

Final Code

ALA Mode (Group 2): Anja Shahu, Anna Wuest, Ligia Flores

12/4/2020

```
library(tidyverse)
library(RColorBrewer)

url <- "https://meps.ahrq.gov/mepsweb/data_files/pufs/h209dat.zip"
download.file(url, temp <- tempfile())
meps_path <- unzip(temp, exdir = tempdir())
source("https://meps.ahrq.gov/mepsweb/data_stats/download_data/pufs/h209/h209ru.txt")
unlink(temp)

# creating a reduced data frame including only the variables that we'll be considering
h209red <- data.frame("pap" = h209$ADPAP42,
  "region" = h209$REGION18,
  "race" = h209$RACETHX,
  "age" = h209$AGE18X,
  "marital_stat" = h209$MARRY18X,
  "educ" = h209$EDUCYR,
  "smoke_freq" = h209$OFTSMK53,
  "income_indiv" = h209$TTLP18X,
  "income_fam" = h209$FAMINC18,
  "income_percpov" = h209$POVLEV18,
  "hrsworked_rd1" = h209$HOUR31H,
  "hrsworked_rd2" = h209$HOUR42H,
  "hrsworked_rd3" = h209$HOUR53H,
  "limitation" = h209$ACTLIM31,
  "menhlth_rd1" = h209$MNHLTH31,
  "menhlth_rd2" = h209$MNHLTH42,
  "menhlth_rd3" = h209$MNHLTH53,
  "genhlth_rd1" = h209$RTHLTH31,
  "genhlth_rd2" = h209$RTHLTH42,
  "genhlth_rd3" = h209$RTHLTH53,
  "totexp" = h209$TOTEXP18,
  "outofpocket_exp" = h209$TOTSLF18,
  "afford_care" = h209$AFRDCA42,
  "have_usc" = h209$HAVEUS42,
  "dist_from_usc" = h209$TMTKUS42,
  "rch_usc_byphn" = h209$PHNREG42,
  "usc_offhrs_nw" = h209$OFFHOU42,
  "usc_asks_abt_trts" = h209$TREATM42,
  "usc_asks_hlp_dec" = h209$DECIDE42,
  "usc_expln_options" = h209$EXPLOP42,
  "inscov_gen_2018" = h209$INSCOV18)

rm(h209) # remove original data set from environment
```

```

h209red <- h209red %>%
  as_tibble() %>%
  filter(pap != -1) %>% # filtering out the people who were not asked pap smear question
  filter(age >= 21 & age <= 65) # filtering to women ages 21-65 (note there was 1 inapplicable person t.

## inputting NAs into hours worked variables
h209red$hrsworked_rd1[h209red$hrsworked_rd1 == -1] <- NA
h209red$hrsworked_rd2[h209red$hrsworked_rd2 == -1] <- NA
h209red$hrsworked_rd3[h209red$hrsworked_rd3 == -1] <- NA

# hours worked (rounded average)
h209red <- h209red %>% rowwise() %>%
  mutate(hrs_worked_avg = round(mean(c(hrsworked_rd1, hrsworked_rd2, hrsworked_rd3), na.rm = TRUE)))

h209red$hrs_worked_avg[is.nan(as.numeric(h209red$hrs_worked_avg))] <- NA

summary(h209red$hrs_worked_avg) # too many missing variables to use

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##      1.00   30.00   40.00   36.36   40.00   168.00   1654

# re-calculating the mental and general health variables

# perceived mental heath adding NA
h209red$menhlth_rd1[h209red$menhlth_rd1 == -8 ]<- NA
h209red$menhlth_rd1[h209red$menhlth_rd1 == -1 ]<- NA

h209red$menhlth_rd2[h209red$menhlth_rd2 == -7 ]<- NA
h209red$menhlth_rd2[h209red$menhlth_rd2 == -8 ]<- NA

h209red$menhlth_rd3[h209red$menhlth_rd3 == -7 ]<- NA
h209red$menhlth_rd3[h209red$menhlth_rd3 == -8 ]<- NA
h209red$menhlth_rd3[h209red$menhlth_rd3 == -1 ]<- NA

# perceived mental health, rounded average for each group
h209red <- h209red %>% rowwise() %>%
  mutate(menhlth_avg = round(mean(c(menhlth_rd1, menhlth_rd2, menhlth_rd3), na.rm=TRUE)))

summary(h209red$menhlth_avg)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000   1.000   2.000   2.115   3.000   5.000

# perceived mental health converted to factor
h209red <- h209red %>%
  mutate(menhlth_avg_f = factor(menhlth_avg,
                                levels = c("5", "4", "3", "2", "1"))) %>%
  mutate(menhlth_avg_f = fct_recode(menhlth_avg_f,
                                    "poor" = "5",
                                    "fair" = "4",
                                    "good" = "3",
                                    "very good" = "2",
                                    "excellent" = "1"))

```

```

# re-calculating the general health variables

# perceived general health NA
h209red$genhlth_rd1[h209red$genhlth_rd1 == -8 ] <- NA
h209red$genhlth_rd1[h209red$genhlth_rd1 == -1 ] <- NA

h209red$genhlth_rd2[h209red$genhlth_rd2 == -8 ] <- NA

h209red$genhlth_rd3[h209red$genhlth_rd3 == -7 ] <- NA
h209red$genhlth_rd3[h209red$genhlth_rd3 == -8 ] <- NA
h209red$genhlth_rd3[h209red$genhlth_rd3 == -1 ] <- NA

# perceived mental health, rounded average of each group
h209red <- h209red %>% rowwise() %>%
  mutate(genhlth_avg = round(mean(c(genhlth_rd1, genhlth_rd2, genhlth_rd3), na.rm=TRUE)))

summary(h209red$genhlth_avg)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000   2.000   2.000   2.351   3.000   5.000

# perceived general health converted to factor
h209red <- h209red %>%
  mutate(genhlth_avg_f = factor(genhlth_avg,
                                levels = c("5", "4", "3", "2", "1"))) %>%
  mutate(genhlth_avg_f = fct_recode(genhlth_avg_f,
                                    "poor" = "5",
                                    "fair" = "4",
                                    "good" = "3",
                                    "very good" = "2",
                                    "excellent" = "1"))

# creating factor versions of other categorical variables

# pap status
h209red <- h209red %>%
  mutate(pap_f = factor(pap,
                        levels = c("1", "2", "-15"))) %>%
  mutate(pap_f = fct_recode(pap_f,
                            "yes" = "1",
                            "no" = "2",
                            NULL = "-15"))

# region
h209red <- h209red %>%
  mutate(region_f = factor(region,
                           levels = c("1", "2", "3", "4"))) %>%
  mutate(region_f = fct_recode(region_f,
                                "northeast" = "1",
                                "midwest" = "2",
                                "south" = "3",
                                "west" = "4"))

# race
h209red <- h209red %>%

```

```

mutate(race_f = factor(race,
                      levels = c("2", "1", "3", "4", "5"))) %>%
mutate(race_f = fct_recode(race_f,
                          "white" = "2",
                          "hispanic" = "1",
                          "black" = "3",
                          "asian" = "4",
                          "other or multiple races" = "5"))

# marital status
h209red <- h209red %>%
  mutate(marital_stat_f = factor(marital_stat,
                                levels = c("5", "1", "2", "3", "4"))) %>%
  mutate(marital_stat_f = fct_recode(marital_stat_f,
                                    "never married" = "5",
                                    "married" = "1",
                                    "widowed" = "2",
                                    "divorced" = "3",
                                    "seperated" = "4"))

# education
h209red <- h209red %>%
  mutate(educ_f = factor(educ)) %>%
  mutate(educ_f = fct_collapse(educ_f,
                              "none or any elementary" = c("0", "1", "2", "3", "4", "5", "6", "7", "8"),
                              "any high school" = c("9", "10", "11", "12"),
                              "any college" = c("13", "14", "15", "16", "17"),
                              NULL = "-15",
                              NULL = c("-8", "-7")))

# smoking frequency
h209red <- h209red %>%
  mutate(smoke_freq_f = factor(smoke_freq,
                              levels = c("3", "2", "1", "-8", "-7", "-1"))) %>%
  mutate(smoke_freq_f = fct_recode(smoke_freq_f,
                                  "never" = "3",
                                  "some days" = "2",
                                  "every day" = "1",
                                  NULL = "-8",
                                  NULL = "-7",
                                  NULL = "-1"))

# limitation
h209red <- h209red %>%
  mutate(limitation_f = factor(limitation,
                              levels = c("2", "1", "-8", "-7", "-1"))) %>%
  mutate(limitation_f = fct_recode(limitation_f,
                                  "no" = "2",
                                  "yes" = "1",
                                  NULL = "-8",
                                  NULL = "-7",
                                  NULL = "-1"))

```

```

# ability to afford care
h209red <- h209red %>%
  mutate(afford_care_f = factor(afford_care,
                                levels = c("2", "1", "-8", "-7"))) %>%
  mutate(afford_care_f = fct_recode(afford_care_f,
    "no" = "2",
    "yes" = "1",
    NULL = "-8",
    NULL = "-7"))

# usual source of care status
h209red <- h209red %>%
  mutate(have_usc_f = factor(have_usc,
                              levels = c("2", "1", "-8", "-7"))) %>%
  mutate(have_usc_f = fct_recode(have_usc_f,
    "no" = "2",
    "yes" = "1",
    NULL = "-8",
    NULL = "-7"))

# distance from provider
h209red <- h209red %>%
  mutate(dist_from_usc = ifelse(have_usc_f == "no",
                                -100,
                                dist_from_usc)) %>% # creating level for not having a provider
  mutate(dist_from_usc_f = factor(dist_from_usc,
    levels = c("1", "2", "3", "4", "5", "6", "-100", "-8", "-7", "-1"))) %>%
  mutate(dist_from_usc_f = fct_recode(dist_from_usc_f,
    "<15" = "1",
    "15 to 30" = "2",
    "31 to 60" = "3",
    "61 to 90" = "4",
    "91 to 120" = "5",
    ">120" = "6",
    "no usc" = "-100",
    NULL = "-8",
    NULL = "-7",
    NULL = "-1"))

# ability to reach provider by phone
h209red <- h209red %>%
  mutate(rch_usc_byphn = ifelse(have_usc_f == "no",
                                -100,
                                rch_usc_byphn)) %>% # creating level for not having a provider
  mutate(rch_usc_byphn_f = factor(rch_usc_byphn,
    levels = c("4", "3", "2", "1", "-100", "-8", "-7", "-1"))) %>%
  mutate(rch_usc_byphn_f = fct_recode(rch_usc_byphn_f,
    "not at all difficult" = "4",
    "not too difficult" = "3",
    "somewhat difficult" = "2",
    "very difficult" = "1",
    "no usc" = "-100",
    NULL = "-8",
    NULL = "-7",
    NULL = "-1"))

```

```

NULL = "-7",
NULL = "-1"))

# provider offers office hours during nights/weekends
h209red <- h209red %>%
  mutate(usc_offhrs_nw = ifelse(have_usc_f == "no",
                                -100,
                                usc_offhrs_nw)) %>% # creating level for not having a provider
  mutate(usc_offhrs_nw_f = factor(usc_offhrs_nw,
                                  levels = c("-100", "2", "1", "-8", "-7", "-1"))) %>%
  mutate(usc_offhrs_nw_f = fct_recode(usc_offhrs_nw_f,
                                     "no usc" = "-100",
                                     "no" = "2",
                                     "yes" = "1",
                                     NULL = "-8",
                                     NULL = "-7",
                                     NULL = "-1"))

# provider asks about treatments
h209red <- h209red %>%
  mutate(usc_asks_abt_trts = ifelse(have_usc_f == "no",
                                    -100,
                                    usc_asks_abt_trts)) %>% # creating level for not having a provider
  mutate(usc_asks_abt_trts_f = factor(usc_asks_abt_trts,
                                      levels = c("-100", "2", "1", "-8", "-7", "-1"))) %>%
  mutate(usc_asks_abt_trts_f = fct_recode(usc_asks_abt_trts_f,
                                         "no usc" = "-100",
                                         "no" = "2",
                                         "yes" = "1",
                                         NULL = "-8",
                                         NULL = "-7",
                                         NULL = "-1"))

# provider asks person to help make decisions
h209red <- h209red %>%
  mutate(usc_asks_hlp_dec = ifelse(have_usc_f == "no",
                                    -100,
                                    usc_asks_hlp_dec)) %>% # creating level for not having a provider
  mutate(usc_asks_hlp_dec_f = factor(usc_asks_hlp_dec,
                                      levels = c("-100", "1", "2", "3", "4", "-8", "-7", "-1"))) %>%
  mutate(usc_asks_hlp_dec_f = fct_recode(usc_asks_hlp_dec_f,
                                         "no usc" = "-100",
                                         "never" = "1",
                                         "sometimes" = "2",
                                         "usually" = "3",
                                         "always" = "4",
                                         NULL = "-8",
                                         NULL = "-7",
                                         NULL = "-1"))

# provider presents and explains all options
h209red <- h209red %>%
  mutate(usc_expln_options = ifelse(have_usc_f == "no",
                                    -100,

```

```

        usc_expln_options)) %>% # creating level for not having a provider
mutate(usc_expln_options_f = factor(usc_expln_options,
        levels = c("-100", "2", "1", "-8", "-7", "-1"))) %>%
mutate(usc_expln_options_f = fct_recode(usc_expln_options_f,
        "no usc" = "-100",
        "no" = "2",
        "yes" = "1",
        NULL = "-8",
        NULL = "-7",
        NULL = "-1"))

# insurance indicator in 2018
h209red <- h209red %>%
  mutate(inscov_gen_2018_f = factor(inscov_gen_2018,
        levels = c("1", "2", "3"))) %>%
  mutate(inscov_gen_2018_f = fct_recode(inscov_gen_2018_f,
        "any private" = "1",
        "public only" = "2",
        "uninsured" = "3"))

# creating combined provider availability variable using 1) distance 2) ability to reach by phone 3) of
h209red <- h209red %>%
  mutate(dist_from_usc = ifelse(have_usc_f == "no", 0, dist_from_usc))
h209red$dist_from_usc[h209red$dist_from_usc == -8] <- NA
h209red$dist_from_usc[h209red$dist_from_usc == -7] <- NA

h209red <- h209red %>%
  mutate(rch_usc_byphn = ifelse(have_usc_f == "no", 0, rch_usc_byphn))
h209red$rch_usc_byphn[h209red$rch_usc_byphn == -8] <- NA
h209red$rch_usc_byphn[h209red$rch_usc_byphn == -7] <- NA

h209red <- h209red %>%
  mutate(usc_offhrs_nw = ifelse(have_usc_f == "no", 0, usc_offhrs_nw))
h209red$usc_offhrs_nw[h209red$usc_offhrs_nw == -8] <- NA
h209red$usc_offhrs_nw[h209red$usc_offhrs_nw == -7] <- NA
h209red$usc_offhrs_nw[h209red$usc_offhrs_nw == -1] <- NA

# creating binary access variables to use for making combined score
# give 0 to those w/o provider
# give 1 to people who have to travel 30+ minutes
# and give 2 to people who are within 30 min
h209red <- h209red %>% mutate(dist_from_usc_bin = ifelse(dist_from_usc %in% c(1, 2), 2,
        ifelse(dist_from_usc %in% c(3, 4, 5, 6), 1,
        dist_from_usc)))

# give 0 to those w/o provider
# and give 1 to people who answer somewhat difficult or very difficult
# give 2 to people who answer not at all difficult or not too difficult
h209red <- h209red %>% mutate(rch_usc_byphn_bin = ifelse(rch_usc_byphn %in% c(1, 2), 1,
        ifelse(rch_usc_byphn %in% c(3, 4, 5, 6), 2,
        rch_usc_byphn)))

# give 0 to those w/o provider
# and give 1 to people whose provider does not offer office hours during night/weekend

```

```

# give 2 to people whose provider does offer
h209red <- h209red %>% mutate(usc_offhrs_nw_bin = ifelse(usc_offhrs_nw == 1, 2,
                                                         ifelse(usc_offhrs_nw == 2, 1,
                                                             usc_offhrs_nw)))

# finally creating combined availability score from binary variables
h209red <- h209red %>%
  mutate(usc_access_score = dist_from_usc_bin + rch_usc_byphn_bin + usc_offhrs_nw_bin)

# creating combined provider satisfaction variable using 1) asking about treatments 2) asks person to h
h209red <- h209red %>%
  mutate(usc_asks_abt_trts = ifelse(have_usc_f == "no", 0, usc_asks_abt_trts))
h209red$usc_asks_abt_trts[h209red$usc_asks_abt_trts == -8] <- NA
h209red$usc_asks_abt_trts[h209red$usc_asks_abt_trts == -7] <- NA

h209red <- h209red %>%
  mutate(usc_asks_hlp_dec = ifelse(have_usc_f == "no", 0, usc_asks_hlp_dec))
h209red$usc_asks_hlp_dec[h209red$usc_asks_hlp_dec == -8] <- NA
h209red$usc_asks_hlp_dec[h209red$usc_asks_hlp_dec == -7] <- NA

h209red <- h209red %>%
  mutate(usc_expln_options = ifelse(have_usc_f == "no", 0, usc_expln_options))
h209red$usc_expln_options[h209red$usc_expln_options == -8] <- NA
h209red$usc_expln_options[h209red$usc_expln_options == -7] <- NA

# creating binary access variables to use for making combined score
# give 0 to those w/o provider
# give 1 to those who answered no
# and give 2 to those who answered yes
h209red <- h209red %>% mutate(usc_asks_abt_trts_bin = ifelse(usc_asks_abt_trts == 1, 2,
                                                             ifelse(usc_asks_abt_trts == 2, 1,
                                                                 usc_asks_abt_trts)))

# give 0 to those w/o provider
# and give 1 to those who answered never or sometimes
# give 2 to those who answered usually or always
h209red <- h209red %>% mutate(usc_asks_hlp_dec_bin = ifelse(usc_asks_hlp_dec %in% c(1, 2), 1,
                                                             ifelse(usc_asks_hlp_dec %in% c(3, 4), 2,
                                                                 usc_asks_hlp_dec)))

# give 0 to those w/o provider
# and give 1 to people who answer no
# give 2 to those who answered yes
h209red <- h209red %>% mutate(usc_expln_options_bin = ifelse(usc_expln_options == 1, 2,
                                                             ifelse(usc_expln_options == 2, 1,
                                                                 usc_expln_options)))

# finally creating combined satisfaction score from binary variables
h209red <- h209red %>%
  mutate(usc_satisf_score = usc_asks_abt_trts_bin + usc_asks_hlp_dec_bin + usc_expln_options_bin)

# creating our data set of variables we'll use in our modeling
df <- h209red %>% dplyr::select(pap_f, age, income_indiv, income_fam, totexp,

```



```

        outofpocket_exp, genhlth_avg_f, region_f,
        race_f, marital_stat_f, educ_f, smoke_freq_f,
        limitation_f, afford_care_f, have_usc_f,
        inscov_gen_2018_f) %>%
mutate(pap_num = as.numeric(pap_f) - 1) # create numeric pap variable

# create complete cases data set for modeling
cc_df <- df %>% drop_na()

# check for high correlation between variables
cc_df_num <- data.frame(cc_df$pap_num, cc_df$age, cc_df$income_indiv,
                        cc_df$income_fam, cc_df$totexp, cc_df$outofpocket_exp,
                        as.numeric(cc_df$genhlth_avg_f) - 1,
                        as.numeric(cc_df$region_f) - 1,
                        as.numeric(cc_df$race_f) - 1,
                        as.numeric(cc_df$marital_stat_f) - 1,
                        as.numeric(cc_df$educ_f) - 1,
                        as.numeric(cc_df$smoke_freq_f) - 1,
                        as.numeric(cc_df$limitation_f) - 1,
                        as.numeric(cc_df$afford_care_f) - 1,
                        as.numeric(cc_df$have_usc_f) - 1,
                        as.numeric(cc_df$inscov_gen_2018_f) - 1)
names <- colnames(cc_df)[2:16]
colnames(cc_df_num) <- c("pap_num", names)
correlation_matrix <- round(cor(cc_df_num), digits = 2) # looks pretty good

library(caret)

## Loading required package: lattice
##
## Attaching package: 'caret'
## The following object is masked from 'package:purrr':
##
##   lift

library(leaps)
library(LogisticDx)
library(ResourceSelection)

## ResourceSelection 0.3-5    2019-07-22

library(MASS)

##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##   select

library(car)

## Loading required package: carData
##
## Attaching package: 'car'

```

```
## The following object is masked from 'package:dplyr':
##
##   recode
## The following object is masked from 'package:purrr':
##
##   some
library(caret)
library(pROC)

## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
## The following objects are masked from 'package:stats':
##
##   cov, smooth, var
```

Associational modeling

Now we will create associational models on the full complete cases data set.

```
# full model
mod_full <- glm(pap_num ~ age + income_indiv + income_fam + totexp + outofpocket_exp + genhlth_avg_f +
summary(mod_full)

##
## Call:
## glm(formula = pap_num ~ age + income_indiv + income_fam + totexp +
##   outofpocket_exp + genhlth_avg_f + region_f + race_f + marital_stat_f +
##   educ_f + smoke_freq_f + limitation_f + afford_care_f + have_usc_f +
##   inscov_gen_2018_f, family = binomial(), data = cc_df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7407  -0.7688  -0.5571   0.8748   2.8002
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.101e+00  3.174e-01  -3.470 0.000521 ***
## age             1.618e-02  3.008e-03   5.379 7.49e-08 ***
## income_indiv   -1.105e-05  1.458e-06  -7.576 3.57e-14 ***
## income_fam      1.289e-06  7.304e-07   1.764 0.077680 .
## totexp         -1.008e-05  3.369e-06  -2.991 0.002780 **
## outofpocket_exp -2.265e-05  2.415e-05  -0.938 0.348393
## genhlth_avg_ffair 2.811e-01  2.742e-01   1.025 0.305198
## genhlth_avg_fggood 1.103e-01  2.715e-01   0.406 0.684451
## genhlth_avg_fvery good 2.502e-02  2.756e-01   0.091 0.927670
## genhlth_avg_fexcellent 1.305e-01  2.839e-01   0.460 0.645778
## region_fmidwest -1.476e-01  1.125e-01  -1.312 0.189669
## region_fsouth    6.767e-02  1.000e-01   0.677 0.498702
## region_fwest    -3.938e-02  1.072e-01  -0.367 0.713478
## race_fhispanic   1.957e-01  8.820e-02   2.219 0.026505 *
## race_fblack      1.177e-01  9.878e-02   1.192 0.233339
## race_fasian      1.173e+00  1.309e-01   8.958 < 2e-16 ***
```

```
## race_fother or multiple races 1.424e-01 1.744e-01 0.816 0.414267
## marital_stat_fmarrried -7.907e-01 8.532e-02 -9.268 < 2e-16 ***
## marital_stat_fwidowed -3.374e-01 1.885e-01 -1.791 0.073365 .
## marital_stat_fdivorced -5.092e-01 1.186e-01 -4.292 1.77e-05 ***
## marital_stat_fseperated -5.457e-01 1.761e-01 -3.100 0.001938 **
## educ_fany high school 4.639e-01 7.220e-02 6.426 1.31e-10 ***
## educ_fnone or any elementary 3.160e-01 1.517e-01 2.083 0.037241 *
## smoke_freq_fsome days 2.536e-01 1.474e-01 1.721 0.085279 .
## smoke_freq_fevery day 3.218e-01 1.021e-01 3.152 0.001621 **
## limitation_fyes 2.347e-01 1.297e-01 1.809 0.070477 .
## afford_care_fyes -3.208e-01 1.186e-01 -2.705 0.006826 **
## have_usc_fyes -6.090e-01 7.143e-02 -8.525 < 2e-16 ***
## inscov_gen_2018_fpublic only 1.124e-02 8.882e-02 0.127 0.899331
## inscov_gen_2018_funinsured 5.956e-01 1.077e-01 5.529 3.22e-08 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 6663.9 on 5862 degrees of freedom
## Residual deviance: 5970.1 on 5833 degrees of freedom
## AIC: 6030.1
##
## Number of Fisher Scoring iterations: 5
```

```
vif(mod_full) # correlation looks good
```

```
##                GVIF Df GVIF^(1/(2*Df))
## age            1.495309 1      1.222828
## income_indiv   1.691205 1      1.300463
## income_fam     1.932126 1      1.390009
## totexp         1.325187 1      1.151168
## outofpocket_exp 1.211220 1      1.100554
## genhlth_avg_f  1.531871 4      1.054758
## region_f       1.253373 3      1.038357
## race_f         1.677527 4      1.066802
## marital_stat_f 1.818178 4      1.077592
## educ_f         1.343687 2      1.076650
## smoke_freq_f   1.156677 2      1.037058
## limitation_f   1.450423 1      1.204335
## afford_care_f  1.097856 1      1.047786
## have_usc_f     1.126836 1      1.061525
## inscov_gen_2018_f 1.739732 2      1.148472
```

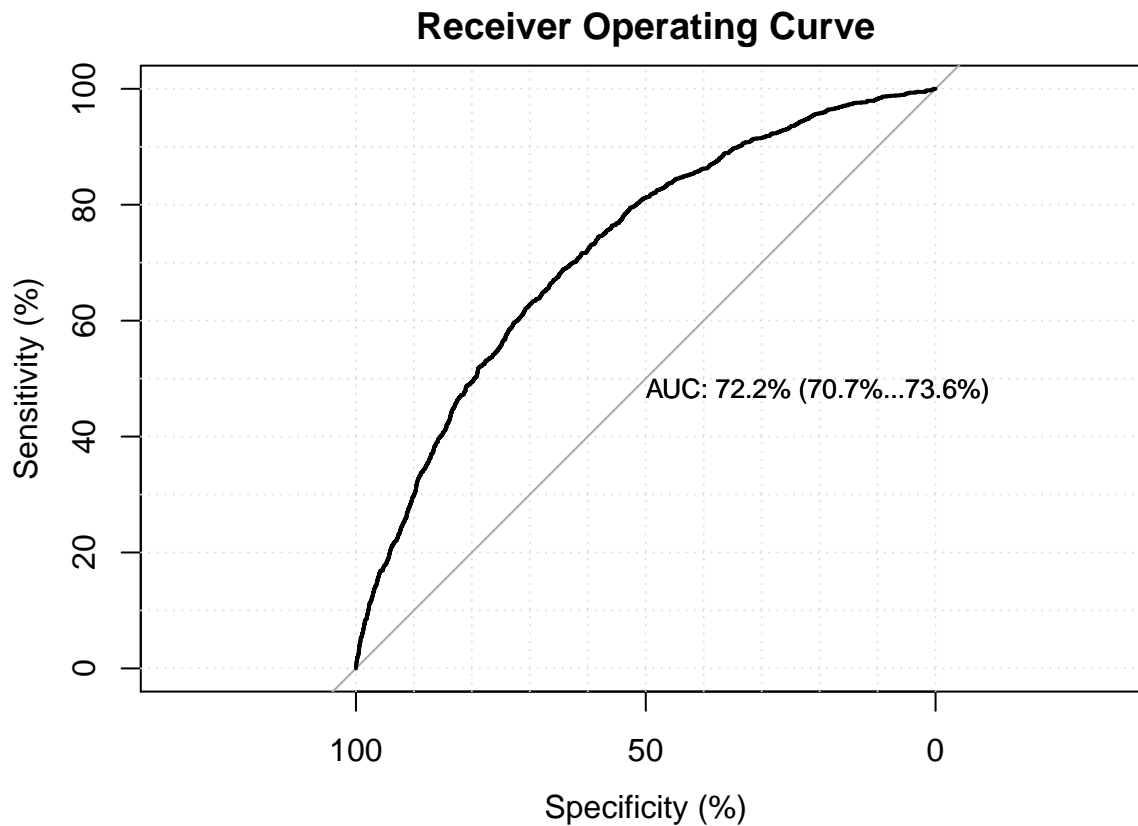
```
hoslem.test(cc_df$pap_num, fitted(mod_full), g = 10) # good fit for full model!
```

```
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: cc_df$pap_num, fitted(mod_full)
## X-squared = 7.3206, df = 8, p-value = 0.5025
```

```
gof(mod_full, g = 10)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



```
##      chiSq df pVal
## PrI      1 2  4
## drI      2 2  2
## PrG      1 1  3
## drG      2 1  1
## PrCT     1 1  3
## drCT     2 1  1
##
##              val df pVal
## HL chiSq      9 3  4
## mHL F         8 4  9
## OsRo Z        7 5  1
## SstPgeq0.5 Z   2 5  7
## SstPl0.5 Z     3 5  3
## SstBoth chiSq  6 2  5
## SllPgeq0.5 chiSq 1 1  8
## SllPl0.5 chiSq  4 1  2
## SllBoth chiSq  5 2  6
```

```
# forward/backward selection model
```

```
# forward model selection
```

```
mod_forw <- step(glm(pap_num ~ 1, data = cc_df, family = binomial()), ~ age + income_indiv + income_fam
```

```
## Start: AIC=6665.86
```

```
## pap_num ~ 1
```

```
##
```

```
##              Df Deviance    AIC
```

```
## + income_indiv      1    6434.5 6438.5
```

```

## + inscov_gen_2018_f  2  6454.4 6460.4
## + educ_f             2  6483.7 6489.7
## + race_f             4  6512.6 6522.6
## + income_fam         1  6519.3 6523.3
## + have_usc_f         1  6521.3 6525.3
## + marital_stat_f     4  6524.0 6534.0
## + outofpocket_exp    1  6609.8 6613.8
## + totexp             1  6629.5 6633.5
## + smoke_freq_f       2  6628.1 6634.1
## + region_f           3  6631.3 6639.3
## + genhlth_avg_f      4  6633.1 6643.1
## + limitation_f       1  6651.8 6655.8
## <none>                6663.9 6665.9
## + age                1  6663.0 6667.0
## + afford_care_f      1  6663.6 6667.6
##
## Step:  AIC=6438.55
## pap_num ~ income_indiv
##
##           Df Deviance    AIC
## + have_usc_f      1  6322.8 6328.8
## + race_f          4  6322.5 6334.5
## + inscov_gen_2018_f 2  6338.3 6346.3
## + marital_stat_f  4  6335.6 6347.6
## + educ_f          2  6356.0 6364.0
## + totexp          1  6404.6 6410.6
## + outofpocket_exp 1  6409.8 6415.8
## + region_f        3  6410.0 6420.0
## + smoke_freq_f    2  6417.0 6425.0
## + income_fam      1  6423.0 6429.0
## + genhlth_avg_f   4  6425.7 6437.7
## <none>            6434.5 6438.5
## + age            1  6433.2 6439.2
## + limitation_f    1  6433.8 6439.8
## + afford_care_f   1  6434.0 6440.0
##
## Step:  AIC=6328.77
## pap_num ~ income_indiv + have_usc_f
##
##           Df Deviance    AIC
## + race_f          4  6229.6 6243.6
## + marital_stat_f  4  6238.0 6252.0
## + educ_f          2  6245.5 6255.5
## + inscov_gen_2018_f 2  6256.5 6266.5
## + smoke_freq_f    2  6303.7 6313.7
## + totexp          1  6307.5 6315.5
## + outofpocket_exp 1  6309.6 6317.6
## + age            1  6312.6 6320.6
## + region_f        3  6309.0 6321.0
## + genhlth_avg_f   4  6307.3 6321.3
## + income_fam      1  6314.9 6322.9
## + limitation_f    1  6316.3 6324.3
## <none>            6322.8 6328.8
## + afford_care_f   1  6321.0 6329.0

```

```

##
## Step: AIC=6243.58
## pap_num ~ income_indiv + have_usc_f + race_f
##
##           Df Deviance    AIC
## + marital_stat_f    4   6151.0 6173.0
## + educ_f            2   6159.5 6177.5
## + inscov_gen_2018_f  2   6172.9 6190.9
## + smoke_freq_f      2   6195.6 6213.6
## + age               1   6217.1 6233.1
## + limitation_f      1   6220.4 6236.4
## + totexp            1   6220.5 6236.5
## + income_fam        1   6220.5 6236.5
## + genhlth_avg_f     4   6216.5 6238.5
## + outofpocket_exp   1   6223.7 6239.7
## + region_f          3   6222.5 6242.5
## <none>              6229.6 6243.6
## + afford_care_f     1   6228.7 6244.7
##
## Step: AIC=6172.96
## pap_num ~ income_indiv + have_usc_f + race_f + marital_stat_f
##
##           Df Deviance    AIC
## + educ_f            2   6083.3 6109.3
## + inscov_gen_2018_f  2   6107.3 6133.3
## + age               1   6111.4 6135.4
## + smoke_freq_f      2   6127.5 6153.5
## + totexp            1   6141.3 6165.3
## + region_f          3   6141.1 6169.1
## + limitation_f      1   6145.2 6169.2
## + genhlth_avg_f     4   6141.1 6171.1
## + outofpocket_exp   1   6147.7 6171.7
## <none>              6151.0 6173.0
## + afford_care_f     1   6149.1 6173.1
## + income_fam        1   6151.0 6175.0
##
## Step: AIC=6109.32
## pap_num ~ income_indiv + have_usc_f + race_f + marital_stat_f +
##      educ_f
##
##           Df Deviance    AIC
## + inscov_gen_2018_f  2   6049.8 6079.8
## + age               1   6053.3 6081.3
## + smoke_freq_f      2   6069.8 6099.8
## + totexp            1   6074.1 6102.1
## + region_f          3   6074.1 6106.1
## + limitation_f      1   6078.9 6106.9
## + outofpocket_exp   1   6081.2 6109.2
## + afford_care_f     1   6081.2 6109.2
## <none>              6083.3 6109.3
## + income_fam        1   6082.0 6110.0
## + genhlth_avg_f     4   6077.3 6111.3
##
## Step: AIC=6079.78

```

```

## pap_num ~ income_indiv + have_usc_f + race_f + marital_stat_f +
##   educ_f + inscov_gen_2018_f
##
##           Df Deviance    AIC
## + age      1   6019.1 6051.1
## + smoke_freq_f  2   6036.5 6070.5
## + totexp    1   6043.3 6075.3
## + limitation_f  1   6043.9 6075.9
## + afford_care_f  1   6044.6 6076.6
## + outofpocket_exp  1   6047.4 6079.4
## + income_fam    1   6047.5 6079.5
## <none>          6049.8 6079.8
## + region_f      3   6044.0 6080.0
## + genhlth_avg_f  4   6043.4 6081.4
##
## Step:  AIC=6051.05
## pap_num ~ income_indiv + have_usc_f + race_f + marital_stat_f +
##   educ_f + inscov_gen_2018_f + age
##
##           Df Deviance    AIC
## + totexp    1   6008.7 6042.7
## + smoke_freq_f  2   6007.4 6043.4
## + afford_care_f  1   6012.7 6046.7
## + outofpocket_exp  1   6014.6 6048.6
## + income_fam    1   6016.2 6050.2
## + limitation_f  1   6016.9 6050.9
## <none>          6019.1 6051.1
## + region_f      3   6013.4 6051.4
## + genhlth_avg_f  4   6014.4 6054.4
##
## Step:  AIC=6042.73
## pap_num ~ income_indiv + have_usc_f + race_f + marital_stat_f +
##   educ_f + inscov_gen_2018_f + age + totexp
##
##           Df Deviance    AIC
## + smoke_freq_f  2   5996.8 6034.8
## + afford_care_f  1   6002.7 6038.7
## + limitation_f  1   6003.2 6039.2
## + income_fam    1   6005.7 6041.7
## <none>          6008.7 6042.7
## + region_f      3   6003.2 6043.2
## + outofpocket_exp  1   6007.5 6043.5
## + genhlth_avg_f  4   6002.7 6044.7
##
## Step:  AIC=6034.84
## pap_num ~ income_indiv + have_usc_f + race_f + marital_stat_f +
##   educ_f + inscov_gen_2018_f + age + totexp + smoke_freq_f
##
##           Df Deviance    AIC
## + afford_care_f  1   5990.0 6030.0
## + limitation_f  1   5992.1 6032.1
## + income_fam    1   5993.1 6033.1
## <none>          5996.8 6034.8
## + region_f      3   5991.4 6035.4

```

```

## + outofpocket_exp 1 5995.7 6035.7
## + genhlth_avg_f 4 5991.4 6037.4
##
## Step: AIC=6030.02
## pap_num ~ income_indiv + have_usc_f + race_f + marital_stat_f +
## educ_f + inscov_gen_2018_f + age + totexp + smoke_freq_f +
## afford_care_f
##
## Df Deviance AIC
## + limitation_f 1 5984.7 6026.7
## + income_fam 1 5986.9 6028.9
## <none> 5990.0 6030.0
## + region_f 3 5984.4 6030.4
## + outofpocket_exp 1 5989.2 6031.2
## + genhlth_avg_f 4 5983.3 6031.3
##
## Step: AIC=6026.67
## pap_num ~ income_indiv + have_usc_f + race_f + marital_stat_f +
## educ_f + inscov_gen_2018_f + age + totexp + smoke_freq_f +
## afford_care_f + limitation_f
##
## Df Deviance AIC
## + income_fam 1 5981.7 6025.7
## <none> 5984.7 6026.7
## + region_f 3 5979.2 6027.2
## + outofpocket_exp 1 5983.9 6027.9
## + genhlth_avg_f 4 5979.5 6029.5
##
## Step: AIC=6025.7
## pap_num ~ income_indiv + have_usc_f + race_f + marital_stat_f +
## educ_f + inscov_gen_2018_f + age + totexp + smoke_freq_f +
## afford_care_f + limitation_f + income_fam
##
## Df Deviance AIC
## <none> 5981.7 6025.7
## + region_f 3 5976.2 6026.2
## + outofpocket_exp 1 5980.9 6026.9
## + genhlth_avg_f 4 5976.6 6028.6
summary(mod_forw)

##
## Call:
## glm(formula = pap_num ~ income_indiv + have_usc_f + race_f +
## marital_stat_f + educ_f + inscov_gen_2018_f + age + totexp +
## smoke_freq_f + afford_care_f + limitation_f + income_fam,
## family = binomial(), data = cc_df)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -1.7658 -0.7686 -0.5623 0.8736 2.8036
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.063e+00 1.395e-01 -7.622 2.50e-14 ***

```



```
## income_indiv          -1.121e-05  1.457e-06  -7.695  1.42e-14 ***
## have_usc_fyes         -6.204e-01  7.081e-02  -8.761  < 2e-16 ***
## race_fhispanic        2.372e-01  8.541e-02   2.777  0.005487 **
## race_fblack           1.890e-01  9.494e-02   1.991  0.046494 *
## race_fasian           1.189e+00  1.292e-01   9.203  < 2e-16 ***
## race_fother or multiple races 1.586e-01  1.741e-01   0.911  0.362171
## marital_stat_fmarrried -7.847e-01  8.498e-02  -9.234  < 2e-16 ***
## marital_stat_fwidowed -3.148e-01  1.878e-01  -1.677  0.093569 .
## marital_stat_fdivorced -4.984e-01  1.182e-01  -4.217  2.48e-05 ***
## marital_stat_fseperated -5.203e-01  1.751e-01  -2.971  0.002966 **
## educ_fany high school  4.722e-01  7.177e-02   6.579  4.73e-11 ***
## educ_fnone or any elementary 3.401e-01  1.509e-01   2.254  0.024204 *
## inscov_gen_2018_fpublic only 1.703e-02  8.750e-02   0.195  0.845693
## inscov_gen_2018_funinsured 6.130e-01  1.067e-01   5.744  9.25e-09 ***
## age                   1.628e-02  2.980e-03   5.461  4.72e-08 ***
## totexp                -1.069e-05  3.135e-06  -3.409  0.000652 ***
## smoke_freq_fsome days  2.531e-01  1.471e-01   1.720  0.085410 .
## smoke_freq_fevery day  3.331e-01  1.014e-01   3.286  0.001016 **
## afford_care_fyes       -2.981e-01  1.161e-01  -2.567  0.010269 *
## limitation_fyes        2.812e-01  1.223e-01   2.299  0.021526 *
## income_fam             1.260e-06  7.246e-07   1.739  0.082110 .
```

```
## ---
```

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

```
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
```

```
## Null deviance: 6663.9 on 5862 degrees of freedom
```

```
## Residual deviance: 5981.7 on 5841 degrees of freedom
```

```
## AIC: 6025.7
```

```
##
```

```
## Number of Fisher Scoring iterations: 5
```

```
vif(mod_forw) # correlation looks good
```

```
##              GVIF Df GVIF^(1/(2*Df))
## income_indiv  1.680824 1      1.296466
## have_usc_f    1.109049 1      1.053114
## race_f        1.434313 4      1.046118
## marital_stat_f 1.791701 4      1.075618
## educ_f        1.323530 2      1.072589
## inscov_gen_2018_f 1.640730 2      1.131773
## age           1.473158 1      1.213737
## totexp        1.153797 1      1.074150
## smoke_freq_f  1.142073 2      1.033769
## afford_care_f 1.058842 1      1.029000
## limitation_f  1.294997 1      1.137979
## income_fam    1.897364 1      1.377448
```

```
# backward model selection
```

```
mod_back <- step(mod_full, direction = "backward")
```

```
## Start: AIC=6030.07
```

```
## pap_num ~ age + income_indiv + income_fam + totexp + outofpocket_exp +
```

```
##      genhlth_avg_f + region_f + race_f + marital_stat_f + educ_f +
```

```
##      smoke_freq_f + limitation_f + afford_care_f + have_usc_f +
```

```

##      inscov_gen_2018_f
##
##              Df Deviance    AIC
## - genhlth_avg_f      4   5975.3 6027.3
## - outofpocket_exp     1   5971.0 6029.0
## - region_f           3   5975.7 6029.7
## <none>                5970.1 6030.1
## - income_fam          1   5973.1 6031.1
## - limitation_f        1   5973.3 6031.3
## - afford_care_f       1   5977.6 6035.6
## - smoke_freq_f        2   5981.6 6037.6
## - totexp              1   5980.5 6038.5
## - age                 1   5999.1 6057.1
## - inscov_gen_2018_f   2   6003.1 6059.1
## - educ_f              2   6011.3 6067.3
## - income_indiv        1   6030.9 6088.9
## - race_f              4   6046.6 6098.6
## - have_usc_f          1   6041.8 6099.8
## - marital_stat_f      4   6059.2 6111.2
##
## Step: AIC=6027.34
## pap_num ~ age + income_indiv + income_fam + totexp + outofpocket_exp +
##      region_f + race_f + marital_stat_f + educ_f + smoke_freq_f +
##      limitation_f + afford_care_f + have_usc_f + inscov_gen_2018_f
##
##              Df Deviance    AIC
## - outofpocket_exp     1   5976.2 6026.2
## - region_f           3   5980.9 6026.9
## <none>                5975.3 6027.3
## - income_fam          1   5978.5 6028.5
## - limitation_f        1   5980.2 6030.2
## - afford_care_f       1   5982.0 6032.0
## - totexp              1   5985.4 6035.4
## - smoke_freq_f        2   5987.6 6035.6
## - age                 1   6005.7 6055.7
## - inscov_gen_2018_f   2   6007.9 6055.9
## - educ_f              2   6018.0 6066.0
## - income_indiv        1   6037.4 6087.4
## - have_usc_f          1   6046.1 6096.1
## - race_f              4   6052.9 6096.9
## - marital_stat_f      4   6065.2 6109.2
##
## Step: AIC=6026.18
## pap_num ~ age + income_indiv + income_fam + totexp + region_f +
##      race_f + marital_stat_f + educ_f + smoke_freq_f + limitation_f +
##      afford_care_f + have_usc_f + inscov_gen_2018_f
##
##              Df Deviance    AIC
## - region_f           3   5981.7 6025.7
## <none>                5976.2 6026.2
## - income_fam          1   5979.2 6027.2
## - limitation_f        1   5981.1 6029.1
## - afford_care_f       1   5983.2 6031.2
## - smoke_freq_f        2   5988.5 6034.5

```

```
## - totexp          1  5989.7 6037.7
## - age             1  6006.1 6054.1
## - inscov_gen_2018_f 2  6008.3 6054.3
## - educ_f          2  6019.2 6065.2
## - income_indiv     1  6038.9 6086.9
## - have_usc_f       1  6048.1 6096.1
## - race_f           4  6054.5 6096.5
## - marital_stat_f   4  6066.4 6108.4
##
## Step: AIC=6025.7
## pap_num ~ age + income_indiv + income_fam + totexp + race_f +
## marital_stat_f + educ_f + smoke_freq_f + limitation_f + afford_care_f +
## have_usc_f + inscov_gen_2018_f
##
##              Df Deviance    AIC
## <none>                5981.7 6025.7
## - income_fam          1  5984.7 6026.7
## - limitation_f         1  5986.9 6028.9
## - afford_care_f        1  5988.5 6030.5
## - smoke_freq_f         2  5994.1 6034.1
## - totexp               1  5995.5 6037.5
## - age                  1  6011.6 6053.6
## - inscov_gen_2018_f    2  6017.6 6057.6
## - educ_f               2  6025.0 6065.0
## - income_indiv         1  6044.6 6086.6
## - race_f               4  6062.7 6098.7
## - have_usc_f           1  6057.4 6099.4
## - marital_stat_f       4  6070.1 6106.1
```

```
summary(mod_back) # forward and backward selection produce the same model here
```

```
##
## Call:
## glm(formula = pap_num ~ age + income_indiv + income_fam + totexp +
## race_f + marital_stat_f + educ_f + smoke_freq_f + limitation_f +
## afford_care_f + have_usc_f + inscov_gen_2018_f, family = binomial(),
## data = cc_df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7658  -0.7686  -0.5623   0.8736   2.8036
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.063e+00  1.395e-01 -7.622 2.50e-14 ***
## age           1.628e-02  2.980e-03  5.461 4.72e-08 ***
## income_indiv  -1.121e-05  1.457e-06 -7.695 1.42e-14 ***
## income_fam     1.260e-06  7.246e-07  1.739 0.082110 .
## totexp        -1.069e-05  3.135e-06 -3.409 0.000652 ***
## race_fhispanic 2.372e-01  8.541e-02  2.777 0.005487 **
## race_fblack    1.890e-01  9.494e-02  1.991 0.046494 *
## race_fasian    1.189e+00  1.292e-01  9.203 < 2e-16 ***
## race_fother or multiple races 1.586e-01  1.741e-01  0.911 0.362171
## marital_stat_fmarrried -7.847e-01  8.498e-02 -9.234 < 2e-16 ***
## marital_stat_fwidowed -3.148e-01  1.878e-01 -1.677 0.093569 .
```

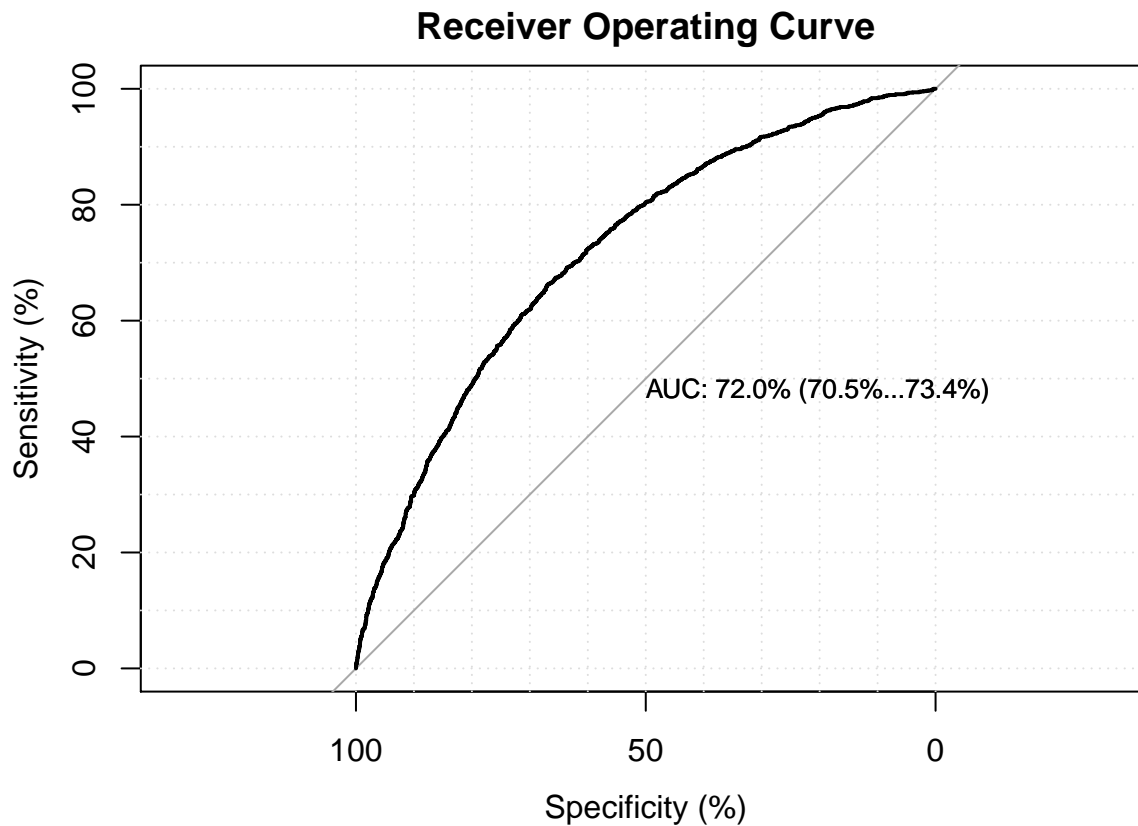
```

## marital_stat_fdivorced      -4.984e-01  1.182e-01  -4.217  2.48e-05 ***
## marital_stat_fseparated     -5.203e-01  1.751e-01  -2.971  0.002966 **
## educ_fany high school       4.722e-01  7.177e-02   6.579  4.73e-11 ***
## educ_fnone or any elementary 3.401e-01  1.509e-01   2.254  0.024204 *
## smoke_freq_fsome days      2.531e-01  1.471e-01   1.720  0.085410 .
## smoke_freq_fevery day      3.331e-01  1.014e-01   3.286  0.001016 **
## limitation_fyes            2.812e-01  1.223e-01   2.299  0.021526 *
## afford_care_fyes           -2.981e-01  1.161e-01  -2.567  0.010269 *
## have_usc_fyes              -6.204e-01  7.081e-02  -8.761  < 2e-16 ***
## inscov_gen_2018_fpublic only 1.703e-02  8.750e-02   0.195  0.845693
## inscov_gen_2018_funinsured   6.130e-01  1.067e-01   5.744  9.25e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 6663.9  on 5862  degrees of freedom
## Residual deviance: 5981.7  on 5841  degrees of freedom
## AIC: 6025.7
##
## Number of Fisher Scoring iterations: 5
hoslem.test(cc_df$pap_num, fitted(mod_back), g = 10) # good fit for forward/backward model!

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  cc_df$pap_num, fitted(mod_back)
## X-squared = 7.1074, df = 8, p-value = 0.5251
gof(mod_back, g = 10)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases

```



```
##      chiSq df pVal
## PrI      1 2  2
## drI      2 2  4
## PrG      1 1  1
## drG      2 1  3
## PrCT     1 1  1
## drCT     2 1  3
##
##              val df pVal
## HL chiSq      9 3  4
## mHL F         8 4  9
## OsRo Z        7 5  1
## SstPgeq0.5 Z  2 5  7
## SstPl0.5 Z    4 5  3
## SstBoth chiSq 6 2  5
## SllPgeq0.5 chiSq 1 1  8
## SllPl0.5 chiSq 3 1  2
## SllBoth chiSq 5 2  6
```

```
# testing if nonlinear terms are needed for age
```

```
# trying quadratic
```

```
mod_age_quad <- glm(pap_num ~ age + I(age^2) + income_indiv + income_fam + totexp + race_f + marital_stat_f +
summary(mod_age_quad)
```

```
##
## Call:
## glm(formula = pap_num ~ age + I(age^2) + income_indiv + income_fam +
##      totexp + race_f + marital_stat_f + educ_f + smoke_freq_f +
```

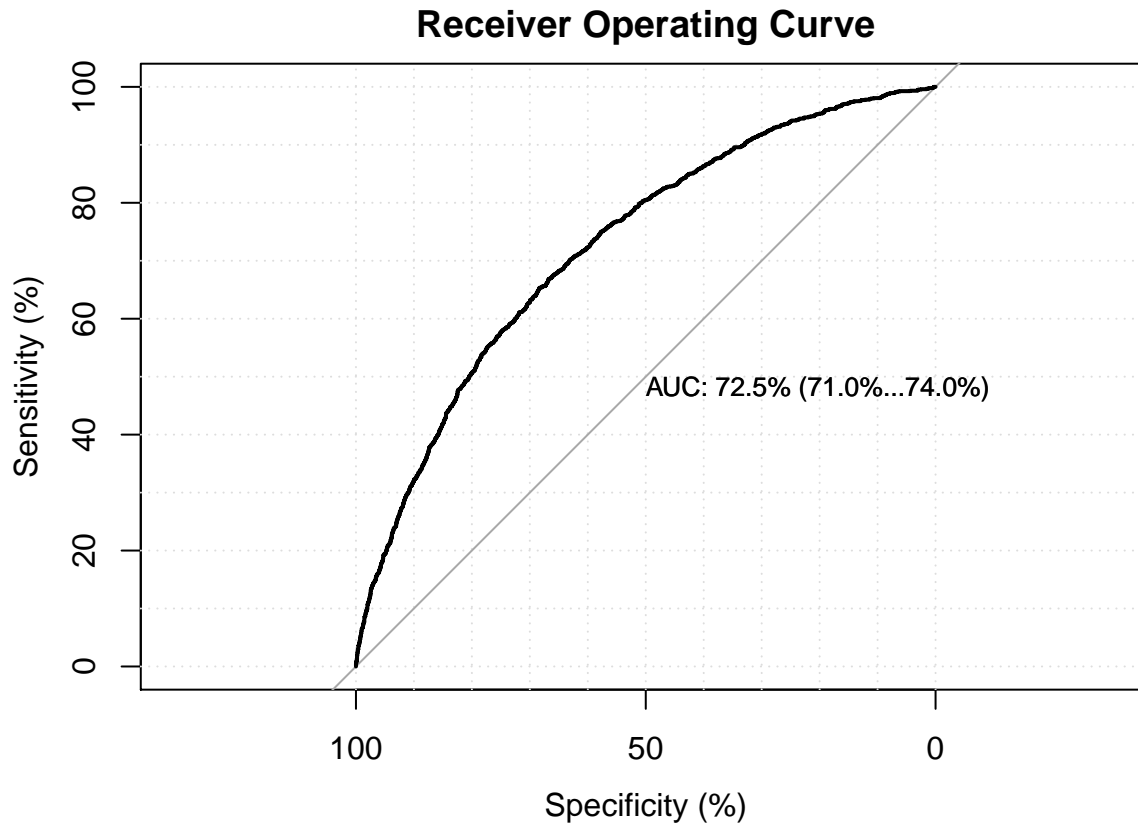
```
##      limitation_f + afford_care_f + have_usc_f + inscov_gen_2018_f,
##      family = binomial(), data = cc_df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q        Max
## -1.7595   -0.7668   -0.5506    0.8249    2.8077
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.405e+00  3.889e-01   3.612 0.000304 ***
## age             -1.168e-01  1.988e-02  -5.873 4.28e-09 ***
## I(age^2)         1.532e-03  2.265e-04   6.766 1.32e-11 ***
## income_indiv     -9.778e-06  1.470e-06  -6.653 2.87e-11 ***
## income_fam        9.854e-07  7.273e-07   1.355 0.175460
## totexp          -1.148e-05  3.199e-06  -3.590 0.000330 ***
## race_fhispanic    2.843e-01  8.611e-02   3.302 0.000961 ***
## race_fblack       2.400e-01  9.560e-02   2.510 0.012071 *
## race_fasian       1.234e+00  1.307e-01   9.442 < 2e-16 ***
## race_fother or multiple races 1.970e-01  1.751e-01   1.125 0.260655
## marital_stat_fmarrried -6.293e-01  8.805e-02  -7.146 8.90e-13 ***
## marital_stat_fwidowed -3.378e-01  1.891e-01  -1.786 0.074049 .
## marital_stat_fdivorced -4.005e-01  1.203e-01  -3.329 0.000870 ***
## marital_stat_fseperated -3.933e-01  1.770e-01  -2.222 0.026268 *
## educ_fany high school  4.502e-01  7.211e-02   6.244 4.27e-10 ***
## educ_fnone or any elementary 3.292e-01  1.515e-01   2.173 0.029789 *
## smoke_freq_fsome days  2.760e-01  1.480e-01   1.865 0.062233 .
## smoke_freq_fevery day  3.865e-01  1.022e-01   3.782 0.000156 ***
## limitation_fyes      2.869e-01  1.230e-01   2.333 0.019641 *
## afford_care_fyes     -2.760e-01  1.167e-01  -2.366 0.017999 *
## have_usc_fyes        -6.220e-01  7.115e-02  -8.742 < 2e-16 ***
## inscov_gen_2018_fpublic only 6.290e-02  8.800e-02   0.715 0.474717
## inscov_gen_2018_funinsured 6.776e-01  1.077e-01   6.290 3.18e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 6663.9  on 5862  degrees of freedom
## Residual deviance: 5936.1  on 5840  degrees of freedom
## AIC: 5982.1
##
## Number of Fisher Scoring iterations: 5
# p < 0.05 for age^2 term so makes sense to keep quadratic term
hoslem.test(cc_df$pap_num, fitted(mod_age_quad), g = 10) # good fit

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  cc_df$pap_num, fitted(mod_age_quad)
## X-squared = 6.3499, df = 8, p-value = 0.6081
```

```
gof(mod_age_quad, g = 10)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



```
##      chiSq df pVal
## PrI      1 2   4
## drI      2 2   2
## PrG      1 1   3
## drG      2 1   1
## PrCT     1 1   3
## drCT     2 1   1
##
##              val df pVal
## HL chiSq      9 3   2
## mHL F         8 4   9
## OsRo Z        7 5   1
## SstPgeq0.5 Z   6 5   3
## SstPl0.5 Z     5 5   6
## SstBoth chiSq  4 2   7
## SllPgeq0.5 chiSq 2 1   4
## SllPl0.5 chiSq 1 1   5
## SllBoth chiSq  3 2   8
```

```
# trying cubic spline
```

```
library(splines)
```

```
library(splines2)
```

```
# fit cubic spline with 3 knots
```

```
mod_age_cubic_spline <- glm(pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv + income_fam + to
summary(mod_age_cubic_spline)
```

```
##
## Call:
## glm(formula = pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv +
##     income_fam + totexp + race_f + marital_stat_f + educ_f +
##     smoke_freq_f + limitation_f + afford_care_f + have_usc_f +
##     inscov_gen_2018_f, family = binomial(), data = cc_df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8377  -0.7559  -0.5417   0.7198   2.6854
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      5.786e-01  1.914e-01   3.023 0.002500 **
## bSpline(age, df = 6, degree = 3)1 -1.700e+00  3.390e-01  -5.014 5.33e-07 ***
## bSpline(age, df = 6, degree = 3)2 -1.641e+00  2.356e-01  -6.965 3.27e-12 ***
## bSpline(age, df = 6, degree = 3)3 -1.514e+00  2.817e-01  -5.374 7.69e-08 ***
## bSpline(age, df = 6, degree = 3)4 -2.980e-01  2.750e-01  -1.084 0.278463
## bSpline(age, df = 6, degree = 3)5 -1.316e+00  2.909e-01  -4.522 6.14e-06 ***
## bSpline(age, df = 6, degree = 3)6 -2.807e-01  2.507e-01  -1.120 0.262750
## income_indiv      -8.352e-06  1.479e-06  -5.647 1.63e-08 ***
## income_fam        2.542e-07  7.449e-07   0.341 0.732902
## totexp           -1.054e-05  3.154e-06  -3.340 0.000837 ***
## race_fhispanic    2.572e-01  8.689e-02   2.960 0.003079 **
## race_fblack       2.337e-01  9.640e-02   2.424 0.015332 *
## race_fasian       1.211e+00  1.315e-01   9.209 < 2e-16 ***
## race_fother or multiple races  1.780e-01  1.782e-01   0.999 0.317934
## marital_stat_fmarrried -5.600e-01  8.997e-02  -6.224 4.83e-10 ***
## marital_stat_fwidowed -3.099e-01  1.895e-01  -1.635 0.102051
## marital_stat_fdivorced -3.945e-01  1.210e-01  -3.259 0.001117 **
## marital_stat_fseperated -3.801e-01  1.781e-01  -2.134 0.032839 *
## educ_fany high school  4.460e-01  7.278e-02   6.128 8.91e-10 ***
## educ_fnone or any elementary  3.137e-01  1.515e-01   2.070 0.038430 *
## smoke_freq_fsome days  3.204e-01  1.489e-01   2.152 0.031379 *
## smoke_freq_fevery day  3.951e-01  1.028e-01   3.842 0.000122 ***
## limitation_fyes     2.657e-01  1.236e-01   2.149 0.031659 *
## afford_care_fyes    -2.930e-01  1.176e-01  -2.492 0.012688 *
## have_usc_fyes      -6.562e-01  7.211e-02  -9.100 < 2e-16 ***
## inscov_gen_2018_fpublic only  1.016e-01  8.891e-02   1.143 0.253172
## inscov_gen_2018_funinsured  6.916e-01  1.085e-01   6.377 1.80e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 6663.9  on 5862  degrees of freedom
## Residual deviance: 5859.6  on 5836  degrees of freedom
## AIC: 5913.6
##
## Number of Fisher Scoring iterations: 5
```

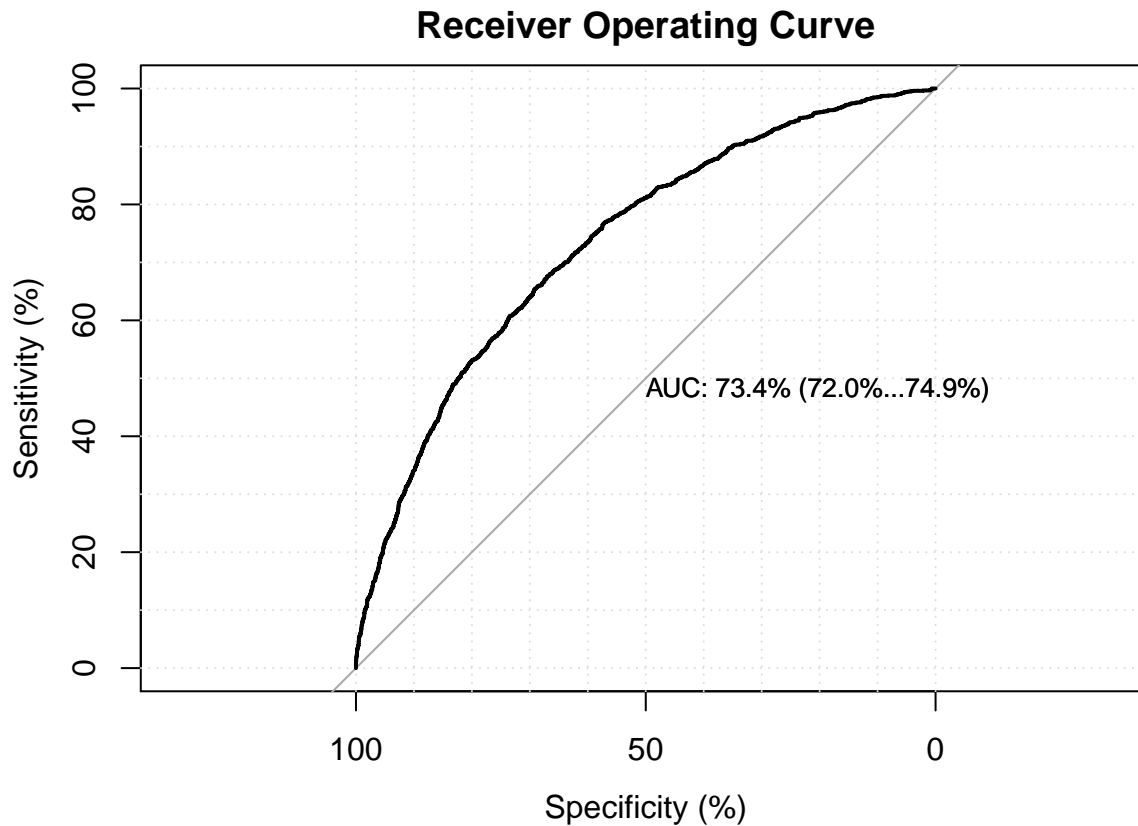


```
hoslem.test(cc_df$pap_num, fitted(mod_age_cubic_spline), g = 10) # good fit
```

```
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: cc_df$pap_num, fitted(mod_age_cubic_spline)
## X-squared = 3.1256, df = 8, p-value = 0.9262
```

```
gof(mod_age_cubic_spline, g = 10)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



```
##      chiSq df pVal
## PrI      1 2  4
## drI      2 2  3
## PrG      1 1  2
## drG      2 1  1
## PrCT     1 1  2
## drCT     2 1  1
##
##              val df pVal
## HL chiSq      8 3  8
## mHL F         9 4  9
## OsRo Z        5 5  3
## SstPgeq0.5 Z  2 5  6
## SstPl0.5 Z    6 5  2
## SstBoth chiSq 4 2  7
## SllPgeq0.5 chiSq 1 1  5
```

```
## S11P10.5 chiSq      3  1    1
## S11Both chiSq      7  2    4
```

```
anova(mod_back, mod_age_cubic_spline, test = "Chisq")
```

```
## Analysis of Deviance Table
```

```
##
```

```
## Model 1: pap_num ~ age + income_indiv + income_fam + totexp + race_f +
## marital_stat_f + educ_f + smoke_freq_f + limitation_f + afford_care_f +
## have_usc_f + inscov_gen_2018_f
```

```
## Model 2: pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv + income_fam +
## totexp + race_f + marital_stat_f + educ_f + smoke_freq_f +
## limitation_f + afford_care_f + have_usc_f + inscov_gen_2018_f
```

```
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
```

```
## 1      5841      5981.7
```

```
## 2      5836      5859.6  5   122.14 < 2.2e-16 ***
```

```
## ---
```

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

```
# LRT p-value < 0.05 so makes sense to have this spline term
```

```
# testing to see if quadratic age is sufficient or we need cubic spline
```

```
# we can do this since quad age model is nested within cubic spline model
```

```
anova(mod_age_quad, mod_age_cubic_spline, test = "Chisq")
```

```
## Analysis of Deviance Table
```

```
##
```

```
## Model 1: pap_num ~ age + I(age^2) + income_indiv + income_fam + totexp +
## race_f + marital_stat_f + educ_f + smoke_freq_f + limitation_f +
## afford_care_f + have_usc_f + inscov_gen_2018_f
```

```
## Model 2: pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv + income_fam +
## totexp + race_f + marital_stat_f + educ_f + smoke_freq_f +
## limitation_f + afford_care_f + have_usc_f + inscov_gen_2018_f
```

```
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
```

```
## 1      5840      5936.1
```

```
## 2      5836      5859.6  4   76.517 9.518e-16 ***
```

```
## ---
```

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

```
# p-value < 0.05 so makes sense to use cubic spline term instead of quad term only
```

```
# Building off the spline model to add interactions
```

```
# interaction model 1
```

```
mod_spline_interaction1 <- glm(pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv + income_fam +
summary(mod_spline_interaction1)
```

```
##
```

```
## Call:
```

```
## glm(formula = pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv +
```

```
## income_fam + totexp + race_f + marital_stat_f + educ_f +
```

```
## smoke_freq_f + limitation_f + afford_care_f + have_usc_f +
```

```
## inscov_gen_2018_f + marital_stat_f * income_fam, family = binomial(),
```

```
## data = cc_df)
```

```
##
```

```
## Deviance Residuals:
```

```

##      Min      1Q   Median      3Q      Max
## -1.9041  -0.7601  -0.5392   0.6994   2.9388
##
## Coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      3.062e-01  2.002e-01   1.529 0.126182
## bSpline(age, df = 6, degree = 3)1 -1.629e+00  3.398e-01  -4.794 1.64e-06 ***
## bSpline(age, df = 6, degree = 3)2 -1.562e+00  2.363e-01  -6.612 3.79e-11 ***
## bSpline(age, df = 6, degree = 3)3 -1.409e+00  2.829e-01  -4.982 6.30e-07 ***
## bSpline(age, df = 6, degree = 3)4 -1.764e-01  2.767e-01  -0.638 0.523626
## bSpline(age, df = 6, degree = 3)5 -1.211e+00  2.924e-01  -4.142 3.45e-05 ***
## bSpline(age, df = 6, degree = 3)6 -1.932e-01  2.516e-01  -0.768 0.442719
## income_indiv      -7.718e-06  1.546e-06  -4.992 5.97e-07 ***
## income_fam         4.182e-06  1.144e-06   3.655 0.000258 ***
## totexp           -1.049e-05  3.146e-06  -3.334 0.000855 ***
## race_fhispanic     2.448e-01  8.715e-02   2.809 0.004973 **
## race_fblack        2.389e-01  9.635e-02   2.479 0.013160 *
## race_fasian        1.197e+00  1.331e-01   8.997 < 2e-16 ***
## race_fother or multiple races  1.804e-01  1.784e-01   1.012 0.311713
## marital_stat_fmarried -2.030e-01  1.173e-01  -1.730 0.083605 .
## marital_stat_fwidowed -2.471e-01  2.611e-01  -0.946 0.343918
## marital_stat_fdivorced -1.678e-01  1.751e-01  -0.958 0.338008
## marital_stat_fseparated -5.044e-01  2.394e-01  -2.107 0.035101 *
## educ_fany high school  4.468e-01  7.281e-02   6.137 8.40e-10 ***
## educ_fnone or any elementary  2.863e-01  1.517e-01   1.887 0.059165 .
## smoke_freq_fsome days  3.280e-01  1.490e-01   2.202 0.027665 *
## smoke_freq_fevery day  3.992e-01  1.027e-01   3.887 0.000101 ***
## limitation_fyes      2.594e-01  1.239e-01   2.095 0.036206 *
## afford_care_fyes     -2.955e-01  1.173e-01  -2.520 0.011747 *
## have_usc_fyes       -6.586e-01  7.224e-02  -9.117 < 2e-16 ***
## inscov_gen_2018_fpublic only  1.300e-01  8.898e-02   1.461 0.143962
## inscov_gen_2018_funinsured   7.012e-01  1.084e-01   6.470 9.80e-11 ***
## income_fam:marital_stat_fmarried -6.252e-06  1.336e-06  -4.679 2.88e-06 ***
## income_fam:marital_stat_fwidowed -1.844e-06  4.455e-06  -0.414 0.678883
## income_fam:marital_stat_fdivorced -5.985e-06  3.179e-06  -1.883 0.059745 .
## income_fam:marital_stat_fseparated 4.274e-06  4.451e-06   0.960 0.336950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 6663.9  on 5862  degrees of freedom
## Residual deviance: 5832.9  on 5832  degrees of freedom
## AIC: 5894.9
##
## Number of Fisher Scoring iterations: 5

```

```

hoslem.test(cc_df$pap_num, fitted(mod_spline_interaction1), g = 10) # good fit

```

```

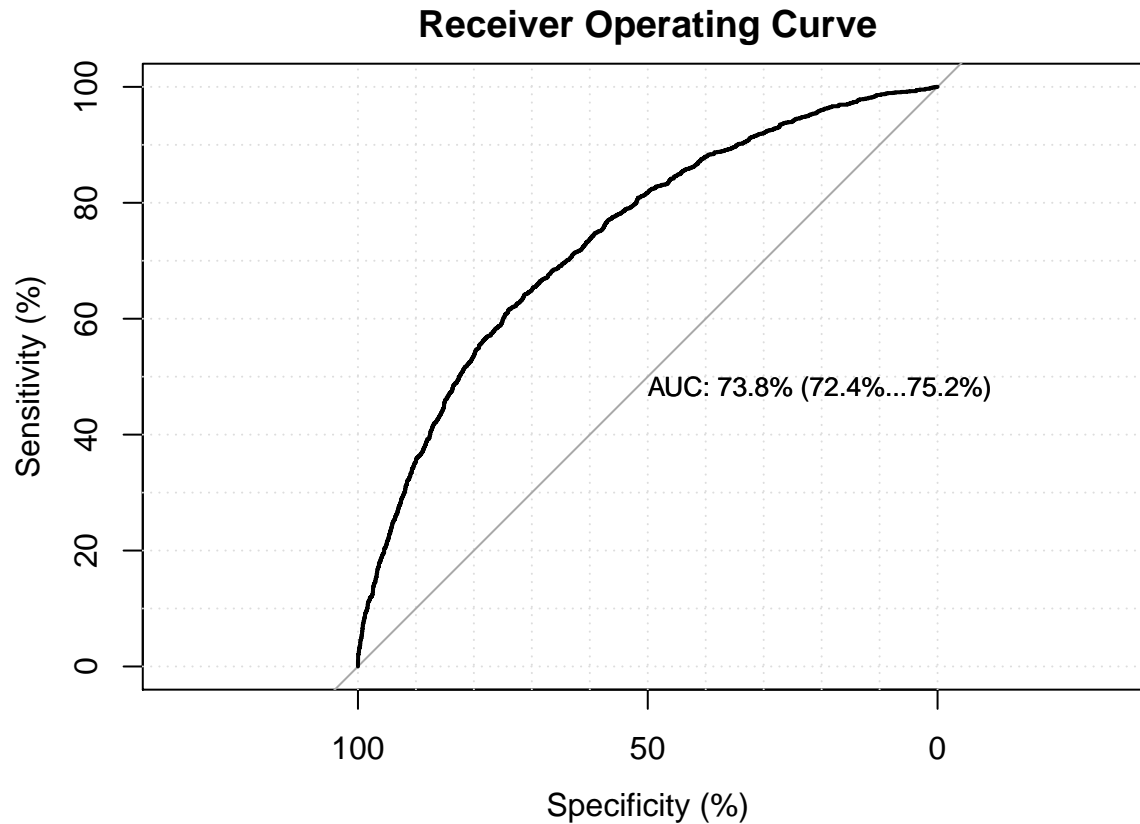
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  cc_df$pap_num, fitted(mod_spline_interaction1)
## X-squared = 4.8074, df = 8, p-value = 0.7779

```

```
gof(mod_spline_interaction1, g = 10)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



```
##      chiSq df pVal
## PrI      2 2    2
## drI      1 2    4
## PrG      2 1    1
## drG      1 1    3
## PrCT     2 1    1
## drCT     1 1    3
##
##              val df pVal
## HL chiSq      8 3    6
## mHL F         9 4    9
## OsRo Z        7 5    1
## SstPgeq0.5 Z   4 5    5
## SstPl0.5 Z     6 5    2
## SstBoth chiSq  5 2    7
## SllPgeq0.5 chiSq 1 1    4
## SllPl0.5 chiSq 2 1    3
## SllBoth chiSq  3 2    8
```

```
anova(mod_age_cubic_spline, mod_spline_interaction1, test = "Chisq")
```

```
## Analysis of Deviance Table
```

```
##
```

```
## Model 1: pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv + income_fam +
```

```
##      totexp + race_f + marital_stat_f + educ_f + smoke_freq_f +
##      limitation_f + afford_care_f + have_usc_f + inscov_gen_2018_f
## Model 2: pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv + income_fam +
##      totexp + race_f + marital_stat_f + educ_f + smoke_freq_f +
##      limitation_f + afford_care_f + have_usc_f + inscov_gen_2018_f +
##      marital_stat_f * income_fam
##      Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1          5836      5859.6
## 2          5832      5832.9  4   26.639 2.351e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# LRT p-value < 0.05 so makes sense to keep this interaction
```

```
# interaction model 2
```

```
mod_spline_interaction2 <- glm(pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv + income_fam +
summary(mod_spline_interaction2)
```

```
##
## Call:
## glm(formula = pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv +
##      income_fam + totexp + race_f + marital_stat_f + educ_f +
##      smoke_freq_f + limitation_f + afford_care_f + have_usc_f +
##      inscov_gen_2018_f + marital_stat_f * income_fam + educ_f *
##      totexp, family = binomial(), data = cc_df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9130  -0.7633  -0.5364   0.6918   3.1283
##
## Coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      3.507e-01  2.009e-01   1.746 0.080847 .
## bSpline(age, df = 6, degree = 3)1 -1.636e+00  3.396e-01  -4.816 1.47e-06 ***
## bSpline(age, df = 6, degree = 3)2 -1.549e+00  2.362e-01  -6.559 5.40e-11 ***
## bSpline(age, df = 6, degree = 3)3 -1.402e+00  2.828e-01  -4.957 7.15e-07 ***
## bSpline(age, df = 6, degree = 3)4 -1.770e-01  2.767e-01  -0.640 0.522448
## bSpline(age, df = 6, degree = 3)5 -1.206e+00  2.925e-01  -4.123 3.74e-05 ***
## bSpline(age, df = 6, degree = 3)6 -1.832e-01  2.519e-01  -0.727 0.467116
## income_indiv      -7.694e-06  1.547e-06  -4.974 6.54e-07 ***
## income_fam         4.141e-06  1.145e-06   3.616 0.000299 ***
## totexp            -2.026e-05  5.670e-06  -3.574 0.000352 ***
## race_fhispanic     2.400e-01  8.712e-02   2.754 0.005879 **
## race_fblack        2.355e-01  9.637e-02   2.444 0.014515 *
## race_fasian        1.181e+00  1.331e-01   8.877 < 2e-16 ***
## race_fother or multiple races  1.866e-01  1.789e-01   1.043 0.296751
## marital_stat_fmarrried -2.107e-01  1.173e-01  -1.796 0.072525 .
## marital_stat_fwidowed -2.539e-01  2.621e-01  -0.969 0.332738
## marital_stat_fdivorced -1.852e-01  1.757e-01  -1.054 0.292005
## marital_stat_fseperated -5.084e-01  2.398e-01  -2.120 0.034003 *
## educ_fany high school  3.617e-01  7.927e-02   4.562 5.06e-06 ***
## educ_fnone or any elementary  3.022e-01  1.633e-01   1.850 0.064269 .
## smoke_freq_fsome days  3.326e-01  1.491e-01   2.230 0.025724 *
## smoke_freq_fevery day  3.941e-01  1.028e-01   3.833 0.000127 ***
## limitation_fyes     2.574e-01  1.245e-01   2.067 0.038724 *
```

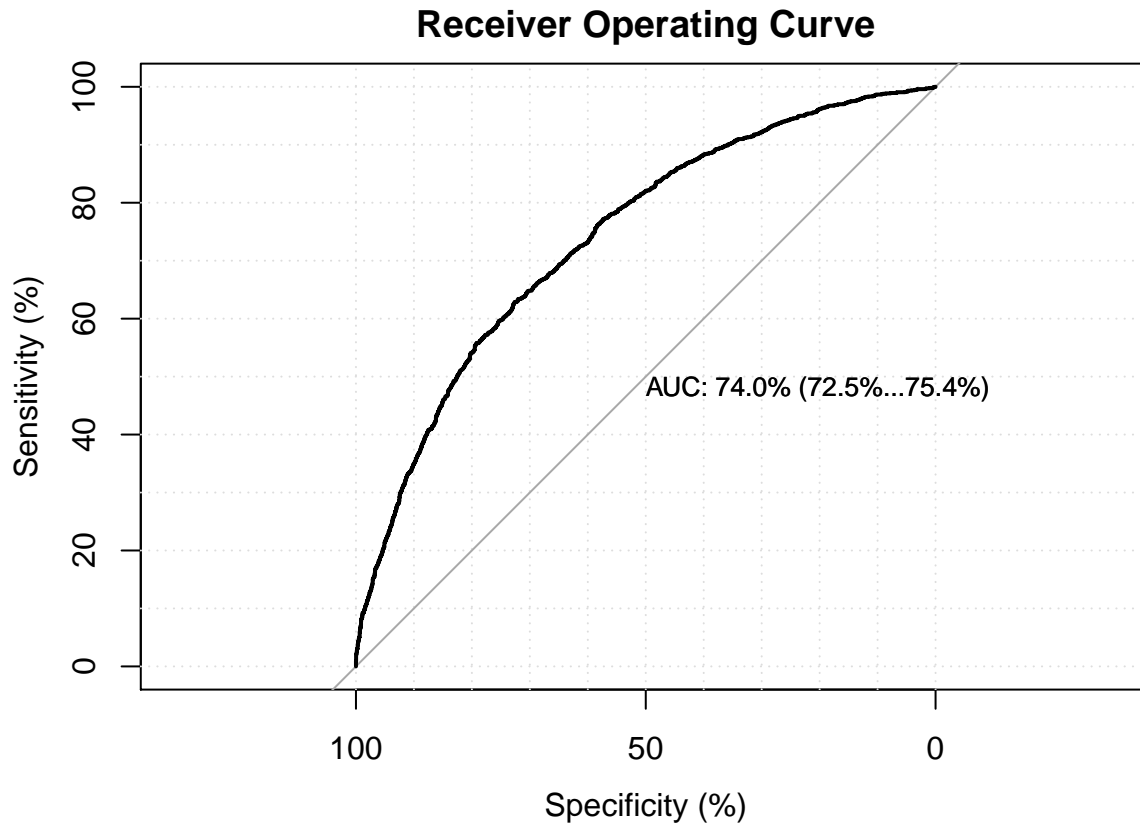
```

## afford_care_fyes          -2.920e-01  1.174e-01  -2.487  0.012873 *
## have_usc_fyes             -6.554e-01  7.228e-02  -9.068  < 2e-16 ***
## inscov_gen_2018_fpublic only  1.336e-01  8.902e-02   1.501  0.133343
## inscov_gen_2018_funinsured    6.999e-01  1.083e-01   6.462  1.03e-10 ***
## income_fam:marital_stat_fmarrried -6.187e-06  1.337e-06  -4.629  3.67e-06 ***
## income_fam:marital_stat_fwidowed -1.756e-06  4.466e-06  -0.393  0.694160
## income_fam:marital_stat_fdivorced -5.770e-06  3.195e-06  -1.806  0.070922 .
## income_fam:marital_stat_fseperated  3.972e-06  4.469e-06   0.889  0.374158
## totexp:educ_fany high school   1.754e-05  6.673e-06   2.628  0.008583 **
## totexp:educ_fnone or any elementary -3.469e-06  1.477e-05  -0.235  0.814280
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 6663.9  on 5862  degrees of freedom
## Residual deviance: 5824.1  on 5830  degrees of freedom
## AIC: 5890.1
##
## Number of Fisher Scoring iterations: 5
hoslem.test(cc_df$pap_num, fitted(mod_spline_interaction2), g = 10) # good fit

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  cc_df$pap_num, fitted(mod_spline_interaction2)
## X-squared = 7.2592, df = 8, p-value = 0.5089
gof(mod_spline_interaction2, g = 10)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases

```



```
##      chiSq df pVal
## PrI      2 2  4
## drI      1 2  2
## PrG      2 1  3
## drG      1 1  1
## PrCT     2 1  3
## drCT     1 1  1
##
##              val df pVal
## HL chiSq      9 3  6
## mHL F         8 4  9
## OsRo Z        7 5  1
## SstPgeq0.5 Z   2 5  8
## SstPl0.5 Z     3 5  2
## SstBoth chiSq  6 2  4
## SllPgeq0.5 chiSq 1 1  7
## SllPl0.5 chiSq 4 1  3
## SllBoth chiSq  5 2  5
```

```
anova(mod_spline_interaction1, mod_spline_interaction2, test = "Chisq")
```

```
## Analysis of Deviance Table
```

```
##
```

```
## Model 1: pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv + income_fam +
## totexp + race_f + marital_stat_f + educ_f + smoke_freq_f +
## limitation_f + afford_care_f + have_usc_f + inscov_gen_2018_f +
## marital_stat_f * income_fam
```

```
## Model 2: pap_num ~ bSpline(age, df = 6, degree = 3) + income_indiv + income_fam +
## totexp + race_f + marital_stat_f + educ_f + smoke_freq_f +
```

```
##      limitation_f + afford_care_f + have_usc_f + inscov_gen_2018_f +
##      marital_stat_f * income_fam + educ_f * totexp
##      Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5832      5832.9
## 2      5830      5824.1  2      8.813   0.0122 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# LRT p-value < 0.05 so makes sense to keep this interaction
```

```
# comparing AIC for all the association models we've built so far
```

```
AIC(mod_full,
    mod_back,
    mod_age_quad,
    mod_age_cubic_spline,
    mod_spline_interaction1,
    mod_spline_interaction2)
```

```
##              df      AIC
## mod_full      30 6030.070
## mod_back      22 6025.698
## mod_age_quad   23 5982.076
## mod_age_cubic_spline 27 5913.560
## mod_spline_interaction1 31 5894.920
## mod_spline_interaction2 33 5890.107
```

```
# pick one of these association models as the final one
```

```
# we should be good with these 6 and don't want to run anything more complicated which will make interp
```

Prediction modeling

Now we're working on creating prediction models using the training set and then testing it on the testing set.

```
# build test and train set
```

```
set.seed(1)
train_index <- createDataPartition(cc_df$pap_num, times = 1, p = 0.7, list = FALSE)
train_set <- cc_df[train_index, ]
```

```
## Warning: The `i` argument of ``[`()`` can't be a matrix as of tibble 3.0.0.
## Convert to a vector.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
test_set <- cc_df[-train_index, ]
```

```
# create function for looking at prediction metrics
```

```
pred_metrics <- function(model, cutoff) {
  p_hat <- predict(model, newdata = test_set, type = "response")
  y_hat <- ifelse(p_hat > cutoff, "no", "yes")
  confusion <- confusionMatrix(data = factor(y_hat, levels = c("yes", "no")),
                               reference = test_set$pap_f)

  roc <- roc(test_set$pap_num, p_hat)
  auc <- auc(roc)

  result <- list(p_hat = p_hat,
                 y_hat = y_hat,
```



```

        confusion = confusion,
        roc = roc,
        auc = auc)

result
}

# full model
fit_full <- glm(pap_num ~ age + income_indiv + income_fam + totexp + outofpocket_exp + genhlth_avg_f +
summary(fit_full)

##
## Call:
## glm(formula = pap_num ~ age + income_indiv + income_fam + totexp +
##      outofpocket_exp + genhlth_avg_f + region_f + race_f + marital_stat_f +
##      educ_f + smoke_freq_f + limitation_f + afford_care_f + have_usc_f +
##      inscov_gen_2018_f, family = binomial(), data = train_set)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7444  -0.7754  -0.5649   0.9213   2.6450
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.436e+00  4.027e-01  -3.565 0.000363 ***
## age           1.696e-02  3.565e-03   4.758 1.96e-06 ***
## income_indiv  -1.154e-05  1.729e-06  -6.670 2.56e-11 ***
## income_fam     1.151e-06  8.581e-07   1.341 0.179775
## totexp        -9.615e-06  4.128e-06  -2.329 0.019838 *
## outofpocket_exp -1.455e-05  2.627e-05  -0.554 0.579752
## genhlth_avg_ffair  4.839e-01  3.539e-01   1.368 0.171462
## genhlth_avg_fggood  3.717e-01  3.506e-01   1.060 0.289032
## genhlth_avg_fvery good  2.836e-01  3.552e-01   0.799 0.424524
## genhlth_avg_fexcellent  4.035e-01  3.643e-01   1.108 0.268038
## region_fmideast  -8.001e-02  1.344e-01  -0.595 0.551532
## region_fsouth    1.143e-01  1.201e-01   0.952 0.341261
## region_fwest     1.944e-03  1.281e-01   0.015 0.987891
## race_fhispanic    1.559e-01  1.057e-01   1.475 0.140344
## race_fblack       1.613e-01  1.160e-01   1.390 0.164434
## race_fasian       1.184e+00  1.558e-01   7.596 3.06e-14 ***
## race_fother or multiple races -1.448e-01  2.198e-01  -0.659 0.510096
## marital_stat_fmarrried -7.774e-01  1.012e-01  -7.685 1.53e-14 ***
## marital_stat_fwidowed -2.791e-01  2.202e-01  -1.267 0.205014
## marital_stat_fdivorced -3.905e-01  1.423e-01  -2.745 0.006059 **
## marital_stat_fseperated -4.198e-01  2.054e-01  -2.043 0.041007 *
## educ_fany high school  4.512e-01  8.639e-02   5.222 1.77e-07 ***
## educ_fnone or any elementary  4.295e-01  1.775e-01   2.420 0.015513 *
## smoke_freq_fsome days  1.343e-01  1.804e-01   0.745 0.456526
## smoke_freq_fevery day  3.300e-01  1.217e-01   2.712 0.006688 **
## limitation_fyes     2.094e-01  1.547e-01   1.353 0.175921
## afford_care_fyes    -2.992e-01  1.390e-01  -2.152 0.031369 *
## have_usc_fyes      -5.510e-01  8.504e-02  -6.479 9.24e-11 ***
## inscov_gen_2018_fpublic only  4.978e-03  1.062e-01   0.047 0.962596
## inscov_gen_2018_funinsured  5.591e-01  1.279e-01   4.372 1.23e-05 ***
## ---

```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```

## Setting direction: controls < cases
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
test_cutoff_df <- data.frame(p_seq, sensitivity, specificity)
# cutoff of 0.26 looks best

# get prediction metrics for full fit
fit_full_pred_metrics <- pred_metrics(fit_full, 0.26)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
fit_full_pred_metrics$confusion

## Confusion Matrix and Statistics
##
##           Reference
## Prediction yes  no
##           yes 870 146
##           no  453 289
##
##           Accuracy : 0.6593
##           95% CI : (0.6366, 0.6814)
##           No Information Rate : 0.7526
##           P-Value [Acc > NIR] : 1
##
##           Kappa : 0.2603
##
## Mcnemar's Test P-Value : <2e-16
##
##           Sensitivity : 0.6576
##           Specificity : 0.6644
##           Pos Pred Value : 0.8563
##           Neg Pred Value : 0.3895
##           Prevalence : 0.7526
##           Detection Rate : 0.4949
##           Detection Prevalence : 0.5779
##           Balanced Accuracy : 0.6610
##
##           'Positive' Class : yes
##
fit_full_pred_metrics$auc

## Area under the curve: 0.7214
# forward/backward selection

```



```

# forward model selection
fit_forw <- step(glm(pap_num ~ 1, data = train_set, family = binomial()), ~ age + income_indiv + income.

## Start:  AIC=4697.78
## pap_num ~ 1
##
##           Df Deviance    AIC
## + income_indiv      1  4531.3 4535.3
## + inscov_gen_2018_f  2  4550.7 4556.7
## + educ_f            2  4566.4 4572.4
## + income_fam        1  4589.4 4593.4
## + race_f            4  4589.9 4599.9
## + marital_stat_f    4  4598.5 4608.5
## + have_usc_f        1  4607.4 4611.4
## + outofpocket_exp   1  4662.1 4666.1
## + totexp            1  4667.5 4671.5
## + smoke_freq_f      2  4672.2 4678.2
## + region_f          3  4673.6 4681.6
## + genhlth_avg_f     4  4672.7 4682.7
## + limitation_f      1  4688.1 4692.1
## <none>              4695.8 4697.8
## + afford_care_f     1  4695.5 4699.5
## + age               1  4695.8 4699.8
##
## Step:  AIC=4535.3
## pap_num ~ income_indiv
##
##           Df Deviance    AIC
## + race_f          4  4448.5 4460.5
## + have_usc_f      1  4463.8 4469.8
## + marital_stat_f  4  4460.8 4472.8
## + inscov_gen_2018_f 2  4468.6 4476.6
## + educ_f          2  4476.4 4484.4
## + totexp          1  4509.8 4515.8
## + outofpocket_exp 1  4518.2 4524.2
## + region_f        3  4516.1 4526.1
## + income_fam       1  4521.9 4527.9
## + smoke_freq_f     2  4520.4 4528.4
## + age              1  4527.9 4533.9
## + genhlth_avg_f    4  4523.2 4535.2
## <none>             4531.3 4535.3
## + limitation_f     1  4530.9 4536.9
## + afford_care_f    1  4531.0 4537.0
##
## Step:  AIC=4460.49
## pap_num ~ income_indiv + race_f
##
##           Df Deviance    AIC
## + marital_stat_f    4  4385.2 4405.2
## + have_usc_f        1  4394.4 4408.4
## + inscov_gen_2018_f 2  4395.5 4411.5
## + educ_f            2  4400.4 4416.4
## + smoke_freq_f      2  4426.6 4442.6
## + totexp            1  4436.5 4450.5

```

```

## + income_fam      1  4438.2 4452.2
## + outofpocket_exp 1  4443.2 4457.2
## + age             1  4443.3 4457.3
## + region_f        3  4440.7 4458.7
## <none>             4448.5 4460.5
## + limitation_f    1  4446.8 4460.8
## + genhlth_avg_f    4  4441.7 4461.7
## + afford_care_f    1  4448.4 4462.4
##
## Step:  AIC=4405.24
## pap_num ~ income_indiv + race_f + marital_stat_f
##
##              Df Deviance    AIC
## + have_usc_f      1  4337.9 4359.9
## + educ_f          2  4338.7 4362.7
## + inscov_gen_2018_f 2  4344.6 4368.6
## + age             1  4365.4 4387.4
## + totexp          1  4372.8 4394.8
## + smoke_freq_f    2  4370.9 4394.9
## + region_f        3  4375.1 4401.1
## + outofpocket_exp 1  4382.2 4404.2
## <none>             4385.2 4405.2
## + afford_care_f    1  4384.7 4406.7
## + limitation_f    1  4384.8 4406.8
## + income_fam      1  4385.1 4407.1
## + genhlth_avg_f    4  4379.9 4407.9
##
## Step:  AIC=4359.87
## pap_num ~ income_indiv + race_f + marital_stat_f + have_usc_f
##
##              Df Deviance    AIC
## + educ_f          2  4290.8 4316.8
## + age             1  4307.1 4331.1
## + inscov_gen_2018_f 2  4309.4 4335.4
## + smoke_freq_f    2  4323.4 4349.4
## + totexp          1  4330.6 4354.6
## + limitation_f    1  4335.1 4359.1
## <none>             4337.9 4359.9
## + region_f        3  4331.9 4359.9
## + outofpocket_exp 1  4336.5 4360.5
## + afford_care_f    1  4336.6 4360.6
## + genhlth_avg_f    4  4330.7 4360.7
## + income_fam      1  4337.9 4361.9
##
## Step:  AIC=4316.81
## pap_num ~ income_indiv + race_f + marital_stat_f + have_usc_f +
##      educ_f
##
##              Df Deviance    AIC
## + age             1  4267.9 4295.9
## + inscov_gen_2018_f 2  4269.7 4299.7
## + totexp          1  4283.8 4311.8
## + smoke_freq_f    2  4282.4 4312.4
## <none>             4290.8 4316.8

```

```

## + region_f          3    4284.9 4316.9
## + limitation_f      1    4289.0 4317.0
## + afford_care_f     1    4289.5 4317.5
## + outofpocket_exp   1    4290.0 4318.0
## + income_fam        1    4290.0 4318.0
## + genhlth_avg_f     4    4286.1 4320.1
##
## Step:  AIC=4295.86
## pap_num ~ income_indiv + race_f + marital_stat_f + have_usc_f +
##      educ_f + age
##
##              Df Deviance    AIC
## + inscov_gen_2018_f  2    4245.7 4277.7
## + totexp             1    4257.9 4287.9
## + smoke_freq_f       2    4260.5 4292.5
## + region_f           3    4261.5 4295.5
## <none>                4267.9 4295.9
## + outofpocket_exp    1    4266.0 4296.0
## + afford_care_f      1    4266.3 4296.3
## + income_fam         1    4266.8 4296.8
## + limitation_f       1    4267.6 4297.6
## + genhlth_avg_f      4    4263.3 4299.3
##
## Step:  AIC=4277.72
## pap_num ~ income_indiv + race_f + marital_stat_f + have_usc_f +
##      educ_f + age + inscov_gen_2018_f
##
##              Df Deviance    AIC
## + totexp             1    4238.5 4272.5
## + smoke_freq_f       2    4238.3 4274.3
## + afford_care_f      1    4241.5 4275.5
## + outofpocket_exp    1    4243.6 4277.6
## <none>                4245.7 4277.7
## + income_fam         1    4244.0 4278.0
## + limitation_f       1    4245.0 4279.0
## + region_f           3    4242.3 4280.3
## + genhlth_avg_f      4    4241.3 4281.3
##
## Step:  AIC=4272.48
## pap_num ~ income_indiv + race_f + marital_stat_f + have_usc_f +
##      educ_f + age + inscov_gen_2018_f + totexp
##
##              Df Deviance    AIC
## + smoke_freq_f       2    4230.9 4268.9
## + afford_care_f      1    4234.5 4270.5
## + limitation_f       1    4236.1 4272.1
## <none>                4238.5 4272.5
## + income_fam         1    4236.6 4272.6
## + outofpocket_exp    1    4238.2 4274.2
## + region_f           3    4235.1 4275.1
## + genhlth_avg_f      4    4234.4 4276.4
##
## Step:  AIC=4268.89
## pap_num ~ income_indiv + race_f + marital_stat_f + have_usc_f +

```

```

##      educ_f + age + inscov_gen_2018_f + totexp + smoke_freq_f
##
##              Df Deviance    AIC
## + afford_care_f    1  4226.4 4266.4
## + income_fam       1  4228.7 4268.7
## <none>              4230.9 4268.9
## + limitation_f     1  4229.0 4269.0
## + outofpocket_exp  1  4230.6 4270.6
## + region_f         3  4227.6 4271.6
## + genhlth_avg_f    4  4227.0 4273.0
##
## Step: AIC=4266.44
## pap_num ~ income_indiv + race_f + marital_stat_f + have_usc_f +
##      educ_f + age + inscov_gen_2018_f + totexp + smoke_freq_f +
##      afford_care_f
##
##              Df Deviance    AIC
## + limitation_f     1  4224.2 4266.2
## <none>              4226.4 4266.4
## + income_fam       1  4224.6 4266.6
## + outofpocket_exp  1  4226.2 4268.2
## + region_f         3  4222.9 4268.9
## + genhlth_avg_f    4  4222.1 4270.1
##
## Step: AIC=4266.19
## pap_num ~ income_indiv + race_f + marital_stat_f + have_usc_f +
##      educ_f + age + inscov_gen_2018_f + totexp + smoke_freq_f +
##      afford_care_f + limitation_f
##
##              Df Deviance    AIC
## <none>              4224.2 4266.2
## + income_fam       1  4222.4 4266.4
## + outofpocket_exp  1  4224.0 4268.0
## + region_f         3  4220.8 4268.8
## + genhlth_avg_f    4  4220.0 4270.0
summary(fit_forw)

##
## Call:
## glm(formula = pap_num ~ income_indiv + race_f + marital_stat_f +
##      have_usc_f + educ_f + age + inscov_gen_2018_f + totexp +
##      smoke_freq_f + afford_care_f + limitation_f, family = binomial(),
##      data = train_set)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7632  -0.7782  -0.5667   0.9275   2.6622
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.042e+00  1.635e-01  -6.370 1.89e-10 ***
## income_indiv  -1.062e-05  1.543e-06  -6.883 5.84e-12 ***
## race_fhispanic  1.922e-01  1.021e-01   1.882  0.05984 .
## race_fblack    2.214e-01  1.106e-01   2.003  0.04521 *

```

```

## race_fasian          1.213e+00  1.526e-01  7.950 1.87e-15 ***
## race_fother or multiple races -1.323e-01  2.191e-01 -0.604 0.54612
## marital_stat_fmarrried -7.305e-01  9.638e-02 -7.579 3.49e-14 ***
## marital_stat_fwidowed -2.486e-01  2.187e-01 -1.137 0.25557
## marital_stat_fdivorced -3.970e-01  1.413e-01 -2.810 0.00496 **
## marital_stat_fseperated -4.134e-01  2.041e-01 -2.025 0.04286 *
## have_usc_fyes        -5.566e-01  8.427e-02 -6.605 3.98e-11 ***
## educ_fany high school  4.446e-01  8.536e-02  5.209 1.90e-07 ***
## educ_fnone or any elementary 4.338e-01  1.760e-01  2.465 0.01371 *
## age                  1.682e-02  3.515e-03  4.783 1.72e-06 ***
## inscov_gen_2018_fpublic only -1.415e-02  1.034e-01 -0.137 0.89110
## inscov_gen_2018_funinsured  5.652e-01  1.261e-01  4.481 7.45e-06 ***
## totexp              -1.031e-05  3.778e-06 -2.728 0.00638 **
## smoke_freq_fsome days  1.413e-01  1.797e-01  0.786 0.43174
## smoke_freq_fevery day  3.271e-01  1.209e-01  2.706 0.00681 **
## afford_care_fyes      -2.926e-01  1.352e-01 -2.164 0.03047 *
## limitation_fyes       2.229e-01  1.478e-01  1.508 0.13146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 4695.8 on 4104 degrees of freedom
## Residual deviance: 4224.2 on 4084 degrees of freedom
## AIC: 4266.2
##
## Number of Fisher Scoring iterations: 5
# backward model selection
fit_back <- step(fit_full, direction = "backward")

## Start: AIC=4274.65
## pap_num ~ age + income_indiv + income_fam + totexp + outofpocket_exp +
## genhlth_avg_f + region_f + race_f + marital_stat_f + educ_f +
## smoke_freq_f + limitation_f + afford_care_f + have_usc_f +
## inscov_gen_2018_f
##
##           Df Deviance    AIC
## - genhlth_avg_f      4  4218.7 4270.7
## - region_f           3  4218.0 4272.0
## - outofpocket_exp     1  4215.0 4273.0
## - income_fam          1  4216.4 4274.4
## - limitation_f        1  4216.5 4274.5
## <none>                4214.6 4274.6
## - afford_care_f       1  4219.4 4277.4
## - smoke_freq_f        2  4222.1 4278.1
## - totexp              1  4221.0 4279.0
## - inscov_gen_2018_f   2  4235.5 4291.5
## - age                 1  4237.4 4295.4
## - educ_f              2  4242.6 4298.6
## - have_usc_f          1  4256.1 4314.1
## - income_indiv        1  4261.8 4319.8
## - race_f              4  4271.9 4323.9
## - marital_stat_f      4  4275.7 4327.7
##

```

```

## Step: AIC=4270.71
## pap_num ~ age + income_indiv + income_fam + totexp + outofpocket_exp +
##      region_f + race_f + marital_stat_f + educ_f + smoke_freq_f +
##      limitation_f + afford_care_f + have_usc_f + inscov_gen_2018_f
##
##           Df Deviance    AIC
## - region_f      3  4222.2 4268.2
## - outofpocket_exp 1  4219.0 4269.0
## - income_fam     1  4220.6 4270.6
## - limitation_f   1  4220.7 4270.7
## <none>           4218.7 4270.7
## - afford_care_f  1  4223.3 4273.3
## - smoke_freq_f   2  4226.4 4274.4
## - totexp         1  4225.4 4275.4
## - inscov_gen_2018_f 2  4239.2 4287.2
## - age           1  4242.6 4292.6
## - educ_f        2  4247.5 4295.5
## - have_usc_f     1  4259.6 4309.6
## - income_indiv   1  4266.5 4316.5
## - race_f        4  4276.2 4320.2
## - marital_stat_f 4  4280.7 4324.7
##
## Step: AIC=4268.22
## pap_num ~ age + income_indiv + income_fam + totexp + outofpocket_exp +
##      race_f + marital_stat_f + educ_f + smoke_freq_f + limitation_f +
##      afford_care_f + have_usc_f + inscov_gen_2018_f
##
##           Df Deviance    AIC
## - outofpocket_exp 1  4222.4 4266.4
## - income_fam     1  4224.0 4268.0
## <none>           4222.2 4268.2
## - limitation_f   1  4224.4 4268.4
## - afford_care_f  1  4226.6 4270.6
## - smoke_freq_f   2  4229.9 4271.9
## - totexp         1  4229.2 4273.2
## - inscov_gen_2018_f 2  4245.5 4287.5
## - age           1  4245.8 4289.8
## - educ_f        2  4251.1 4293.1
## - have_usc_f     1  4265.5 4309.5
## - income_indiv   1  4270.3 4314.3
## - race_f        4  4281.4 4319.4
## - marital_stat_f 4  4282.6 4320.6
##
## Step: AIC=4266.44
## pap_num ~ age + income_indiv + income_fam + totexp + race_f +
##      marital_stat_f + educ_f + smoke_freq_f + limitation_f + afford_care_f +
##      have_usc_f + inscov_gen_2018_f
##
##           Df Deviance    AIC
## - income_fam     1  4224.2 4266.2
## <none>           4222.4 4266.4
## - limitation_f   1  4224.6 4266.6
## - afford_care_f  1  4226.9 4268.9
## - smoke_freq_f   2  4230.1 4270.1

```

```

## - totexp          1  4231.3 4273.3
## - inscov_gen_2018_f 2  4245.6 4285.6
## - age             1  4245.8 4287.8
## - educ_f          2  4251.4 4291.4
## - have_usc_f       1  4266.0 4308.0
## - income_indiv     1  4270.9 4312.9
## - race_f           4  4282.1 4318.1
## - marital_stat_f   4  4282.9 4318.9
##
## Step: AIC=4266.19
## pap_num ~ age + income_indiv + totexp + race_f + marital_stat_f +
##      educ_f + smoke_freq_f + limitation_f + afford_care_f + have_usc_f +
##      inscov_gen_2018_f
##
##              Df Deviance    AIC
## <none>                4224.2 4266.2
## - limitation_f        1  4226.4 4266.4
## - afford_care_f        1  4229.0 4269.0
## - smoke_freq_f         2  4231.6 4269.6
## - totexp               1  4233.0 4273.0
## - inscov_gen_2018_f    2  4247.0 4285.0
## - age                  1  4247.2 4287.2
## - educ_f               2  4252.1 4290.1
## - have_usc_f           1  4267.2 4307.2
## - marital_stat_f       4  4284.0 4318.0
## - income_indiv         1  4278.5 4318.5
## - race_f               4  4286.7 4320.7

```

```
summary(fit_back)
```

```

##
## Call:
## glm(formula = pap_num ~ age + income_indiv + totexp + race_f +
##      marital_stat_f + educ_f + smoke_freq_f + limitation_f + afford_care_f +
##      have_usc_f + inscov_gen_2018_f, family = binomial(), data = train_set)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7632  -0.7782  -0.5667   0.9275   2.6622
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.042e+00  1.635e-01  -6.370 1.89e-10 ***
## age             1.682e-02  3.515e-03   4.783 1.72e-06 ***
## income_indiv   -1.062e-05  1.543e-06  -6.883 5.84e-12 ***
## totexp         -1.031e-05  3.778e-06  -2.728  0.00638 **
## race_hispanic   1.922e-01  1.021e-01   1.882  0.05984 .
## race_black      2.214e-01  1.106e-01   2.003  0.04521 *
## race_asian      1.213e+00  1.526e-01   7.950 1.87e-15 ***
## race_other or multiple races -1.323e-01  2.191e-01  -0.604  0.54612
## marital_stat_married -7.305e-01  9.638e-02  -7.579 3.49e-14 ***
## marital_stat_widowed -2.486e-01  2.187e-01  -1.137  0.25557
## marital_stat_divorced -3.970e-01  1.413e-01  -2.810  0.00496 **
## marital_stat_separated -4.134e-01  2.041e-01  -2.025  0.04286 *
## educ_fany high school  4.446e-01  8.536e-02   5.209 1.90e-07 ***

```

```

## educ_fnone or any elementary    4.338e-01  1.760e-01  2.465  0.01371 *
## smoke_freq_fsome days          1.413e-01  1.797e-01  0.786  0.43174
## smoke_freq_fevery day          3.271e-01  1.209e-01  2.706  0.00681 **
## limitation_fyes                 2.229e-01  1.478e-01  1.508  0.13146
## afford_care_fyes               -2.926e-01  1.352e-01 -2.164  0.03047 *
## have_usc_fyes                  -5.566e-01  8.427e-02 -6.605  3.98e-11 ***
## inscov_gen_2018_fpublic only   -1.415e-02  1.034e-01 -0.137  0.89110
## inscov_gen_2018_funinsured     5.652e-01  1.261e-01  4.481  7.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 4695.8 on 4104 degrees of freedom
## Residual deviance: 4224.2 on 4084 degrees of freedom
## AIC: 4266.2
##
## Number of Fisher Scoring iterations: 5
# forward and backward selection produce the same model
hoslem.test(train_set$pap_num, fitted(fit_back), g = 10) # good fit to training for forward/backward mo

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: train_set$pap_num, fitted(fit_back)
## X-squared = 8.5181, df = 8, p-value = 0.3846
# now we can see how forward/backward model performs on testing data
fit_back_pred_metrics <- pred_metrics(fit_back, 0.26)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
fit_back_pred_metrics$confusion # better accuracy than full fit

## Confusion Matrix and Statistics
##
##           Reference
## Prediction yes  no
##           yes 874 142
##           no  449 293
##
##           Accuracy : 0.6638
##           95% CI : (0.6412, 0.6859)
##           No Information Rate : 0.7526
##           P-Value [Acc > NIR] : 1
##
##           Kappa : 0.2702
##
## Mcnemar's Test P-Value : <2e-16
##
##           Sensitivity : 0.6606
##           Specificity : 0.6736
##           Pos Pred Value : 0.8602
##           Neg Pred Value : 0.3949

```



```

##           Prevalence : 0.7526
##           Detection Rate : 0.4972
##           Detection Prevalence : 0.5779
##           Balanced Accuracy : 0.6671
##
##           'Positive' Class : yes
##
fit_back_pred_metrics$auc # slightly worse auc than full fit

## Area under the curve: 0.7205
# going to use auc as the main comparison metric for picking the best prediction model since it looks o

# going forward with building off of full model since it has better auc

# add nonlinear terms for age

# try quadratic first
fit_age_quad <- glm(formula = pap_num ~ age + I(age^2) + income_indiv + income_fam + totexp + outofpock
hoslem.test(train_set$pap_num, fitted(fit_age_quad), g = 10) # good fit to training

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: train_set$pap_num, fitted(fit_age_quad)
## X-squared = 3.6041, df = 8, p-value = 0.891
# now we can see how it performs on testing data
fit_age_quad_pred_metrics <- pred_metrics(fit_age_quad, 0.26)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
fit_age_quad_pred_metrics$auc # slightly better auc than full fit model

## Area under the curve: 0.7266
# try spline
fit_age_cubic_spline <- glm(formula = pap_num ~ age + bSpline(age, df = 6, degree = 3) + income_indiv +
hoslem.test(train_set$pap_num, fitted(fit_age_quad), g = 10) # good fit to training

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: train_set$pap_num, fitted(fit_age_quad)
## X-squared = 3.6041, df = 8, p-value = 0.891
# now we can see how it performs on testing data
fit_age_cubic_spline_pred_metrics <- pred_metrics(fit_age_cubic_spline, 0.26)

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
fit_age_cubic_spline_pred_metrics$auc # slightly lower auc than age quad model

## Area under the curve: 0.7265

```

```

# building off of quad age model since it has the best auc so far

# add interaction terms

# interaction 1
fit_age_quad_interaction1 <- glm(formula = pap_num ~ age + I(age^2) + income_indiv + income_fam + totexp,
                                data = train_set, family = binomial)

hoslem.test(train_set$pap_num, fitted(fit_age_quad_interaction1), g = 10) # good fit to training

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: train_set$pap_num, fitted(fit_age_quad_interaction1)
## X-squared = 8.2218, df = 8, p-value = 0.4121

# now we can see how it performs on testing data
fit_age_quad_interaction1_pred_metrics <- pred_metrics(fit_age_quad_interaction1, 0.26)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
fit_age_quad_interaction1_pred_metrics$auc # slightly better auc than quad age model

## Area under the curve: 0.7304

# interaction 2
fit_age_quad_interaction2 <- glm(formula = pap_num ~ age + I(age^2) + income_indiv + income_fam + totexp,
                                data = train_set, family = binomial)

hoslem.test(train_set$pap_num, fitted(fit_age_quad_interaction2), g = 10) # good fit to training

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: train_set$pap_num, fitted(fit_age_quad_interaction2)
## X-squared = 5.6043, df = 8, p-value = 0.6915

# now we can see how it performs on testing data
fit_age_quad_interaction2_pred_metrics <- pred_metrics(fit_age_quad_interaction2, 0.26)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
fit_age_quad_interaction2_pred_metrics$auc # no difference in auc from interaction 1 so this interaction is not significant

## Area under the curve: 0.7317

# build and compare more prediction models
# since we don't need to worry about interpretation for our prediction model, feel free to make models

```