Nowacek HW 4

```r
d <- read_csv("https://www2.stat.duke.edu/courses/Fall24/sta490.01/data/hw4.csv")
```

```
Rows: 432 Columns: 3
-- Column specification ---------------------------------------------------------
Delimiter: ","
dbl (3): week, arrest, educ

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.
```
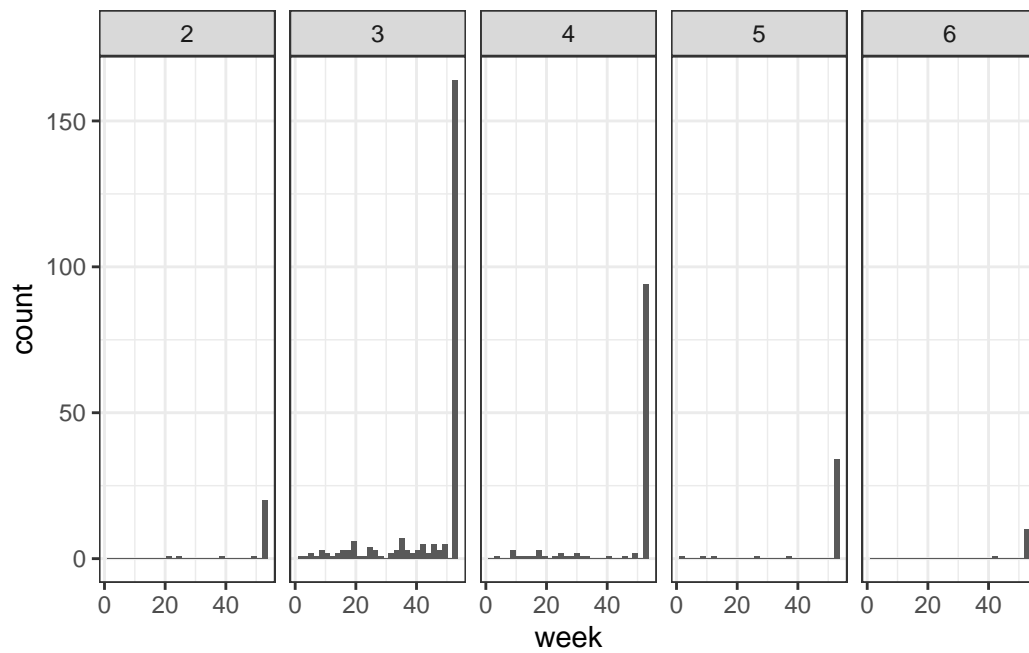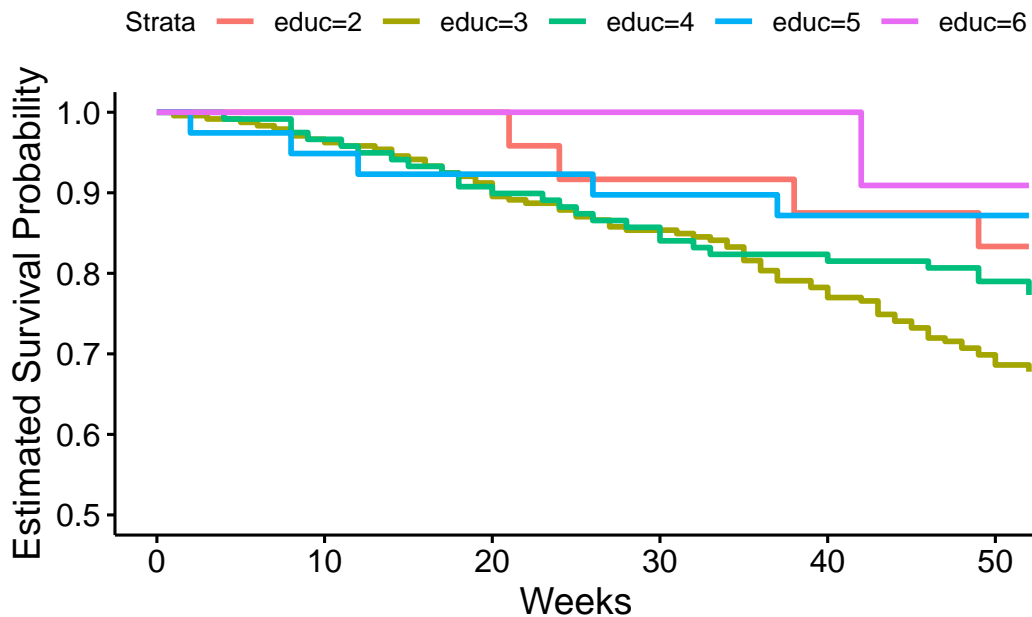
```r
ggplot(aes(x = week), data = d) +
  geom_histogram() +
  facet_grid(~ educ) +
  theme_bw()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
p <- survfit(Surv(week, arrest) ~ educ, data = d)

ggsurvplot(p,
           main = "Kaplan-Meier Estimates by Education Status",
           xlab = "Weeks",
           ylab = "Estimated Survival Probability",
           ylim = c(0.5, 1), conf.int = FALSE, censor = FALSE)
```

```
model <- survdiff(Surv(week, arrest) ~ educ, data = d)
model
```

```
Call:
survdiff(formula = Surv(week, arrest) ~ educ, data = d)

          N Observed Expected (O-E)^2/E (O-E)^2/V
educ=2   24        4     6.81     1.157      1.24
educ=3  239       77    61.54     3.884      8.49
educ=4  119       27    31.59     0.668      0.93
educ=5   39        5    10.78     3.102      3.45
educ=6   11        1     3.28     1.581      1.64

 Chisq= 10.5  on 4 degrees of freedom, p= 0.03
```

This model shows us that there are significant differences in the survival of groups based on education. To get a better idea of this effect, I will condense the education variable into two levels: no high school (groups 2 and 3) and at least some high school (groups 4-6).

```
d <- d |>
  mutate(educ2 = ifelse(educ %in% c(2, 3), 0, 1))

model2 <- survdiff(Surv(week, arrest) ~ educ2, data = d)
model2
```

```
Call:
survdiff(formula = Surv(week, arrest) ~ educ2, data = d)

          N Observed Expected (O-E)^2/E (O-E)^2/V
educ2=0 263       81     68.3      2.34      5.88
educ2=1 169       33     45.7      3.51      5.88

 Chisq= 5.9  on 1 degrees of freedom, p= 0.02
```

This p-value tells us that there is a significant difference between these two groups, indicating that some high-school education vs non-high school education makes a significant impact on time to recidivism.