Pairs Assignment 02/21

Kerem Tuncer and Nikhil Gopal

Question 15.18

Nikhil Gopal uni: nsg2127 Kerem Tuncer uni: kt2716

We will be using a data set from a Canadian study of mortality by age and smoking status.

```
rm(list = ls())
library(rstanarm)
dat <- read.table("https://data.princeton.edu/wws509/datasets/smoking.dat")
```

```
set.seed(1907)
dat$smoke <- as.factor(dat$smoke)
dat <- within(dat, smoke <- relevel(smoke, ref = "no"))
fit_1 <- stan_glm(dead ~ age + smoke, family=poisson(link="log"),
                  offset = log(pop), data = dat, refresh = 0)
print(fit_1)
```

```
## stan_glm
##  family:       poisson [log]
##  formula:      dead ~ age + smoke
##  observations: 36
##  predictors:   12
## ------
##                     Median MAD_SD
## (Intercept)         -3.7    0.1
## age45-59             0.6    0.1
## age50-54             1.0    0.1
## age55-59             1.4    0.1
## age60-64             1.7    0.1
## age65-69             2.0    0.1
## age70-74             2.3    0.1
## age75-79             2.6    0.1
## age80+               2.8    0.1
## smokecigarPipeOnly   0.0    0.0
## smokecigarretteOnly  0.4    0.0
## smokecigarrettePlus  0.2    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

We have modeled the relationship between age and smoking status to death rate per individual. We fit a poisson regression using log(population) as an offset because the population size will effect the amount of

people who die. The regression coefficients tell us the percentage change in death rate for an individual who falls into a certain category.

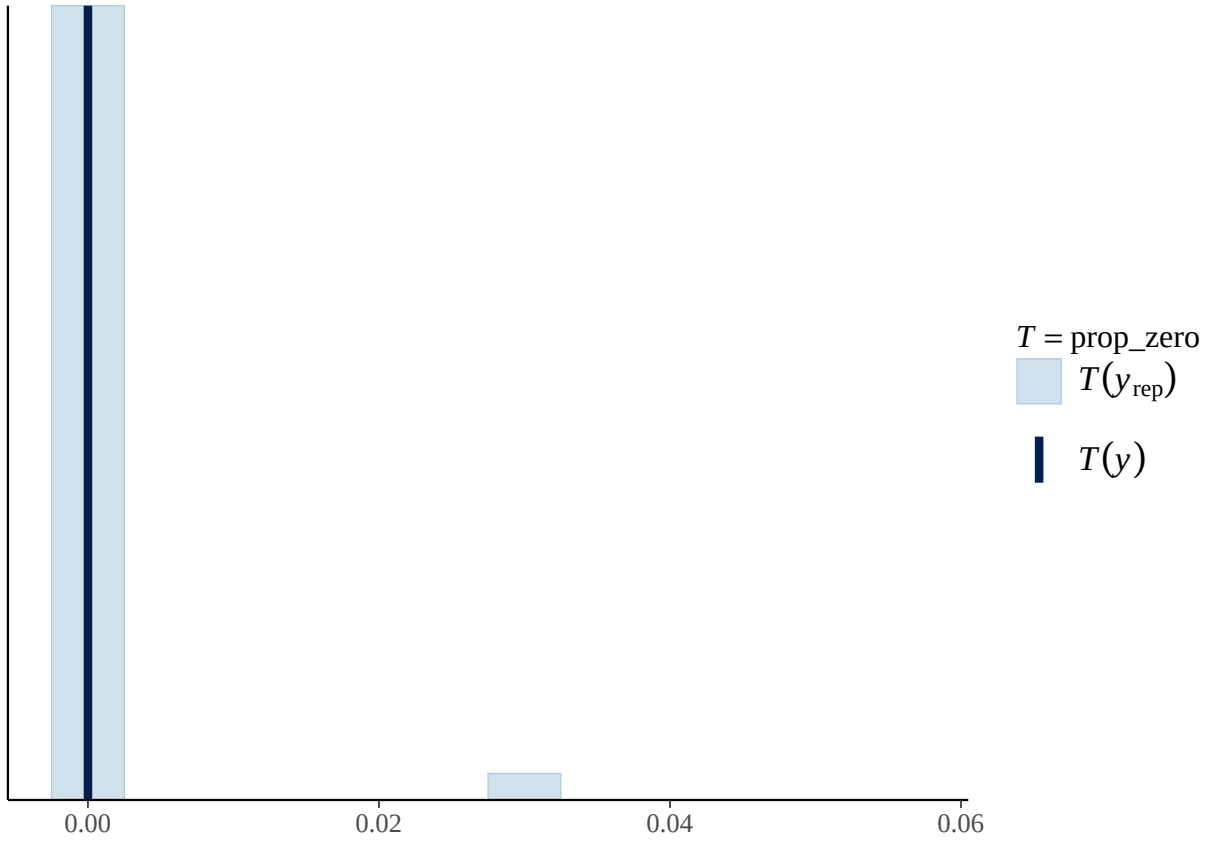
The coefficient estimates of the age categories all have positive signs. Likewise, the coefficient estimate is increasing as the age group increases. This means that the death rate from lung cancer per individual increases as one gets older.

On the other hand, the smokecigarPipeOnly has a coefficient estimate of zero, implying that it causes no change in the death rate per individual from lung cancer.

The coefficient estimate for cigarette-only smokers is 0.4. Thus, the expected multiplicative increase is e^0.4 = 1.49, or a 49% positive difference in the death rate between none smokers and cigarette-only smokers. In other words, the death rate from lung cancer per individual is 49% higher for a cigarette smoker compared to a none smoker.

The coefficient estimate for cigarette and pipe/cigar smokers is 0.4. Thus, the expected multiplicative increase is e^0.2 = 1.22, or a 22% positive difference in the death rate between none smokers and cigarette and cigar/pipe smokers. In other words, the death rate from lung cancer per individual is 22% higher for a cigarette and cigar/pipe smoker compared to a none smoker.

```
yrep <- posterior_predict(fit_1)
prop_zero <- function(y) mean(y == 0)
(prop_zero_test1 <- pp_check(fit_1, plotfun = "stat", stat = "prop_zero", binwidth = .005))
```



The proportion of zeros computed from the sample is the dark blue vertical line and the proportion of zeros from the replicated data sets is shown with the light blue histogram. Given that the proportion of zeroes are similar, we can assume that our model has a good fit.

```
posterior_interval(
  fit_1,
  prob = 0.95,
  type = "central",
  pars = NULL,
  regex_pars = NULL
)
```

```
##                             2.5%      97.5%
## (Intercept)          -3.82112131 -3.5485606
## age45-59              0.39559994  0.7155512
## age50-54              0.82513888  1.1398099
## age55-59              1.24820024  1.5116475
## age60-64              1.52974589  1.7814171
## age65-69              1.87238110  2.1259391
## age70-74              2.14484712  2.4049007
## age75-79              2.42518679  2.6980854
## age80+                2.70735647  2.9981181
## smokecigarPipeOnly   -0.04801075  0.1435653
## smokecigarretteOnly   0.33840126  0.4974590
## smokecigarrettePlus   0.14318531  0.2982731
```

Finally, let's take a look at the uncertainty related to the estimates of smoking status.

The 95% confidence interval for cigar/pipe smokers is [-0.04801075, 0.1435653]. Thus, the confidence interval for the expected multiplicative change is [$e^{-0.04801075,e}$0.1435653] = [0.9531235, 1.1543822]. Therefore, we are 95% confident that the true multiplicative change in the death rate from lung cancer per individual is between -5% to 15% higher for a cigar/pipe smoker compared to a none smoker.

The 95% confidence interval for cigarette smokers is [0.33840126, 0.4974590]. Thus, the confidence interval for the expected multiplicative change is [e^0.33840126, e^0.4974590] = [1.402703, 1.644537]. Therefore, we are 95% confident that the true multiplicative change in the death rate from lung cancer per individual is between 40% to 64% higher for a cigarette smoker compared to a none smoker.

The 95% confidence interval for cigarette and pipe/cigar smokers is [0.14318531, 0.2982731]. Thus, the confidence interval for the expected multiplicative change is [e^0.14318531, e^0.2982731] = [1.153944, 1.347530]. Therefore, we are 95% confident that the true multiplicative change in the death rate from lung cancer per individual is between 15% to 35% higher for a cigarette and pipe/cigar smoker compared to a none smoker.