432 Homework 2 Answer Sketch

Thomas E. Love

Due 2018-02-02. Version: 2018-02-01

Contents

0.1 Setup and Data Ingest

```
library(skimr)
library(broom)
library(leaps)
library(modelr)
```

```
Attaching package: 'modelr'

The following object is masked from 'package:broom':

    bootstrap
```

```
library(tidyverse)
```

```
-- Attaching packages ---------------------------------------------------------------------------------
v ggplot2 2.2.1     v purrr   0.2.4
v tibble  1.4.2     v dplyr   0.7.4
v tidyr   0.7.2     v stringr 1.2.0
v readr   1.1.1     v forcats 0.2.0

-- Conflicts ------------------------------------------------------------------------------------------
x modelr::bootstrap()  masks broom::bootstrap()
x dplyr::contains()    masks skimr::contains()
x dplyr::ends_with()   masks skimr::ends_with()
```

```
x dplyr::everything()  masks skimr::everything()
x dplyr::filter()      masks stats::filter()
x dplyr::lag()         masks stats::lag()
x dplyr::matches()     masks skimr::matches()
x dplyr::num_range()   masks skimr::num_range()
x dplyr::one_of()      masks skimr::one_of()
x dplyr::starts_with() masks skimr::starts_with()
```

```
skim_with(numeric = list(hist = NULL), integer = list(hist = NULL))
```

Note: I loaded the data for this assignment into a subfolder of my R Project directory for Homework 2 called `data`. Hence, I use the following command to load in the `hbp330.csv` data.

```
hbp330 <- read.csv("data/hbp330.csv") %>% tbl_df
```

1 Question 1 (30 points)

> Consider the `hbp330` data used in Homework 1. Fit and interpret an ANOVA model to evaluate the effect of race on income. What conclusions can you draw? In developing an answer, please decide whether collapsing the race factor into a smaller number of levels would be sensible in this case. Be sure to provide a written explanation of your findings, in complete sentences.

1.1 A Smaller Data Set

We'll select the variables we need for questions 1-3 in this homework, and then look over that new data set.

```
hw2_small <- hbp330 %>%
  select(subject, income, race, sex, insurance)

skim(hw2_small)
```

```
Skim summary statistics
 n obs: 330
 n variables: 5

Variable type: factor
  variable missing complete   n n_unique
 insurance       0      330 330        4
      race       2      328 330        4
       sex       0      330 330        2
   subject       0      330 330      330
                          top_counts ordered
 Med: 134, Med: 130, Com: 53, Uni: 13   FALSE
  Bla: 180, Whi: 131, Asi: 10, Mul: 7   FALSE
                F: 203, M: 127, NA: 0   FALSE
       A00: 1, A00: 1, A00: 1, A00: 1   FALSE

Variable type: integer
 variable missing complete   n     mean       sd  p0   p25 median   p75
   income       0      330 330 35243.33 16056.44 100 25600  30600 42475
   p100
 147400
```

We have two missing values in the `race` variable, out of a total of 330 people, and given that this only affects less than 1% of the subjects in all, I think we'll just omit those cases for questions 1-3.

```
hw2_q13 <- hw2_small %>% na.omit()

hw2_q13
```

```
# A tibble: 328 x 5
   subject income race         sex   insurance
   <fct>    <int> <fct>        <fct> <fct>
 1 A169     42900 Black/AA     F     Commercial
 2 B036     67300 White        M     Medicare
 3 B103     26100 Black/AA     M     Medicaid
 4 A090     23900 Black/AA     M     Medicaid
 5 B118     25300 Multi-Racial M     Medicaid
 6 B105     25900 White        F     Medicaid
 7 B078     28700 White        F     Medicare
 8 B018     30500 White        F     Medicaid
 9 B108     45200 Asian/PI     F     Medicaid
10 A009     28000 Black/AA     F     Commercial
# ... with 318 more rows
```

1.2 Should we collapse the `race` categories?

```
hw2_q13 %>% count(race)
```

```
# A tibble: 4 x 2
  race             n
  <fct>        <int>
1 Asian/PI        10
2 Black/AA       180
3 Multi-Racial     7
4 White          131
```

The Asian/Pacific Islander and Multi-Racial categories are quite small. Perhaps it would make sense to collapse them together. We'll do so, into a new factor called race_3 (for three categories) and we'll also reorder the categories in order of median income.

```
hw2_q13 <- hw2_q13 %>%
  mutate(race_3 = fct_collapse(race,
                    Other = c("Asian/PI", "Multi-Racial"))) %>%
  mutate(race_3 = fct_reorder(race_3, income, median))
```

and, as a sanity check ...

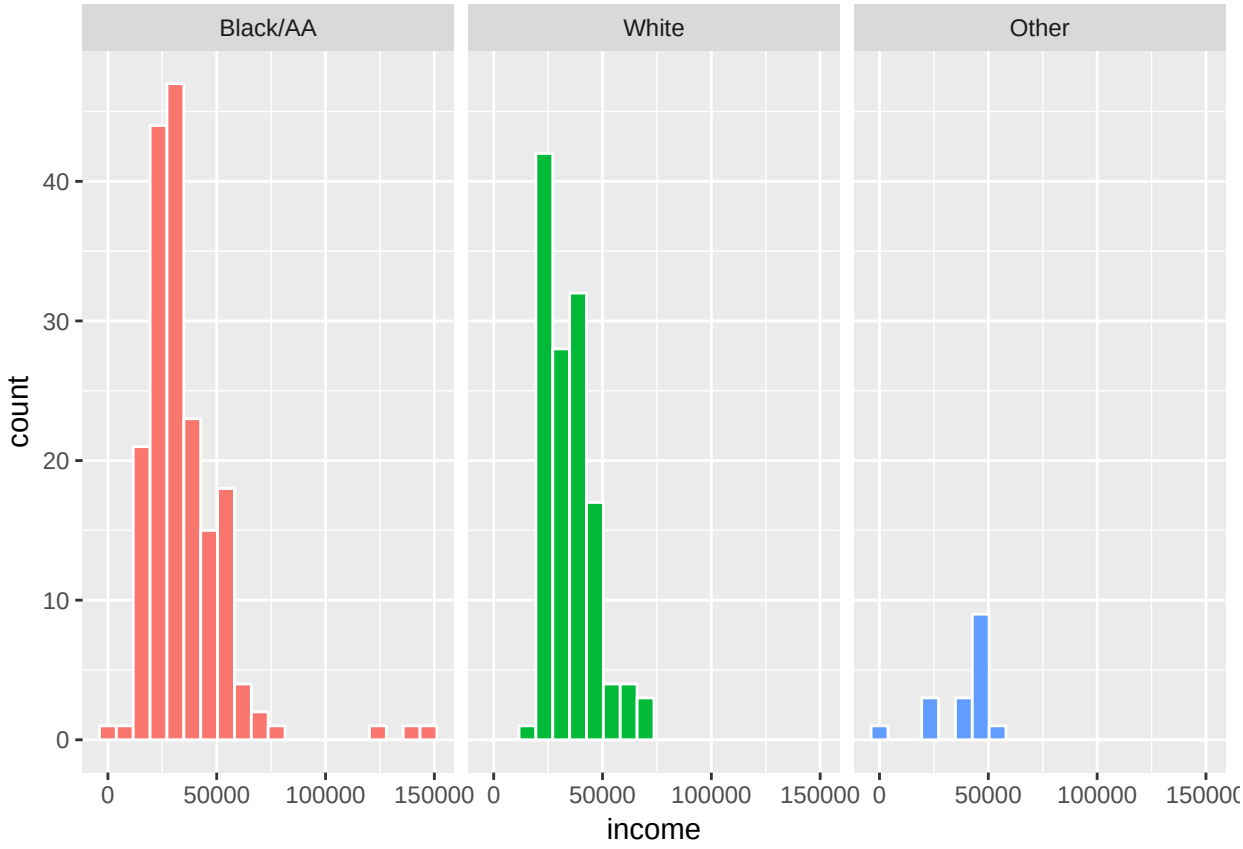
```
hw2_q13 %>% group_by(race_3, race) %>%
  summarize(n = n(), median(income))
```

```
# A tibble: 4 x 4
# Groups:   race_3 [?]
  race_3   race             n `median(income)`
  <fct>    <fct>        <int>            <dbl>
1 Black/AA Black/AA       180            29300
2 White    White          131            30700
3 Other    Asian/PI        10            44600
4 Other    Multi-Racial     7            25600
```

1.3 EDA for `income` by `race_3` group

We need to do some exploratory data analysis. Let's look at the `income` data within the three `race_3` categories.

```
ggplot(hw2_q13, aes(x = income, fill = race_3)) +
  geom_histogram(bins = 20, col = "white") +
  guides(fill = FALSE) +
  facet_wrap(~ race_3)
```



There are three large outliers in the "Black/AA" group, which is a bit surprising, although otherwise there's at most a modest skew apparent in each group. These data look a little right-skewed in each case, but generally sufficiently well-approximated by Normal distributions to let me feel comfortable summarizing them with means and standard deviations, at least to start. Our numerical summaries are:

```
hw2_q13 %>% group_by(race_3) %>%
  skim(income)
```

```
Skim summary statistics
 n obs: 328
 n variables: 6
 group variables: race_3

Variable type: integer
   race_3 variable missing complete   n     mean       sd    p0   p25
 Black/AA   income       0      180 180 34710    19117.89   200 24700
    White   income       0      131 131 35456.49 11222.64 15800 25850
    Other   income       0       17  17 38600    13010.91   100 37800
```

```
 median   p75   p100
  29300 41550 147400
  30700 42100  71400
  44300 45200  54000
```

1.4 Building the ANOVA model

This is a one-way analysis of variance model.

```
hw2_model1 <- lm(income ~ race_3, data = hw2_q13)
anova(hw2_model1)
```

```
Analysis of Variance Table

Response: income
           Df     Sum Sq   Mean Sq F value Pr(>F)
race_3      2 2.4832e+08 124162398  0.4775 0.6208
Residuals 325 8.4505e+10 260015766
```

```
summary(hw2_model1)
```

```
Call:
lm(formula = income ~ race_3, data = hw2_q13)

Residuals:
   Min     1Q Median     3Q    Max
-38500  -9822  -4783   6636 112690

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34710.0     1201.9  28.880   <2e-16 ***
race_3White    746.5     1851.9   0.403    0.687
race_3Other   3890.0     4091.4   0.951    0.342
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16130 on 325 degrees of freedom
Multiple R-squared:  0.00293,   Adjusted R-squared:  -0.003206
F-statistic: 0.4775 on 2 and 325 DF,  p-value: 0.6208
```

```
TukeyHSD(aov(income ~ race_3, data = hw2_q13))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = income ~ race_3, data = hw2_q13)

$race_3
                     diff       lwr       upr     p adj
White-Black/AA  746.4885 -3613.720  5106.697 0.9143603
Other-Black/AA 3890.0000 -5743.217 13523.217 0.6085467
Other-White    3143.5115 -6643.941 12930.964 0.7300825
```

Our conclusion from the Tukey HSD comparisons, and from the ANOVA F test in the `anova` and `summary` output for the linear model is that there are no statistically significant differences in income across our three

race groups. This is still true (see below) even if we don't separate out the two small groups in the original `race` variable.

```
anova(lm(income ~ race, data = hw2_q13))
```

```
Analysis of Variance Table

Response: income
           Df     Sum Sq   Mean Sq F value Pr(>F)
race        3 1.2618e+09 420602170  1.6322 0.1818
Residuals 324 8.3492e+10 257690254
```

2 Question 2 (20 points)

> Now fit a two-factor ANOVA model to evaluate the effects of `race` and `sex` on `income`. What can you conclude? Be sure to provide a written explanation of your findings, in complete sentences.

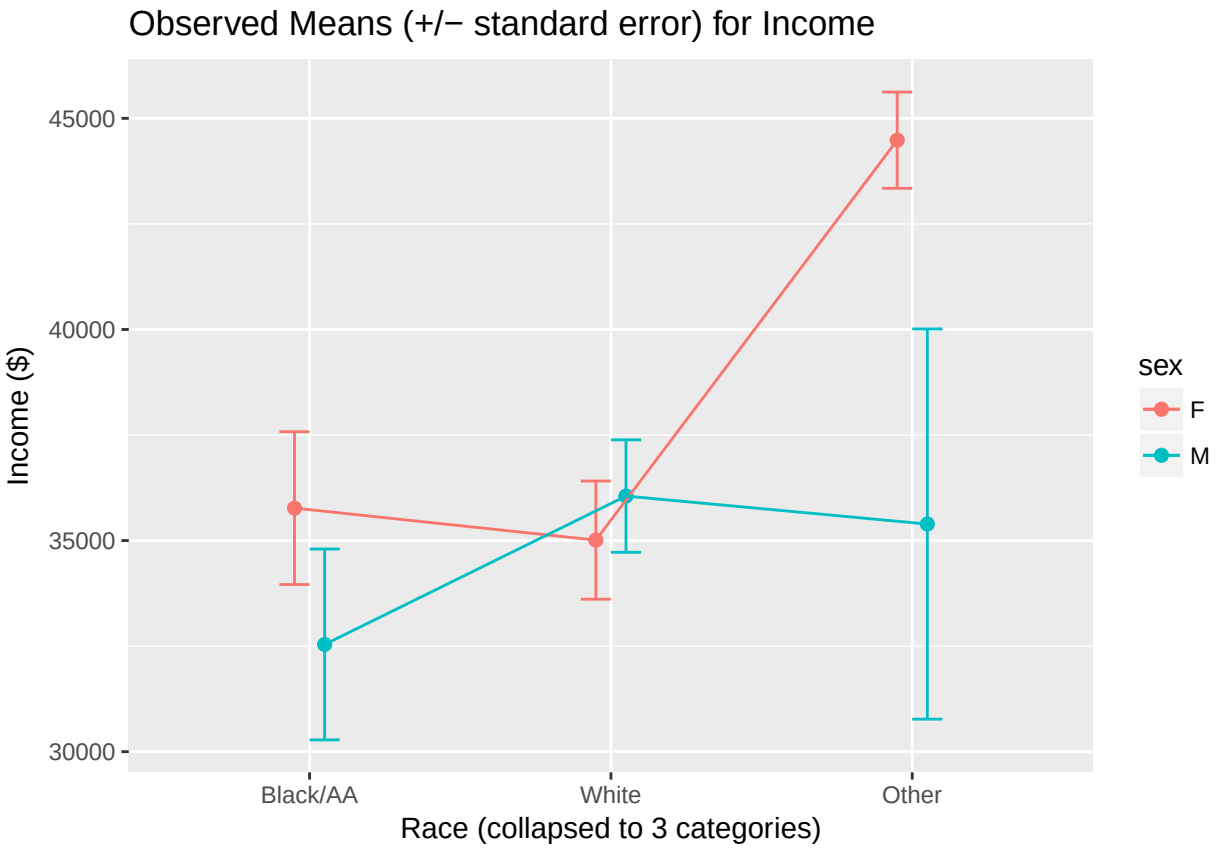
2.1 The ANOVA model with interaction

2.1.1 A Means Plot to look for meaningful interaction

```
hw2q2_summary <- hw2_q13 %>%
  group_by(race_3, sex) %>%
  summarize(meaninc = mean(income), seinc = sd(income)/sqrt(n()) )

pd <- position_dodge(0.2)

ggplot(hw2q2_summary, aes(x = race_3, y = meaninc, color = sex)) +
  geom_errorbar(aes(ymin = meaninc - seinc,
                    ymax = meaninc + seinc),
                width = 0.2, position = pd) +
  geom_point(size = 2, position = pd) +
  geom_line(aes(group = sex), position = pd) +
  labs(y = "Income ($)",
       x = "Race (collapsed to 3 categories)",
       title = "Observed Means (+/- standard error) for Income")
```



Note that if you fail to collapse the Race groups, then the Multi-Racial group will throw an error when you try to plot error bars, because a standard deviation (and thus a standard error) cannot be estimated.

It looks like an interaction might be useful in this situation, as the lines are not parallel, but it's not clear that the Other group is providing a lot of useful information.

2.1.2 ANOVA test for the model

```
hw2_model2_with_int <- lm(income ~ race_3*sex, data = hw2_q13)
anova(hw2_model2_with_int)
```

```
Analysis of Variance Table

Response: income
            Df     Sum Sq   Mean Sq F value Pr(>F)
race_3       2 2.4832e+08 124162398  0.4775 0.6208
sex          1 2.2337e+08 223371203  0.8590 0.3547
race_3:sex   2 5.4614e+08 273071802  1.0501 0.3511
Residuals  322 8.3736e+10 260048476
```

It doesn't look like the interaction term is significant, however, although it does account for more variation than the `race_3` or `sex` main effects within this model. The conclusion would be that there aren't any statistically significant differences in `income` attributable to either `race_3` or `sex`.

```
summary(hw2_model2_with_int)
```

```
Call:
lm(formula = income ~ race_3 * sex, data = hw2_q13)

Residuals:
   Min     1Q Median     3Q    Max
-35569  -9611  -4625   6793 111631

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       35768.6     1466.0  24.399   <2e-16 ***
race_3White        -757.9     2369.9  -0.320    0.749
race_3Other        8714.7     6744.7   1.292    0.197
sexM              -3229.6     2560.6  -1.261    0.208
race_3White:sexM   4272.5     3829.9   1.116    0.265
race_3Other:sexM  -5862.8     8575.5  -0.684    0.495
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16130 on 322 degrees of freedom
Multiple R-squared:  0.01201,   Adjusted R-squared:  -0.003332
F-statistic: 0.7828 on 5 and 322 DF,  p-value: 0.5627
```

2.2 The ANOVA model without interaction

A model without interaction also finds no statistically significant differences in `income` by either `race_3` or `sex`.

```
hw2_model2_without <- lm(income ~ race + sex, data = hw2_q13)
anova(hw2_model2_without)
```

```
Analysis of Variance Table

Response: income
           Df     Sum Sq   Mean Sq F value Pr(>F)
race        3 1.2618e+09 420602170  1.6301 0.1822
sex         1 1.5216e+08 152155701  0.5897 0.4431
Residuals 323 8.3339e+10 258016986
```

3 Question 3 (20 points)

> Now attempt to fit a two-factor ANOVA model to evaluate the effect of (uncollapsed) `race` and `insurance` on `income`. A problem should occur when you fit this `race` and `insurance` model, that doesn't happen, for instance, when you evaluate the effects of both `race` and `sex` on `income`. So what happens when you fit the `race-insurance` model, exactly, and why does it happen?

```
hw2_model3 <- lm(income ~ race_3*insurance, data = hw2_q13)
anova(hw2_model3)
```

```
Analysis of Variance Table

Response: income
                 Df     Sum Sq   Mean Sq F value Pr(>F)
race_3            2 2.4832e+08 124162398  0.4736 0.6232
```

```
insurance          3 1.0934e+09 364478131  1.3902 0.2457
race_3:insurance   4 3.7922e+07   9480464  0.0362 0.9975
Residuals        318 8.3374e+10 262181660
```

That p value for the interaction term looks a little high. What's happening?

```
summary(hw2_model3)
```

```
Call:
lm(formula = income ~ race_3 * insurance, data = hw2_q13)

Residuals:
   Min     1Q Median     3Q    Max
-37885  -9316  -3923   6638 113107

Coefficients: (2 not defined because of singularities)
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   36990.91    2818.67  13.124   <2e-16 ***
race_3White                    1074.09    4588.46   0.234    0.815
race_3Other                    6306.76    8311.94   0.759    0.449
insuranceMedicaid             -3393.81    3427.04  -0.990    0.323
insuranceMedicare             -2697.67    3389.38  -0.796    0.427
insuranceUninsured             5809.09    8572.65   0.678    0.498
race_3White:insuranceMedicaid  -998.12    5467.69  -0.183    0.855
race_3Other:insuranceMedicaid -1919.24    9646.55  -0.199    0.842
race_3White:insuranceMedicare   -95.33    5462.66  -0.017    0.986
race_3Other:insuranceMedicare        NA         NA      NA       NA
race_3White:insuranceUninsured -2885.20   10757.82  -0.268    0.789
race_3Other:insuranceUninsured       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16190 on 318 degrees of freedom
Multiple R-squared:  0.01628,   Adjusted R-squared:  -0.01156
F-statistic: 0.5847 on 9 and 318 DF,  p-value: 0.8096
```

Aha - we've got some terms that the model cannot fit - `NA` values in the estimates are a big problem.

3.1 Exploring the Data - Why Can't We Estimate all of our Coefficients?

As to why this happens, a little more exploratory data analysis would tell us...

```
hw2_q13 %>% count(race_3, insurance)
```

```
# A tibble: 10 x 3
   race_3   insurance      n
   <fct>    <fct>      <int>
 1 Black/AA Commercial    33
 2 Black/AA Medicaid      69
 3 Black/AA Medicare      74
 4 Black/AA Uninsured      4
 5 White    Commercial    20
 6 White    Medicaid      52
 7 White    Medicare      50
 8 White    Uninsured      9
```

```
 9 Other    Medicaid      13
10 Other    Medicare       4
```

We see that for the "Other" `race_3` group, we only observe subjects with Medicaid and Medicare insurance. So the model cannot fit the interaction of `race_3` with `insurance`, because it cannot make either a "Other race, Commercial" or "Other race, Uninsured" estimate.

- Note that the `NA` values don't correspond to the counts of 0. That's because of the order in which the models are estimated. If, instead of running `race_3 * insurance` you instead run `insurance * race_3` you get the following...

```
hw2_model3a <- lm(income ~ insurance*race_3, data = hw2_q13)
summary(hw2_model3a)
```

```
Call:
lm(formula = income ~ insurance * race_3, data = hw2_q13)

Residuals:
   Min     1Q Median     3Q    Max
-37885  -9316  -3923   6638 113107

Coefficients: (2 not defined because of singularities)
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   36990.91    2818.67  13.124   <2e-16 ***
insuranceMedicaid             -3393.81    3427.04  -0.990    0.323
insuranceMedicare             -2697.67    3389.38  -0.796    0.427
insuranceUninsured             5809.09    8572.65   0.678    0.498
race_3White                    1074.09    4588.46   0.234    0.815
race_3Other                    6306.76    8311.94   0.759    0.449
insuranceMedicaid:race_3White  -998.12    5467.69  -0.183    0.855
insuranceMedicare:race_3White   -95.33    5462.66  -0.017    0.986
insuranceUninsured:race_3White -2885.20   10757.82  -0.268    0.789
insuranceMedicaid:race_3Other  -1919.24    9646.55  -0.199    0.842
insuranceMedicare:race_3Other         NA         NA      NA       NA
insuranceUninsured:race_3Other        NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16190 on 318 degrees of freedom
Multiple R-squared:  0.01628,   Adjusted R-squared:  -0.01156
F-statistic: 0.5847 on 9 and 318 DF,  p-value: 0.8096
```

Now, at least one of the two NAs corresponds to a count of zero. Changing the order of the levels in the `race_3` and/or insurance factors which also have an impact on which estimates are missing in this output.

- There's no doubt about it. You really do need to look at the data closely.

4 Question 4 (30 points)

Again, consider the `hbp330` data used in Homework 1. Build your best model for the prediction of body-mass index, considering the following 14 predictors: `practice`, `age`, `race`, `eth_hisp`, `sex`, `insurance`, `income`, `hsgrad`, `tobacco`, `depdiag`, `sbp`, `dbp`, `statin` and `bpmed`. Use an appropriate best subsets procedure to aid in your search, and use a cross-validation strategy to assess and compare potential models.

- Feel free to omit the cases with missing values in the variables you are considering (these 14 predictors, plus the `bmi` outcome) before proceeding. This should not materially affect your sample size very much.
- Use the `nvmax = 7` command within your call to `regsubsets` to limit your investigation to models containing no more than seven of these candidate predictors.
- Do not transform any variables, and consider models with main effects only so that no product terms are used.
- A 5-fold cross-validation strategy would be very appropriate. Another reasonable choice would involve partitioning the data once (prior to fitting any models) into training and test samples, as we did in 431.

  Be sure to provide a written explanation of your conclusions and specify the variables in your final model, in complete sentences.

```
hw2q4 <- hbp330 %>%
  mutate( bmi = weight / (height*height) ) %>%
  select(subject, bmi, practice, age, race, eth_hisp, sex,
                   insurance, income, hsgrad, tobacco,
                   depdiag, sbp, dbp, statin, bpmed) %>%
  drop_na

skim(hw2q4)
```

```
Skim summary statistics
 n obs: 325
 n variables: 16

Variable type: factor
  variable missing complete   n n_unique
   depdiag       0      325 325        2
  eth_hisp       0      325 325        2
 insurance       0      325 325        4
  practice       0      325 325        2
      race       0      325 325        4
       sex       0      325 325        2
   subject       0      325 325      325
   tobacco       0      325 325        3
                          top_counts ordered
             No: 211, Yes: 114, NA: 0   FALSE
              No: 261, Yes: 64, NA: 0   FALSE
 Med: 131, Med: 128, Com: 53, Uni: 13   FALSE
                A: 176, B: 149, NA: 0   FALSE
  Bla: 178, Whi: 131, Asi: 10, Mul: 6   FALSE
                F: 201, M: 124, NA: 0   FALSE
       A00: 1, A00: 1, A00: 1, A00: 1   FALSE
   nev: 138, for: 115, cur: 72, NA: 0   FALSE

Variable type: integer
 variable missing complete   n     mean       sd  p0   p25 median   p75
      age       0      325 325    55.5     11.53  23    48     57    65
    bpmed       0      325 325     0.66     0.48   0     0      1     1
      dbp       0      325 325    74.73    10.24  41    68     74    82
   hsgrad       0      325 325    81.71     9.66   0    75     81    89
   income       0      325 325 35430.46 15987.45 200 25600  30600 42600
      sbp       0      325 325   128.28    17.39  84   116    128   138
   statin       0      325 325     0.7      0.46   0     0      1     1
   p100
```

```
     77
      1
    106
    100
 147400
    194
      1

Variable type: numeric
 variable missing complete   n  mean   sd    p0   p25 median   p75  p100
      bmi       0      325 325 34.83 8.05 16.73 29.73  33.91 39.22 64.04
```

We lose a total of five observations by dropping missing values. Next, we'll establish the "best subsets" groups.

```
q4_preds <- with(hw2q4,
              cbind(practice, age, race, eth_hisp, sex,
                    insurance, income, hsgrad, tobacco,
                    depdiag, sbp, dbp, statin, bpmed))

q4_subs <- regsubsets(q4_preds, y = hw2q4$bmi, nvmax = 7)

q4_rs <- summary(q4_subs)

q4_rs
```

```
Subset selection object
14 Variables  (and intercept)
          Forced in Forced out
practice      FALSE      FALSE
age           FALSE      FALSE
race          FALSE      FALSE
eth_hisp      FALSE      FALSE
sex           FALSE      FALSE
insurance     FALSE      FALSE
income        FALSE      FALSE
hsgrad        FALSE      FALSE
tobacco       FALSE      FALSE
depdiag       FALSE      FALSE
sbp           FALSE      FALSE
dbp           FALSE      FALSE
statin        FALSE      FALSE
bpmed         FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         practice age race eth_hisp sex insurance income hsgrad tobacco
1  ( 1 ) " "      " " " "  " "      "*" " "       " "    " "    " "
2  ( 1 ) " "      "*" " "  " "      "*" " "       " "    " "    " "
3  ( 1 ) " "      "*" " "  " "      "*" " "       " "    " "    " "
4  ( 1 ) " "      "*" " "  " "      "*" " "       " "    " "    "*"
5  ( 1 ) "*"      "*" "*"  " "      "*" " "       " "    " "    " "
6  ( 1 ) "*"      "*" "*"  " "      "*" " "       " "    " "    "*"
7  ( 1 ) "*"      "*" "*"  " "      "*" " "       " "    " "    "*"
         depdiag sbp dbp statin bpmed
1  ( 1 ) " "     " " " " " "    " "
2  ( 1 ) " "     " " " " " "    " "
```

```
3  ( 1 ) " "     " " " " " " " "    "*"
4  ( 1 ) " "     " " " " " " " "    "*"
5  ( 1 ) " "     " " " " " " " "    "*"
6  ( 1 ) " "     " " " " " " " "    "*"
7  ( 1 ) "*"     " " " " " " " "    "*"
```

```
round(q4_rs$adjr2, 4)
```

```
[1] 0.0373 0.0476 0.0634 0.0750 0.0889 0.1006 0.1008
```

```
round(q4_rs$cp, 1)
```

```
[1] 19.1 16.4 11.8  8.7  4.8  1.8  2.7
```

```
round(q4_rs$bic, 1)
```

```
[1] -1.8 -0.5 -1.2 -0.5 -0.6  0.0  4.6
```

```
# since n for hw2q4 is 325, and we are looking at 2-8 inputs
q4_rs$aic.corr <- 325*log(q4_rs$rss / 325) + 2*(2:8) +
                  (2 * (2:8) * ((2:8)+1) / (325 - (2:8) - 1))

round(q4_rs$aic.corr, 1)
```

```
[1] 1345.3 1342.8 1338.4 1335.4 1331.6 1328.5 1329.4
```

So, here are our "best subsets" models:

Inputs	Predictors Included	Adj. r^2	C_p	BIC	corr. AIC
2	sex	0.0373	19.1	**-1.8**	1345.3
3	sex, age	0.0476	16.4	-0.5	1342.8
4	sex, age, bpmed	0.0634	11.8	-1.2	1338.4
5	sex, age, bpmed, tobacco	0.0750	8.7	-0.5	1335.4
6	sex, age, bpmed, practice, race	0.0889	**4.8**	-0.6	1331.6
7	sex, age, bpmed, tobacco, practice, race	0.1006	1.8	0.0	**1328.5**
8	sex, age, bpmed, tobacco, practice, race, depdiag	**0.1008**	2.7	4.6	1329.4

4.1 Building our 4 Plots

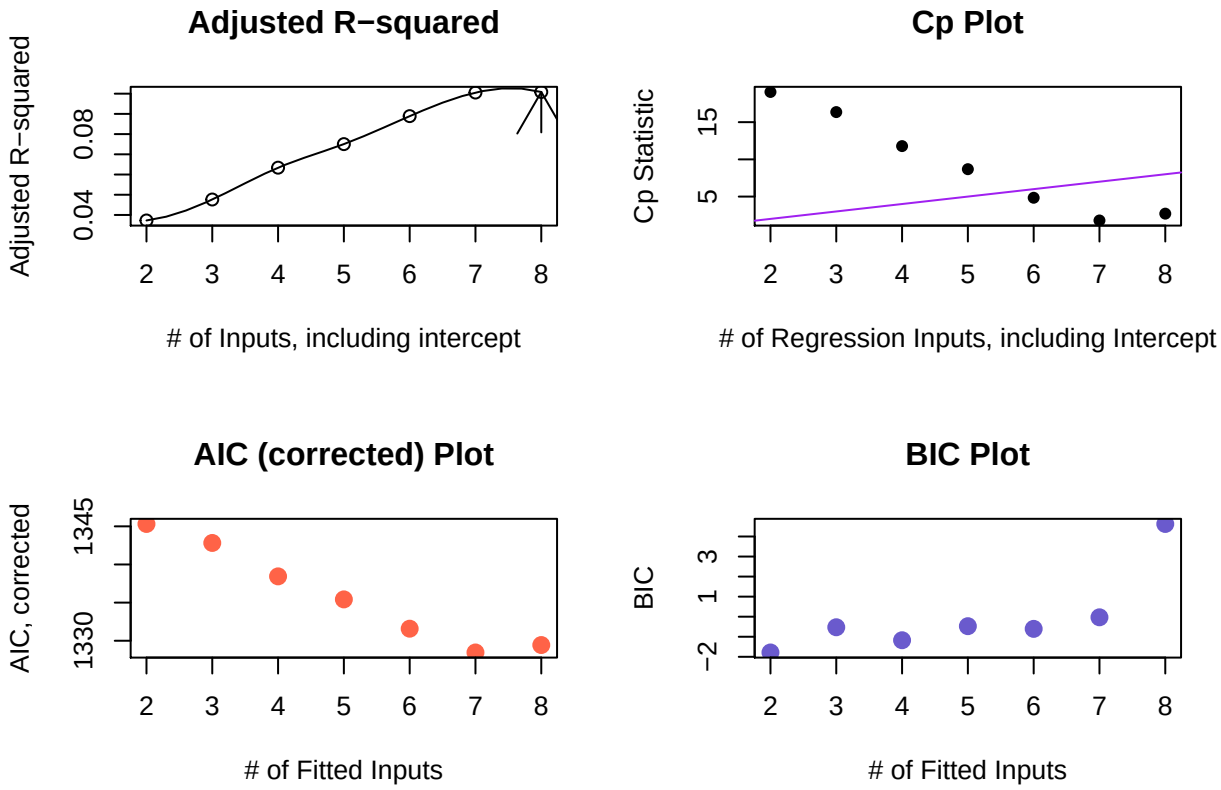
```
par(mfrow = c(2,2))
m2 <- max(q4_rs$adjr2)
m1 <- which.max(q4_rs$adjr2) + 1
plot(q4_rs$adjr2 ~ I(2:8), ylab="Adjusted R-squared",
     xlab="# of Inputs, including intercept",
     main = "Adjusted R-squared")
lines(spline(q4_rs$adjr2 ~ I(2:8)))
arrows(m1, m2-0.02, m1, m2)

plot(q4_rs$cp ~ I(2:8),
     ylab="Cp Statistic",
     xlab="# of Regression Inputs, including Intercept",
     pch=16, main="Cp Plot")
abline(0,1, col = "purple")

plot(q4_rs$aic.corr ~ I(2:8), ylab="AIC, corrected", xlab="# of Fitted Inputs",
```

```
     pch=16, cex=1.5, col="tomato", main="AIC (corrected) Plot")

plot(q4_rs$bic ~ I(2:8), ylab="BIC", xlab="# of Fitted Inputs",
     pch=16, cex=1.5, col="slateblue", main="BIC Plot")
```



4.2 Selecting a Winner

The models we'll consider are:

Inputs	Predictors Included	Reason
2	`sex`	lowest BIC
6	`sex`, `age`, `bpmed`, `practice`, `race`	suggested by C_p
7	`sex`, `age`, `bpmed`, `tobacco`, `practice`, `race`	lowest AIC (corr.)
8	`sex`, `age`, `bpmed`, `tobacco`, `practice`, `race`, `depdiag`	highest adj. R^2

We'll fit each of these four models in turn, and then perform a 5-fold cross validation for each, then compare results. In each case, we'll calculate the root mean squared error of the predictions, and the mean absolute prediction error across the complete samples.

4.2.1 Model 2 cross-validation

```r
set.seed(4320142)

q4m2 <- hw2q4 %>%
  crossv_kfold(k = 5) %>%
  mutate(model = map(train, ~ lm(bmi ~ sex, data = .)))

q4m2_pred <- q4m2 %>%
  unnest(map2(model, test, ~ augment(.x, newdata = .y)))

res2 <- q4m2_pred %>%
  summarize(Model = "2",
            RMSE = sqrt(mean((bmi - .fitted) ^2)),
            MAE = mean(abs(bmi - .fitted)))
```

4.2.2 Model 6 cross-validation

```r
set.seed(4320146)

q4m6 <- hw2q4 %>%
  crossv_kfold(k = 5) %>%
  mutate(model = map(train,
                     ~ lm(bmi ~ sex + age + bpmed +
                            practice + race, data = .)))

q4m6_pred <- q4m6 %>%
  unnest(map2(model, test, ~ augment(.x, newdata = .y)))

res6 <- q4m6_pred %>%
  summarize(Model = "6",
            RMSE = sqrt(mean((bmi - .fitted) ^2)),
            MAE = mean(abs(bmi - .fitted)))
```

4.2.3 Model 7 cross-validation

```r
set.seed(4320147)

q4m7 <- hw2q4 %>%
  crossv_kfold(k = 5) %>%
  mutate(model = map(train,
                     ~ lm(bmi ~ sex + age + bpmed + tobacco +
                            practice + race, data = .)))

q4m7_pred <- q4m7 %>%
  unnest(map2(model, test, ~ augment(.x, newdata = .y)))

res7 <- q4m7_pred %>%
  summarize(Model = "7",
            RMSE = sqrt(mean((bmi - .fitted) ^2)),
            MAE = mean(abs(bmi - .fitted)))
```

4.2.4 Model 8 cross-validation

```
set.seed(4320148)

q4m8 <- hw2q4 %>%
  crossv_kfold(k = 5) %>%
  mutate(model = map(train,
                     ~ lm(bmi ~ sex + age + bpmed + tobacco +
                            practice + race + depdiag, data = .)))

q4m8_pred <- q4m8 %>%
  unnest(map2(model, test, ~ augment(.x, newdata = .y)))

res8 <- q4m8_pred %>%
  summarize(Model = "8",
            RMSE = sqrt(mean((bmi - .fitted) ^2)),
            MAE = mean(abs(bmi - .fitted)))
```

4.2.5 Summary Table

```
bind_rows(res2, res6, res7, res8)
```

```
# A tibble: 4 x 3
  Model  RMSE   MAE
  <chr> <dbl> <dbl>
1 2      7.92  6.12
2 6      7.83  6.02
3 7      7.59  5.85
4 8      7.59  5.85
```

Model 7 yielded slightly better predictions in terms of RMSE or MAE than the other options here. So that's the model including `sex`, `age`, `bpmed`, `tobacco`, `practice`, and `race`.

Refitting this model to the complete case sample of 325 people, we have the following summary results.

```
summary(lm(bmi ~ sex + age + bpmed + tobacco + practice +
             race, data = hw2q4))
```

```
Call:
lm(formula = bmi ~ sex + age + bpmed + tobacco + practice + race,
    data = hw2q4)

Residuals:
     Min       1Q   Median       3Q      Max
-19.2223  -4.9598  -0.9021   4.2354  26.9390

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     36.11158    3.65461   9.881  < 2e-16 ***
sexM            -3.37685    0.88860  -3.800 0.000174 ***
age             -0.14015    0.03849  -3.641 0.000317 ***
bpmed            2.31824    0.91406   2.536 0.011689 *
tobaccoformer    4.19974    1.16306   3.611 0.000355 ***
```

```
tobacconever      3.08996    1.11643   2.768 0.005979 **
practiceB        -2.62935    1.61177  -1.631 0.103818
raceBlack/AA      3.62214    2.87113   1.262 0.208036
raceMulti-Racial  2.02264    4.02215   0.503 0.615403
raceWhite         6.57284    2.49496   2.634 0.008844 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.554 on 315 degrees of freedom
Multiple R-squared:  0.1437,    Adjusted R-squared:  0.1192
F-statistic: 5.873 on 9 and 315 DF,  p-value: 1.4e-07
```