Week 7: Interactions and Multinomial regression

Monica Alexander

04/03/2021

Overview

We are going to go through two of the models mentioned in the lecture.

Packages:

```
library(tidyverse)
library(here)
library(nnet) # for multinomial
```

Interactions

Using the country indicators dataset again.

```
country_ind <- read_csv(here("data/country_indicators.csv"))
```

Question of interest: in 2017, how is TFR associated with life expectancy and whether or not a country is in a developed region?

Filter the data and create and indicator variable:

```
country_ind_2017 <- country_ind %>%
  filter(year==2017) %>%
  mutate(dev_region = ifelse(region=="Developed regions", "yes", "no"))
```

Run a model with interaction:

```
mod <- lm(tfr ~ life_expectancy + dev_region + life_expectancy*dev_region, data = country_ind_2017)
summary(mod)
```

```
##
## Call:
## lm(formula = tfr ~ life_expectancy + dev_region + life_expectancy *
##     dev_region, data = country_ind_2017)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23326 -0.29618 -0.02426  0.28744  2.54832
```

```
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    13.52646    0.52158  25.933  < 2e-16 ***
## life_expectancy                -0.14454    0.00722 -20.019  < 2e-16 ***
## dev_regionyes                 -12.95159    2.91594  -4.442 1.59e-05 ***
## life_expectancy:dev_regionyes   0.15711    0.03557   4.417 1.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6164 on 172 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7745
## F-statistic: 201.4 on 3 and 172 DF,  p-value: < 2.2e-16
```

Visualizing interactions

Grab coefficients:

```
intercept_non_dev <- coef(mod)[[1]]
slope_non_dev <- coef(mod)[[2]]
intercept_dev <- intercept_non_dev+ coef(mod)[[3]]
slope_dev <- slope_non_dev+ coef(mod)[[4]]
```

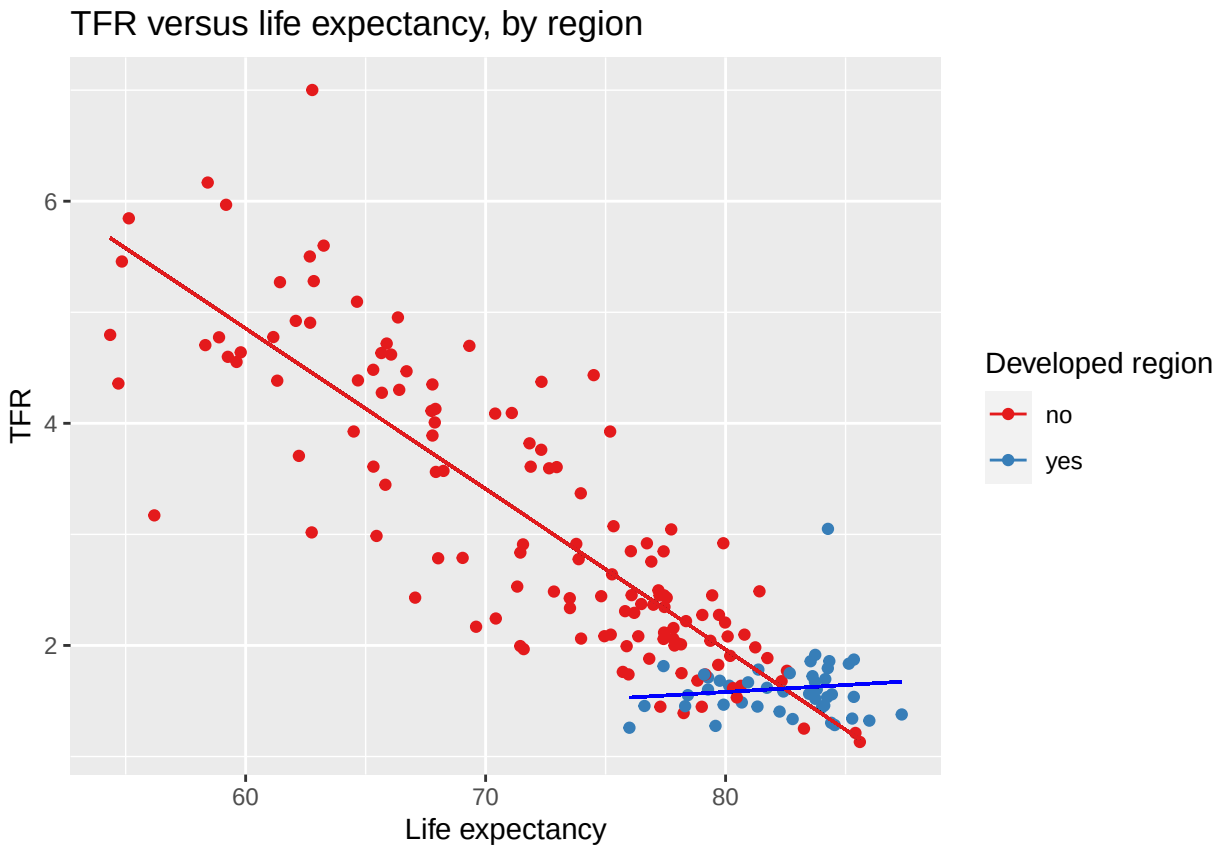
Also need min and max life expectancies by region

```
min_max <- country_ind_2017 %>%
  group_by(dev_region) %>%
  summarize(min = min(life_expectancy), max = max(life_expectancy))
min_max
```

```
## # A tibble: 2 x 3
##   dev_region   min   max
## * <chr>      <dbl> <dbl>
## 1 no          54.4  85.6
## 2 yes         76.0  87.3
```

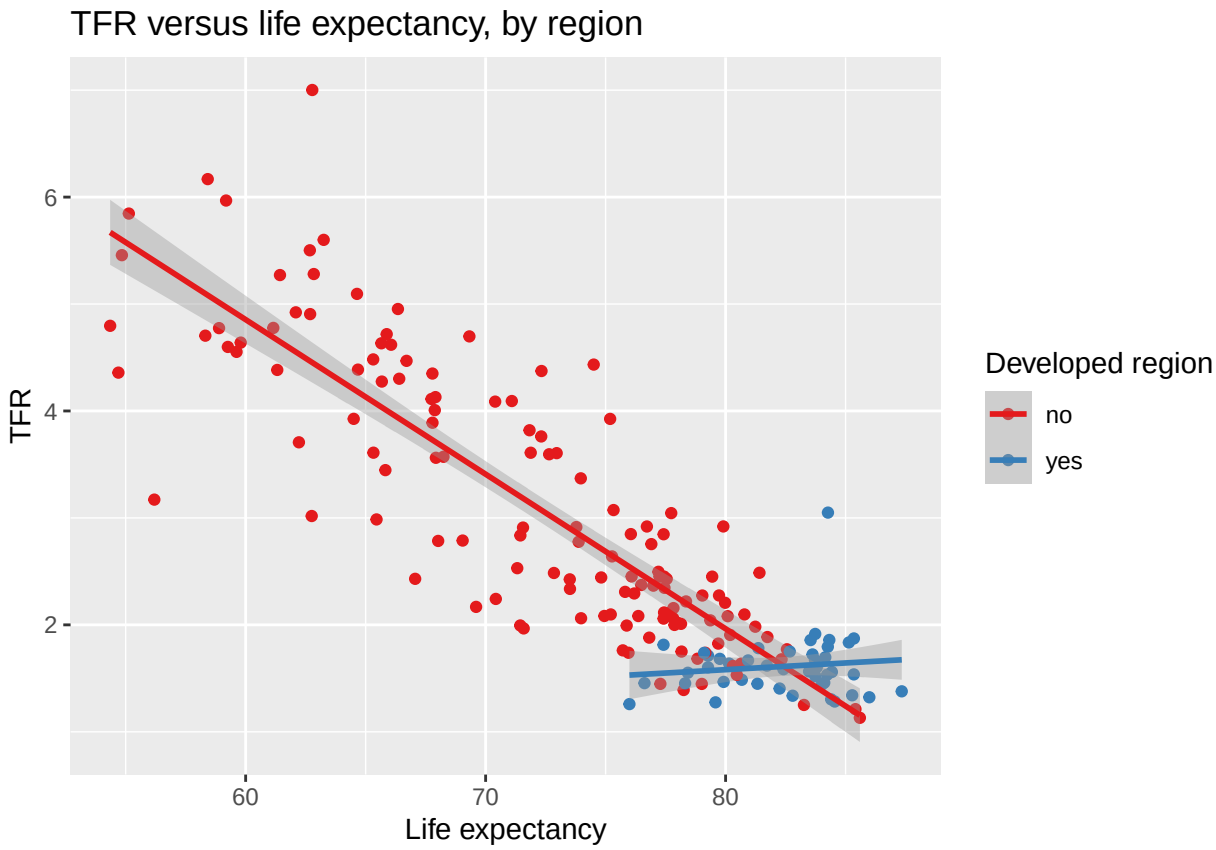
Plot:

```
ggplot(aes(life_expectancy, tfr, color = dev_region), data = country_ind_2017) +
  geom_point() +
  ggtitle("TFR versus life expectancy, by region") +
  ylab("TFR") + xlab("Life expectancy") +
  scale_color_brewer(name = "Developed region", palette = "Set1")+
  geom_segment(aes(x = min_max$min[1], xend = min_max$max[1],
                   y = intercept_non_dev+ slope_non_dev*min_max$min[1],
                   yend = intercept_non_dev+ slope_non_dev*min_max$max[1])) +
  geom_segment(aes(x = min_max$min[2], xend = min_max$max[2],
                   y = intercept_dev+ slope_dev*min_max$min[2],
                   yend = intercept_dev+ slope_dev*min_max$max[2]), color = "blue")
```



A quicker way:

```
ggplot(aes(life_expectancy, tfr, color = dev_region), data = country_ind_2017) +
  geom_point() + geom_smooth(method = "lm") +
  ggtitle("TFR versus life expectancy, by region") +
  ylab("TFR") + xlab("Life expectancy") +
  scale_color_brewer(name = "Developed region", palette = "Set1")
```



Multinomial

Question of interest: how does infant mortality cause of death vary by race, mother's age and prematurity?

Data prep

Read in infant data and do some cleaning:

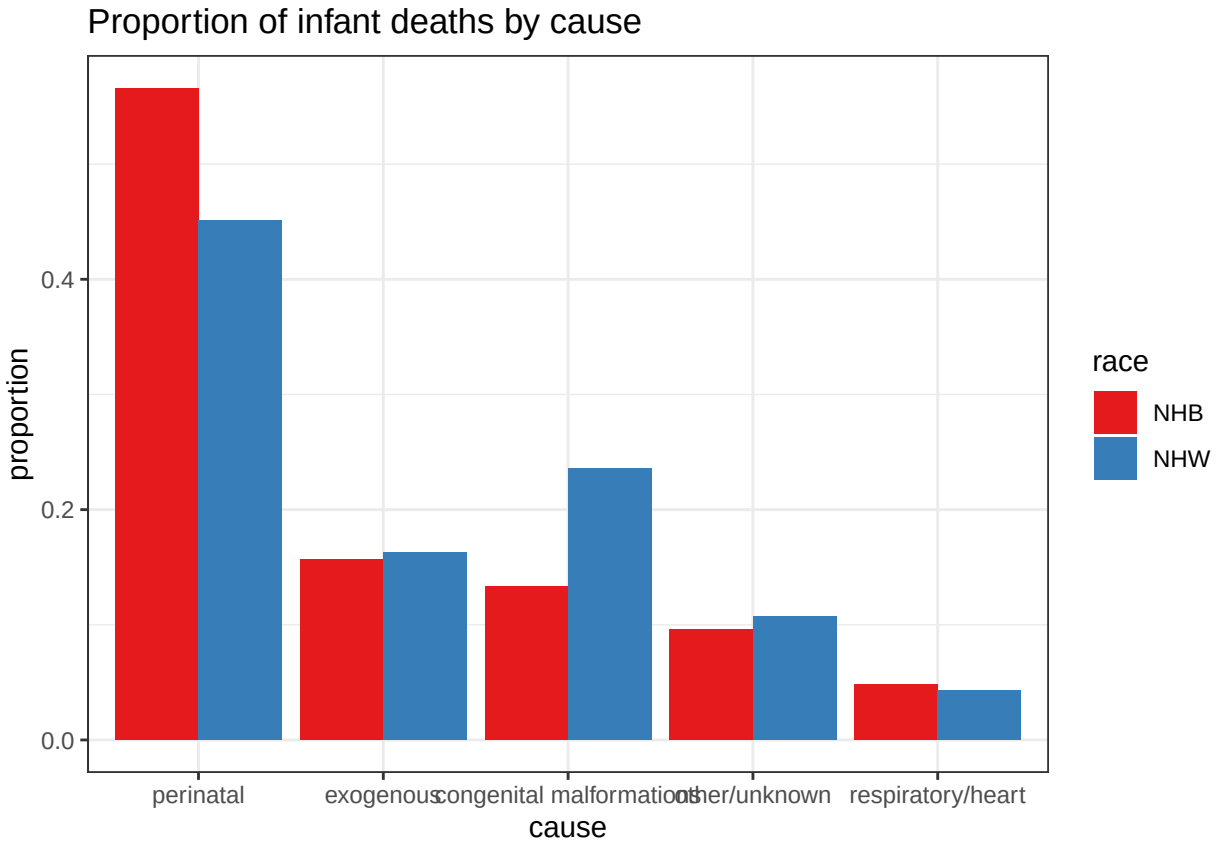
```
d <- read_rds(here("data/infant.RDS"))
d <- d %>%
  mutate(neo_death = ifelse(aged<=28, 1, 0),
         cod_group = case_when(
           str_starts(cod, "peri") ~ "perinatal",
           cod %in% c("other", "unknown") ~ "other/unknown",
           cod %in% c("sids", "maltreatment", "infection") ~ "exogenous",
           cod %in% c("resp", "heart") ~ "respiratory/heart",
           TRUE ~ "congenital malformations"
         ),
         preterm = ifelse(gest<37, 1, 0)) %>%
  filter(gest<99, !is.na(mom_age_group))

infant <- d %>% select(race, mom_age, gest, preterm, cod_group)
infant
```

```
## # A tibble: 17,024 x 5
##    race  mom_age  gest preterm cod_group
##    <chr>   <dbl> <dbl>   <dbl> <chr>
##  1 NHW        30    27       1 perinatal
##  2 NHW        32    36       1 congenital malformations
##  3 NHW        25    44       0 perinatal
##  4 NHB        29    21       1 perinatal
##  5 NHB        23    26       1 perinatal
##  6 NHW        34    39       0 congenital malformations
##  7 NHB        27    26       1 perinatal
##  8 NHB        24    40       0 exogenous
##  9 NHB        39    36       1 congenital malformations
## 10 NHB        30    39       0 congenital malformations
## # ... with 17,014 more rows
```

Graph by race from lecture:

```
infant %>%
  group_by(race, cod_group) %>%
  tally() %>%
  group_by(race) %>%
  mutate(prop = n/sum(n)) %>%
  mutate(cod_group = fct_reorder(cod_group, -prop)) %>%
  ggplot(aes(cod_group, prop, fill  = race)) + geom_bar(stat = "identity", position = 'dodge') +
  labs(title = "Proportion of infant deaths by cause", x = "cause", y = "proportion") +
  theme_bw() +
  scale_fill_brewer(palette = "Set1")
```



Making into wide format

```
infant_wide <- infant %>%
  group_by(race, mom_age, gest, preterm, cod_group) %>%
  tally(name = "deaths") %>%
  pivot_wider(names_from = cod_group, values_from = deaths) %>%
  mutate_all(.funs = funs(ifelse(is.na(.), 0, .)))
head(infant_wide)
```

```
## # A tibble: 6 x 9
## # Groups:   race, mom_age, gest, preterm [6]
##   race  mom_age  gest preterm perinatal exogenous `other/unknown`
##   <chr>   <dbl> <dbl>   <dbl>     <dbl>     <dbl>           <dbl>
## 1 NHB        14    19       1         1         0               0
## 2 NHB        14    21       1         1         0               0
## 3 NHB        14    22       1         1         0               0
## 4 NHB        14    23       1         1         0               0
## 5 NHB        14    24       1         3         1               1
## 6 NHB        14    25       1         1         0               0
## # ... with 2 more variables: `congenital malformations` <dbl>,
## #   `respiratory/heart` <dbl>
```

Making the Y variable:

```
infant_wide$Y <- as.matrix(infant_wide[,c("perinatal",
                                          "exogenous",
                                          "congenital malformations",
                                          "respiratory/heart", "other/unknown")])
head(infant_wide$Y)
```

```
##      perinatal exogenous congenital malformations respiratory/heart
## [1,]         1         0                         0                 0
## [2,]         1         0                         0                 0
## [3,]         1         0                         0                 0
## [4,]         1         0                         0                 0
## [5,]         3         1                         0                 0
## [6,]         1         0                         0                 0
##      other/unknown
## [1,]             0
## [2,]             0
## [3,]             0
## [4,]             0
## [5,]             1
## [6,]             0
```

Regression

```
library(nnet)
mod_mn <- multinom(Y ~ race+ mom_age+ preterm, data = infant_wide)
```

```
## # weights:  25 (16 variable)
## initial  value 27399.071021
## iter  10 value 20149.661320
## iter  20 value 19437.349750
## final  value 19436.462463
## converged
```

```
summary(mod_mn)
```

```
## Call:
## multinom(formula = Y ~ race + mom_age + preterm, data = infant_wide)
##
## Coefficients:
##                           (Intercept)      raceNHW     mom_age   preterm
## exogenous                  2.56320808  0.088345261 -0.05692035 -3.429460
## congenital malformations -0.01647076  0.621524245  0.01916732 -2.423940
## respiratory/heart        -0.15823646 -0.004845986 -0.01780013 -2.251658
## other/unknown             1.10771251  0.145290756 -0.02245255 -3.137589
##
## Std. Errors:
##                           (Intercept)    raceNHW     mom_age    preterm
## exogenous                   0.1235975 0.05354744 0.004365804 0.06000498
## congenital malformations    0.1093744 0.04840430 0.003501902 0.05449309
## respiratory/heart           0.1811151 0.07928810 0.006287607 0.08451183
## other/unknown               0.1361523 0.06022037 0.004717569 0.06546394
##
## Residual Deviance: 38872.92
## AIC: 38904.92
```

Pull out coefficients:

```
coef(mod_mn)
```

```
##                           (Intercept)      raceNHW     mom_age   preterm
## exogenous                  2.56320808  0.088345261 -0.05692035 -3.429460
## congenital malformations -0.01647076  0.621524245  0.01916732 -2.423940
## respiratory/heart        -0.15823646 -0.004845986 -0.01780013 -2.251658
## other/unknown             1.10771251  0.145290756 -0.02245255 -3.137589
```

```
exp(coef(mod_mn))
```

```
##                           (Intercept)   raceNHW   mom_age    preterm
## exogenous                  12.9773831 1.0923652 0.9446693 0.03240443
## congenital malformations    0.9836641 1.8617637 1.0193522 0.08857195
## respiratory/heart           0.8536479 0.9951657 0.9823574 0.10522463
## other/unknown               3.0274253 1.1563757 0.9777976 0.04338727
```

Exercise: plot coefficient estimates and standard errors.

Predicted probabilities

```
predict_df <- tibble(race = rep(c("NHW", "NHB"), each = 2),
       mom_age = 30,
       preterm = rep(c(0,1),2))
preds <- bind_cols(predict_df, as_tibble(predict(mod_mn, newdata = predict_df, type = 'probs')))
preds
```

```
## # A tibble: 4 x 8
##   race  mom_age preterm perinatal exogenous `congenital mal~ `respiratory/he~
##   <chr>   <dbl>   <dbl>     <dbl>     <dbl>            <dbl>            <dbl>
## 1 NHW        30       0     0.110    0.282             0.357           0.0547
## 2 NHW        30       1     0.666    0.0555            0.192           0.0349
## 3 NHB        30       0     0.140    0.329             0.245           0.0700
## 4 NHB        30       1     0.740    0.0564            0.115           0.0390
## # ... with 1 more variable: `other/unknown` <dbl>
```

Plot:

```
preds %>%
  pivot_longer(`perinatal`:`other/unknown`, names_to = "cod_group", values_to = "probability") %>%
  mutate(preterm = ifelse(preterm==1, "pre-term", "full-term")) %>%
  ggplot(aes(race, probability, fill = cod_group)) +
  geom_bar(stat = "identity")+
  facet_grid(~preterm) +
  ggtitle("Predicted probabilities of infant death by race, prematurity and cause\nMothers aged 30")
```