Nowacek HW 4-5

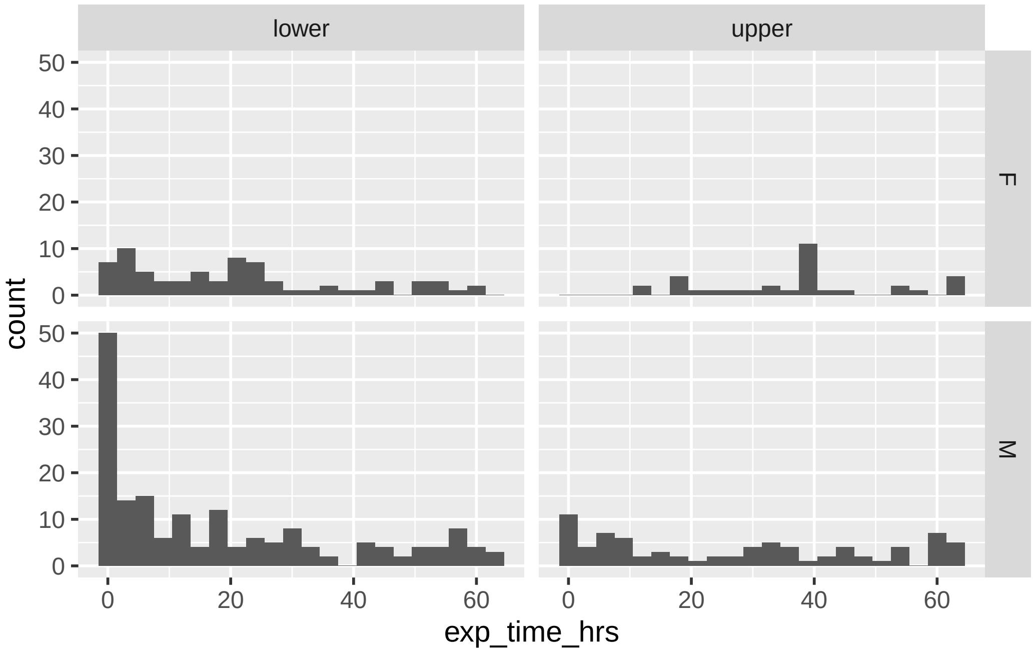
```r
d <- read_csv("https://www2.stat.duke.edu/courses/Fall24/sta490.01/data/got_dat.csv")
```

```
Rows: 359 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): name
dbl (7): id, sex, social_status, intro_time_hrs, dth_flag, event_time_hrs, e...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
d_mod <- d |>
  mutate(
    sex = recode(sex, `1` = "M", `2` = "F"),
    social_status = recode(social_status, `1` = "upper", `2` = "lower"))
```

```r
ggplot(data = d_mod, aes(x = exp_time_hrs)) +
  geom_histogram(binwidth = 3) +
  facet_grid(sex ~ social_status)
```



Problem 2

```
cph <- coxph(Surv(exp_time_hrs, dth_flag) ~ intro_time_hrs + social_status + sex,
             data = d)
summary(cph)
```

```
Call:
coxph(formula = Surv(exp_time_hrs, dth_flag) ~ intro_time_hrs + 
    social_status + sex, data = d)

  n= 359, number of events= 212 

                    coef exp(coef)  se(coef)      z Pr(>|z|)    
intro_time_hrs  0.008965  1.009005  0.004467  2.007 0.044746 *  
social_status   0.332408  1.394322  0.153343  2.168 0.030178 *  
sex            -0.662136  0.515749  0.171495 -3.861 0.000113 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

               exp(coef) exp(-coef) lower .95 upper .95
intro_time_hrs    1.0090     0.9911    1.0002    1.0179
```

```
social_status     1.3943     0.7172     1.0324     1.8832
sex               0.5157     1.9389     0.3685     0.7218

Concordance= 0.627  (se = 0.021 )
Likelihood ratio test= 25.86  on 3 df,   p=1e-05
Wald test            = 24  on 3 df,   p=2e-05
Score (logrank) test = 24.54  on 3 df,   p=2e-05
```

The exponentiated coefficient of sex indicates that female characters are expected to have a hazard rate about 0.544 times as high as male characters, controlling for social status and intro time. The corresponding hypothesis test of that conclusion is:

H_0 : There is no difference between the expected hazard rates of male and female characters with the same social status and intro time.

H_a : There is a difference between the expected hazard rates of male and female characters with the same social status and intro time.

As the coefficient of sex is significant at the $\alpha = 0.05$ level, we can conclude reject the null hypothesis in favor of the alternative that there is a difference between the expected hazard rates of male and female characters with the same social status and intro time.

This conclusion is consistent with that of the AFT model used in the midterm, as that model suggested that:

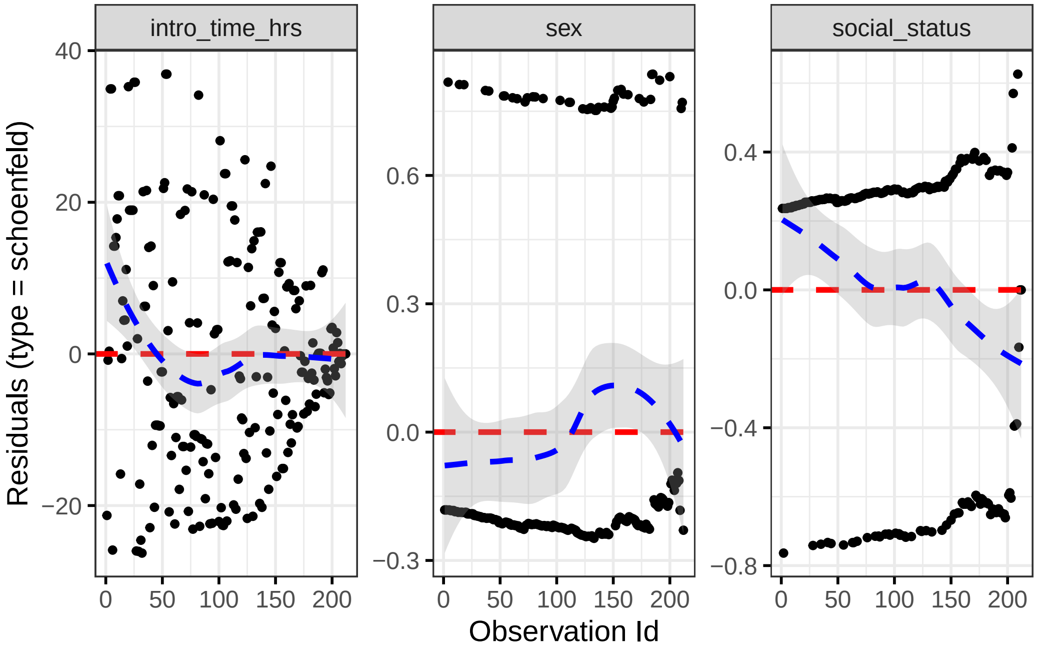
"The coefficient of sex is also positive and significant at the $\alpha = 0.05$ level, indicating that women are expected to live longer than men, controlling for social status and when a character first appears on screen." - from my midterm.

Problem 3

```r
ggcoxdiagnostics(cph, type = "schoenfeld")
```

```
Warning: `gather_()` was deprecated in tidyr 1.2.0.
i Please use `gather()` instead.
i The deprecated feature was likely used in the survminer package.
  Please report the issue at <https://github.com/kassambara/survminer/issues>.
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```
cox.zph(cph)
```

```
                chisq df       p
intro_time_hrs  3.92  1 0.04780
social_status  13.81  1 0.00020
sex             3.76  1 0.05251
GLOBAL         19.77  3 0.00019
```

From the Grambsch and Therneau test, at an alpha = 0.05 level, we do not have significant evidence to reject the null hypothesis of proportional hazards for sex, though it is marginal. So, our test suggests that the assumption of proportional hazards holds for sex, and does not hold for intro time or social status.

Problem 4

```
# Generate baseline hazards
breslow <- basehaz(cph, d, centered = TRUE)

# Pull out the time column
time_col <- breslow$time
```

```r
# Remove the time column using select
breslow <- breslow |>
  select(-time)

# Rename columns using d$name
colnames(breslow) <- d$name

# Bind the time column back as a new column
breslow <- bind_cols(breslow, time = time_col)
```

```
New names:
* `Willa` -> `Willa...268`
* `Willa` -> `Willa...340`
* `Lannister Soldier` -> `Lannister Soldier...348`
* `Lannister Soldier` -> `Lannister Soldier...353`
```

```r
# Filter for the specific time value
filtered_b <- breslow |>
  filter(time == "1.02")

# Transform to long format and create final structure
f_b_long <- filtered_b |>
  pivot_longer(
    cols = -time,
    names_to = "name",  # Create a new column for the names of the original columns
    values_to = "haz"
  ) |>
  mutate(time = filtered_b$time[1]) |>
  select(time, name, haz)  # Include the new 'variable' column
```

```r
f_b_long <- f_b_long |>
  mutate(surv_prob = exp(-(haz)))
```

The dataset f_b_long displays the estimated hazard rate of each individual at time = 1.02.

The three lowest estimated survival probabilities belong to Lannister Soldier (intro time = 62.89),
Unsullied Lieutenant/Captain, and Crying Man in that order. ALso in order, their survival probabilities are 0.7308369, 0.7326807, and 0.7327012 respectively. These slight differences are due to different intro times, as they are all male lower class.

The three highest to Catelyn, Sansa, and Arya Stark. These three each have the same estimated survival probability at 0.9365556 as all of their covariates are the same: intro time of 0.13, females, and highborn.