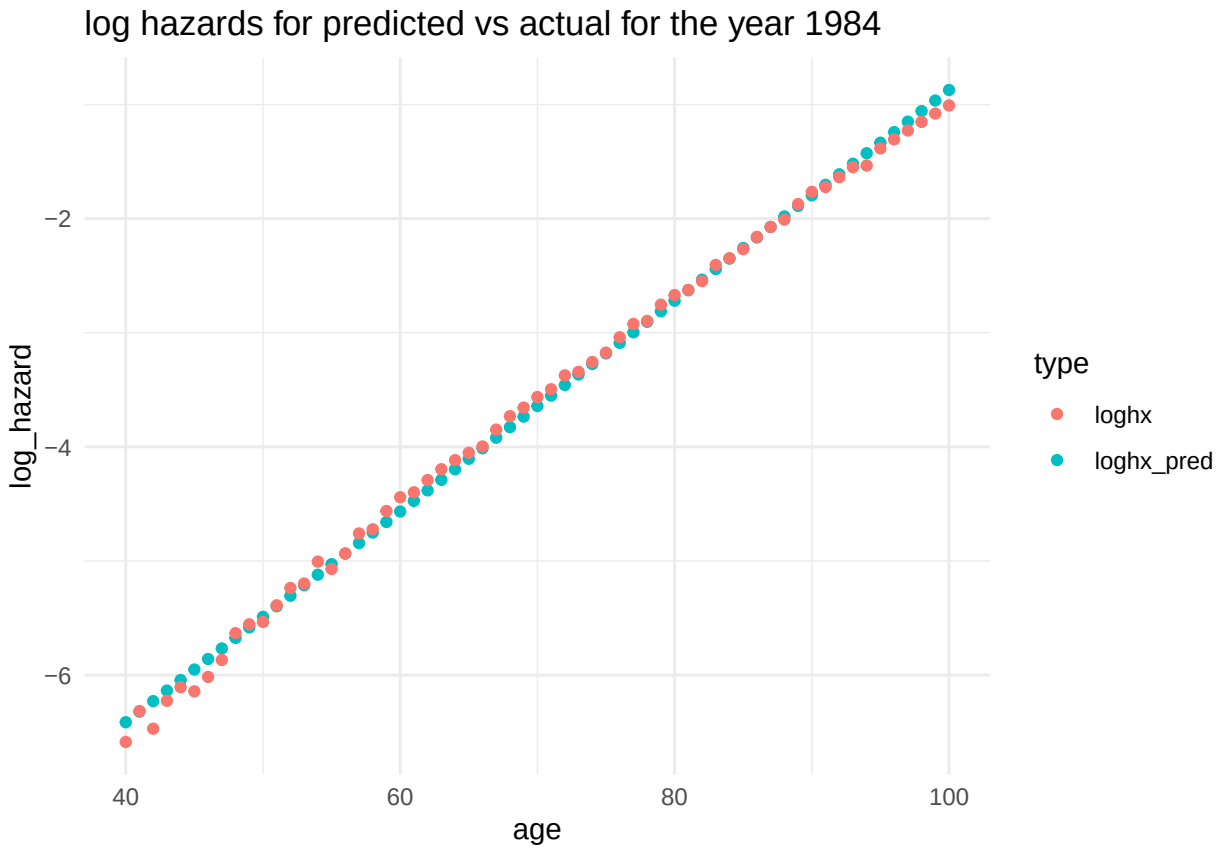

log hazards for predicted vs actual for the year 1979
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

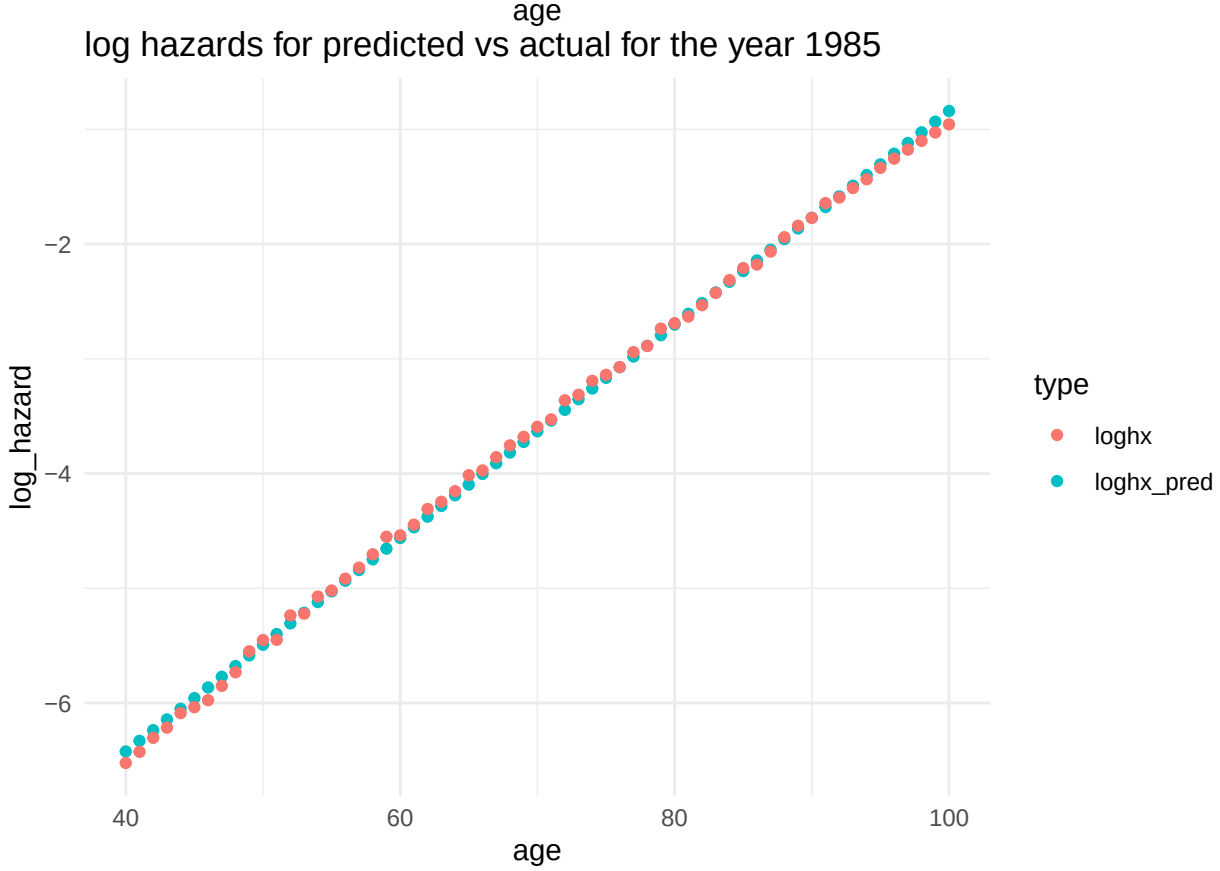

log hazards for predicted vs actual for the year 1980
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

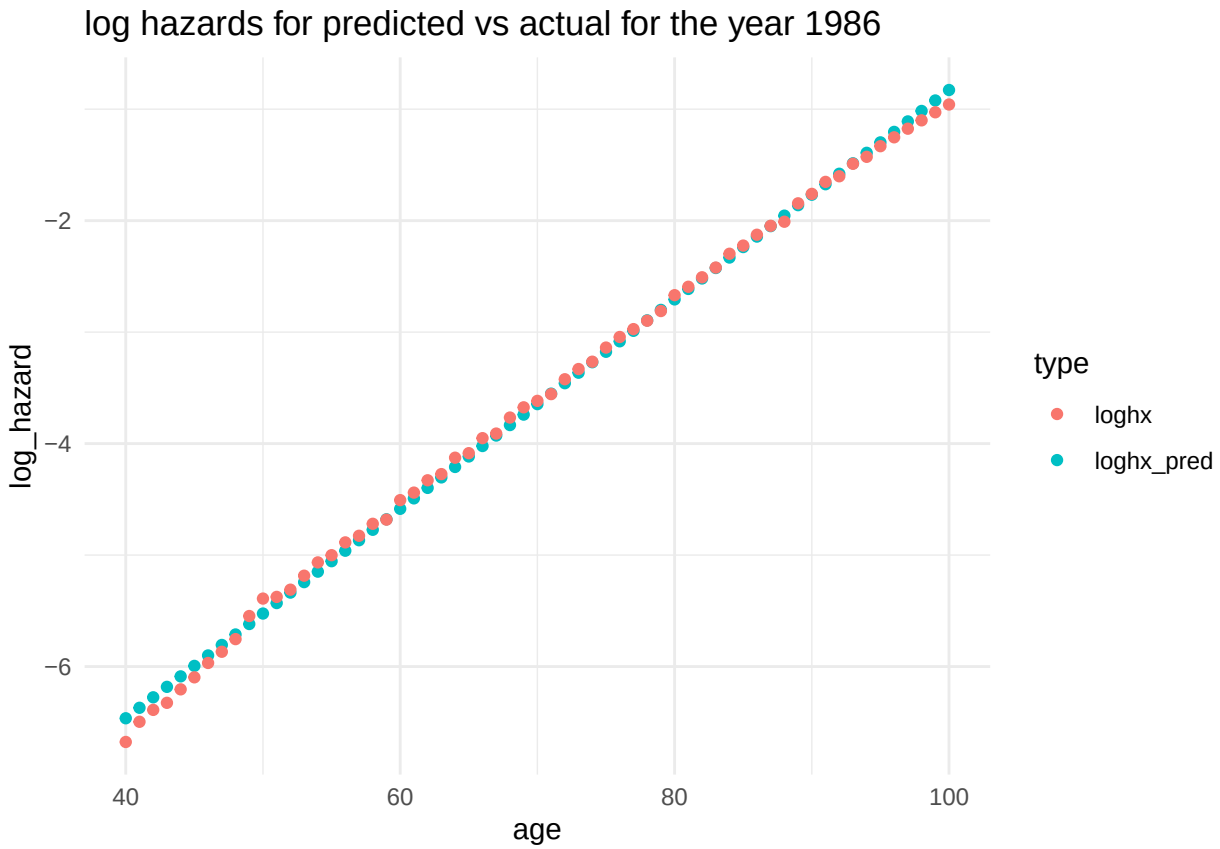

log hazards for predicted vs actual for the year 1981
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

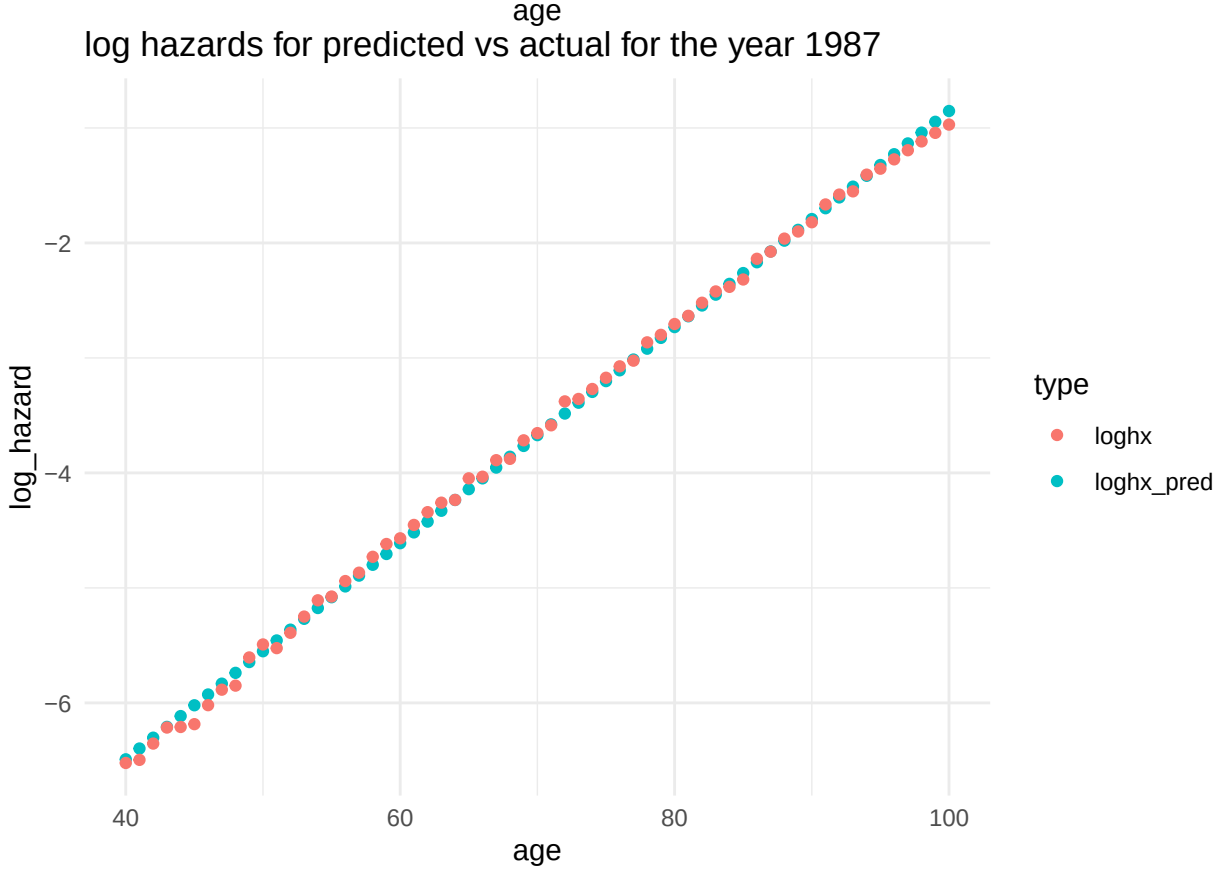

log hazards for predicted vs actual for the year 1982
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

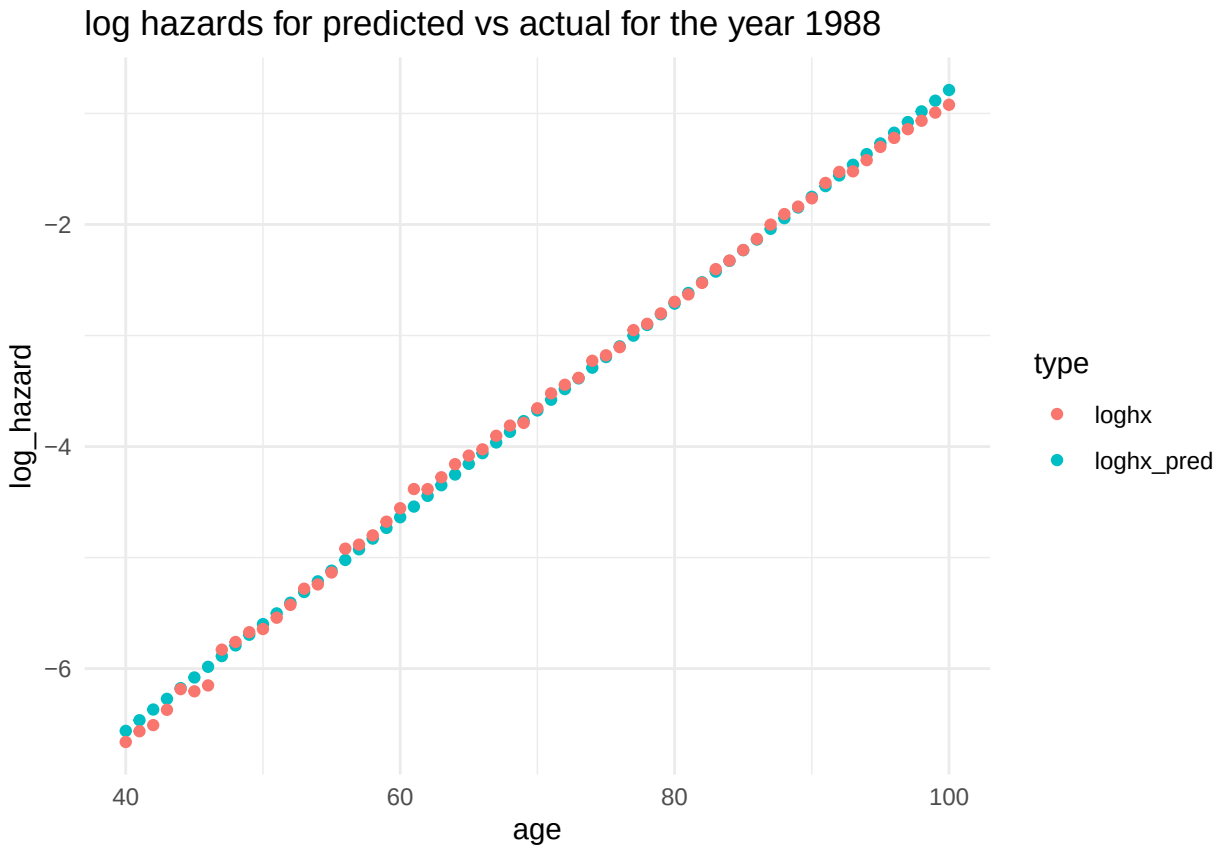

log hazards for predicted vs actual for the year 1983
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

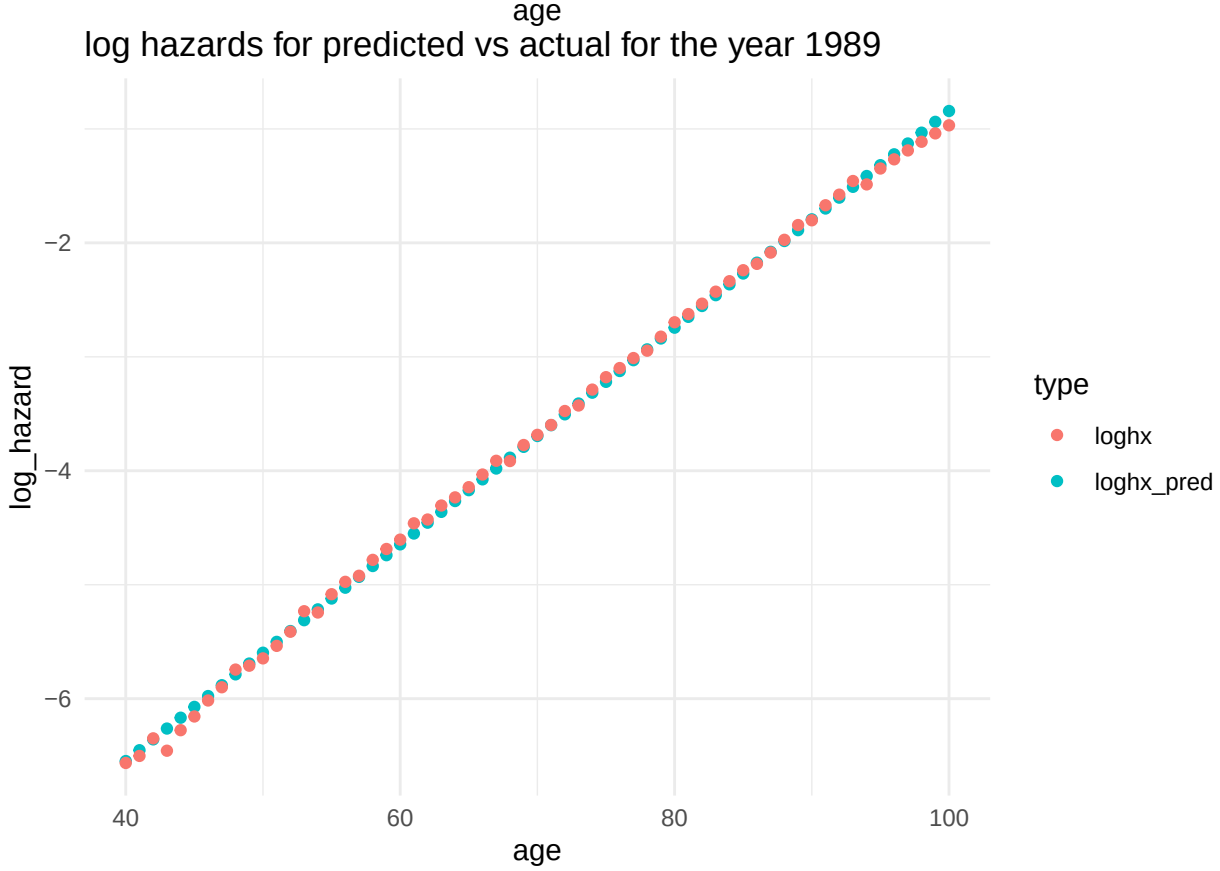

log hazards for predicted vs actual for the year 1984
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

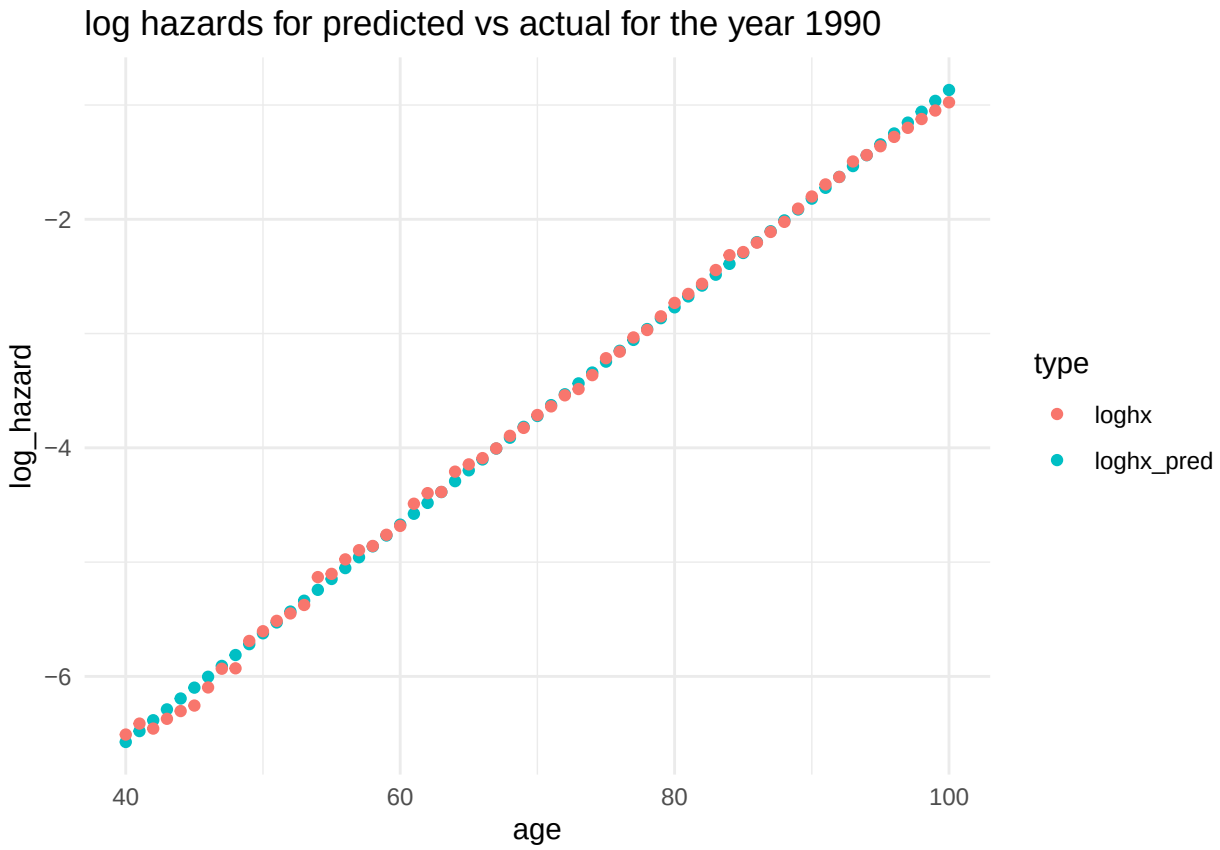

log hazards for predicted vs actual for the year 1985
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

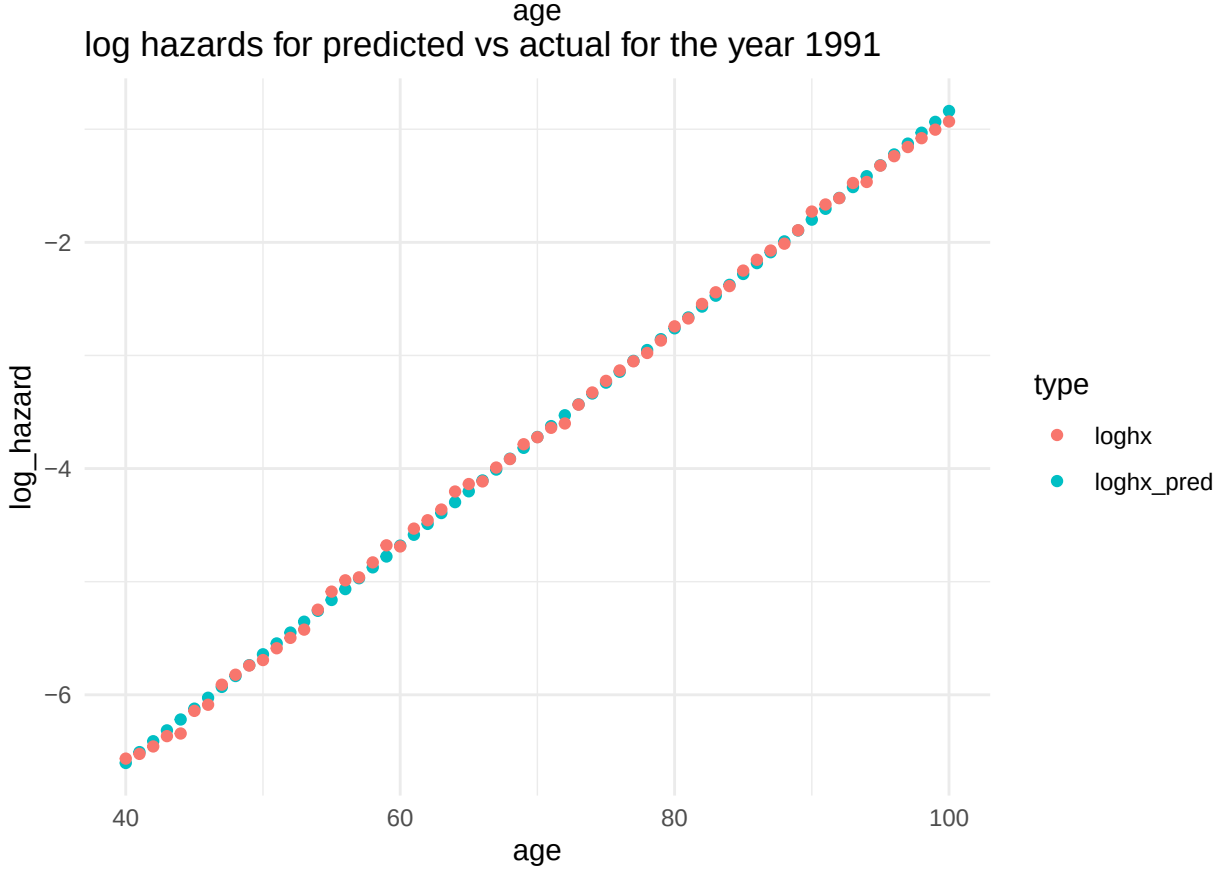

log hazards for predicted vs actual for the year 1986
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

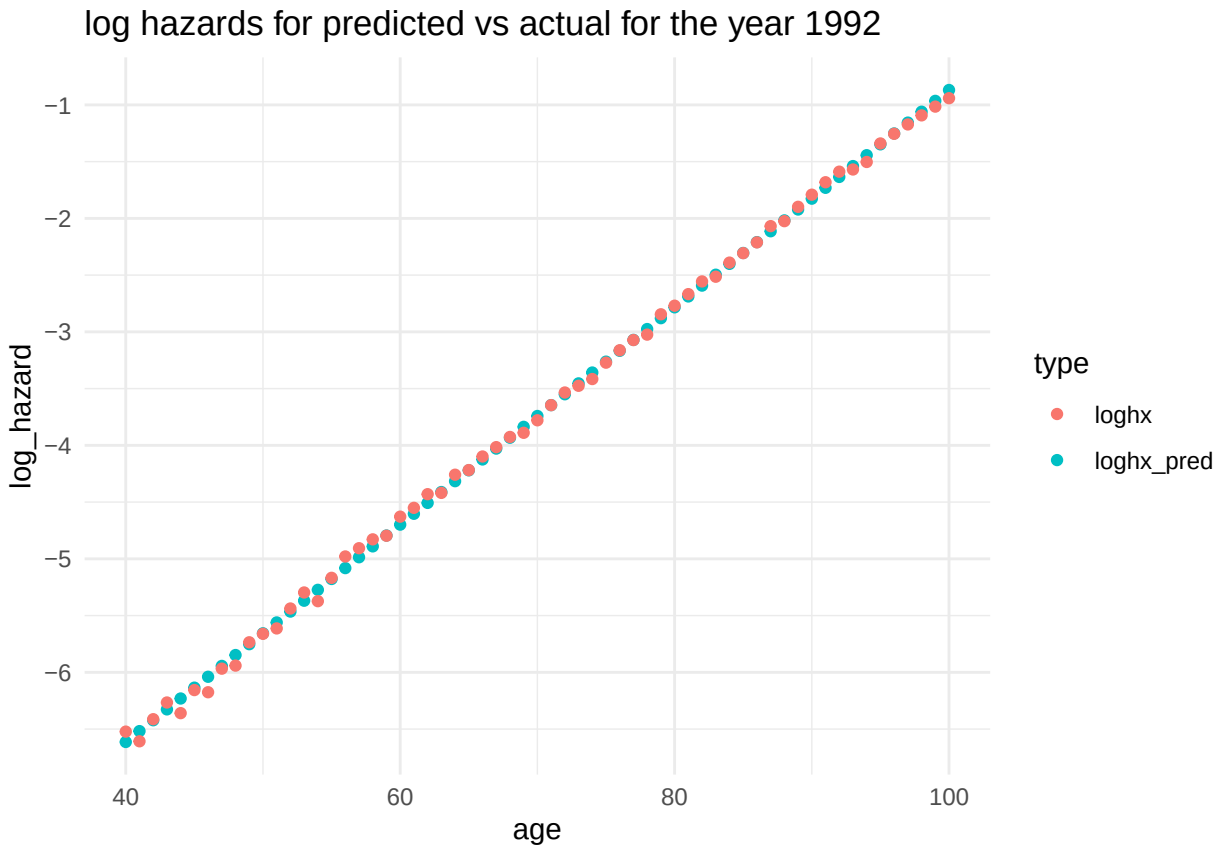

log hazards for predicted vs actual for the year 1987
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

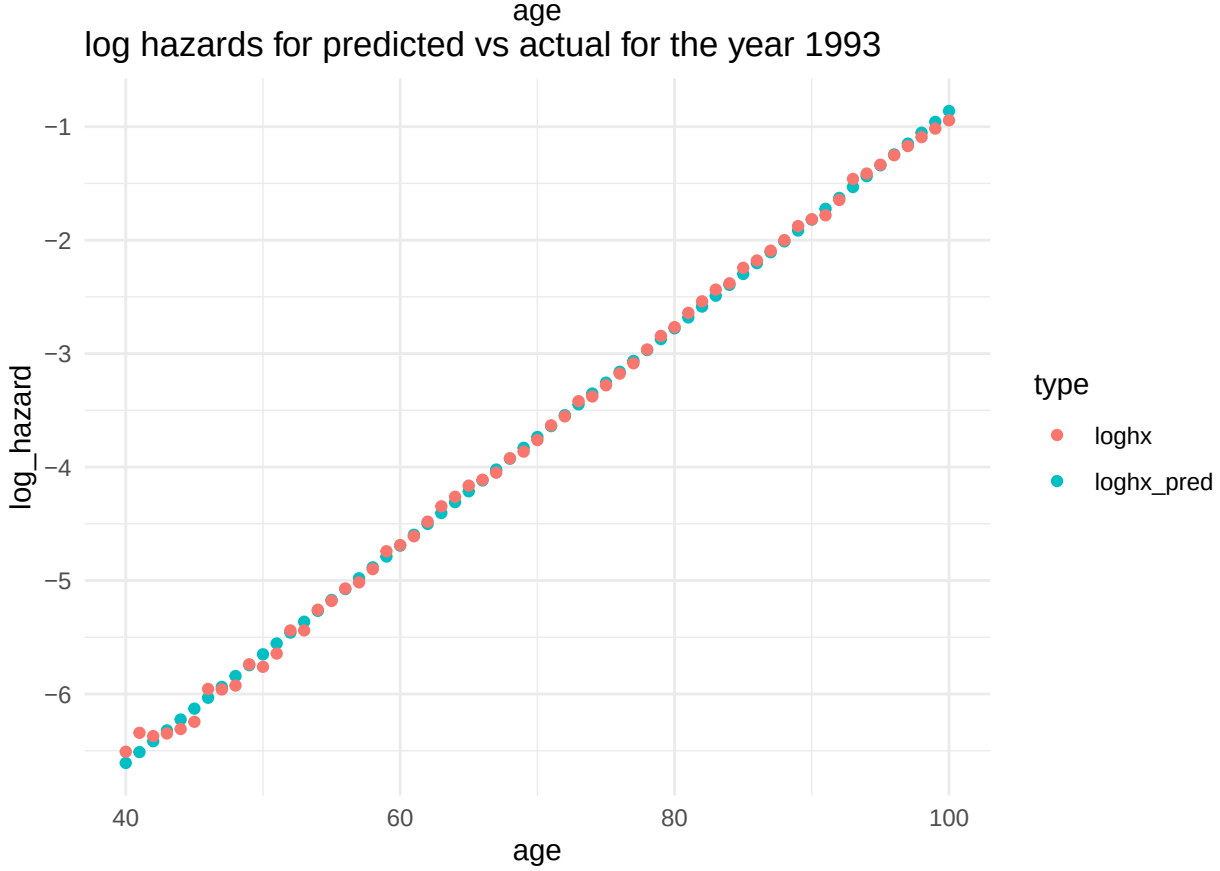

log hazards for predicted vs actual for the year 1988
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

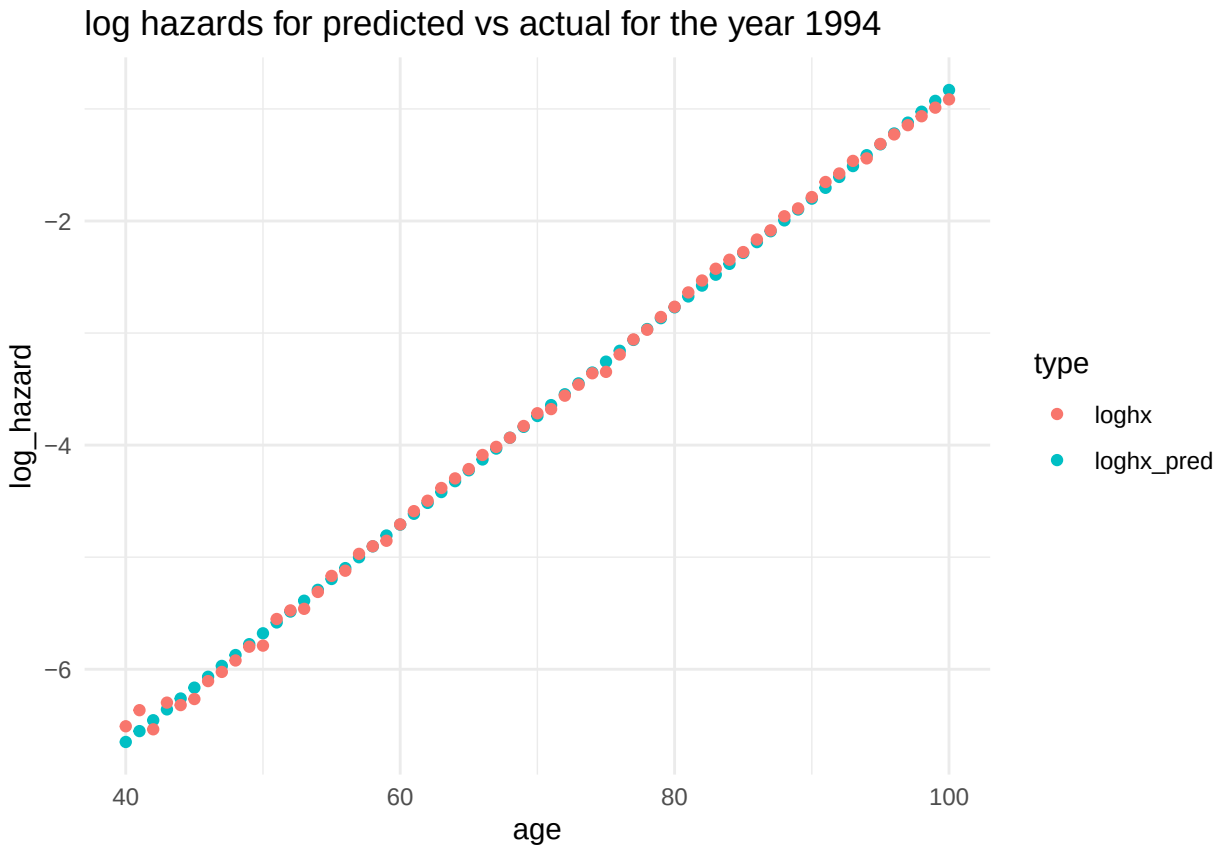

log hazards for predicted vs actual for the year 1989
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

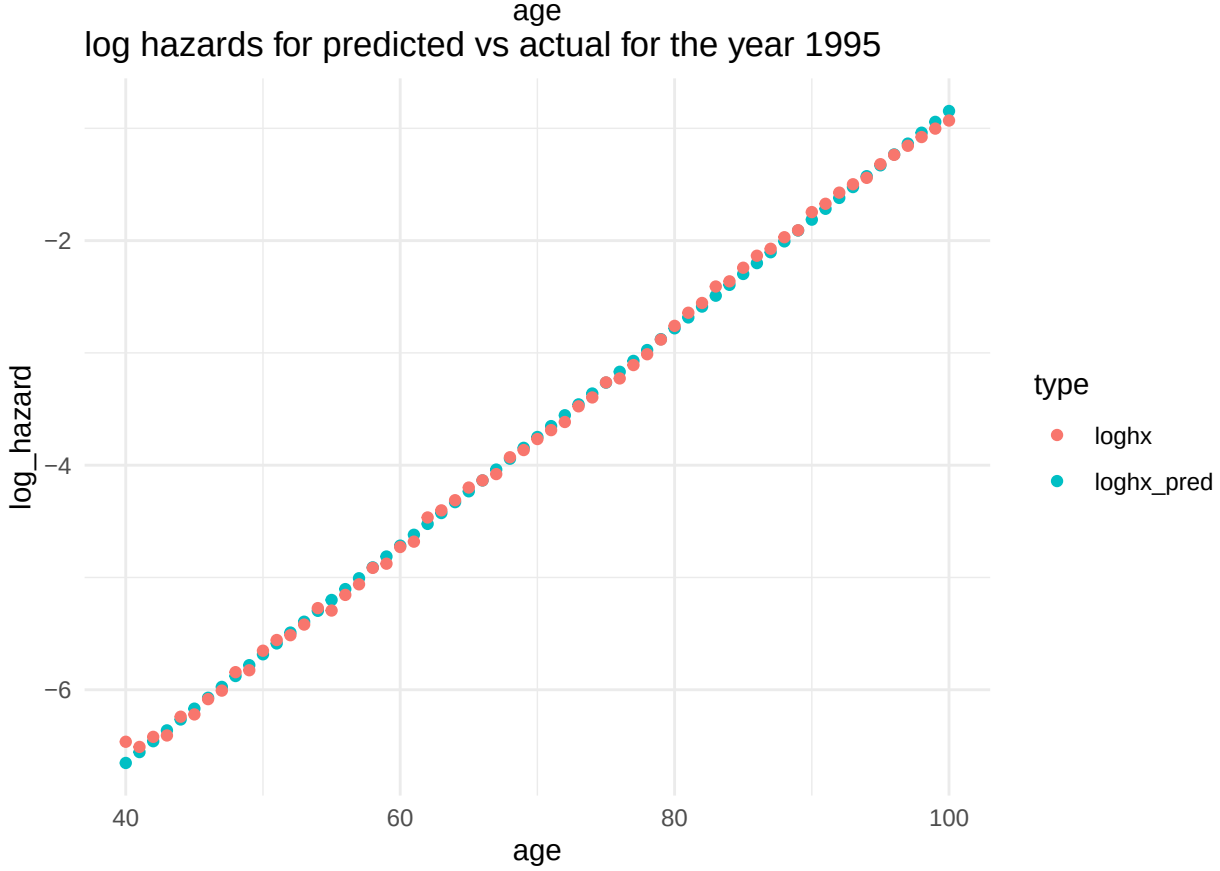

log hazards for predicted vs actual for the year 1990
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

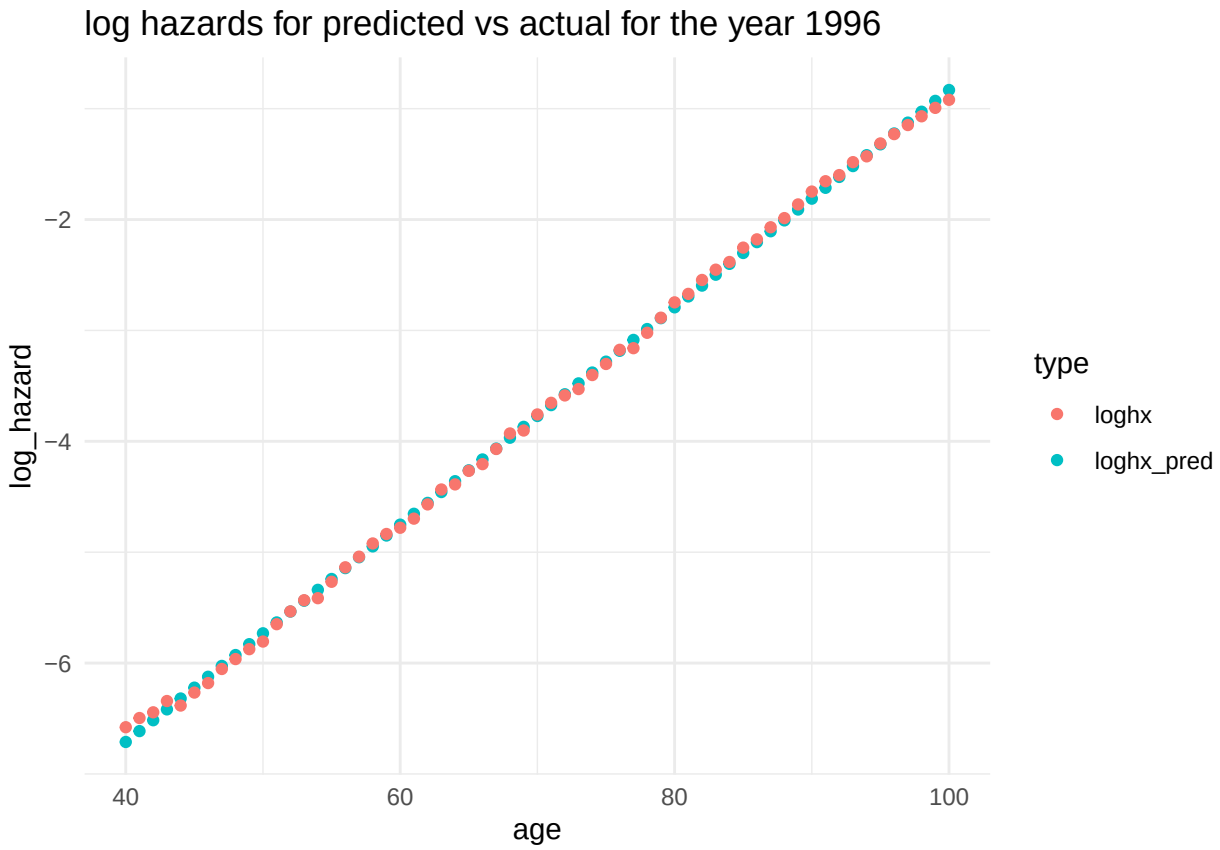

log hazards for predicted vs actual for the year 1991
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

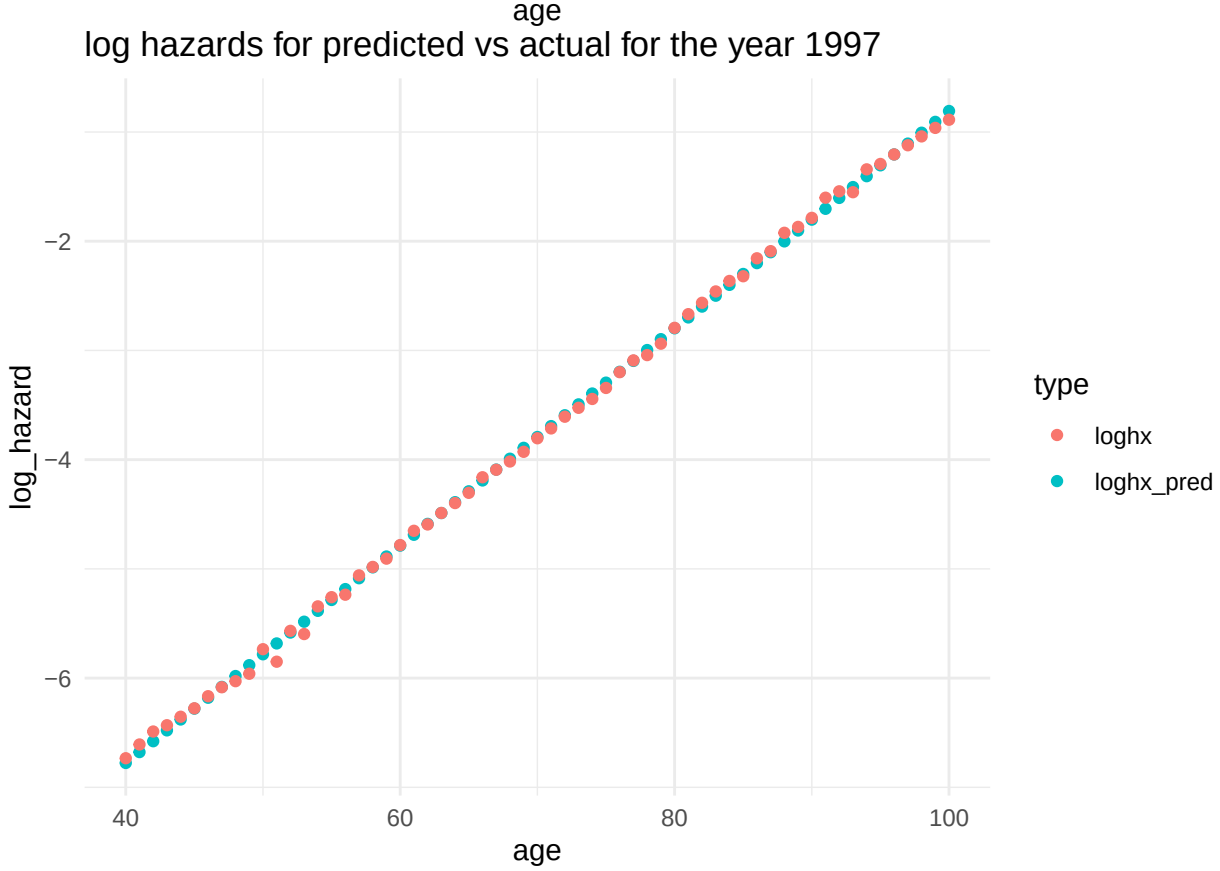

log hazards for predicted vs actual for the year 1992
log_hazard
−1
−2
−3
−4
−5
−6
40
60
80
100
age
type
loghx
loghx_pred

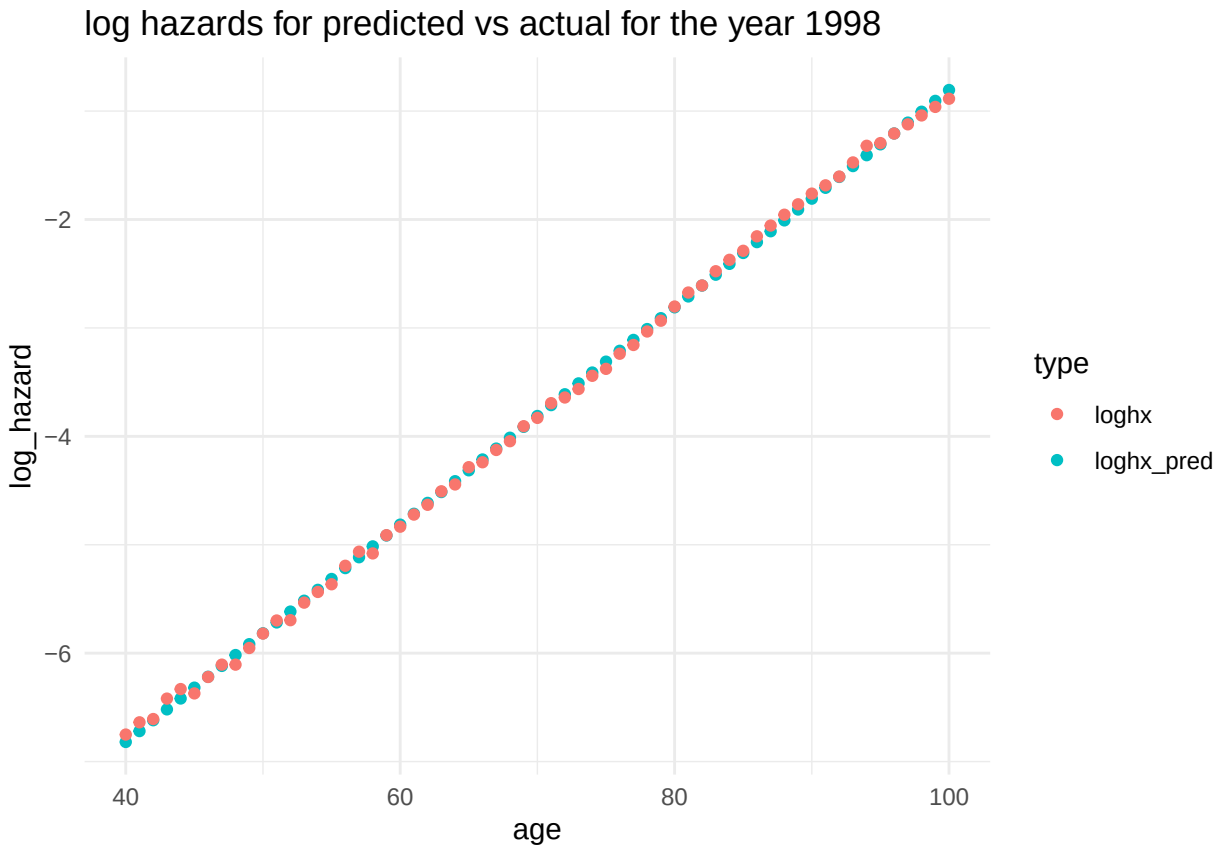

log hazards for predicted vs actual for the year 1993
log_hazard
−1
−2
−3
−4
−5
−6
40
60
80
100
age
type
loghx
loghx_pred

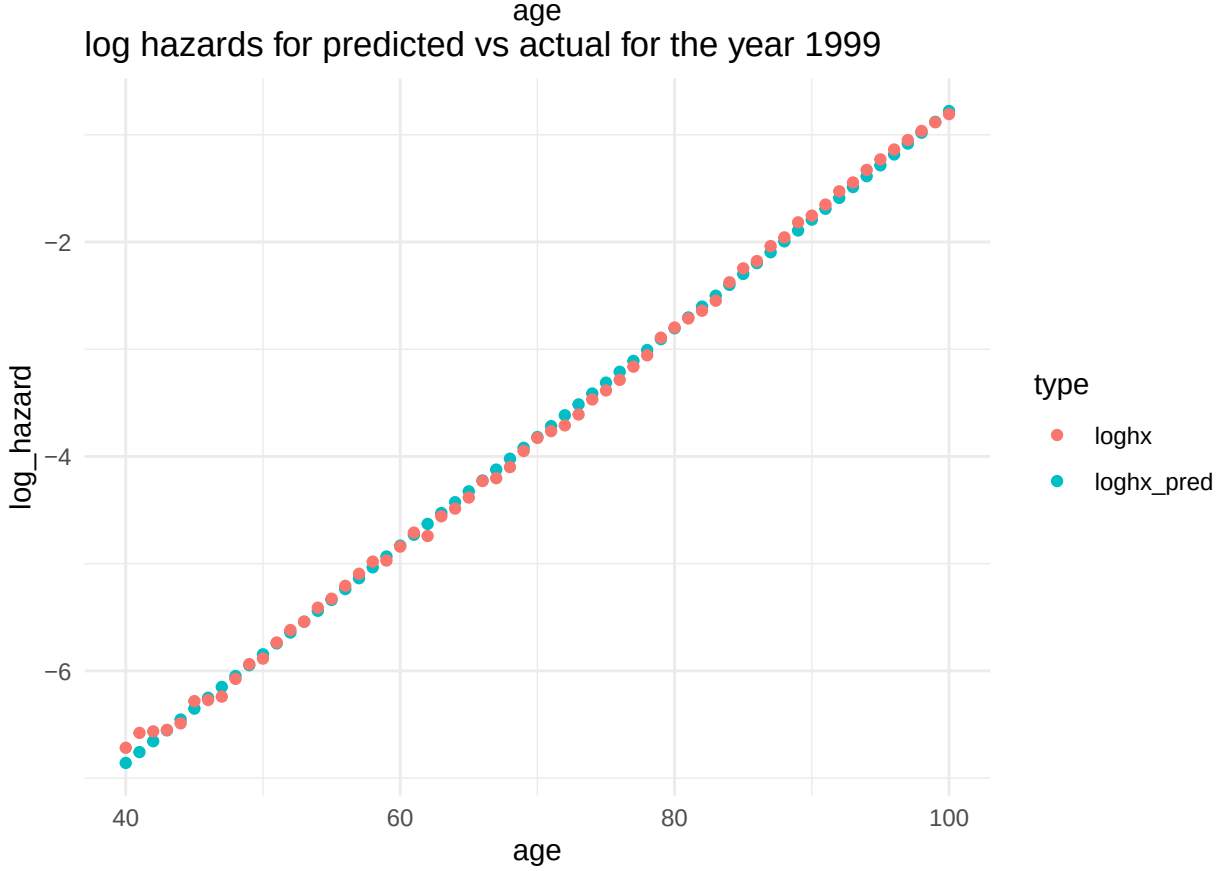

log hazards for predicted vs actual for the year 1994
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

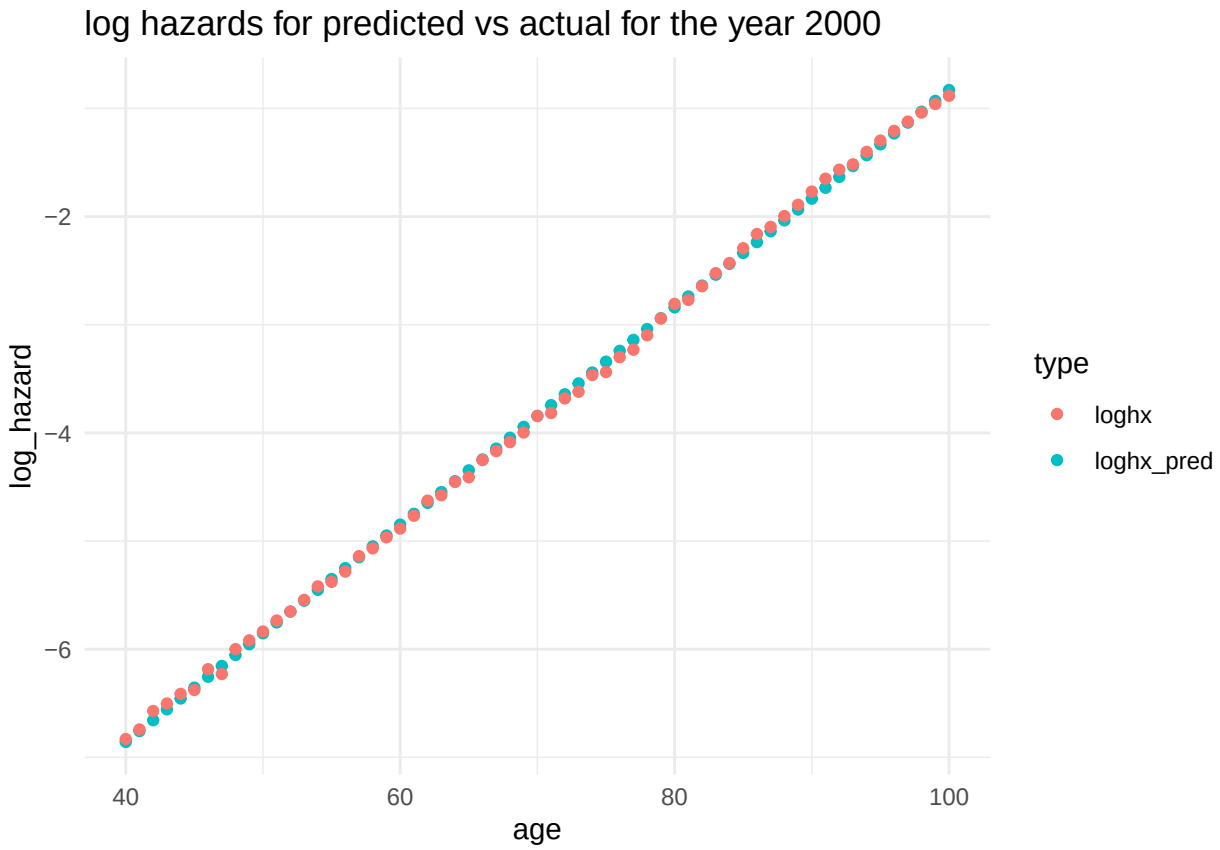

log hazards for predicted vs actual for the year 1995
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

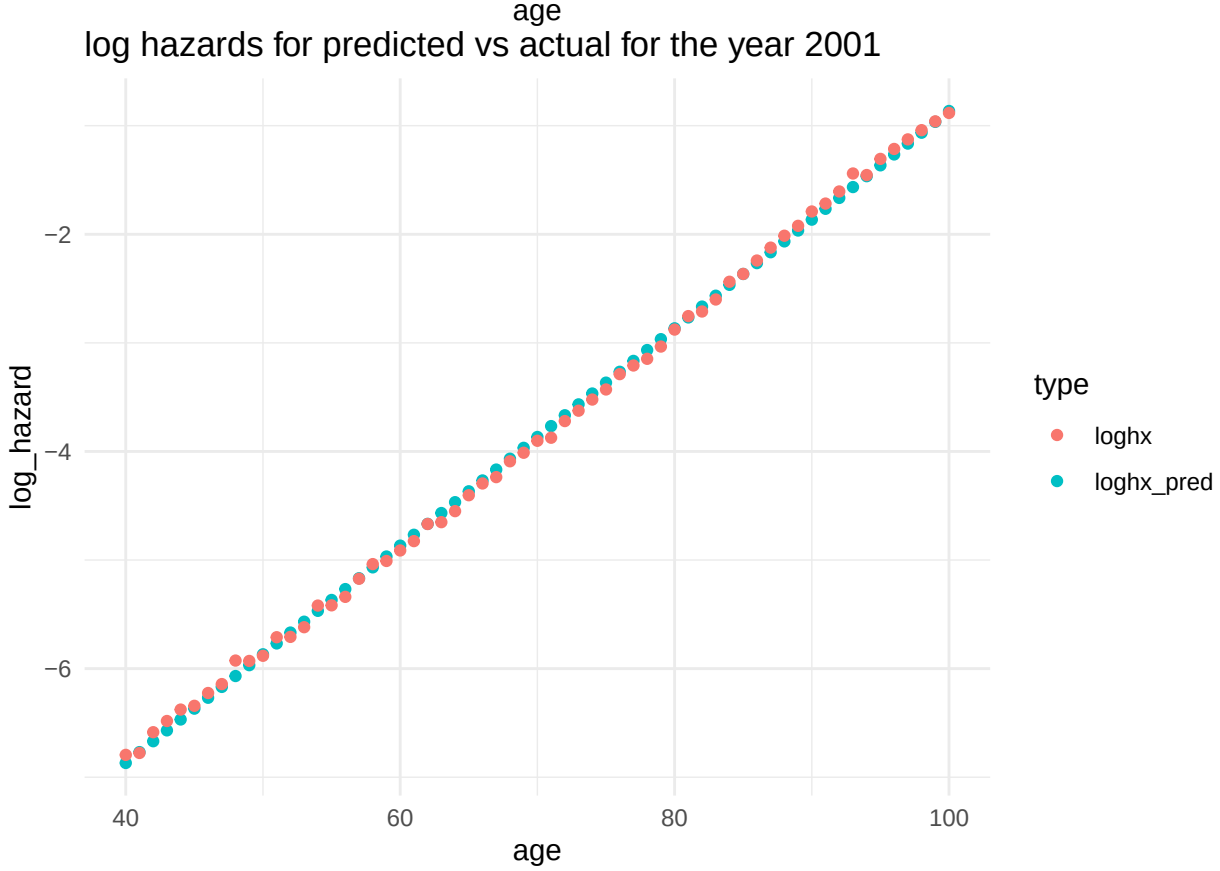

log hazards for predicted vs actual for the year 1996
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

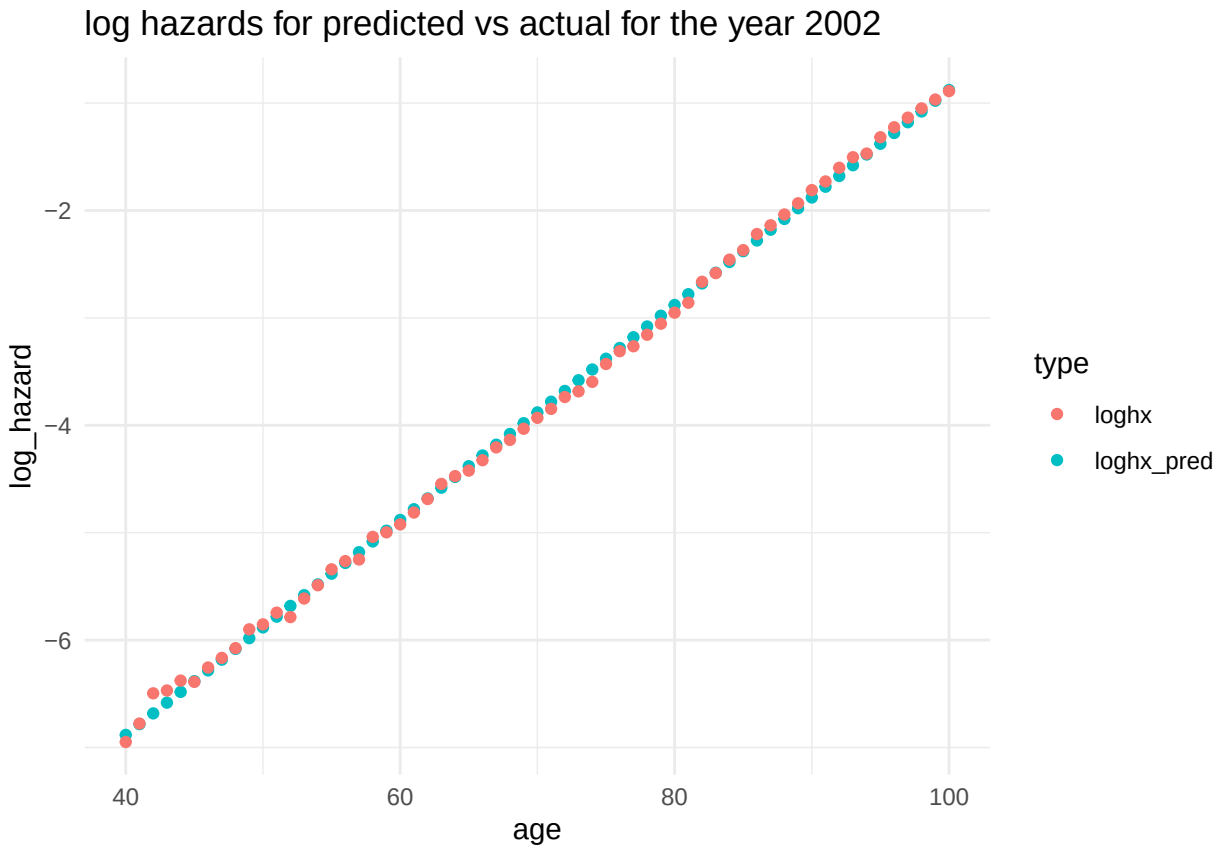

log hazards for predicted vs actual for the year 1997
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

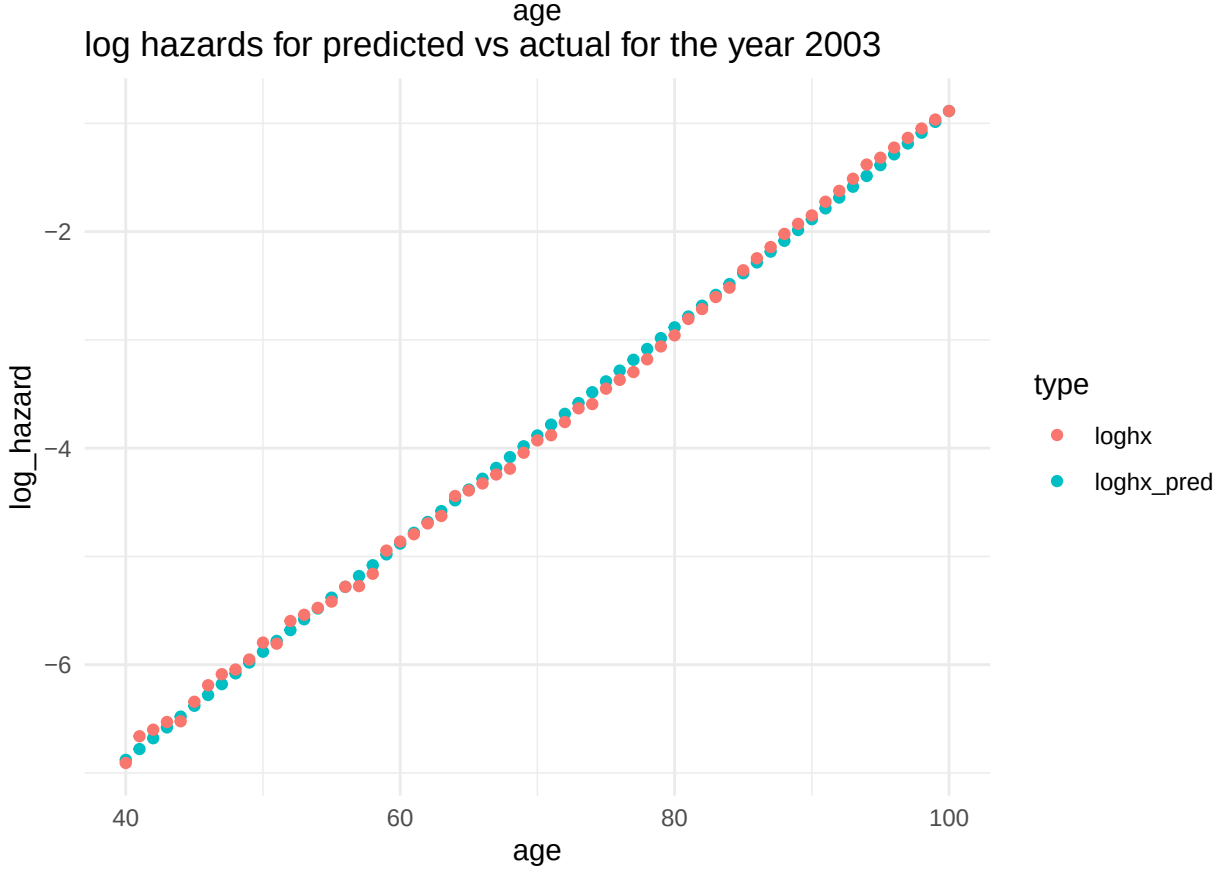

log hazards for predicted vs actual for the year 1998
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

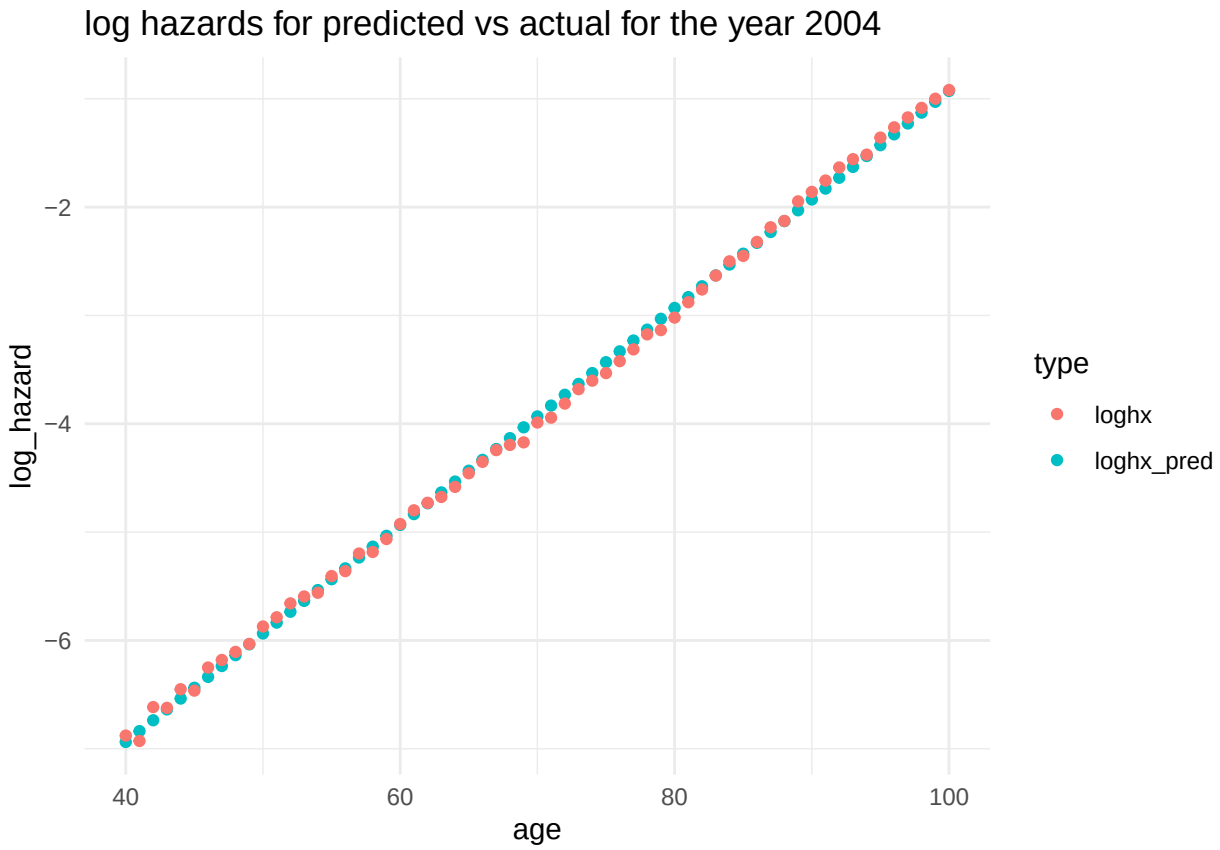

log hazards for predicted vs actual for the year 1999
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

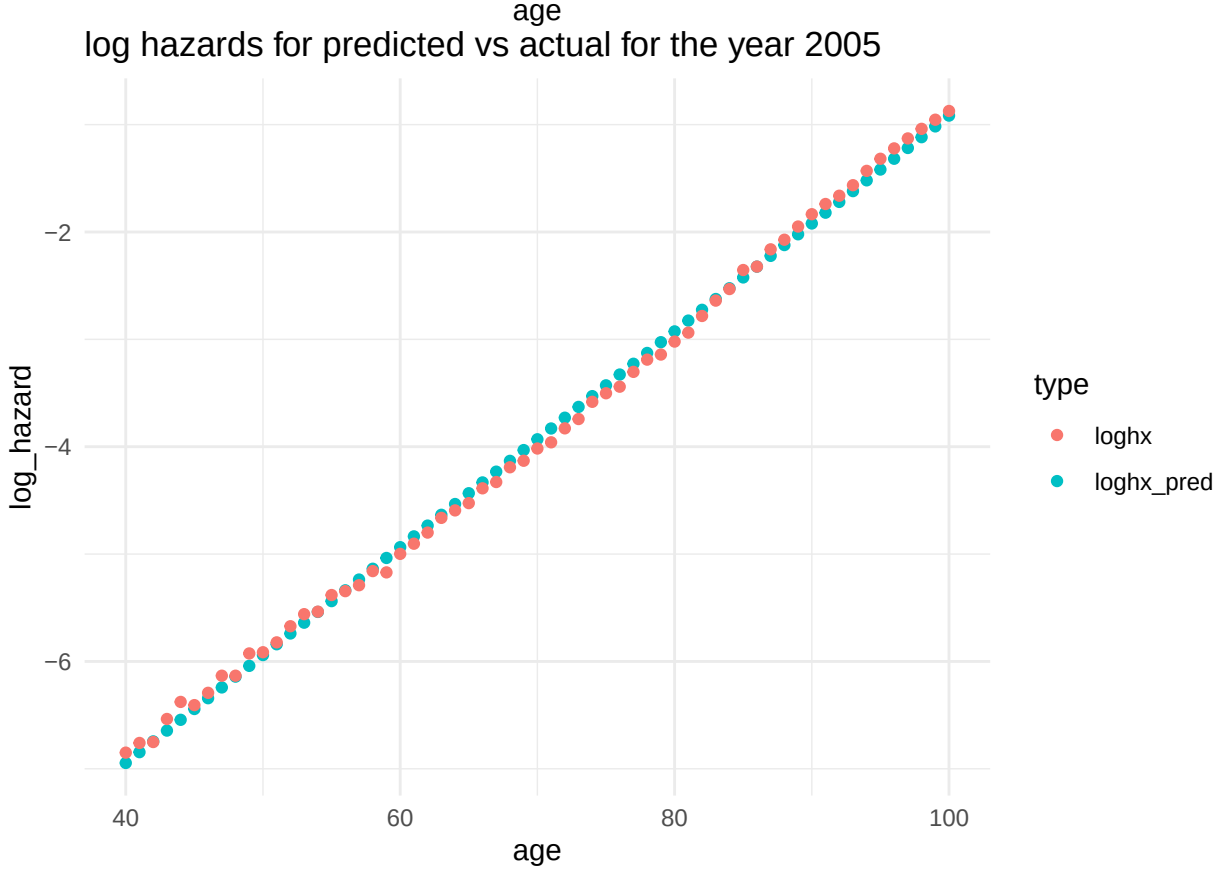

log hazards for predicted vs actual for the year 2000
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

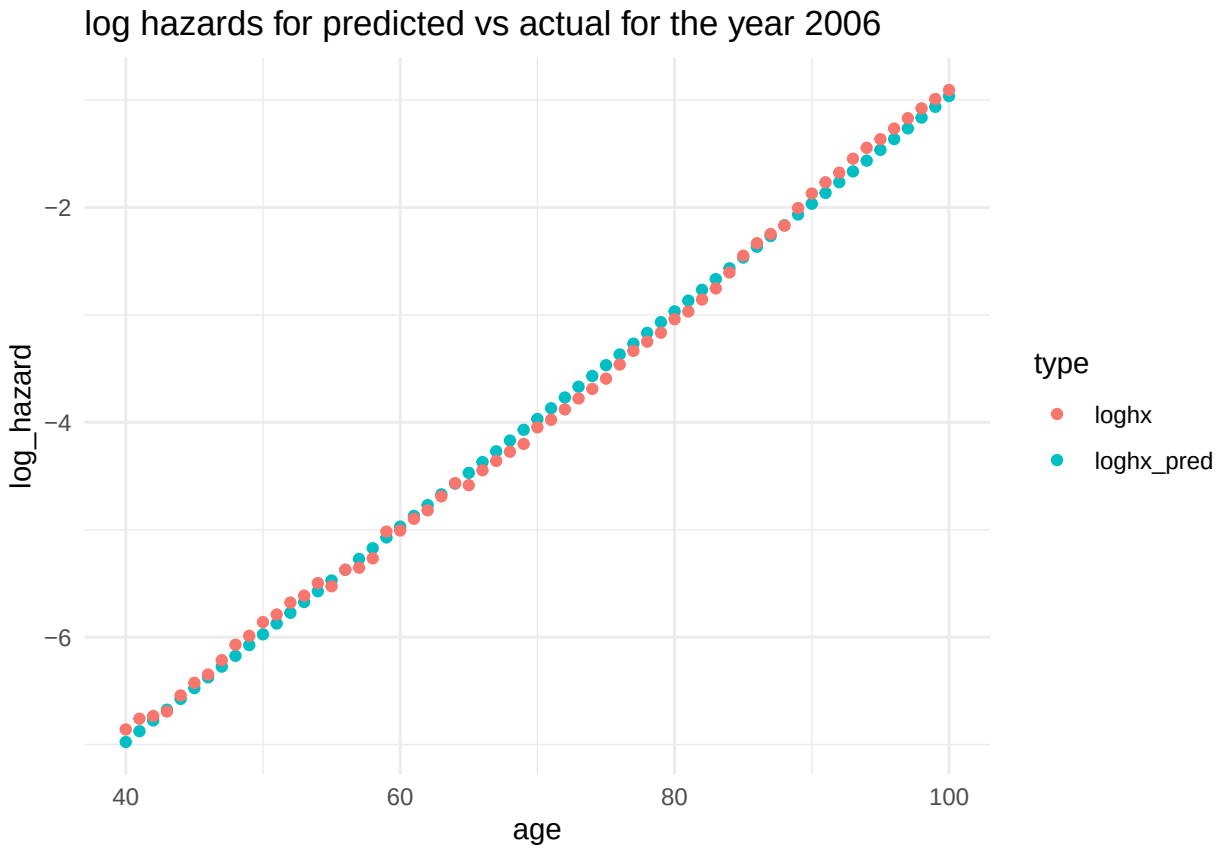

log hazards for predicted vs actual for the year 2001
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

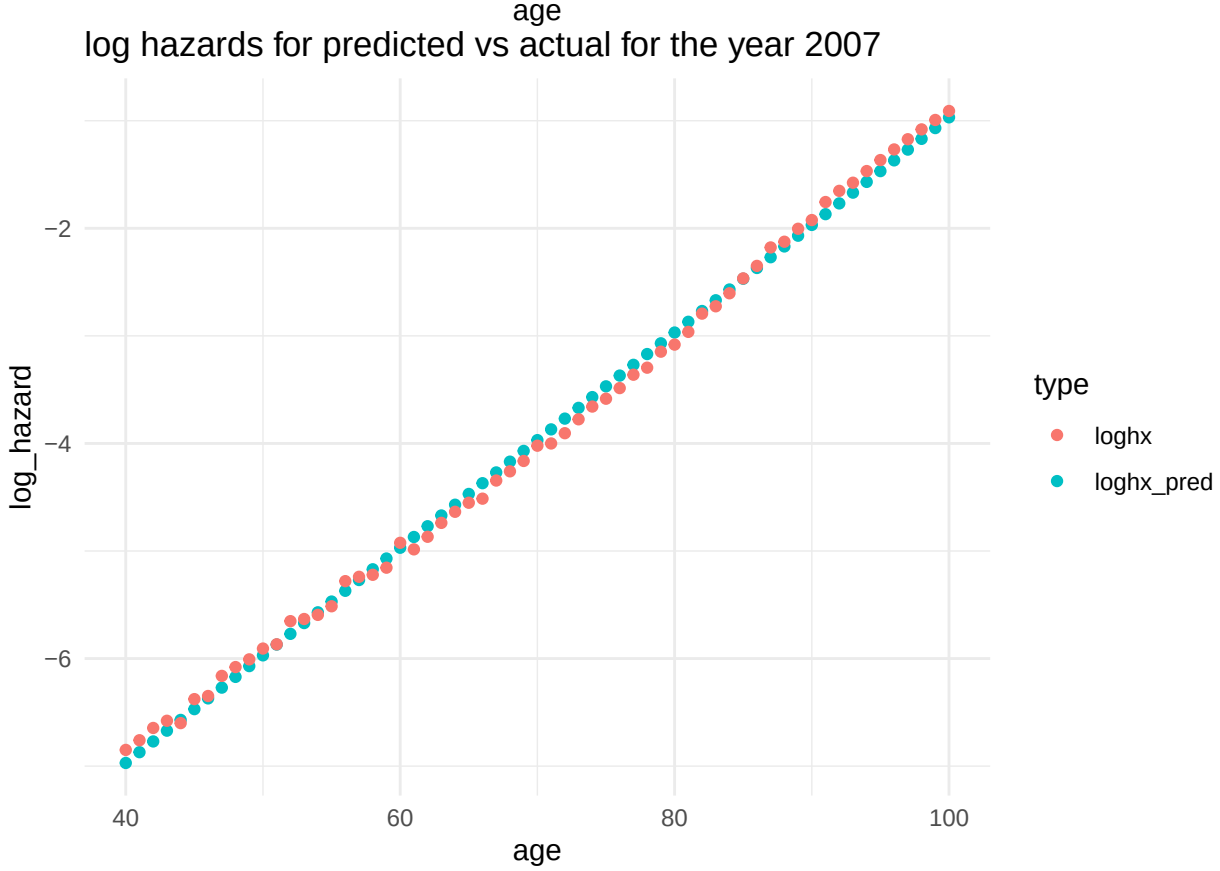

log hazards for predicted vs actual for the year 2002
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

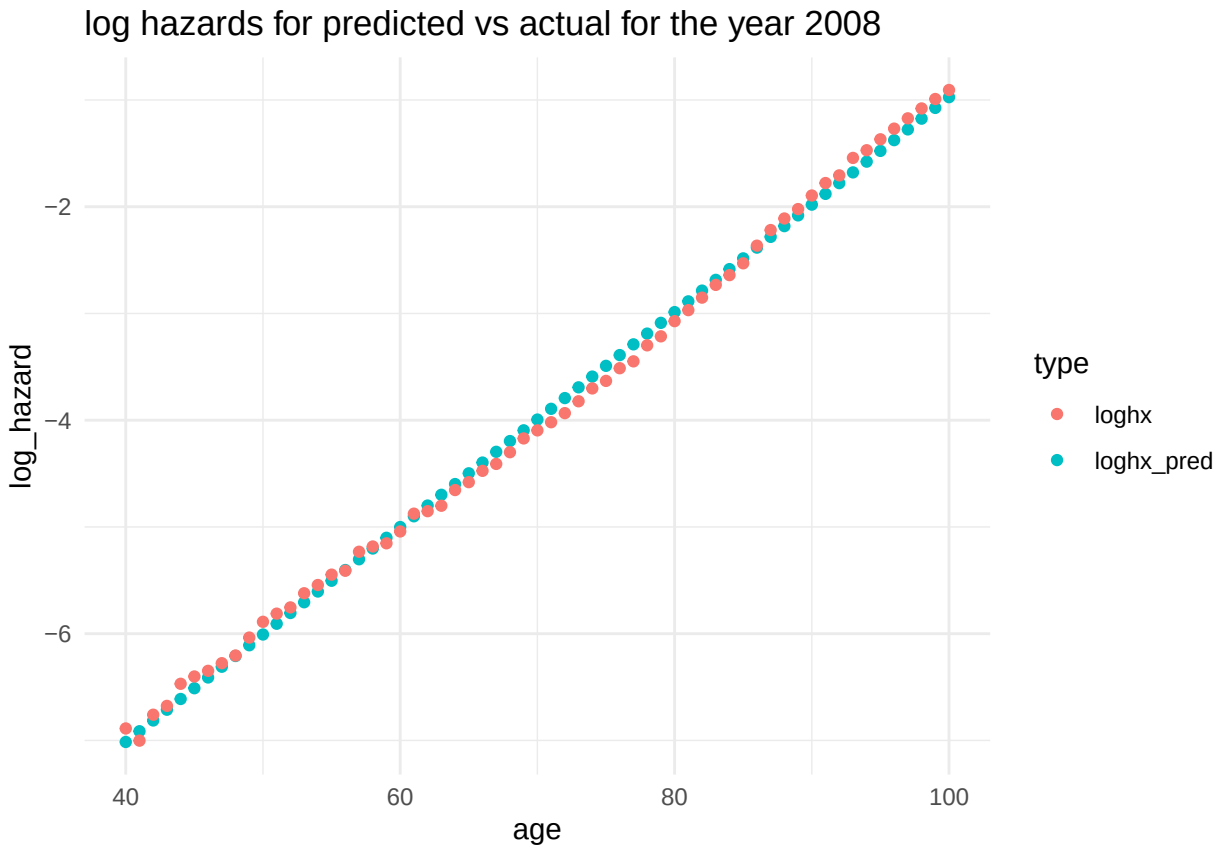

log hazards for predicted vs actual for the year 2003
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

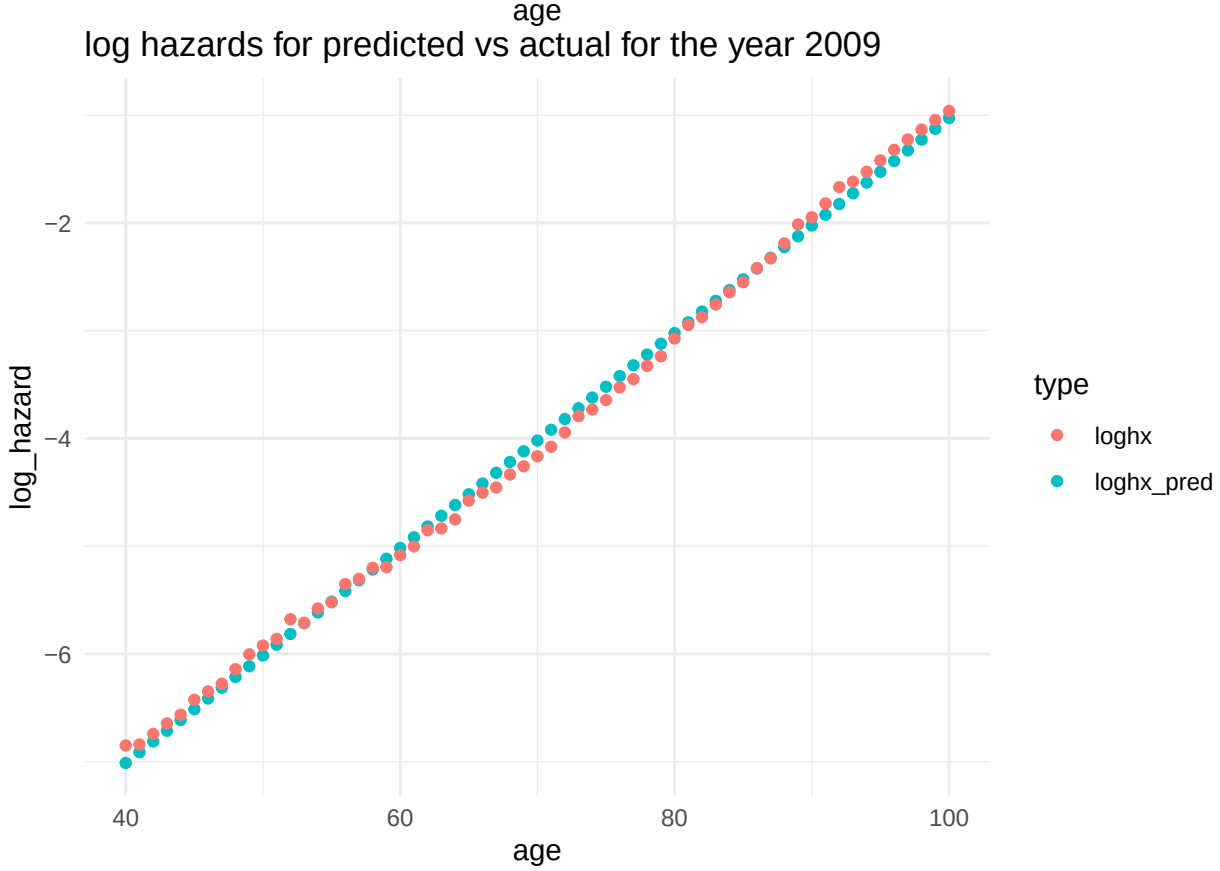

log hazards for predicted vs actual for the year 2004
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

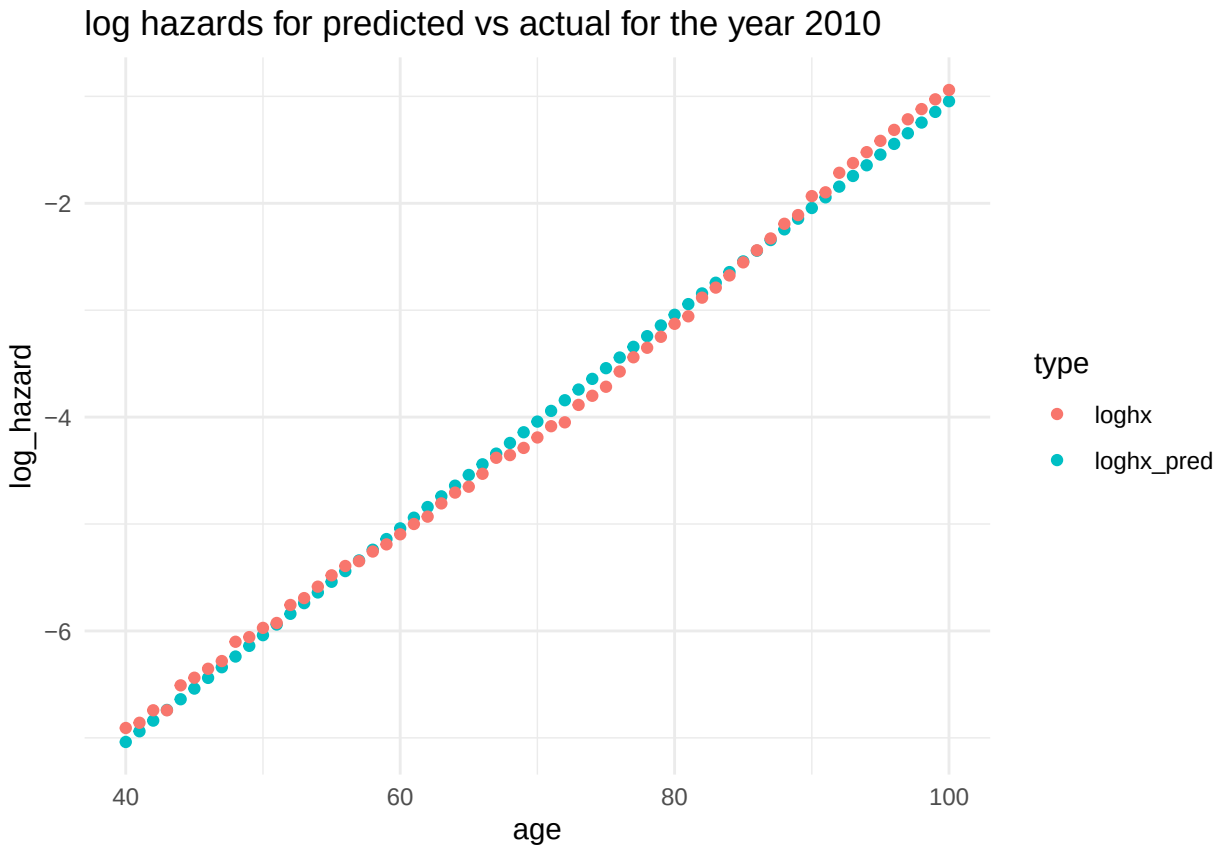

log hazards for predicted vs actual for the year 2005
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

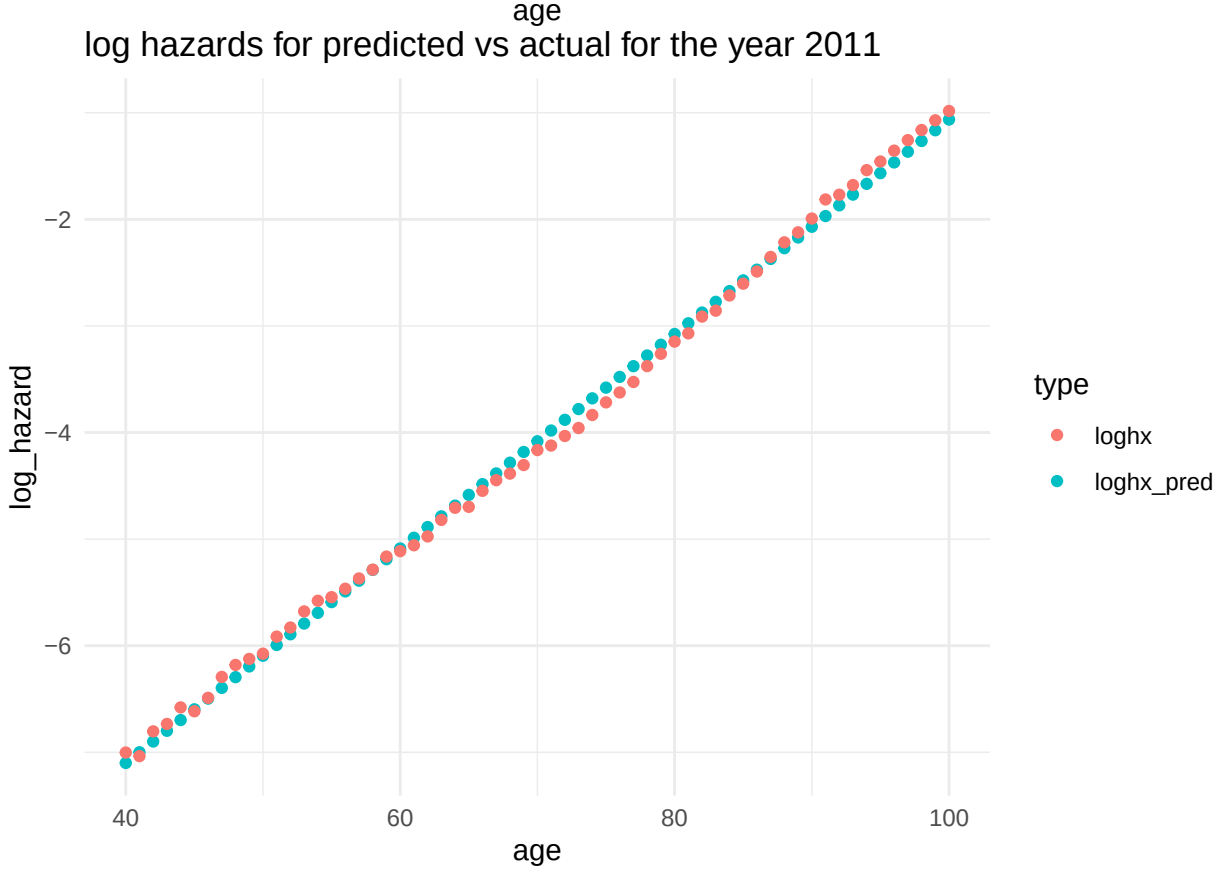

log hazards for predicted vs actual for the year 2006
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred


log hazards for predicted vs actual for the year 2007
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

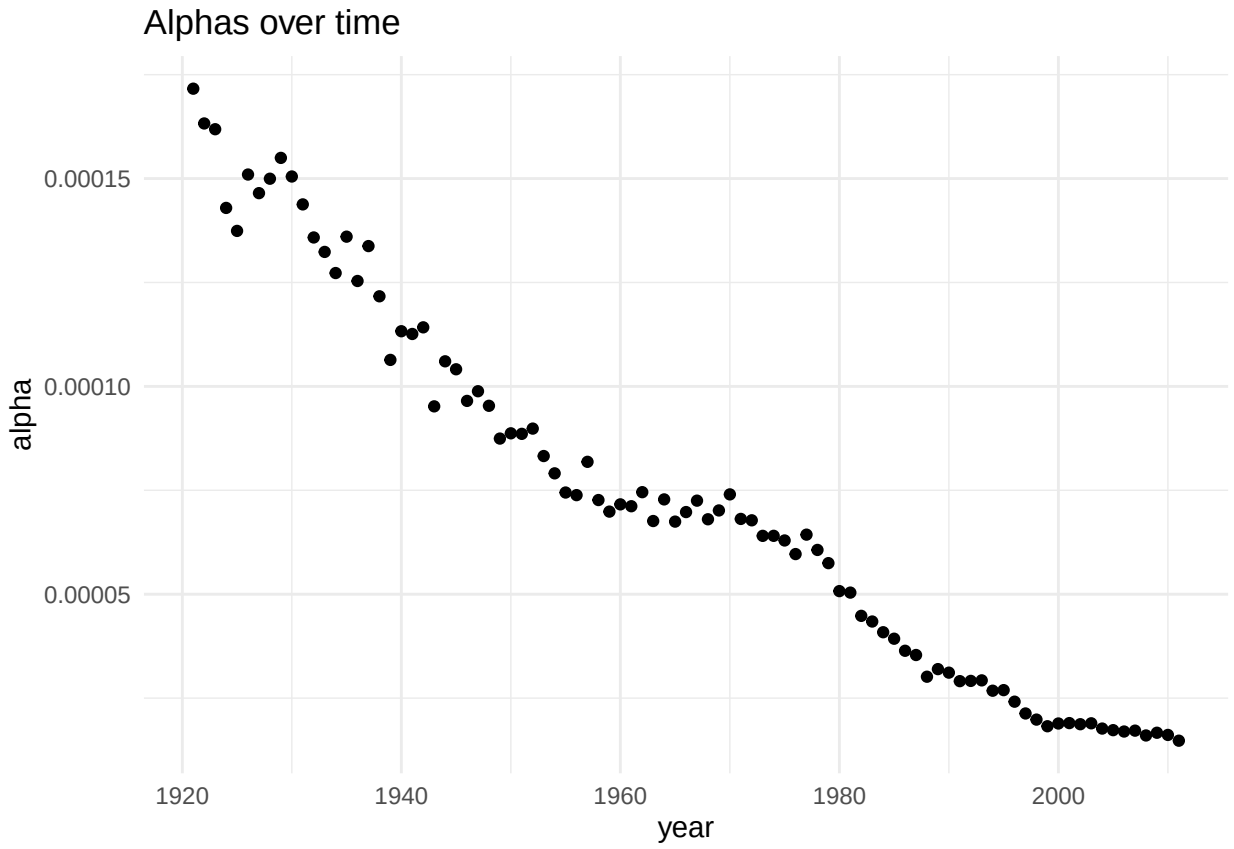

log hazards for predicted vs actual for the year 2008
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred


log hazards for predicted vs actual for the year 2009
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

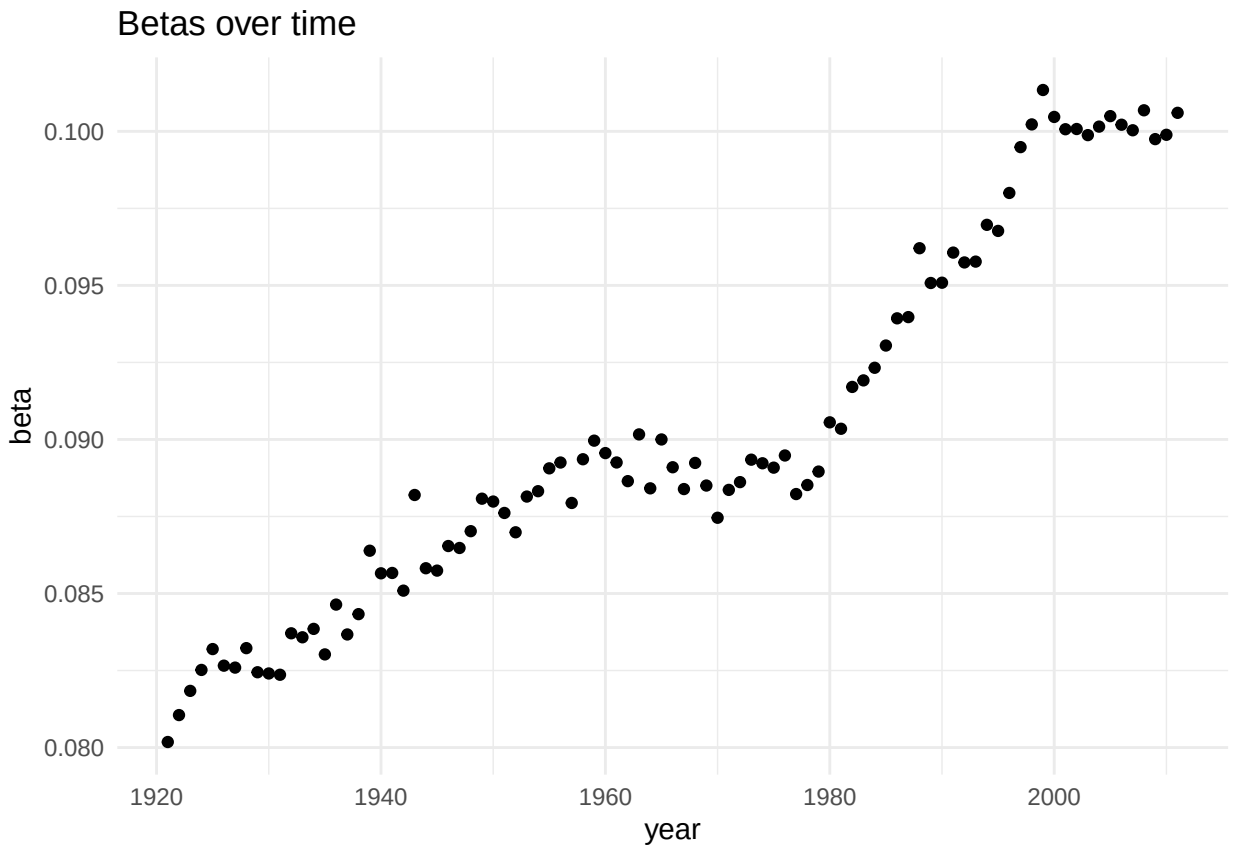

log hazards for predicted vs actual for the year 2010
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred


log hazards for predicted vs actual for the year 2011
log_hazard
−2
−4
−6
40
60
80
100
age
type
loghx
loghx_pred

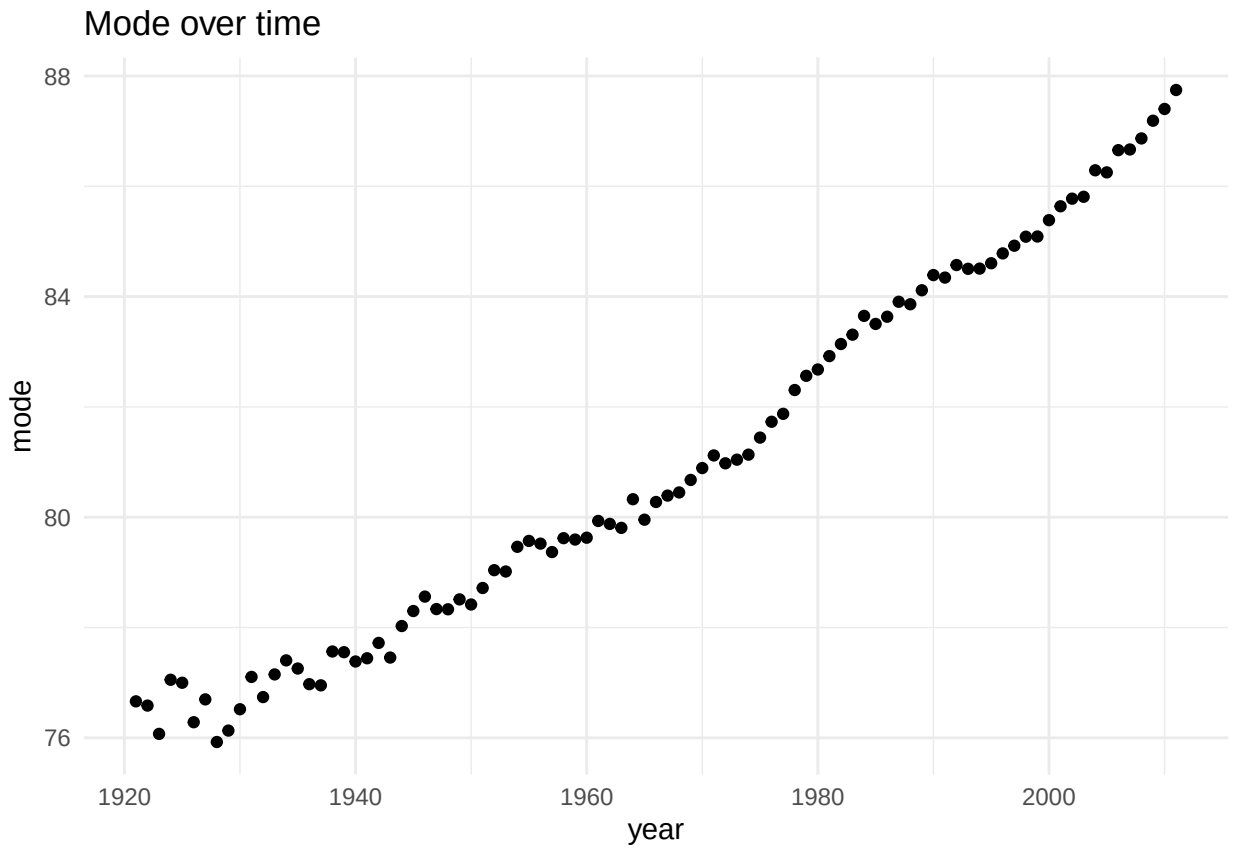
```r
df_all_years <- tibble(year = years, alpha = alphas,beta = betas) %>%
    mutate(mode = log(beta / alpha) / beta)

df_all_years %>%
    ggplot(aes(x = year, y = alpha)) +
    geom_point() +
    theme_minimal() +
    labs(title = "Alphas over time")
```




```r
df_all_years %>%
    ggplot(aes(x = year, y = beta)) +
    geom_point() +
    theme_minimal() +
    labs(title = "Betas over time")
```




```
df_all_years %>%
    ggplot(aes(x = year, y = mode)) +
    geom_point() +
    theme_minimal() +
    labs(title = "Mode over time")
```



Alphas are decreasing over time quite steadily and Betas are increasing over time almost as steadily. Mode over time increases very steadily.

This means that people now have a lower starting off mortality that increases a bit faster with age. Overall modal age of death is increasing so people live longer.

Question 2

```
deaths <- read_rds(here("data", "infant.RDS"))
births <- read_rds(here("data", "births.RDS"))
```

a

```
tot_deaths <- deaths %>% group_by(race) %>% summarize(deaths = n())
tot_births <- births %>% group_by(race) %>% summarize(births = sum(births))

summary_tbl <- left_join(tot_deaths, tot_births) %>%
    mutate(IMR = deaths / births)
```

```
## Joining, by = "race"
```

```
kableExtra::kable(summary_tbl, format = "html")
```

race

deaths

births

IMR

NHB

6407

582587

0.0109975

NHW

10617

2132442

0.0049788

```
mortalities <- summary_tbl %>% pull(IMR)
ratio <- mortalities[1] / mortalities[2]
```

So we have that the ratio of black to white mortality is 2.2088659

b

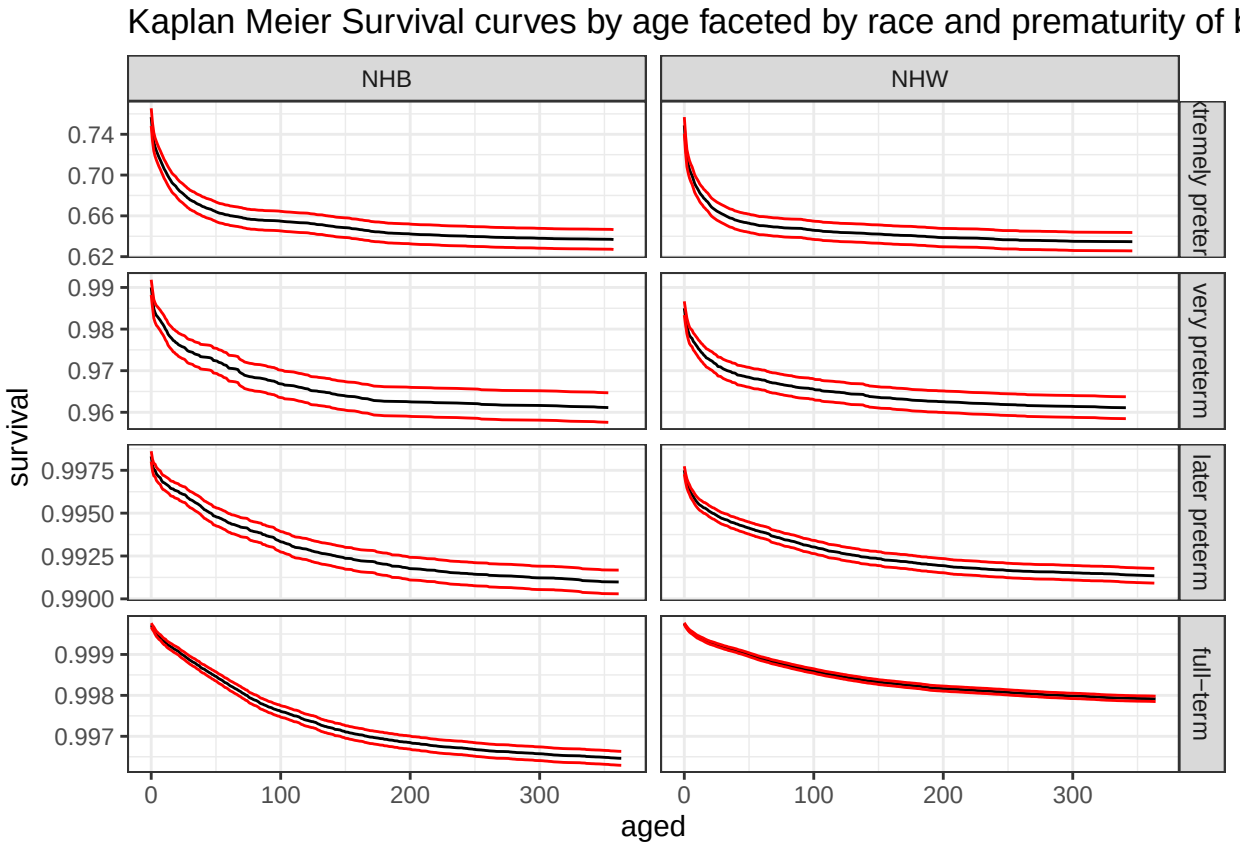
```
library(survival)

deaths_sum <- deaths %>% group_by(race, prematurity, aged) %>% summarize(deaths = n())

km_df <- deaths_sum %>% mutate(deaths_tot = cumsum(deaths)) %>%
    left_join(births) %>%
    mutate(exposure = births - deaths_tot + deaths,
           hazard = deaths / exposure,
           survival = cumprod(1 - hazard),
           var_comp = hazard / (exposure - deaths),
           variance  = survival^2 * cumsum(var_comp),
           survival_plus_two = survival + 2 * sqrt(variance),
           survival_minus_two = survival - 2 * sqrt(variance)
    )
```

```
## Joining, by = c("race", "prematurity")
```

c

```
# km_df %>% filter(race == "NHB") %>%
#   ggplot() +
#   aes(x = aged, y = survival, m = survival_plus_two, n = survival_minus_two,  facet = prematurity) +
#   geom_path() +
#   facet_grid(prematurity~., scales = "free_y") +
#   geom_line(aes(y = survival_plus_two), color = "red") +
#   geom_line(aes(y = survival_minus_two),color = "red") +
#   theme_bw() +
#   labs(title = "NHB")
#
# km_df %>% filter(race != "NHB") %>%
#   ggplot() +
#   aes(x = aged, y = survival, m = survival_plus_two, n = survival_minus_two,  facet = prematurity) +
#   geom_path() +
```

```
#   facet_grid(prematurity~., scales = "free_y") +
#   geom_line(aes(y = survival_plus_two), color = "red") +
#   geom_line(aes(y = survival_minus_two),color = "red") +
#   theme_bw() +
#   labs(title = "NHW")

km_df %>%
    ggplot() +
    aes(x = aged, y = survival, m = survival_plus_two, n = survival_minus_two,  facet = prematurity) +
    geom_path() +
    facet_grid(prematurity~race, scales = "free_y") +
    geom_line(aes(y = survival_plus_two), color = "red") +
    geom_line(aes(y = survival_minus_two),color = "red") +
    theme_bw() +
    labs(title = "Kaplan Meier Survival curves by age faceted by race and prematurity of birth")
```



d

Because there are just so god damn many full-term babies. The curve for NHB for those is not too much lower than for the NHW babies but there are just so many of them that this probably skews the entire statistic.

e

Because the births are not equal so higher mortality is not the same as high number of deaths. Therefore since there are more NHW in general the number of deaths will be much higher.

f

```
pch <- deaths %>% mutate(event = 1,
                         aged = if_else(aged == 0, 1e-20, aged))
cuts <- c(1, 7, 14, 28, 60, 90, 120)
cuts <- cuts - 0.000000000001

pch <- survSplit(Surv(time = pch$aged, event = pch$event)~race + prematurity + 1,data = pch, cut = cuts
    mutate(interval = factor(tstart), interval_len = tstop - tstart)

E_d <- pch %>% group_by(race, prematurity, interval) %>%
    summarize(E_d = sum(interval_len)) %>% pull(E_d)

pch_agg <- pch %>% ungroup() %>% group_by(race, prematurity, interval) %>%
    summarize(deaths = sum(event),
              interval_len = max(interval_len)) %>%
    mutate(deaths_sum = cumsum(deaths))

exposures_df <- left_join(pch_agg, births) %>% group_by(race, prematurity) %>%
    summarise(alive_at_end = max(births) - sum(deaths))
```

```
## Joining, by = c("race", "prematurity")
```

```
pch_agg <- left_join(pch_agg, exposures_df) %>%
    mutate(E_a = alive_at_end * interval_len) %>%
    left_join(births) %>%
    rename(total = births)
```

```
## Joining, by = c("race", "prematurity")
## Joining, by = c("race", "prematurity")
```

```
pch_agg$E_d <- E_d

pch_agg <- pch_agg %>%
    mutate(E_k = E_d + E_a) %>%
    rename(D_k = deaths)
```

This was absolutely harrowing

```
#
# t_start <- c(0, 1, 7, 14, 28, 60, 90, 120, 365)
# lengths <- c(1, 6, 7, 14, 32, 30, 30, 245)
#
# km_df$interval <- cut(km_df$aged, t_start, right = FALSE)
#
# km_df <- km_df %>% ungroup() %>%
#   mutate(interval_lenght = case_when(aged < 1 ~ 1,
#                                      aged < 7 ~ 6,
#                                      aged < 14 ~ 7,
#                                      aged < 28 ~ 14,
#                                      aged < 60 ~ 32,
#                                      aged < 90 ~ 30,
#                                      aged < 120 ~ 30,
#                                      TRUE ~ 244))
#
# interval_df <- km_df %>% group_by(race, prematurity, interval, interval_lenght, births) %>%
```

```
#   summarize(interval_deaths = sum(deaths))
#
# exposures_df <- interval_df %>% group_by(race, prematurity) %>%
#   summarise(alive = max(births) - sum(interval_deaths))
#
# interval_df <- left_join(interval_df, exposures_df) %>%
#   mutate(E_a = alive * interval_lenght) %>%
#   left_join(births) %>%
#   rename(total = births) %>%
#   mutate(E_d = total - interval_deaths,
#          E_k = E_d + E_a)
#
```

```
glm_godplzno <- glm(formula = D_k ~ race + prematurity + interval + race*interval + prematurity*interval
summary(glm_godplzno)
```

```
##
## Call:
## glm(formula = D_k ~ race + prematurity + interval + race * interval +
##     prematurity * interval - 1, family = "poisson", data = pch_agg,
##     offset = (log(E_k)))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.78820  -0.89474  -0.00277   0.94833   2.94402
##
## Coefficients:
##                                                  Estimate Std. Error  z value
## raceNHB                                          -1.15397    0.01956  -59.007
## raceNHW                                          -1.07678    0.01771  -60.788
## prematurityvery preterm                          -3.20993    0.04960  -64.717
## prematuritylater preterm                         -4.99018    0.04361 -114.438
## prematurityfull-term                             -7.18830    0.04331 -165.982
## interval0.999999999999                           -3.47106    0.04627  -75.013
## interval6.999999999999                           -4.37403    0.06405  -68.288
## interval13.999999999999                          -4.96323    0.06170  -80.445
## interval27.999999999999                          -5.80939    0.06224  -93.344
## interval59.999999999999                          -7.05884    0.11401  -61.917
## interval89.999999999999                          -7.35071    0.12776  -57.536
## interval119.999999999999                         -8.00514    0.06690 -119.657
## raceNHW:interval0.999999999999                    0.02397    0.05366    0.447
## raceNHW:interval6.999999999999                   -0.14696    0.07168   -2.050
## raceNHW:interval13.999999999999                  -0.31906    0.06759   -4.721
## raceNHW:interval27.999999999999                  -0.65440    0.05766  -11.349
## raceNHW:interval59.999999999999                  -0.75305    0.06908  -10.901
## raceNHW:interval89.999999999999                  -0.54337    0.07992   -6.799
## raceNHW:interval119.999999999999                 -0.53785    0.05096  -10.555
## prematurityvery preterm:interval0.999999999999    1.08307    0.08723   12.417
## prematuritylater preterm:interval0.999999999999   1.15304    0.07645   15.083
## prematurityfull-term:interval0.999999999999       1.58488    0.06965   22.754
## prematurityvery preterm:interval6.999999999999    1.18617    0.11722   10.119
## prematuritylater preterm:interval6.999999999999   1.22197    0.10426   11.720
## prematurityfull-term:interval6.999999999999       2.04903    0.08623   23.762
## prematurityvery preterm:interval13.999999999999   1.21639    0.11633   10.457
```

```
## prematuritylater preterm:interval13.999999999999   1.27886    0.10302   12.414
## prematurityfull-term:interval13.999999999999       2.35064    0.08243   28.518
## prematurityvery preterm:interval27.999999999999    1.42286    0.11852   12.005
## prematuritylater preterm:interval27.999999999999   2.02027    0.09321   21.674
## prematurityfull-term:interval27.999999999999       3.24332    0.07929   40.907
## prematurityvery preterm:interval59.999999999999    2.37472    0.16547   14.351
## prematuritylater preterm:interval59.999999999999   2.98244    0.13976   21.339
## prematurityfull-term:interval59.999999999999       4.36743    0.12533   34.848
## prematurityvery preterm:interval89.999999999999    2.09730    0.19460   10.778
## prematuritylater preterm:interval89.999999999999   2.98084    0.15326   19.450
## prematurityfull-term:interval89.999999999999       4.22254    0.13853   30.481
## prematurityvery preterm:interval119.999999999999   1.74403    0.11684   14.927
## prematuritylater preterm:interval119.999999999999  2.49010    0.09082   27.418
## prematurityfull-term:interval119.999999999999      3.84919    0.08041   47.867
##                                                   Pr(>|z|)
## raceNHB                                            < 2e-16 ***
## raceNHW                                            < 2e-16 ***
## prematurityvery preterm                            < 2e-16 ***
## prematuritylater preterm                           < 2e-16 ***
## prematurityfull-term                               < 2e-16 ***
## interval0.999999999999                             < 2e-16 ***
## interval6.999999999999                             < 2e-16 ***
## interval13.999999999999                            < 2e-16 ***
## interval27.999999999999                            < 2e-16 ***
## interval59.999999999999                            < 2e-16 ***
## interval89.999999999999                            < 2e-16 ***
## interval119.999999999999                           < 2e-16 ***
## raceNHW:interval0.999999999999                      0.6551
## raceNHW:interval6.999999999999                      0.0403 *
## raceNHW:interval13.999999999999                   2.35e-06 ***
## raceNHW:interval27.999999999999                    < 2e-16 ***
## raceNHW:interval59.999999999999                    < 2e-16 ***
## raceNHW:interval89.999999999999                   1.06e-11 ***
## raceNHW:interval119.999999999999                   < 2e-16 ***
## prematurityvery preterm:interval0.999999999999     < 2e-16 ***
## prematuritylater preterm:interval0.999999999999    < 2e-16 ***
## prematurityfull-term:interval0.999999999999        < 2e-16 ***
## prematurityvery preterm:interval6.999999999999     < 2e-16 ***
## prematuritylater preterm:interval6.999999999999    < 2e-16 ***
## prematurityfull-term:interval6.999999999999        < 2e-16 ***
## prematurityvery preterm:interval13.999999999999    < 2e-16 ***
## prematuritylater preterm:interval13.999999999999   < 2e-16 ***
## prematurityfull-term:interval13.999999999999       < 2e-16 ***
## prematurityvery preterm:interval27.999999999999    < 2e-16 ***
## prematuritylater preterm:interval27.999999999999   < 2e-16 ***
## prematurityfull-term:interval27.999999999999       < 2e-16 ***
## prematurityvery preterm:interval59.999999999999    < 2e-16 ***
## prematuritylater preterm:interval59.999999999999   < 2e-16 ***
## prematurityfull-term:interval59.999999999999       < 2e-16 ***
## prematurityvery preterm:interval89.999999999999    < 2e-16 ***
## prematuritylater preterm:interval89.999999999999   < 2e-16 ***
## prematurityfull-term:interval89.999999999999       < 2e-16 ***
## prematurityvery preterm:interval119.999999999999   < 2e-16 ***
## prematuritylater preterm:interval119.999999999999  < 2e-16 ***
```

```
## prematurityfull-term:interval119.999999999999      < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1.9620e+09  on 64  degrees of freedom
## Residual deviance: 1.2211e+02  on 24  degrees of freedom
## AIC: 636.62
##
## Number of Fisher Scoring iterations: 4
```

```
coef_effects <- lapply(split(exp(coef(glm_godplzno)), names(exp(coef(glm_godplzno)))), unname)

fi_ep_nhb <- coef_effects$raceNHB

fi_ep_nhw <- coef_effects$raceNHW

fi_ft_nhb <- coef_effects$raceNHB * coef_effects$`prematurityfull-term`

f120_ft_nhb <- coef_effects$interval119.999999999999 * coef_effects$`prematurityfull-term` * coef_effec

f120_ep_nhb <- coef_effects$interval119.999999999999

f120_ft_nhw <- coef_effects$interval119.999999999999 * coef_effects$`prematurityfull-term` * coef_effec
    coef_effects$`raceNHW:interval119.999999999999`
```

Hazard of dying in first interval for extremely preterm and NHB mother is just 0.3153815

a) Relative hazard is 1.0802449

b) Relative hazard is 7.5536964×10^{-4}

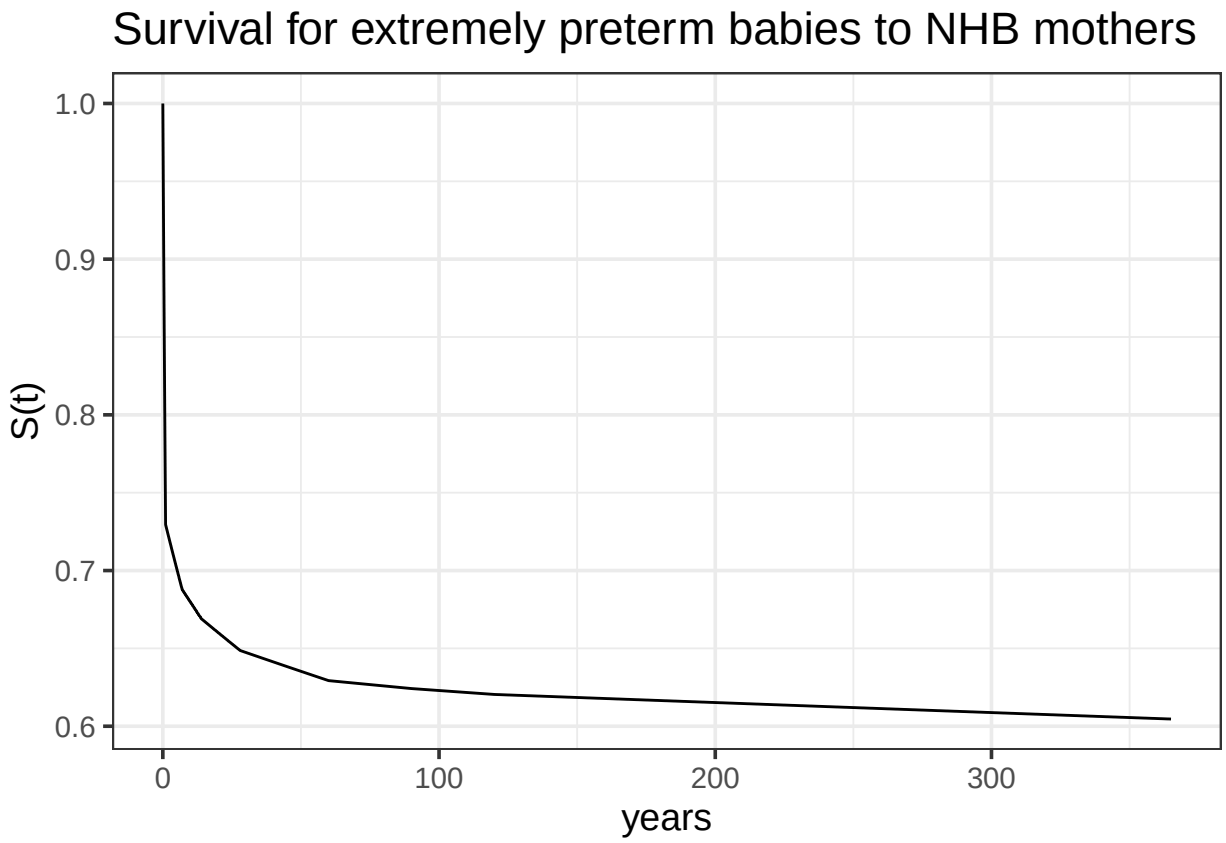
c) Relative hazard is 0.0354684

d) Relative hazard is 0.1989626

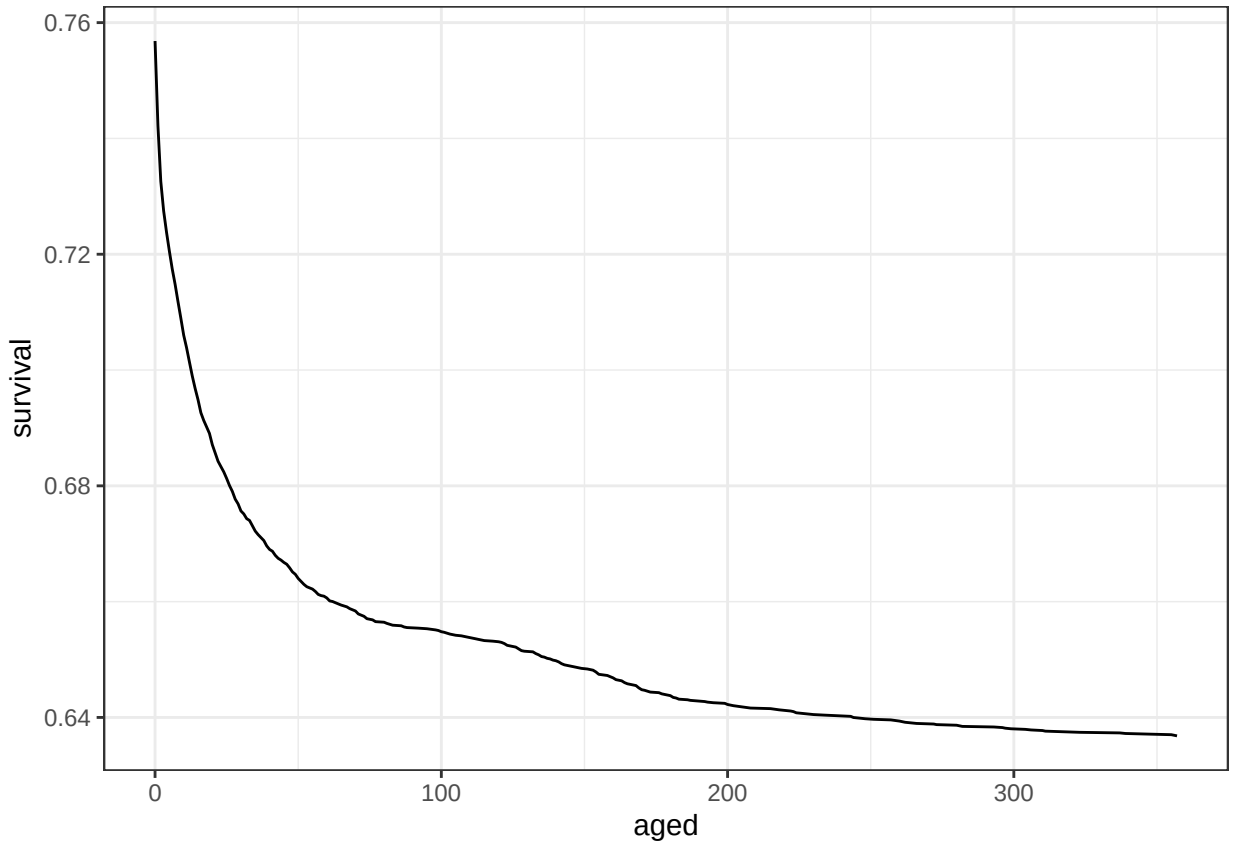
g

```
survival_prob <- function(lambdas,
                     cuts, # start and end times that lambdas refers to, starting at 0 and ending at ma
                     ## observation time of interest,
                     ## thus length is one more than length of lambda
                     neval = 1000 # at how many points do you want to evaluate S(t) within each interva
                     ){
  lengthintervals <- rep((cuts[-1] - cuts[-length(cuts)])/neval, each = neval)
  t_seq <- c(0, cumsum(lengthintervals))
  cumulative_hazard <- cumsum(lengthintervals*rep(lambdas, each  = neval))
  surv_probs <- c(1, exp(-cumulative_hazard))
  return(tibble(time = t_seq, surv = surv_probs ))
}

lambdas <- coef_effects$raceNHB * c(1, coef_effects$interval0.999999999999, coef_effects$interval6.9999
                                    coef_effects$interval27.999999999999,
                                    coef_effects$interval59.999999999999,
                                    coef_effects$interval89.999999999999,
```

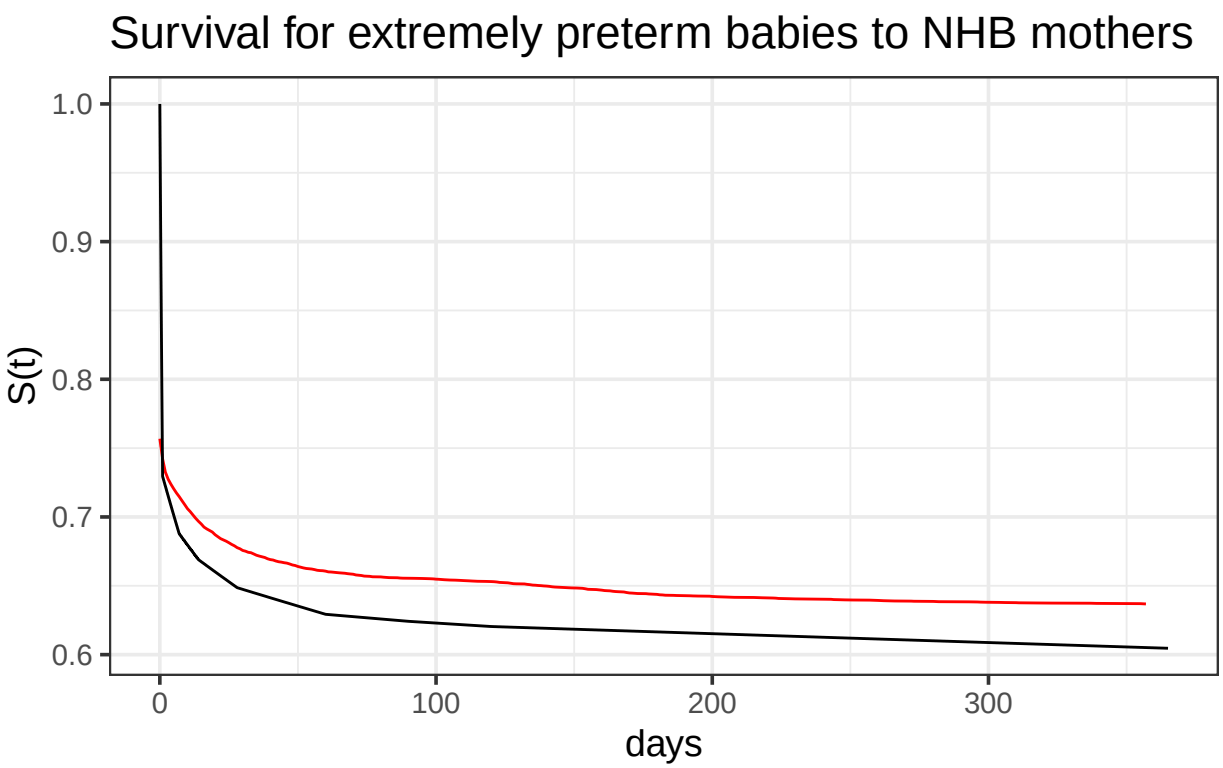
```
                                    coef_effects$interval119.999999999999)
new_cuts = c(0, cuts, 365)

df_surv <- survival_prob(lambdas, new_cuts)

ggplot(aes(time, surv), data = df_surv) +
    geom_line() +
    ggtitle("Survival for extremely preterm babies to NHB mothers") +
    xlab("years") + ylab("S(t)") +
    theme_bw(base_size = 14)
```




```
km_df %>% filter(race == "NHB" & prematurity == "extremely preterm") %>%
    ggplot() +
    aes(x = aged, y = survival, m = survival_plus_two, n = survival_minus_two,  facet = prematurity) +
    geom_line() +
    theme_bw()
```




```
km_df %>%
    filter(race == "NHB" & prematurity == "extremely preterm") %>%
    ggplot() +
    geom_line(aes(x = aged, y = survival), color = "red") +
    geom_line(data = df_surv, aes(x = time, y = surv),  color = "black") +
    ggtitle("Survival for extremely preterm babies to NHB mothers") +
    xlab("days") + ylab("S(t)") +
    theme_bw(base_size = 14) +
    labs(caption = "Kaplan-Meier in red \nPiecewice Constant Hazards in black")
```



Kaplan−Meier in red
Piecewice Constant Hazards in black

BONUS

```
pch_tot <- glm(formula = D_k ~ offset(log(E_k))-1 + interval, data = pch_agg, family = "poisson")
pch_tot2 <- glm(formula = D_k ~ interval-1, data = pch_agg, offset = (log(E_k)), family = "poisson")
summary(pch_tot)
```

```
##
## Call:
## glm(formula = D_k ~ offset(log(E_k)) - 1 + interval, family = "poisson",
##     data = pch_agg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -81.178    1.786    8.452   13.943  149.419
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## interval0                  -5.97395    0.01205  -495.9   <2e-16 ***
## interval0.999999999999     -8.94757    0.02176  -411.2   <2e-16 ***
## interval6.999999999999     -9.80969    0.03101  -316.4   <2e-16 ***
## interval13.999999999999   -10.39066    0.02932  -354.4   <2e-16 ***
## interval27.999999999999   -10.87741    0.02475  -439.6   <2e-16 ***
## interval59.999999999999   -11.25306    0.03085  -364.8   <2e-16 ***
## interval89.999999999999   -11.52811    0.03540  -325.7   <2e-16 ***
## interval119.999999999999 -12.54633    0.02069  -606.5   <2e-16 ***
## ---
```

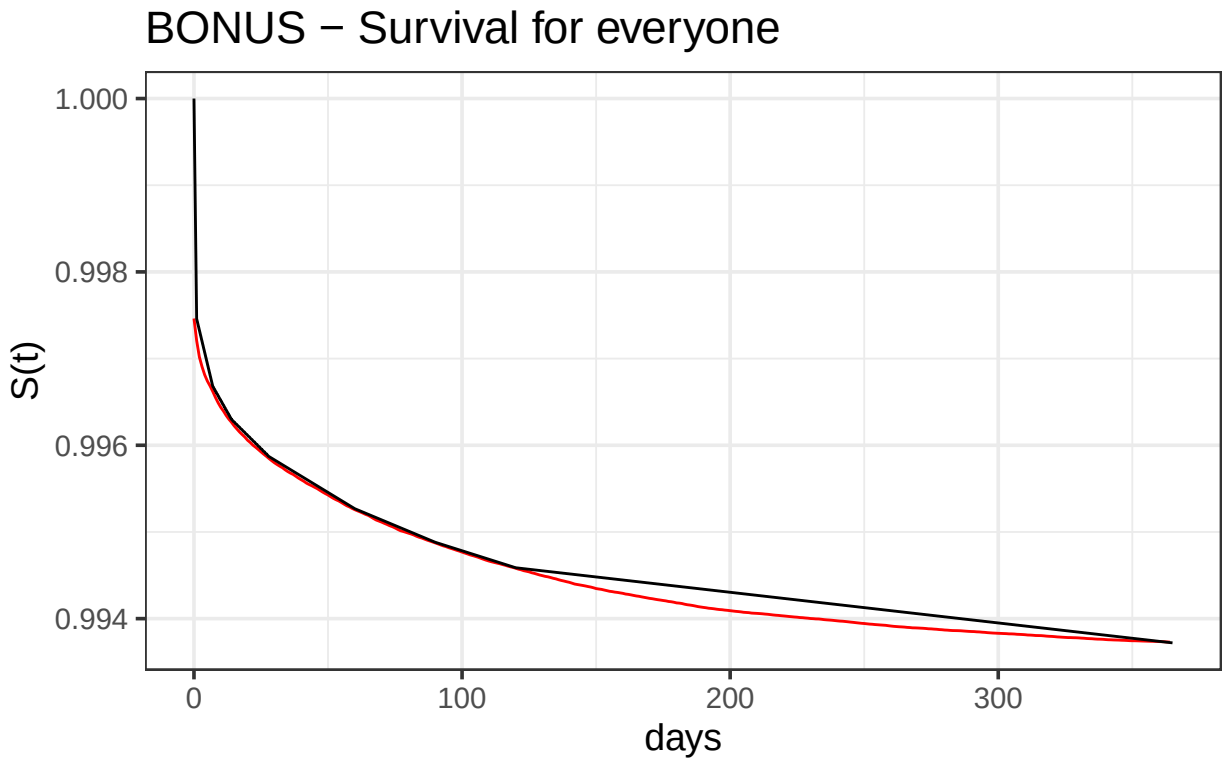
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1962031031  on 64  degrees of freedom
## Residual deviance:      69989  on 56  degrees of freedom
## AIC: 70439
##
## Number of Fisher Scoring iterations: 9
```

```
summary(pch_tot2)
```

```
##
## Call:
## glm(formula = D_k ~ interval - 1, family = "poisson", data = pch_agg,
##     offset = (log(E_k)))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -81.178    1.786    8.452   13.943  149.419
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## interval0                 -5.97395    0.01205  -495.9   <2e-16 ***
## interval0.999999999999    -8.94757    0.02176  -411.2   <2e-16 ***
## interval6.999999999999    -9.80969    0.03101  -316.4   <2e-16 ***
## interval13.999999999999  -10.39066    0.02932  -354.4   <2e-16 ***
## interval27.999999999999  -10.87741    0.02475  -439.6   <2e-16 ***
## interval59.999999999999  -11.25306    0.03085  -364.8   <2e-16 ***
## interval89.999999999999  -11.52811    0.03540  -325.7   <2e-16 ***
## interval119.999999999999 -12.54633    0.02069  -606.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1962031031  on 64  degrees of freedom
## Residual deviance:      69989  on 56  degrees of freedom
## AIC: 70439
##
## Number of Fisher Scoring iterations: 9
```

```
ce <- lapply(split(exp(coef(pch_tot2)), names(exp(coef(pch_tot2)))), unname)

lambdas <- exp(coef(pch_tot2))
df_surv <- survival_prob(lambdas, new_cuts)

deaths_sum <- deaths %>% group_by(aged) %>% summarize(deaths = n())

tot_births <- sum(births$births)

km_df <- deaths_sum %>% mutate(deaths_tot = cumsum(deaths)) %>%
    mutate(births = tot_births,
           exposure = births - deaths_tot + deaths,
           hazard = deaths / exposure,
```

```
            survival = cumprod(1 - hazard),
            var_comp = hazard / (exposure - deaths),
            variance  = survival^2 * cumsum(var_comp),
            survival_plus_two = survival + 2 * sqrt(variance),
            survival_minus_two = survival - 2 * sqrt(variance)
    )



km_df %>%
    ggplot() +
    geom_line(aes(x = aged, y = survival), color = "red") +
    geom_line(data = df_surv, aes(x = time, y = surv),  color = "black") +
    ggtitle("BONUS - Survival for everyone") +
    xlab("days") + ylab("S(t)") +
    theme_bw(base_size = 14) +
    labs(caption = "Kaplan-Meier in red \nPiecewice Constant Hazards in black")
```



Kaplan−Meier in red
Piecewice Constant Hazards in black