Assignment 5 Answer Sketch

432 Staff

Due 2018-03-02. Sketch developed 2018-03-01

Contents

```r
library(skimr)
library(simputation)
library(Epi)
library(broom)
library(rms)
library(tidyverse)

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

1 Question 1 (25 points)

We don't write answer sketches for essay questions.

1.1 Instructions for Data for Questions 2-9

> The data come from the NHANES National Youth Fitness Survey. Data collected in the `nnyf1.csv` file above and on the [Data and Code] page from our site come from the **Demographics** files, and from the **Medical Conditions** and **Physical Activity** files, which are each part of the **Questionnaire** data.

I merged files on the basis of the respondent sequence number (SEQN). The variables available to you are:

1. `SEQN` - the *respondent sequence number* (there are 1,576 subject in the `nnyfs1.csv` file made available to you)
2. `RIASEX` (from the Demographics files) - *sex of subject (1 = male, 2 = female)*
3. `RIDAGEYR` (from the Demographics files) - *age in years at screening (3-15)*
4. `RIDRETH1` (from the Demographics files) - *race/hispanic origin (1 = Mexican-American, 2 = Other Hispanic, 3 = Non-Hispanic White, 4 = Non-Hispanic Black, 5 = Other Race including Multi-Racial)*
5. `INDFMPIR` (from the Demographics files; **impute** all subjects with missing values on the basis of `RIDRETH1` and `RIDAGEYR`) - *ratio of family income to poverty (data show 0-4.99, and then truncated as 5 for all who are in fact greater than or equal to 5)*
6. `MCQ010` (from the Medical Conditions files; all subjects should have values of 1 [Yes] or 2 [No]) - *has the child ever been told they have asthma*
7. `PAQ706` (from the Physical Activity files; **drop** all subjects with values other than 0, 1, 2, 3, 4, 5, 6, or 7) - *days (in the past 7) physically active at least 60 minutes*

1.2 Data Management for Questions 2-9

NOTE I'm going to make a meal out of this here, getting every little scrap of adjustment to the data I will use in any of the questions that follow done here. You probably did some management of data as you went through, and probably were smart enough to drop some of the details (like sanity checks, and rechecking with tables, counts, summaries, skims, etc.) That's totally fine with me. So long as you explain what you actually did to modify the data, that's the main thing.

First, let's look at the data immediately after import, to see if any of the values we observe don't match what we've been led to believe from the description above. Specifically, we want to verify that we have 1,576 observations at the start.

```
nnyfs1.raw <- read.csv("nnyfs1.csv") %>% tbl_df

nnyfs1.raw
```

```
# A tibble: 1,576 x 7
    SEQN RIASEX RIDAGEYR RIDRETH1 INDFMPIR MCQ010 PAQ706
   <dbl>  <int>    <int>    <int>    <dbl>  <int>  <int>
 1 71917      2       15        4    0.210      2      3
 2 71918      2        8        4    5.00       1      5
 3 71919      2       14        3    5.00       2      3
 4 71920      2       15        3    0.870      1      3
 5 71921      1        3        3    4.34       2      7
 6 71922      1       12        2    5.00       2      2
 7 71923      1       12        3    5.00       2      5
 8 71924      2        8        5    2.74       2      3
 9 71925      1        7        1    0.460      2      7
10 71926      1        8        4    1.57       2      7
# ... with 1,566 more rows
```

and we do have 1,576 observations.

We have also been told that:

1. `RIASEX` and `MCQ010` has values of 1 or 2 and nothing else. Is that true?

```
nnyfs1.raw %>% count(RIASEX, MCQ010)
```

```
# A tibble: 4 x 3
```

```
  RIASEX MCQ010     n
   <int>  <int> <int>
1      1      1   121
2      1      2   666
3      2      1   151
4      2      2   638
```

Yes, it seems correct, We see only values of 1 and 2 for each variable.

2. RIDRETH1 is supposed to have only integer values between 1 and 5. Is that true?

```
nnyfs1.raw %>% count(RIDRETH1)
```

```
# A tibble: 5 x 2
  RIDRETH1     n
     <int> <int>
1        1   242
2        2   238
3        3   619
4        4   345
5        5   132
```

Again, looks good.

3. RIDAGEYR has values between 3 and 15. Is that the case? Since the data are actually stored in integer years, we can still use count to check this, or we could use a simple summary or skim to look at the maximum and minimum values. Let's do both here, to show you what I mean.

```
nnyfs1.raw %>% count(RIDAGEYR)
```

```
# A tibble: 13 x 2
   RIDAGEYR     n
      <int> <int>
 1        3   115
 2        4   116
 3        5   121
 4        6   133
 5        7   131
 6        8   123
 7        9   104
 8       10   128
 9       11   113
10       12   138
11       13   122
12       14   133
13       15    99
```

```
nnyfs1.raw %>% skim
```

```
Skim summary statistics
 n obs: 1576
 n variables: 7

Variable type: integer
 variable missing complete    n mean   sd p0 p25 median p75 p100
   MCQ010       0     1576 1576 1.83 0.38  1   2      2   2    2
   PAQ706       3     1573 1576 5.58 4.6   0   4      7   7   99
   RIASEX       0     1576 1576 1.5  0.5   1   1      2   2    2
 RIDAGEYR       0     1576 1576 8.99 3.69  3   6      9  12   15
 RIDRETH1       0     1576 1576 2.93 1.15  1   2      3   4    5

Variable type: numeric
 variable missing complete    n     mean     sd    p0      p25   median
 INDFMPIR     100     1476 1576     2.23   1.61     0     0.83     1.74
     SEQN       0     1576 1576 72704.5  455.1  71917 72310.75 72704.5
      p75  p100
     3.47     5
 73098.25 73492
```

We can see from either approach that the minimum `RIDAGEYR` is in fact 3 and that the maximum is 15.

4. `INDFMPIR` should have values between 0 and 5 (probably with many values of 5)

We can see from the `skim` and its presentation of the minimum (0) and maximum (5) that we're probably all right, though we don't yet know how many 5s we actually have. Let's find out.

```
nnyfs1.raw %>% count(INDFMPIR == 5)
```

```
# A tibble: 3 x 2
  `INDFMPIR == 5`     n
  <lgl>           <int>
1 F                1280
2 T                 196
3 NA                100
```

We have 196 values of 5, and 100 missing values, with the remaining values falling between 0 and 4.99. So we'll need to impute soon.

5. `PAQ706` has a series of values, but we're going to drop anything that isn't an integer between 0 and 7

The minimum (from the skim) is 0 and the maximum is 99. From the skim, we see these are integers, so we can count them.

```
nnyfs1.raw %>% count(PAQ706)
```

```
# A tibble: 10 x 2
   PAQ706     n
    <int> <int>
 1      0    72
 2      1    35
 3      2    94
 4      3   126
 5      4   105
 6      5   211
 7      6    78
 8      7   849
 9     99     3
10     NA     3
```

It turns out we have 3 missing and 3 more 99 values that we'll need to drop. The remaining observation should stay.

So, our cleanup tasks are:

1. Drop the values of `PAQ706` above 7, including the three missing and three 99 cases.
2. Impute the remaining missing `INDFMPIR` values (100 now, may be less after we drop the problematic `PAQ706` cases.)

Once that is done, we can answer question 2, but we'll do some additional cleanup anticipating what we'll need in Questions 3-9.

1.2.1 Drop the values of PAQ706 outside of the 0-7 range

There are six subjects who need to be removed here. We can do that by retaining only those subjects with PAQ706 < 8.

```
nnyfs1.new <- nnyfs1.raw %>%
    dplyr::filter(PAQ706 < 8)

nnyfs1.new %>% count(PAQ706)
```

```
# A tibble: 8 x 2
  PAQ706     n
   <int> <int>
1      0    72
2      1    35
3      2    94
4      3   126
5      4   105
6      5   211
7      6    78
8      7   849
```

1.2.2 Dealing with missing values in INDFMPIR

Since all observed INDFMPIR values are in 0-5 (note the minimum and maximum above), we need only impute the missing values. I'll use simple imputation, and predictive mean matching using all of the other variables in the data.

```
set.seed(4321243)

nnyfs1.new <- nnyfs1.new %>%
    impute_pmm(INDFMPIR ~ RIASEX + RIDAGEYR +
                   RIDRETH1 + MCQ010 + PAQ706)

# check to ensure that all missing values are successfully imputed
nnyfs1.new %>% select(INDFMPIR) %>% skim
```

```
Skim summary statistics
 n obs: 1570
 n variables: 1

Variable type: numeric
 variable missing complete    n mean   sd p0  p25 median  p75 p100
 INDFMPIR       0     1570 1570 2.22 1.57  0 0.87   1.78 3.47    5
```

1.2.3 Re-specifying and re-naming variables

I'd like these variables to be more useful to me in modeling work. So, I will:

1. Create 1/0 well-named representations (called `female` and `asthma`, respectively) of the two binary variables, RIASEX and MCQ010, to make Questions 4-9 easier.

2. Create well-named descriptive factor representations (called `sex` and `asthma_f`) of the two binary variables, `RIASEX` and `MCQ010`, to make Question 3 easier. I'll also move the FEMALES to the front of the list of levels for `sex`.
3. Create a new 1/0 representation of whether `RIDRETH1` is 3 to help in Question 9.

```
nnyfs1.new <- nnyfs1.new %>%
    mutate(female = ifelse(RIASEX == 2, 1, 0),
           asthma = ifelse(MCQ010 == 1, 1, 0),
           sex = fct_recode(factor(RIASEX),
                            M = "1", F = "2"),
           sex = fct_relevel(sex, "F"),
           asthma_f = fct_recode(factor(MCQ010),
                                Yes = "1", No = "2"),
           nonh_white = ifelse(RIDRETH1 == 3, 1, 0))
```

1.2.4 Some Sanity Checks

Do the results in our three different versions of the response about asthma match up?

```
nnyfs1.new %>% count(MCQ010, asthma, asthma_f)
```

```
# A tibble: 2 x 4
  MCQ010 asthma asthma_f     n
   <int>  <dbl> <fct>    <int>
1      1   1.00 Yes        271
2      2   0    No        1299
```

Yes.

Do the results in our three different versions of the response about sex match up?

```
nnyfs1.new %>% count(sex, RIASEX, female)
```

```
# A tibble: 2 x 4
  sex   RIASEX female     n
  <fct>  <int>  <dbl> <int>
1 F          2   1.00   787
2 M          1   0      783
```

Yes.

Do the results in our `nonh_white` match what we were trying to get out of `RIDRETH1`?

```
nnyfs1.new %>% count(RIDRETH1, nonh_white)
```

```
# A tibble: 5 x 3
  RIDRETH1 nonh_white     n
     <int>      <dbl> <int>
1        1       0      242
2        2       0      238
3        3       1.00   616
4        4       0      343
5        5       0      131
```

Yes, that's right.

1.2.5 One Last skim to see what I've done

```
nnyfs1.new %>% skim
```

```
Skim summary statistics
 n obs: 1570 
 n variables: 12 

Variable type: factor 
 variable missing complete    n n_unique                 top_counts ordered
 asthma_f       0     1570 1570        2 No: 1299, Yes: 271, NA: 0   FALSE
      sex       0     1570 1570        2     F: 787, M: 783, NA: 0   FALSE

Variable type: integer 
 variable missing complete    n mean   sd p0 p25 median p75 p100
   MCQ010       0     1570 1570 1.83 0.38  1   2      2   2    2
   PAQ706       0     1570 1570 5.41 2.12  0   4      7   7    7
   RIASEX       0     1570 1570 1.5  0.5   1   1      2   2    2
 RIDAGEYR       0     1570 1570 8.98 3.69  3   6      9  12   15
 RIDRETH1       0     1570 1570 2.93 1.15  1   2      3   4    5

Variable type: numeric 
   variable missing complete    n     mean     sd    p0      p25   median
     asthma       0     1570 1570     0.17   0.38     0     0        0   
     female       0     1570 1570     0.5    0.5      0     0        1   
   INDFMPIR       0     1570 1570     2.22   1.57     0     0.87     1.78
 nonh_white       0     1570 1570     0.39   0.49     0     0        0   
       SEQN       0     1570 1570 72704.25 455.13 71917 72310.25 72703.5
      p75  p100
     0        1
     1        1
     3.47     5
     1        1
 73097.75 73492
```

Looks good.

- The `asthma`, `female` and `nonh_white` binary variables are numeric, as planned, and take the values 0 and 1.
- The `asthma_f` and `sex` variables are factors, as expected.
- and we have no missingness remaining. I think we're all set.

2 Question 2 (5 points)

> How many of those subjects wind up in your final data set, after applying the inclusion and exclusion criteria described above?

```
nrow(nnyfs1.new)
```

```
[1] 1570
```

3 Question 3 (10 points)

> Find the cross-product odds ratio and an appropriate 95% confidence interval for that odds ratio for being told you have asthma for females as compared to males within this sample. Specify the relevant cross-tabulation (contingency table).

```
twoby2(nnyfs1.new$sex, nnyfs1.new$asthma_f)
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Yes
Comparing : F vs. M

  Yes  No    P(Yes) 95% conf. interval
F 150 637    0.1906    0.1647   0.2196
M 121 662    0.1545    0.1309   0.1816

                                   95% conf. interval
             Relative Risk: 1.2334    0.9917   1.5340
         Sample Odds Ratio: 1.2883    0.9903   1.6760
Conditional MLE Odds Ratio: 1.2881    0.9821   1.6919
    Probability difference: 0.0361   -0.0014   0.0734

             Exact P-value: 0.0616
        Asymptotic P-value: 0.0591
------------------------------------------------------
```

The cross-product odds ratio is 1.29, with 95% CI (0.99, 1.68). The odds of a female being told they have asthma are estimated to be 1.29 times that of a male in this sample, but the confidence interval still includes 1, so the effect size doesn't meet the 5% significance level standard.

Note that this question was substantially easier to do with the respecified and (in the case of `sex` also reordered) factor variables to display the information about sex and asthma.

4 Question 4 (5 points)

> Use a logistic regression model to predict: `MCQ010` "Ever been told you have asthma" = YES [1] on the basis of the following variables: on the basis of the following variables: sex (captured in an indicator of `female`), subject's age at screening, Ratio of family income to poverty, and number of days physically active in the past 7. Specify the equation of the model you have fit.

Note that we expected you to treat the number of days physically active as a quantitative predictor, rather than as a factor. Some of you may have instead treated it as a factor, which is a really bad idea (using a factor in this case for this variable ignores the fact that it is a count, adds a lot of complexity and chews up a lot of degrees of freedom, for no meaningful gain. And it makes question 7 enormously more complicated than it needs to be.) I didn't absolutely prevent you from treating it as a factor, but doing so makes your life much harder, though. In this sketch, I'll do the simpler and sensible thing.

Note: Your answers will differ from ours because of the imputation of `INDFMPIR`. This will affect questions 4-9, a little bit.

4.1 Approach 1: Using `glm` to fit the model

`PAQ706` is treated here as quantitative…

```
(m4 <- glm(asthma ~ female + RIDAGEYR + INDFMPIR + PAQ706,
          data = nnyfs1.new, family="binomial"))
```

```
Call:  glm(formula = asthma ~ female + RIDAGEYR + INDFMPIR + PAQ706,
    family = "binomial", data = nnyfs1.new)

Coefficients:
(Intercept)       female     RIDAGEYR     INDFMPIR       PAQ706
   -1.66376      0.22039      0.07167     -0.12461     -0.08075

Degrees of Freedom: 1569 Total (i.e. Null);  1565 Residual
Null Deviance:      1444
Residual Deviance: 1406     AIC: 1416
```

Model `m4` reads as follows:

- The log odds of asthma = -1.66 + 0.22 `female` + 0.07 `RIDAGEYR` - 0.12 `INDFMPIR` - 0.08 `PAQ706`

if you used our random seed to help with the imputation. Your answer will be a little different, perhaps.

4.2 Approach 2: Using `lrm` to fit the model

```
d <- datadist(nnyfs1.new)
options(datadist="d")

(m4_lrm <- lrm(asthma ~ female + RIDAGEYR + INDFMPIR +
                   PAQ706, data = nnyfs1.new, x=T, y=T))
```

```
Logistic Regression Model

 lrm(formula = asthma ~ female + RIDAGEYR + INDFMPIR + PAQ706,
     data = nnyfs1.new, x = T, y = T)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           1570    LR chi2      38.78    R2       0.041    C       0.616
  0            1299    d.f.             4    g        0.475    Dxy     0.231
  1             271    Pr(> chi2) <0.0001    gr       1.609    gamma   0.231
 max |deriv| 3e-10                           gp       0.067    tau-a   0.066
                                             Brier    0.139

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -1.6638 0.3206 -5.19  <0.0001
 female     0.2204 0.1365  1.62  0.1063
 RIDAGEYR   0.0717 0.0198  3.61  0.0003
 INDFMPIR  -0.1246 0.0448 -2.78  0.0054
 PAQ706    -0.0807 0.0319 -2.54  0.0112
```

This is, of course, the same model we displayed before as `m4`.

5 Question 5 (10 points)

> Specify and interpret the model's odds ratio estimate for being told you have asthma for females as compared to males, after adjusting for the other variables included in the model you fit in Question 4. Provide a 95% confidence interval for this odds ratio.

We can use either `glm` or `lrm` to accomplish this, since the predictor we're studying is binary.

With `glm`, we'd use…

```
exp(coef(m4))
```

```
(Intercept)      female    RIDAGEYR    INDFMPIR      PAQ706 
  0.1894245   1.2465648   1.0742981   0.8828431   0.9224284 
```

```
exp(confint(m4, level = 0.95))
```

```
Waiting for profiling to be done...

                2.5 %    97.5 %
(Intercept) 0.1002142 0.3524837
female      0.9545348 1.6303766
RIDAGEYR    1.0335109 1.1171976
INDFMPIR    0.8077369 0.9628785
PAQ706      0.8670404 0.9824699
```

And the odds ratio estimate is 1.247, with 95% CI (0.95, 1.63).

With `lrm`, we'd use…

```
summary(m4_lrm)
```

```
             Effects              Response : asthma 

 Factor      Low  High  Diff. Effect   S.E.     Lower 0.95 Upper 0.95
 female      0.00  1.00 1.0    0.22039 0.136460 -0.047056   0.487840 
  Odds Ratio 0.00  1.00 1.0    1.24660       NA  0.954030   1.628800 
 RIDAGEYR    6.00 12.00 6.0    0.43000 0.119090  0.196600   0.663410 
  Odds Ratio 6.00 12.00 6.0    1.53730       NA  1.217300   1.941400 
 INDFMPIR    0.87  3.47 2.6   -0.32398 0.116420 -0.552170  -0.095793 
  Odds Ratio 0.87  3.47 2.6    0.72326       NA  0.575700   0.908650 
 PAQ706      4.00  7.00 3.0   -0.24224 0.095555 -0.429520  -0.054952 
  Odds Ratio 4.00  7.00 3.0    0.78487       NA  0.650820   0.946530 
```

And again, get an odds ratio for `female`'s effect of 1.247, with 95% CI (0.95, 1.63).

Note that for the non-binary variables, the effect estimates look different, as they should.

- The `glm` approach presents confidence intervals for the effect of increasing a predictor by 1, holding everything else constant.
- The `lrm` approach presents confidence intervals for the effect of increasing a predictor from what is listed in the summary output as Low to High (usually the 25th to 75th percentiles) for quantitative predictors.
- Either way, the model estimates that a female will have 1.24 [95% CI 0.96-1.63] times the odds of being told they have asthma compared to a male with the same values of the age, ratio of family income to poverty, and days physically active in the past 7. Again, the difference attributable to sex is not enough for us to call it significant at the 5% level.

6 Question 6 (10 points)

> Compare your result in Question 3 to your result in Question 5. Are they different? If so, why?

Question	Odds Ratio Estimate	95% Confidence Interval
3 (unadj.)	1.29	(0.99, 1.68)
5 (adj.)	1.24	(0.95, 1.63)

The results are slightly different, because Question 5's model adjusts for age at screening, ratio of family income to poverty and days physically active in the past, but those adjustments don't make an enormous practical difference in the estimate. In either case, we would (barely) fail to reject a null hypothesis of no `sex` effect, at the 5% significance level.

7 Question 7 (10 points)

> Specify and interpret the Question 4 model's odds ratio estimate (and a 95% confidence interval around that point estimate) associated with the "days physically active in the past 7" predictor.

We have to choose what we're showing here, based on the modeling strategy, as we noted in the answer to Question 5.

- If we show the `glm` result, we will be estimating the odds ratio associated with a change of 1 day of exercise. That turns out to be 0.92, with 95% CI (0.87, 0.98).
- If we show the `lrm` result, we will be estimating the odds ratio associated with a change of 3 days in exercise, from 4 to 7, specifically, which is a much larger change. So it's not surprising that odds ratio is further from 1. Specifically it is 0.78, with 95% CI (0.65, 0.95).

8 Question 8 (10 points)

> Use the model you fit in Question 4 to provide a prediction for the probability that a 10-year-old male child will have been told they have asthma, if they are active 3 days in the past 7, and have a ratio of family income to poverty of 2.5.

To answer this question, we'll create a little data frame (called `question8`) containing the data for this new subject.

```r
question8 <- data_frame(female = 0, RIDAGEYR = 10, PAQ706 = 3, INDFMPIR = 2.5)
```

With the `glm` approach, our predicted probability is easily obtained:

```r
predict(m4, newdata = question8, type = "response")
```

```
        1 
0.1822994 
```

With the `lrm` approach, we can run …

```r
predict(m4_lrm, newdata = question8, type="fitted.ind")
```

```
        1 
0.1822994 
```

The predicted probability is 0.18

9 Question 9 (15 points)

Refit the model you fit in Question 4 but now, add in an additional predictor variable that indicates if the subject's race/Hispanic origin value is Non-Hispanic White (i.e. RIDRETH1 = 3), or not. Decide whether or not an interaction term between age and race/ethnicity is required (but do not consider other interaction terms or other types of non-linearity). Specify the logistic regression equation you wind up fitting.

9.1 Using glm

We'll fit the models with and without the interaction term, and then assess them.

```
(m9_int <- glm(asthma ~ female + INDFMPIR + PAQ706 +
                  nonh_white*RIDAGEYR,
              data = nnyfs1.new, family="binomial"))
```

```

Call:  glm(formula = asthma ~ female + INDFMPIR + PAQ706 + nonh_white *
    RIDAGEYR, family = "binomial", data = nnyfs1.new)

Coefficients:
        (Intercept)               female             INDFMPIR
          -1.644122             0.220925            -0.132596
             PAQ706           nonh_white             RIDAGEYR
          -0.081896             0.016053             0.069182
nonh_white:RIDAGEYR
           0.005493

Degrees of Freedom: 1569 Total (i.e. Null);  1563 Residual
Null Deviance:      1444
Residual Deviance: 1405     AIC: 1419
```

```
(m9_noint <- glm(asthma ~ female + INDFMPIR + PAQ706 +
                  nonh_white + RIDAGEYR,
                data = nnyfs1.new, family="binomial"))
```

```

Call:  glm(formula = asthma ~ female + INDFMPIR + PAQ706 + nonh_white +
    RIDAGEYR, family = "binomial", data = nnyfs1.new)

Coefficients:
(Intercept)       female     INDFMPIR       PAQ706   nonh_white
   -1.66525      0.22117     -0.13224     -0.08193      0.06939
   RIDAGEYR
    0.07133

Degrees of Freedom: 1569 Total (i.e. Null);  1564 Residual
Null Deviance:      1444
Residual Deviance: 1405     AIC: 1417
```

We could compare the models using:

- Analysis of Deviance
- AIC
- BIC

```r
anova(m9_int)
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: asthma

Terms added sequentially (first to last)

                    Df Deviance Resid. Df Resid. Dev
NULL                                 1569     1444.4
female               1   3.5802      1568     1440.8
INDFMPIR             1   5.8215      1567     1435.0
PAQ706               1  16.1256      1566     1418.9
nonh_white           1   0.3509      1565     1418.5
RIDAGEYR             1  13.1171      1564     1405.4
nonh_white:RIDAGEYR  1   0.0208      1563     1405.4
```

```r
anova(m9_noint, m9_int)
```

```
Analysis of Deviance Table

Model 1: asthma ~ female + INDFMPIR + PAQ706 + nonh_white + RIDAGEYR
Model 2: asthma ~ female + INDFMPIR + PAQ706 + nonh_white * RIDAGEYR
  Resid. Df Resid. Dev Df Deviance
1      1564     1405.4            
2      1563     1405.4  1 0.020752
```

Either way, the drop in deviance is 0.021, using up 1 df, which is nowhere near statistically significant. The p value is about 0.89 for the interaction term.

```r
pchisq(0.02075, 1, lower.tail = FALSE)
```

```
[1] 0.8854621
```

```r
glance(m9_noint)
```

```
  null.deviance df.null   logLik      AIC      BIC deviance df.residual
1      1444.409    1569 -702.707 1417.414 1449.567 1405.414        1564
```

```r
glance(m9_int)
```

```
  null.deviance df.null    logLik      AIC      BIC deviance df.residual
1      1444.409    1569 -702.6966 1419.393 1456.905 1405.393        1563
```

The AIC and BIC are lower for the model without the interaction, as well.

So, using any of those methods, the model wiout the interaction seems more appropriate. That model is:

- log odds of `asthma` = -1.66 + 0.22 `female` - 0.13 `INDFMPIR` - 0.08 `PAQ706` + 0.07 `nonh_white` + 0.07 `RIDAGEYR`

9.2 Using the `lrm` approach

We can again fit the two models, and then compare them with ANOVA, AIC or BIC, and probably also several other measures.

```
(m9_int_lrm <- lrm(asthma ~ female + INDFMPIR + PAQ706 +
                  nonh_white*RIDAGEYR,
                  data = nnyfs1.new, x=TRUE, y=TRUE))
```

```
Logistic Regression Model

 lrm(formula = asthma ~ female + INDFMPIR + PAQ706 + nonh_white *
     RIDAGEYR, data = nnyfs1.new, x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           1570    LR chi2      39.02    R2       0.041    C       0.616
  0            1299    d.f.             6    g        0.477    Dxy     0.232
  1             271    Pr(> chi2) <0.0001    gr       1.612    gamma   0.232
 max |deriv| 4e-10                           gp       0.067    tau-a   0.066
                                             Brier    0.139

                       Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept             -1.6441 0.3522 -4.67  <0.0001
 female                 0.2209 0.1365  1.62  0.1055
 INDFMPIR              -0.1326 0.0478 -2.77  0.0055
 PAQ706                -0.0819 0.0320 -2.56  0.0104
 nonh_white             0.0161 0.3995  0.04  0.9679
 RIDAGEYR               0.0692 0.0248  2.79  0.0053
 nonh_white * RIDAGEYR  0.0055 0.0381  0.14  0.8855
```

```
(m9_noint_lrm <- lrm(asthma ~ female + INDFMPIR + PAQ706 +
                  nonh_white + RIDAGEYR,
                  data = nnyfs1.new, x=TRUE, y=TRUE))
```

```
Logistic Regression Model

 lrm(formula = asthma ~ female + INDFMPIR + PAQ706 + nonh_white +
     RIDAGEYR, data = nnyfs1.new, x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           1570    LR chi2      39.00    R2       0.041    C       0.616
  0            1299    d.f.             5    g        0.476    Dxy     0.233
  1             271    Pr(> chi2) <0.0001    gr       1.610    gamma   0.233
 max |deriv| 3e-10                           gp       0.067    tau-a   0.067
                                             Brier    0.139

            Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept  -1.6653 0.3206 -5.19  <0.0001
 female      0.2212 0.1365  1.62  0.1051
 INDFMPIR   -0.1322 0.0477 -2.77  0.0056
 PAQ706     -0.0819 0.0320 -2.56  0.0104
 nonh_white  0.0694 0.1491  0.47  0.6417
 RIDAGEYR    0.0713 0.0199  3.59  0.0003
```

For instance, the models with and without interaction have the same C statistic, and Nagelkerke R^2, so the interaction cannot be doing much, and the p value for the `nonh_white * RIDAGEYR` interaction term is, again, 0.89. I could run AIC and BIC again for the lrm fit, but they will yield the same results we saw previously.

Any way you look at it, the interaction term doesn't add anything of importance to the model, so a model without interaction seems more sensible. That model is:

- log odds of asthma = -1.66 + 0.22 female - 0.13 INDFMPIR - 0.08 PAQ706 + 0.07 nonh_white + 0.07 RIDAGEYR