Homework 5 Answer Sketch and Grading Rubric

432 Staff

Due 2018-04-12. Version 2019-04-18

Preliminaries

```
library(skimr)
library(ggrepel)
library(viridis)
library(GGally)
library(nnet)
library(countreg)
library(rms)
library(MASS)
library(broom)
library(tidyverse)

skim_with(numeric = list(hist = NULL),
          integer = list(hist = NULL)) # drop histograms
```

```
ohc <- read.csv("oh_counties_2017.csv") %>% tbl_df
```

1 Question 1 (10 points)

We don't write answer sketches for essay questions.

1.1 Grading Question 1

Award 10 points for an essay that:

1. is focused and responsive to all parts of the prompt
2. says something interesting and thoughtful about the topic
3. is completely clear to you as a reader
4. contains an actual example from the writer's experience that helps explain their idea
5. contains no typographical, syntax or grammar errors
6. is 5-12 sentences long

If an essay fails to meet standard 1, 2, 3, or 4, it should: - lose 3 points for each such standard that is badly missed (for instance, there is no focus to the essay, or the whole thing is unclear.) - lose 2 points for each standard that is partially missed (for instance, a part of the essay is unclear, but most of it is clear, or if part of the prompt is not well covered by the response.) - Deduct 4 points if the essay is less than 5 sentences long, or 1 point for each sentence beyond 12. - Deduct a point for any minor errors in English, and 2-3 points for more egregious errors.

Most students should receive scores between 7 and 9, likely losing points related to clarity, focus or including a meaningful example. I expect no more than 4-5 essays will meet the standard for a "10".

TAs, please provide brief comments indicating strong/weak points on all essays. Thank you.

2 Question 2 (10 points)

Build a reasonable linear or generalized linear model in your development sample (86 counties) to predict one of the outcomes in the `oh_counties_2017.csv` data set that describes a percentage (that must fall between 0 and 100) effectively using at least three and no more than 6 other variables from the list above. Demonstrate how well the model fits as well as the conclusions you draw from the model carefully. Be sure to discuss model assumptions. Then use the model to predict Cuyahoga County and Monroe County results, and assess the quality of those predictions.

2.1 Creating The Development and Test Samples

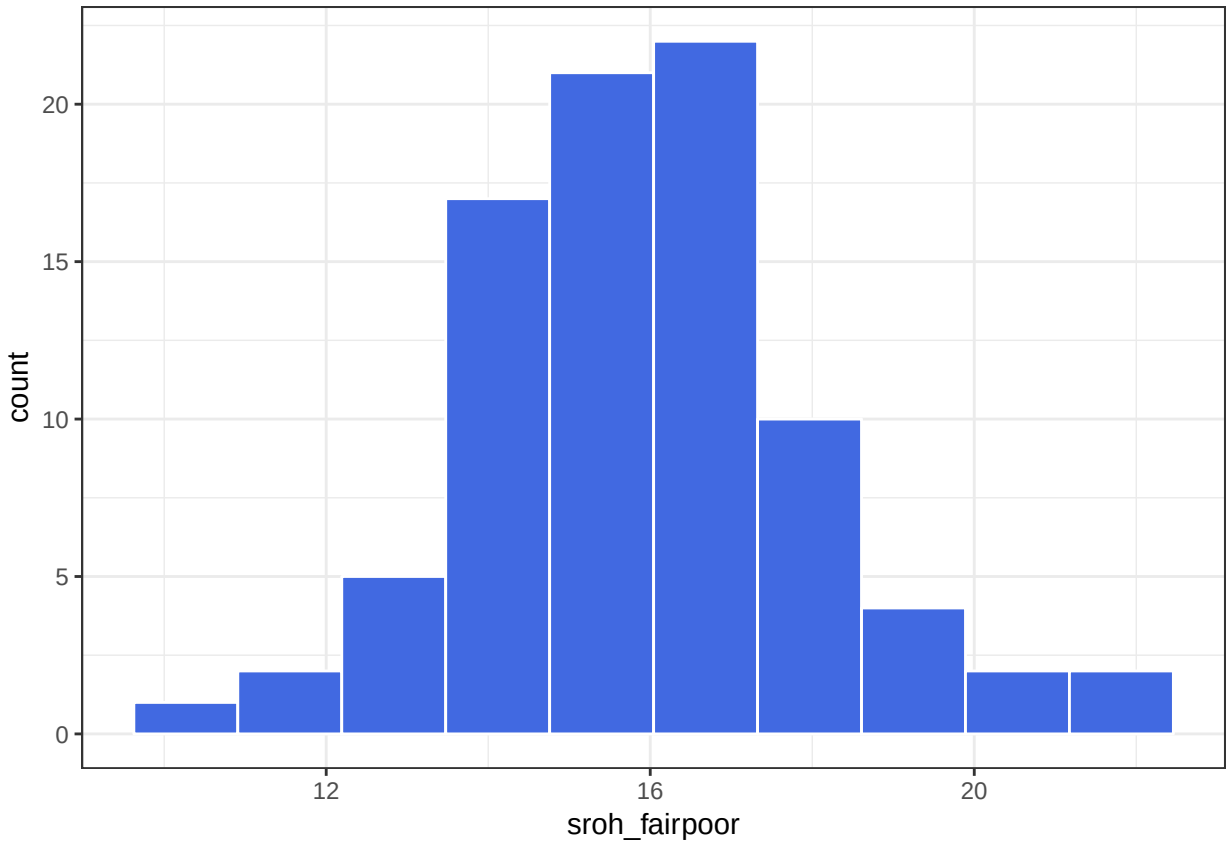
To start, we'll create the two partitions of the original `ohc` data set. There are several potential approaches, but we used:

```
ohc_86 <- ohc %>%
    filter(county != "Cuyahoga" & county != "Monroe")

ohc_2 <- ohc %>%
  filter(county %in% c("Cuyahoga", "Monroe"))
```

2.2 Select and graph an outcome. We pick `sroh_fairpoor`.

Now, using the development sample (`ohc_86`) we are to fit a percentage outcome (between 0 and 100) using 3-6 predictors. We chose `sroh_fairpoor` as our outcome, which is the percentage of adults in each county who report fair or poor health (via BRFSS). One nice feature of this outcome is that we have no values near the floor (0) or ceiling (100) of a percentage outcome, and that it's reasonably fit by a Normal model.

```
ggplot(ohc_86, aes(x = sroh_fairpoor)) +
    geom_histogram(fill = "royalblue", col = "white",
                   bins = 10) +
    theme_bw()
```



2.3 Select and graph/summarize predictors.

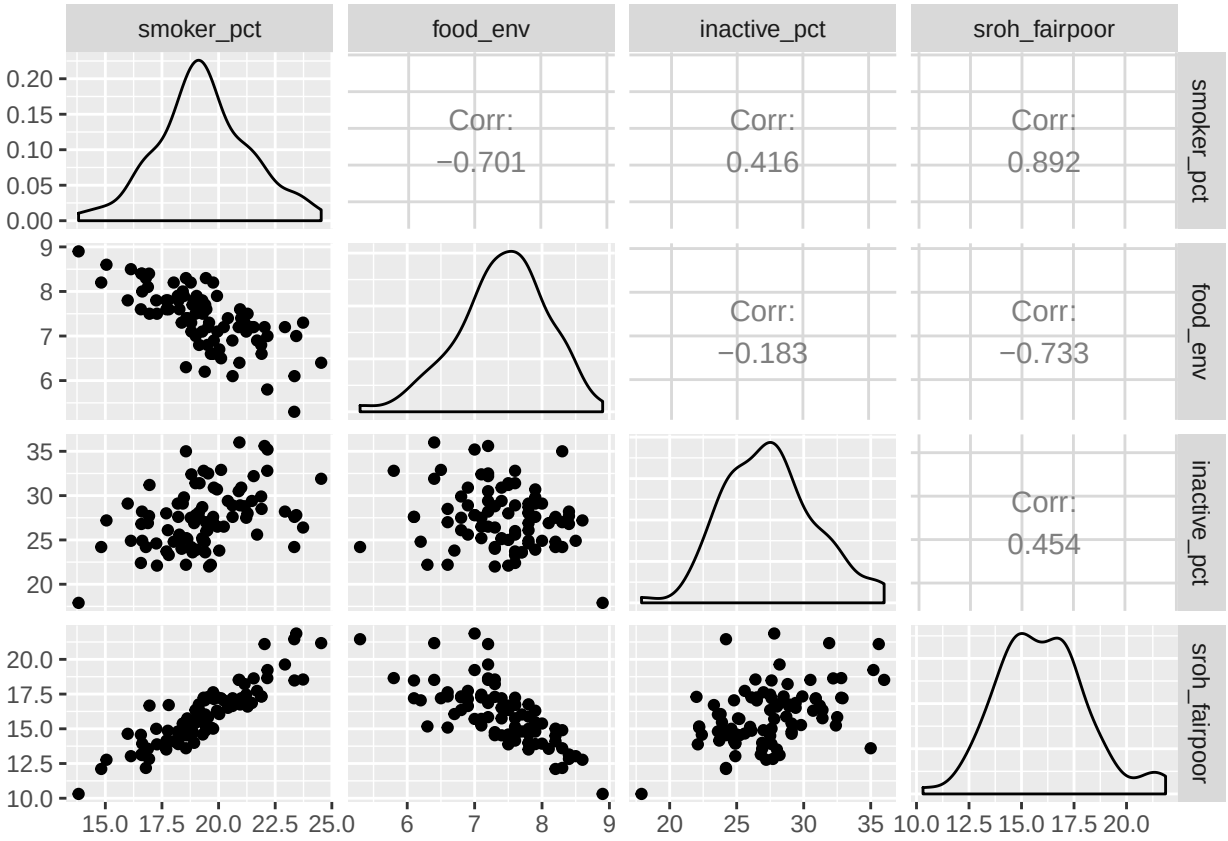
We're going to build a model using these three predictors:

- `smoker_pct`
- `food_env`
- `inactive_pct`

although I'll note that in a few minutes of looking around, you could do slightly better (at least in the training sample) using `median_income` instead of `inactive_pct`.

Here's a scatterplot matrix.

```
ggpairs(data = dplyr::select(ohc_86,
        smoker_pct, food_env, inactive_pct, sroh_fairpoor))
```



Each of the variables seems reasonably symmetric, and the correlations between variables are `sroh_poor` are fairly strong. I would also like to look at some descriptive statistics across the 86 counties in the development sample. One of the reasons to do this is to help me understand whether the two counties I've held out (Cuyahoga and Monroe) look unusual on these predictors.

```
ohc_86 %>% dplyr::select(sroh_fairpoor, smoker_pct,
                         food_env, inactive_pct) %>% skim
```

```
Skim summary statistics
 n obs: 86
 n variables: 4

-- Variable type:numeric ------------------------------------------------------------------------------------------
      variable missing complete  n  mean   sd    p0   p25   p50   p75
      food_env       0       86 86  7.41 0.67  5.3   7.03  7.5   7.8
  inactive_pct       0       86 86 27.49 3.46 17.9  24.9  27.5  29.32
    smoker_pct       0       86 86 19.33 2.07 13.82 18.22 19.28 20.62
 sroh_fairpoor       0       86 86 15.97 2.16 10.31 14.57 15.8  17.23
  p100
  8.9
 36
 24.53
 21.86
```

2.4 Fit a linear regression model

Now, let's fit the model.

```
model_q2 <- lm(sroh_fairpoor ~ smoker_pct + food_env +
                    inactive_pct, data = ohc_86)

summary(model_q2)
```

```
Call:
lm(formula = sroh_fairpoor ~ smoker_pct + food_env + inactive_pct,
    data = ohc_86)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7648 -0.6813 -0.1332  0.4528  2.6848

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.93260    2.57168   2.307 0.023583 *
smoker_pct    0.69797    0.07244   9.636 3.93e-15 ***
food_env     -0.76884    0.20646  -3.724 0.000359 ***
inactive_pct  0.08169    0.03149   2.594 0.011220 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8989 on 82 degrees of freedom
Multiple R-squared:  0.8327,    Adjusted R-squared:  0.8265
F-statistic:   136 on 3 and 82 DF,  p-value: < 2.2e-16
```

Each of the predictors adds statistically significant predictive value to the model (at the 5% level) given all of the others, and the overall R^2 looks pretty good at about 83%. We conclude here that poorer outcomes at the county level (i.e. higher values of `sroh_fairpoor` are associated with higher rates of smoking, worse food environments and higher rates of inactivity (or, if you prefer, lower rates of activity.)

2.5 A look at collinearity

We saw in the scatterplot matrix that there is some potentially meaningful correlation between predictors (collinearity) especially between `smoker_pct` and `food_env`, but by VIF, this doesn't seem to be a severe problem.

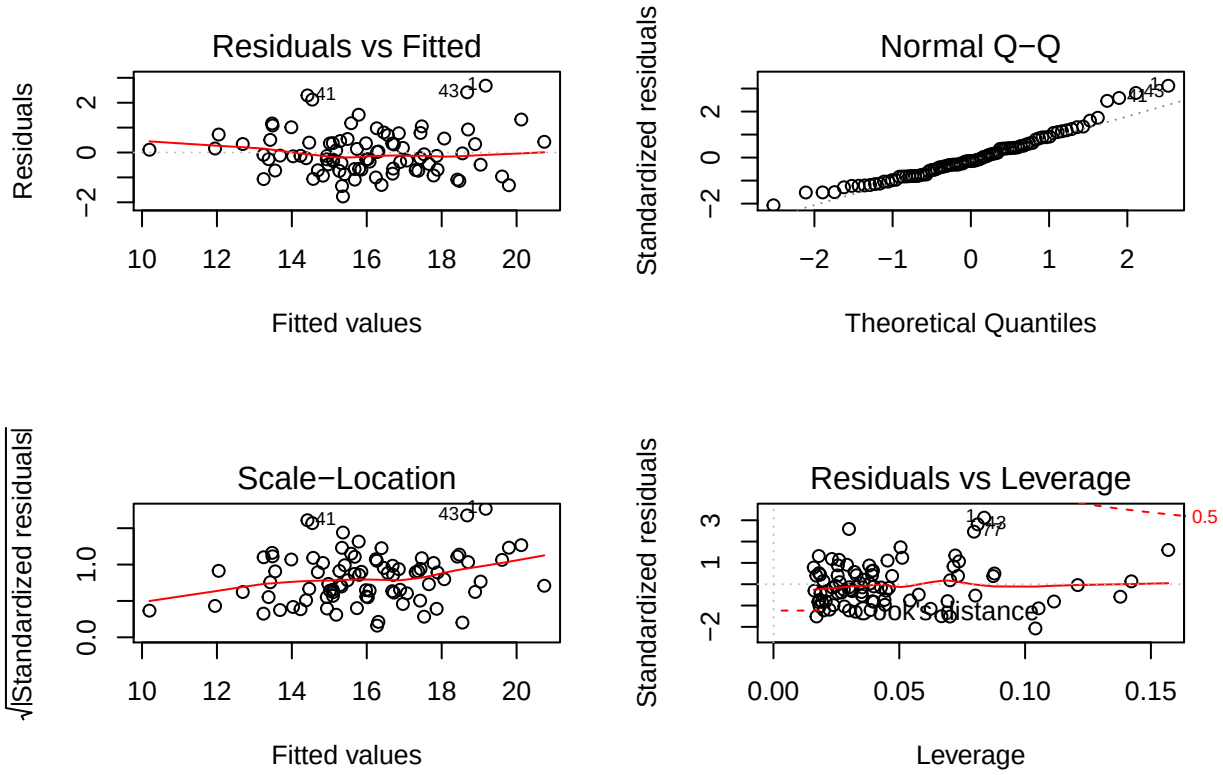
```
rms::vif(lm(sroh_fairpoor ~ smoker_pct + food_env + inactive_pct, data = ohc_86))
```

```
  smoker_pct     food_env inactive_pct
    2.368424     2.025559     1.245054
```

2.6 Considering Regression Assumptions

Looking at regression assumptions, then, we examine the residual plots, where I see no serious problems with assumptions, although a few counties aren't especially well fit.

```
par(mfrow = c(2,2))
plot(model_q2)
```




```
par(mfrow = c(1,1))
```

2.7 Make Predictions for Cuyahoga and Monroe counties

So, now we'll use the model to make predictions for Cuyahoga County and Monroe County. First, let's look at the data from those counties on the variables in our model.

```
ohc_2 %>% dplyr::select(county, sroh_fairpoor, smoker_pct,
                        food_env, inactive_pct)
```

```
# A tibble: 2 x 5
  county   sroh_fairpoor smoker_pct food_env inactive_pct
  <fct>            <dbl>      <dbl>    <dbl>        <dbl>
1 Cuyahoga          17.1       18.7      6.5         24.2
2 Monroe            16.2       19.6      7.2         35.4
```

We note that:

- Cuyahoga and Monroe fall between the median and 75th percentiles of the other 86 counties on `sroh_fairpoor`
- Cuyahoga and Monroe each fall in the middle half (between the 25th and 75th percentiles) of the other 86 counties on `smoker_pct`
- Cuyahoga has a `food_env` in the bottom quarter of the distribution of the other 86 counties, and
- Monroe has an `inactive_pct` in the top quarter of the distribution of the other 86 counties.

Now, using our `model_q2` to make predictions in Cuyahoga and Monroe, we have the following point estimates and 95% prediction intervals.

```
predict(model_q2, newdata = ohc_2, interval = "prediction")
```

```
       fit      lwr      upr
1 15.96415 14.11190 17.81640
2 16.99697 15.13146 18.86249
```

- For Cuyahoga County, the model predicts 15.96% will have a fair or poor self-reported overall health, which is 1.12 percentage points lower than the observed value of 17.08%.
- For Monroe County, the model predicts 17.00%, and the observed value is 16.22%, an error of 0.78 percentage points too high.
- The RMSE within the training sample was 0.89 percentage points, so these errors are of similar magnitude.
- In each case, the observed value is well within the prediction interval from the model.

On the whole, these seem like potentially useful predictions, but it is a shame that the order (Cuyahoga worse than Monroe) is the opposite of what was predicted.

2.8 A Note from Dr. Love

Lots and lots of people (including the teaching assistants) used things like "best subsets" to choose the model from some larger pool of potential predictors. It's not clear to me why. The goal here was to build and evaluate a single model. All we did here was to select a few predictors of interest, and run the model. The goal was not, for instance, to find the best possible model (by some criterion) out of all of the possible models that could be fit.

2.9 Grading Rubric: Question 2

- Award 4 points for fitting a linear regression model to an outcome that is, in fact, a percentage of something.
- Award another 2 points for obtaining predictions using that model for the two held-out counties.
- Take off 1 point from the total if you find two or three typographical or syntax/grammar errors in this response.
- Take off 2 points from the total if you find an especially large number (more than 3, let's say) such errors.
- Award up to 4 additional points if the student did a good job explaining what they did, so that a good answer to the question should get the full 10 points, and a correct but poorly explained response should receive 6 points.

3 Question 3 (15 points)

Divide the 86 counties in your development sample into three groups (low, middle and high) in a rational way in terms of the `years_lost_rate` outcome. Make that new grouping your outcome for an ordinal logistic regression model. Fit a model (using a carefully selected group of no more than 5 predictor variables) and assess its performance carefully. Do not include the `age65plus` variable as a predictor, as the `years_lost_rate` data are age-adjusted already. Demonstrate how well the model fits as well as the conclusions you draw from the model carefully. Then use the model to predict Cuyahoga County and Monroe County results, and assess the quality of those predictions.

3.1 Explore `years_lost_rate` and divide into 3 categories

We could take a numerical approach, where we identify the tertiles (the points which divide the `years_lost_rate` into three groups of equal size.)

```
quantile(ohc_86$years_lost_rate, c(0.33, 0.67))
```

```
   33%    67%
6849.1 8273.6
```

Rounding a little, we might use 6850 and 8275 as cutoffs for our low/middle/high outcome. Let's try that.

```
ohc_86 <- ohc_86 %>%
    mutate(yrslostcat0 =
            Hmisc::cut2(years_lost_rate,
                        cuts = c(6850, 8275)))
ohc_86 %>% count(yrslostcat0)
```

```
# A tibble: 3 x 2
  yrslostcat0       n
  <fct>         <int>
1 [ 4129, 6850)    29
2 [ 6850, 8275)    29
3 [ 8275,12091]    28
```

Note that this factor specifies that the levels are:

- Low: less than 6850
- Middle: at least 6850 and less than 8275
- High: at least 8275

We could use this factor as it is, or recode it to (low/middle/high) but in either case, we want to be sure this is ordered.

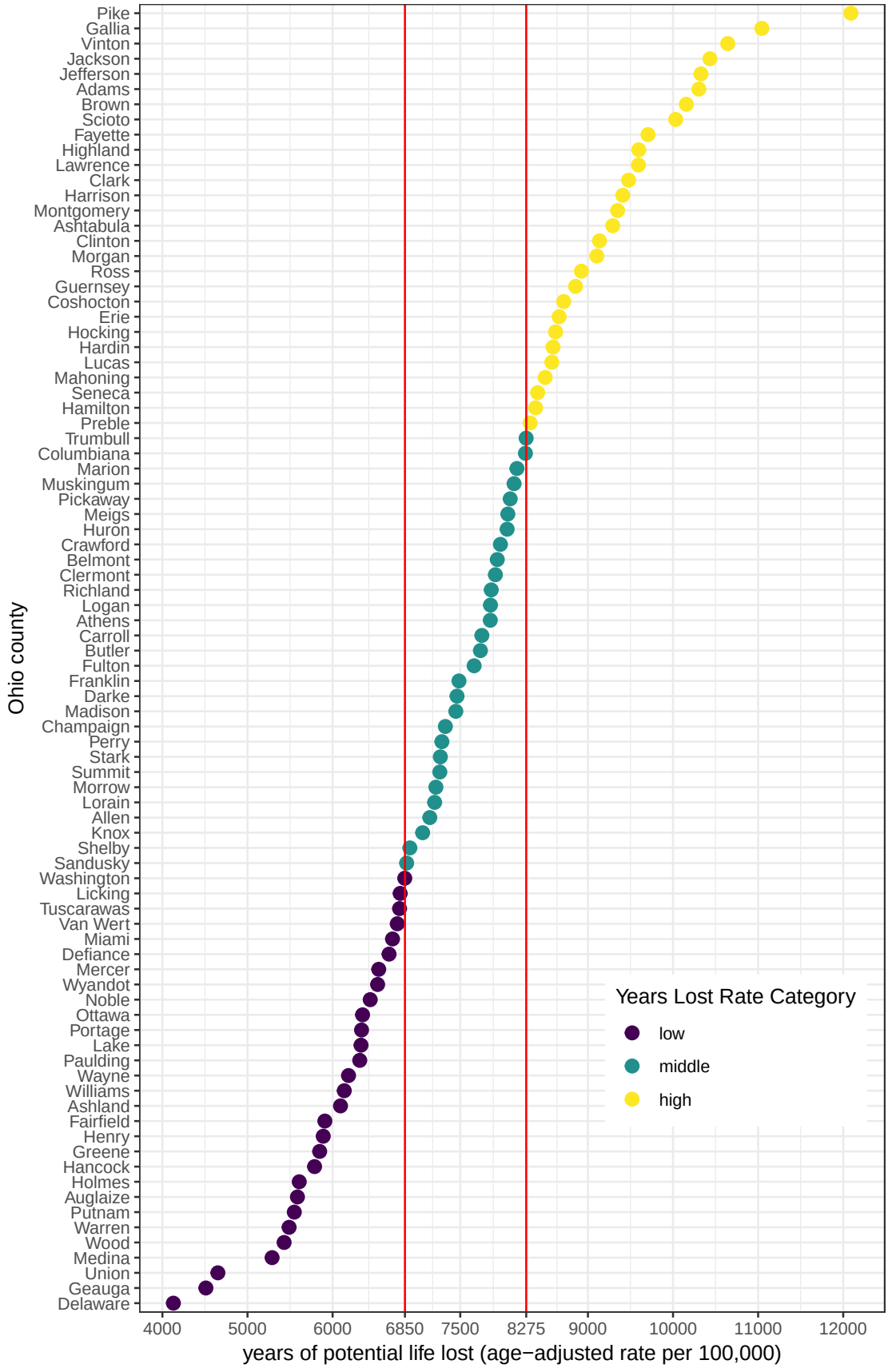
```
ohc_86 <- ohc_86 %>%
    mutate(yrslostcat = fct_recode(yrslostcat0,
                                   low = "[ 4129, 6850)",
                                   middle = "[ 6850, 8275)",
                                   high = "[ 8275,12091]"),
           yrslostcat = factor(yrslostcat, ordered = TRUE))
ohc_86 %>% count(yrslostcat, yrslostcat0)
```

```
# A tibble: 3 x 3
  yrslostcat yrslostcat0       n
  <ord>      <fct>         <int>
1 low        [ 4129, 6850)    29
2 middle     [ 6850, 8275)    29
3 high       [ 8275,12091]    28
```

We could apply these to the data in a *Cleveland dot plot*, for example.

```
ohc_86 %>% arrange(years_lost_rate) %>%
    mutate(county = factor(county, levels = .$county)) %>%
    ggplot(., aes(x = county, y = years_lost_rate,
                  col = yrslostcat)) +
    geom_point(size = 3) +
    geom_hline(yintercept = 6850, col = "red") +
    geom_hline(yintercept = 8275, col = "red") +
    scale_y_continuous(
        breaks = c(4000, 5000, 6000, 6850, 7500, 8275,
                   9000, 10000, 11000, 12000)) +
    guides(color = guide_legend("Years Lost Rate Category")) +
    coord_flip() +
    theme_bw() +
    theme(legend.position = c(0.8, 0.2)) +
```

```
    labs(x = "Ohio county",
         y = "years of potential life lost (age-adjusted rate per 100,000)")
```


Pike
Gallia
Vinton
Jackson
Jefferson
Adams
Brown
Scioto
Fayette
Highland
Lawrence
Clark
Harrison
Montgomery
Ashtabula
Clinton
Morgan
Ross
Guernsey
Coshocton
Erie
Hocking
Hardin
Lucas
Mahoning
Seneca
Hamilton
Preble
Trumbull
Columbiana
Marion
Muskingum
Pickaway
Meigs
Huron
Crawford
Belmont
Clermont
Richland
Logan
Athens
Carroll
Butler
Fulton
Franklin
Darke
Madison
Champaign
Perry
Stark
Summit
Morrow
Lorain
Allen
Knox
Shelby
Sandusky
Washington
Licking
Tuscarawas
Van Wert
Miami
Defiance
Mercer
Wyandot
Noble
Ottawa
Portage
Lake
Paulding
Wayne
Williams
Ashland
Fairfield
Henry
Greene
Hancock
Holmes
Auglaize
Putnam
Warren
Wood
Medina
Union
Geauga
Delaware
Ohio county
4000
5000
6000
6850
7500
8275
9000
10000
11000
12000
years of potential life lost (age−adjusted rate per 100,000)
Years Lost Rate Category
low
middle
high

3.2 Select up to 5 predictor variables

We'll select

- `lbw_pct`, % of births with low birth weight (< 2500 g)
- `smoker_pct`, % of adults that report currently smoking
- `exc_drink`, % of adults that report excessive drinking
- `teen_births`, Teen births / females ages 15-19 x 1,000
- `associations`, social associations / population x 10,000

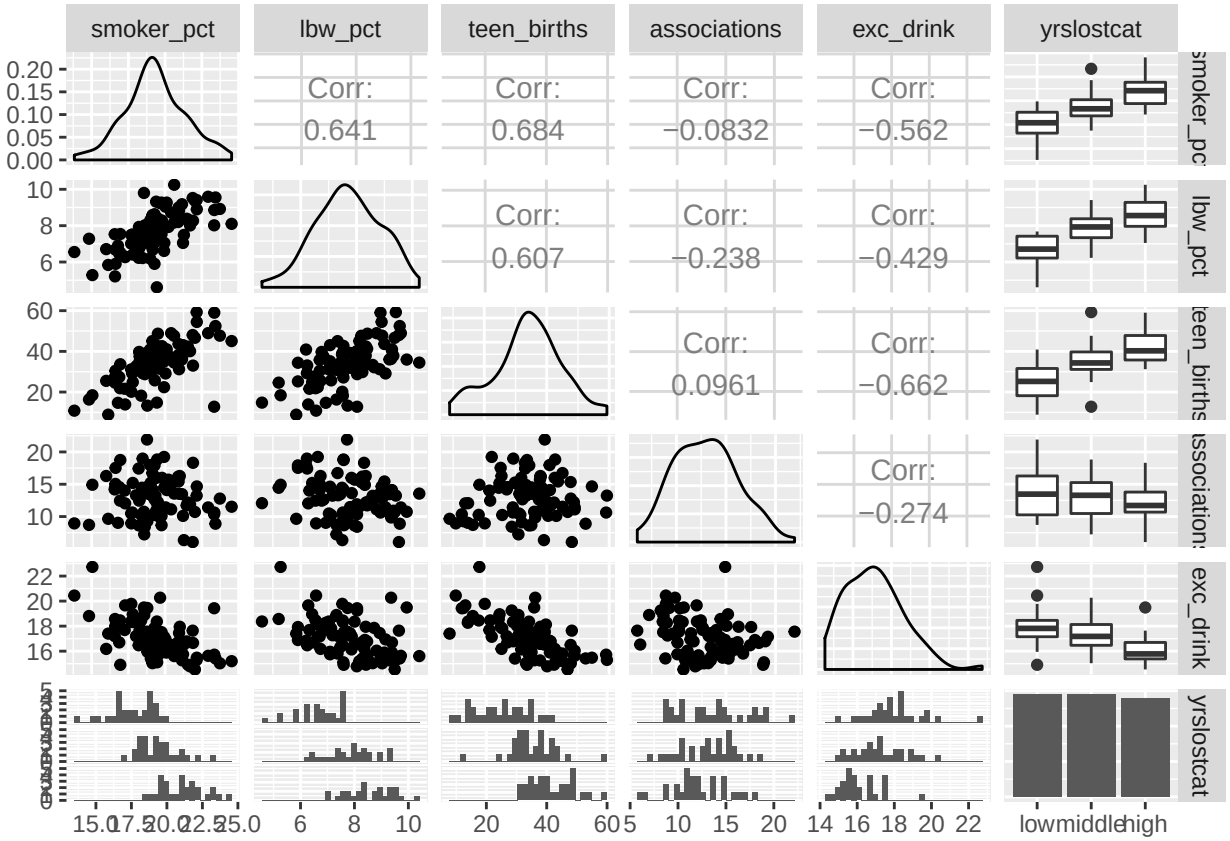
Here are the values of those predictors in the development sample of 86 counties, summarized numerically:

```
ohc_86 %>% select(county, smoker_pct, lbw_pct, teen_births,
                  associations, exc_drink) %>% skim()
```

```
Skim summary statistics
 n obs: 86
 n variables: 6

-- Variable type:factor ---------------------------------------------------------------------------
 variable missing complete  n n_unique                     top_counts
   county       0       86 86       86 Ada: 1, All: 1, Ash: 1, Ash: 1
 ordered
   FALSE

-- Variable type:numeric --------------------------------------------------------------------------
     variable missing complete  n  mean    sd    p0   p25   p50   p75
 associations       0       86 86 12.97  3.22  6.05 10.51 13    15.02
    exc_drink       0       86 86 17.12  1.52 14.54 15.9  17.01 18.01
      lbw_pct       0       86 86  7.69  1.15  4.61  6.93  7.58  8.39
   smoker_pct       0       86 86 19.33  2.07 13.82 18.22 19.28 20.62
  teen_births       0       86 86 33.92 10.75  8.88 28.87 34.34 40.36
  p100
 21.92
 22.73
 10.25
 24.53
 59.28
```

3.3 Scatterplot Matrix

```
ggpairs(data = dplyr::select(ohc_86,
        smoker_pct, lbw_pct, teen_births,
                  associations, exc_drink, yrslostcat))
```



The plot suggests that there is some overlap between the values of all five predictors across levels of our categorical *years lost* outcome, but that each predictor is reasonably symmetric and shows a relationship that is either modest or in the direction we'd guess in advance.

3.4 Fit proportional odds logistic regression model

```
model_q3 <- polr(yrslostcat ~ smoker_pct + lbw_pct +
                   teen_births + associations + exc_drink,
                 data = ohc_86, Hess = TRUE)

summary(model_q3)
```

```
Call:
polr(formula = yrslostcat ~ smoker_pct + lbw_pct + teen_births +
    associations + exc_drink, data = ohc_86, Hess = TRUE)

Coefficients:
                Value Std. Error t value
smoker_pct    0.58635    0.21694   2.703
lbw_pct       0.63035    0.34291   1.838
teen_births   0.09085    0.04068   2.233
associations -0.23641    0.09681  -2.442
exc_drink    -0.40361    0.26176  -1.542

Intercepts:
            Value   Std. Error t value
```

```
low|middle   7.7690  7.1827     1.0816
middle|high 10.8668  7.2846     1.4917

Residual Deviance: 106.7771
AIC: 120.7771
```

3.5 Interpreting the Model

In terms of the coefficients, our interpretation is pretty straightforward, but only after we exponentiate.

```
exp(coef(model_q3))
```

```
  smoker_pct      lbw_pct  teen_births associations    exc_drink
   1.7974107    1.8782740    1.0951000    0.7894593    0.6679019
```

```
exp(confint(model_q3))
```

```
                2.5 %    97.5 %
smoker_pct   1.1978600 2.8164790
lbw_pct      0.9726042 3.7702579
teen_births  1.0138352 1.1906127
associations 0.6440804 0.9467275
exc_drink    0.3912678 1.1041096
```

If we compare two counties with the same values of the other predictors in the model, but:

- County A has `smoker_pct` that is one percentage point higher than County B, the model predicts that County A will have 80% higher odds (1.80 times the odds of County B) of being in a higher (worse) category for years lost, and at the 5% level, this is a statistically significant difference, according to the confidence interval.
- County A has a `teen_births` rate that is one more birth per 1000 females age 15-19 in the county higher than the rate in County B, the model predicts that County A will have 10% higher odds (1.095 times the odds of county B) of falling in a higher (worse) category for years lost, and this is also a statistically significant effect.
- County A has a `associations` value that is one association per 10,000 population higher than County B, then County A will have lower odds (specifically 79% of the odds) of County B of falling in a higher (worse) category for years lost, and this, too, is statistically significant.
- The effects for `lbw_pct` and `exc_drink` move in unsurprising directions, although they don't quite meet the standard for statistical significance, as the confidence intervals for their odds ratios include 1.

3.6 How well does the model fit?

Let's build a cross-tabulation of the predictions made by this model, against the actual classifications for the 86 counties in the development sample.

```
addmargins(table(predict(model_q3), ohc_86$yrslostcat,
      dnn = c("Predicted", "Observed Values")))
```

```
         Observed Values
Predicted low middle high Sum
   low     23      4    0  27
   middle   6     20   10  36
   high     0      5   18  23
   Sum     29     29   28  86
```

Our model predicts the category correctly for 23 + 20 + 18 = 61/88 or 71% of counties in the development sample. The model gets

- 23/29 or 79% correct of the counties that are actually *low*,

- 20/29 or 69% correct for the counties that are *middle*, and
- 18/28 or 64% correct for the counties that are really *high*.

3.7 Other Summary Statistics for this model

To get some other summaries, I'd recommend fitting the model with the `lrm` function from the `rms` package.

```
d <- datadist(ohc_86)
options(datadist = "d")

model_q3_lrm <- lrm(yrslostcat ~ smoker_pct + lbw_pct +
                        teen_births + associations +
                        exc_drink,
                    data = ohc_86, x = T, y = T)

model_q3_lrm
```

```
Logistic Regression Model

 lrm(formula = yrslostcat ~ smoker_pct + lbw_pct + teen_births +
     associations + exc_drink, data = ohc_86, x = T, y = T)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs            86    LR chi2      82.16    R2       0.692    C       0.916
  low           29    d.f.             5    g        3.507    Dxy     0.833
  middle        29    Pr(> chi2) <0.0001    gr      33.347    gamma   0.834
  high          28                          gp       0.381    tau-a   0.562
 max |deriv| 1e-05                          Brier    0.084

              Coef     S.E.   Wald Z Pr(>|Z|)
 y>=middle     -7.7687 7.1825 -1.08  0.2794
 y>=high      -10.8665 7.2844 -1.49  0.1358
 smoker_pct     0.5863 0.2169  2.70  0.0069
 lbw_pct        0.6304 0.3429  1.84  0.0660
 teen_births    0.0908 0.0407  2.23  0.0255
 associations  -0.2364 0.0968 -2.44  0.0146
 exc_drink     -0.4036 0.2617 -1.54  0.1231
```

The Nagelkerke R^2 is quite high, at 0.69, and the C statistic is excellent at 0.92. If you like, we could validate those summary statistics with `validate` but I won't do that here.

3.8 Assessing the Proportional Odds Assumption

To test the proportional odds assumption, I'll fit the analogous multinomial logit and see if I detect a substantial improvement in fit quality.

```
m4_multi <- multinom(yrslostcat ~ smoker_pct + lbw_pct +
                        teen_births + associations +
                        exc_drink,
                    data = ohc_86)
```

```
# weights:  21 (12 variable)
initial  value 94.480657
iter  10 value 61.242360
```

```
iter  20 value 51.074390
iter  30 value 50.486157
iter  40 value 50.225501
iter  50 value 50.044138
iter  50 value 50.044138
iter  50 value 50.044138
final  value 50.044138
converged
```

```
m4_multi
```

```
Call:
multinom(formula = yrslostcat ~ smoker_pct + lbw_pct + teen_births +
    associations + exc_drink, data = ohc_86)

Coefficients:
       (Intercept) smoker_pct   lbw_pct teen_births associations
middle   -18.72835  0.5433114 0.9355231   0.1713063   -0.2290821
high     -18.56955  1.0199946 1.2983690   0.1822937   -0.4193514
         exc_drink
middle -0.01901173
high   -0.66444803

Residual Deviance: 100.0883
AIC: 124.0883
```

Now, I'll obtain the log likelihood values for each of the models I'm going to compare:

```
logLik(model_q3)
```

```
'log Lik.' -53.38853 (df=7)
```

```
logLik(m4_multi)
```

```
'log Lik.' -50.04414 (df=12)
```

To build a test, I calculate G = -2 × the difference in log likelihoods, and compare it to a χ^2 distribution with appropriate (12 - 7) degrees of freedom.

```
G <- as.numeric(-2 * (logLik(model_q3) - logLik(m4_multi)))

G
```

```
[1] 6.688778
```

```
pchisq(G, 5, lower.tail = FALSE)
```

```
[1] 0.2448341
```

The p value is not significant here, so this is an indication that the proportional odds model fits about as well as the more complex multinomial logit, so there's some evidence that our `modelq4` may be adequate.

3.9 Observed Classification of Cuyahoga and Monroe Counties

First, it would be helpful to know which category Cuyahoga and Monroe actually belong to.

```
ohc_2 %>% select(county, years_lost_rate)
```

```
# A tibble: 2 x 2
  county   years_lost_rate
```

```
  <fct>              <int>
1 Cuyahoga            7828
2 Monroe              6903
```

Remember that our cutoffs were:

- Low: less than 6850
- Middle: at least 6850 and below 8275
- High: at least 8275

So both Cuyahoga and Monroe would actually be in the Middle category.

3.10 Values of the Predictors in Cuyahoga and Monroe

```
ohc_2 %>% select(county, smoker_pct, lbw_pct, teen_births,
                 associations, exc_drink)
```

```
# A tibble: 2 x 6
  county   smoker_pct lbw_pct teen_births associations exc_drink
  <fct>         <dbl>   <dbl>       <dbl>        <dbl>     <dbl>
1 Cuyahoga       18.7   10.5         35.5          9.2      18.1
2 Monroe         19.6    8.36        34.9         20.0      15.7
```

So we see that:

- Cuyahoga and Monroe fall in the middle half of `smoker_pct` and `teen_births` across Ohio's counties.
- Cuyahoga has an unusually high `lbw_pct`, higher than any other county in Ohio, while Monroe is in the middle half of the main set of 86 counties on that measure.
- Cuyahoga is in the bottom quarter while Monroe is in the top quarter of the `associations` predictor.
- Cuyahoga is in the top quarter while Monroe is in the bottom quarter of the `exc_drink` predictor.

3.11 Predictions for Cuyahoga and Monroe Counties

We can either predict the actual classification (with `type = "class"`) or the model probabilities (with 'type = "probs") of these new counties (Cuyahoga then Monroe) actually falling into each of the three classifications.

```
predict(model_q3, newdata = ohc_2, type = "probs")
```

```
         low    middle      high
1 0.02761907 0.3585409 0.6138400
2 0.24997894 0.6307167 0.1193044
```

As you can see, the model predicts that Cuyahoga will be in the high group (with probability 0.61), but also has (according to the model) a 36% probability of falling into the category we did observe (the middle.) Monroe is most likely, according to the model, to fall in the middle group (with probability 0.63), and in fact, that's where it fell. Not too bad, especially with Cuyahoga being a serious outlier on the `lbw_pct` predictor.

3.12 Grading Rubric: Question 3

- Award 5 points for fitting an ordinal logistic regression model.
- Award another 2 points for creating the outcome in a rational way and explaining what they did.
- Award another 3 points for obtaining predictions using that model for the two held-out counties.
- Take off 1 point from the total if you find two or three typographical or syntax/grammar errors in this response.
- Take off 2 points from the total if you find an especially large number (more than 3, let's say) such errors.

- Award up to 5 additional points if the student did a good job explaining what they did, so that a good answer to the question with no more than one typo should get the full 15 points, and a correct response with very poor explanations would receive 10.

4 Question 4 (15 points)

Build a new outcome variable that is a count (possible values = 0-4) of whether the county meets each of the following standards:

- the county has a `smoker_pct` value **below** the Ohio-wide average of 22
- the county has an `obese_pct` value **below** the Ohio-wide average of 31
- the county has an `exer_access` value **above** the Ohio-wide average of 83
- the county has **NOT** had a water violation in the past year (as shown by `h2oviol` = No)

Your job is to fit **two** possible regression models in your development sample to predict this count, using the same predictors (at least 3 and no more than 6 of those not used in the calculation of standards) available in the data set. Demonstrate how well each model fits the counts by developing a rootogram and other summaries that you deem useful, then select the model you prefer, specifying your reasons. Next, use your preferred model to predict Cuyahoga County and Monroe County results, and assess the quality of those predictions.

4.1 Create the count outcome variable `q5count`

We will create the count variable as follows:

```
ohc_86 <- ohc_86 %>%
    mutate(q5count = 0 + (smoker_pct < 22) +
                (obese_pct < 31) + (exer_access > 83) +
                (h2oviol == "No"))
ohc_86 %>% count(q5count)
```

```
# A tibble: 5 x 2
  q5count     n
    <dbl> <int>
1       0     4
2       1    21
3       2    45
4       3    12
5       4     4
```

Given the relatively small number of zeros in the data, we're going to fit Poisson and Negative Binomial regression models to this outcome.

4.1.1 Sanity Check - do these counts match Dr. Love's counts?

To check our work, we can look at the values of `q5count` for the five counties Dr. Love listed in the assignment. Since Cuyahoga was one of those counties, we'll hold off on that for a moment and look at the other four first.

```
ohc_86 %>%
    filter(county %in% c("Highland", "Erie", "Ashland", "Athens")) %>%
    select(county, q5count, smoker_pct, obese_pct, exer_access, h2oviol)
```

```
# A tibble: 4 x 6
  county   q5count smoker_pct obese_pct exer_access h2oviol
  <fct>      <dbl>      <dbl>     <dbl>       <dbl> <fct>
1 Ashland        2       19.9      27.9        64.6 Yes
2 Athens         3       23.3      28.3        85   No
```

```
3 Erie            1       19.0      35.4        80.0 Yes
4 Highland        0       23.7      32.1        57.7 Yes
```

Those match. We should run this for the other two counties, as well, so we can check Cuyahoga.

```
ohc_2 <- ohc_2 %>%
    mutate(q5count = 0 + (smoker_pct < 22) +
                (obese_pct < 31) + (exer_access > 83) +
                (h2oviol == "No"))
ohc_2 %>% select(county, q5count, smoker_pct, obese_pct, exer_access, h2oviol)
```

```
# A tibble: 2 x 6
  county   q5count smoker_pct obese_pct exer_access h2oviol
  <fct>      <dbl>      <dbl>     <dbl>       <dbl> <fct>
1 Cuyahoga       4       18.7      30           95.6 No
2 Monroe         2       19.6      36.8         71.3 No
```

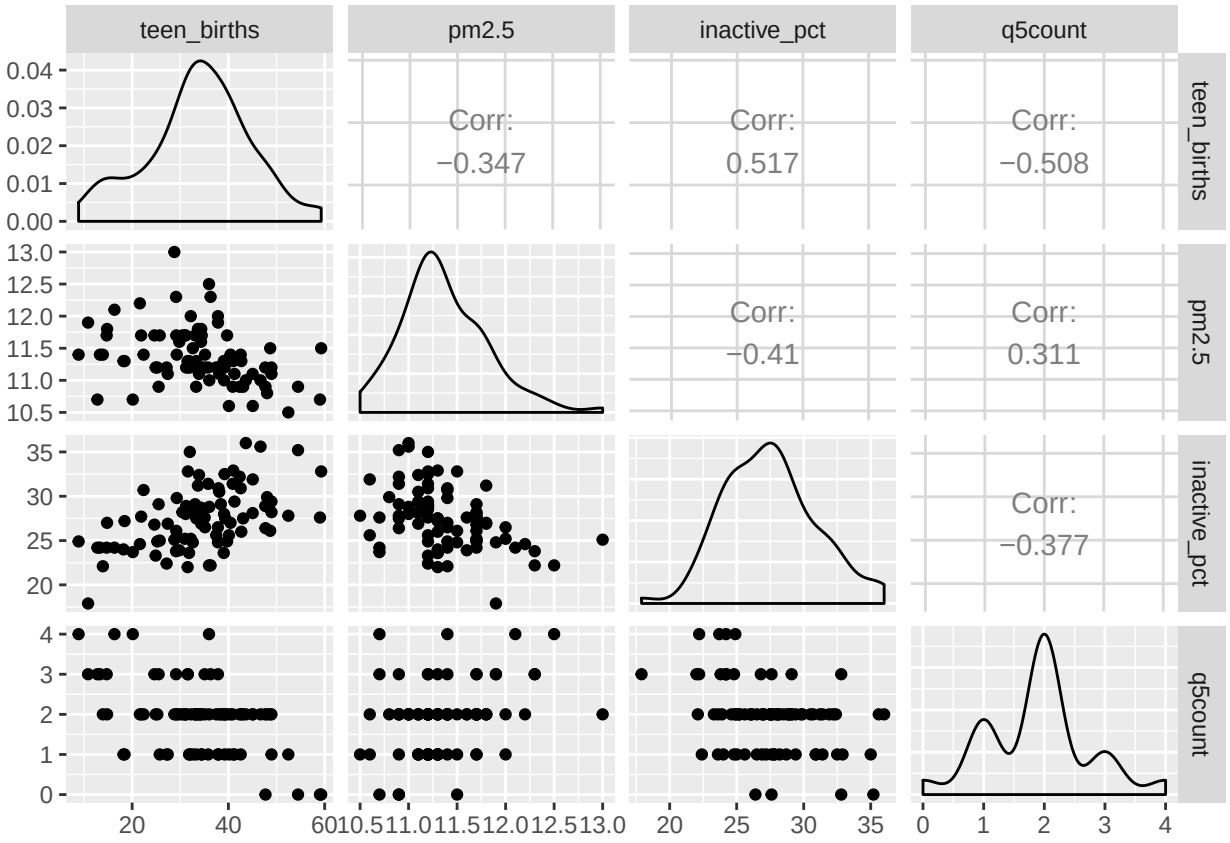
And Cuyahoga's count is 4, as was specified in the Homework. Good.

4.2 Select a set of predictors for our models

We need to select 3-6 predictors for this count outcome, and we're not allowed to use any of the variables that went into the count. There aren't many predictors that look like they do much. We'll choose:

- `teen_births`
- `pm2.5` and
- `inactive_pct`

```
ggpairs(data = ohc_86 %>% select(teen_births, pm2.5, inactive_pct, q5count))
```

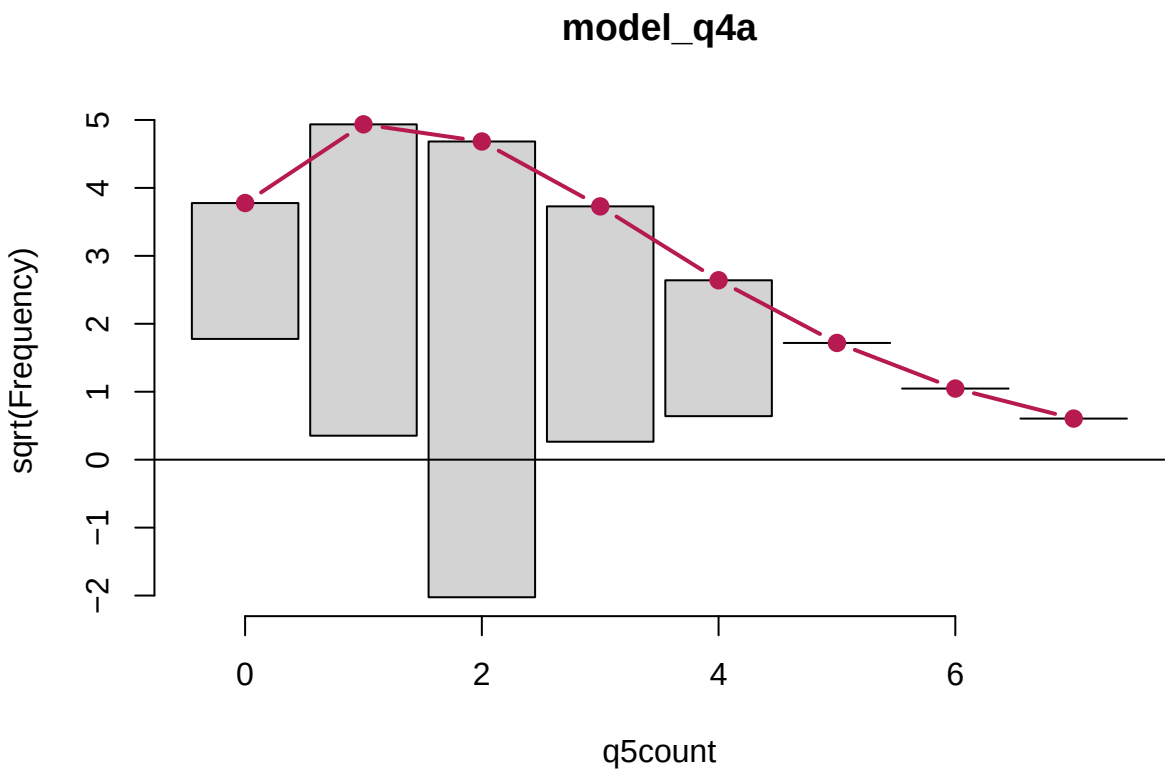


4.3 Fit Model 4A, a Poisson regression model

```
model_q4a <- glm(q5count ~ teen_births + pm2.5 + inactive_pct,
                 family=poisson(), data = ohc_86)

summary(model_q4a)
```

```

Call:
glm(formula = q5count ~ teen_births + pm2.5 + inactive_pct, family = poisson(),
    data = ohc_86)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66509  -0.45363   0.07632   0.39608   1.06027

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.068150   2.491168   0.027   0.9782
teen_births  -0.017132   0.008642  -1.982   0.0474 *
pm2.5         0.134457   0.188020   0.715   0.4745
inactive_pct -0.014763   0.028606  -0.516   0.6058
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 39.783  on 85  degrees of freedom
Residual deviance: 30.092  on 82  degrees of freedom
AIC: 246.67

Number of Fisher Scoring iterations: 4
```

```
rootogram(model_q4a)
```



The Poisson model doesn't really fit that well. We have too many predicted values of 2, and not enough 0s or 4s, and the Poisson model would also suggest that we'd have considerable values out into the tails, which isn't actually possible.

4.4 Fit Model 4B, a Poisson Regression on only one predictor

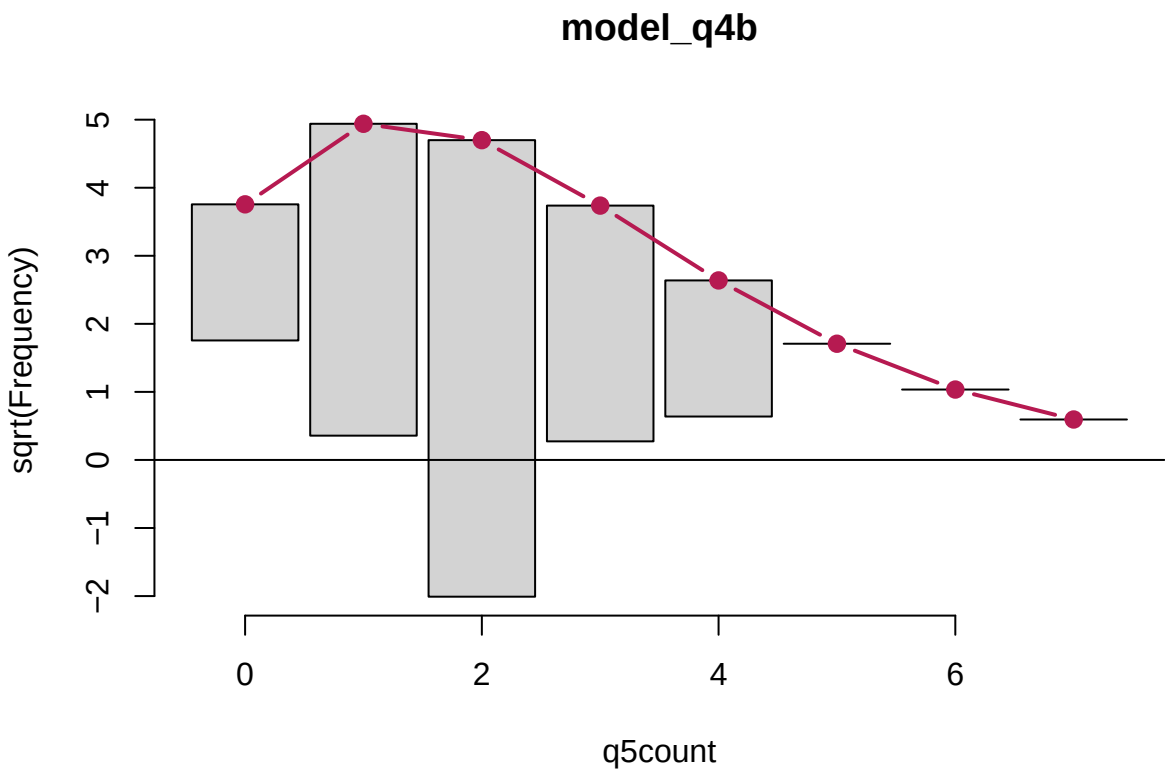
The only predictor that seems to do anything here is `teen_births`. Would building a model on that predictor alone show a meaningful improvement?

```
model_q4b <- glm(q5count ~ teen_births,
                 family=poisson(), data = ohc_86)

summary(model_q4b)
```

```
Call:
glm(formula = q5count ~ teen_births, family = poisson(), data = ohc_86)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6596  -0.4994   0.0719   0.3584   1.4383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.334363   0.239937   5.561 2.68e-08 ***
teen_births -0.021261   0.007212  -2.948   0.0032 **
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 39.783  on 85  degrees of freedom
Residual deviance: 31.178  on 84  degrees of freedom
AIC: 243.76

Number of Fisher Scoring iterations: 4
```

```
rootogram(model_q4b)
```



The AIC is better, but the rootogram isn't much better if at all.

Would a negative binomial model be any better? Is this just a dispersion problem?

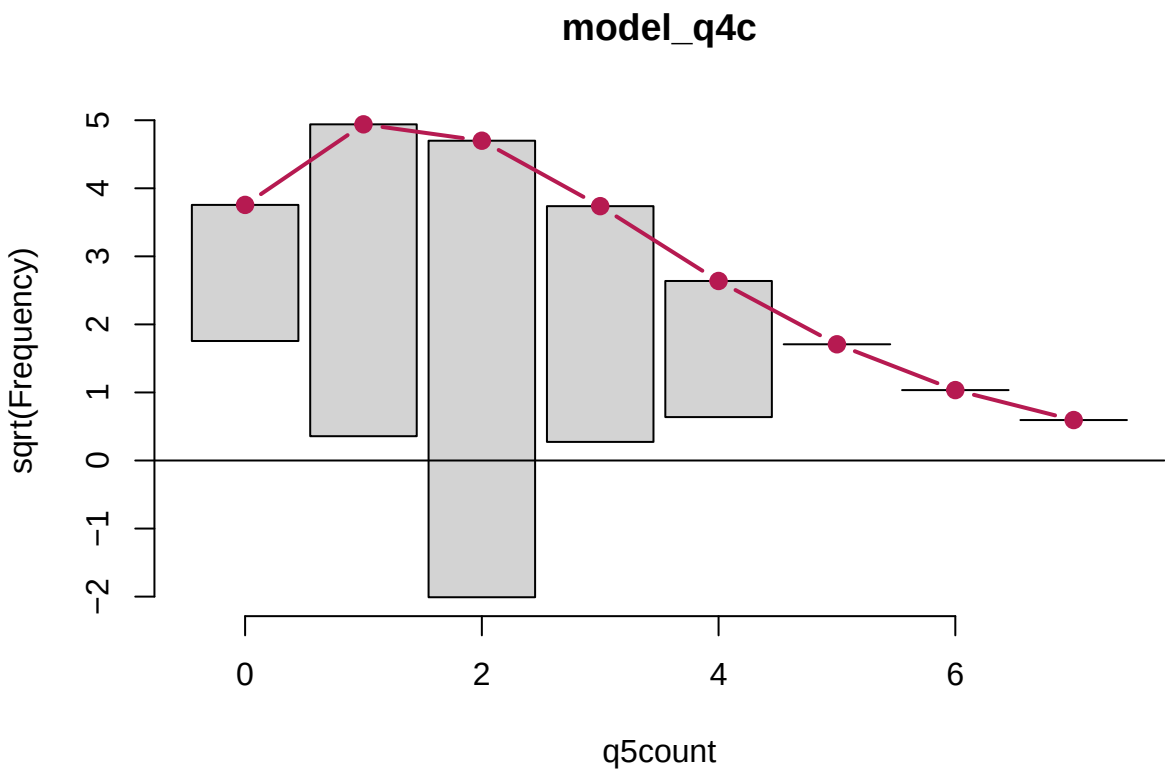
4.5 Fit Model 4C, a Negative Binomial regression model

```
model_q4c <- glm.nb(q5count ~ teen_births,
                    link = log, data = ohc_86)

summary(model_q4c)
```

```
Call:
glm.nb(formula = q5count ~ teen_births, data = ohc_86, link = log,
    init.theta = 101537.898)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6596  -0.4994   0.0719   0.3584   1.4382

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.334364   0.239941   5.561 2.68e-08 ***
teen_births -0.021261   0.007212  -2.948   0.0032 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(101537.9) family taken to be 1)

    Null deviance: 39.782  on 85  degrees of freedom
Residual deviance: 31.178  on 84  degrees of freedom
AIC: 245.76

Number of Fisher Scoring iterations: 1

              Theta:  101538
          Std. Err.:  1644413
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -239.761
```

```
rootogram(model_q4c)
```



The Negative Binomial model's rootogram looks about the same as our Poisson models. No real improvement, and in fact, the AIC is a bit worse. I also ran the negative binomial regression on all of the predictors we included in Model 4A, and that's no help, either.

So I'll stick with Model 4B, as bad as it is.

4.6 Other Options

- We could fit a censored regression (tobit) model that forces all of the values to fall between 0 and 4, or
- We could treat the outcome not as a count, but instead as an ordered factor, and then fit a proportional odds logistic regression model.

But in either case, we wouldn't have a rootogram to study, so I won't do that.

4.7 Predicting Cuyahoga County and Monroe County

While I prefer Model 4B, I'll show the predictions with Model 4A, since that does include at least 3 predictors, and thus seems like a model someone might have actually chosen.

Remember the actual values of `q5count` for our two held-out counties were:

```
ohc_2 %>% select(county, q5count)
```

```
# A tibble: 2 x 2
  county   q5count
  <fct>      <dbl>
1 Cuyahoga       4
2 Monroe         2
```

To obtain our predictions with Model 4A, we need:

```
predict(model_q4a, newdata = ohc_2, type = "response")
```

```
       1        2
2.311445 1.492484
```

The predicted count for Cuyahoga County is 2.3 and the observed count was 4, which seems like a big miss. The predicted count for Monroe County is 1.5 and the observed count was 2, which seems like a smaller miss.

Predictions with our models 5B and 5C are essentially identical, and a little better for Monroe, but worse for Cuyahoga.

```
predict(model_q4b, newdata = ohc_2, type = "response")
```

```
       1        2
1.786869 1.809808
```

```
predict(model_q4c, newdata = ohc_2, type = "response")
```

```
       1        2
1.786868 1.809808
```

4.8 Grading Rubric: Question 4

- Award 5 points for fitting two different count regression models.
- Award another 3 points for creating the outcome correctly.
- Award another 3 points for obtaining predictions using that model for the two held-out counties.
- Take off 1 point from the total if you find two or three typographical or syntax/grammar errors in this response.
- Take off 2 points from the total if you find an especially large number (more than 3, let's say) such errors.
- Award up to 4 additional points if the student did a good job explaining what they did, so that a good answer to the question with no more than one typo should get the full 15 points, and one that is correct but very poorly explained would get 11.