432 Class 09

https://thomaselove.github.io/432-2024

2024-02-12

## Today’s Agenda, Part I

* A New NHANES Example
* [Logistic Regression Analyses in Project A](https://thomaselove.github.io/432-2024/projA.html#new-section-9.-logistic-regression-analyses)
  + Establishing a Research Question
  + Identifying / Tidying Outcome and Candidate Predictors
  + Dealing with Missing Data
  + Building a “Main Effects” Model Y and Plotting Effects
  + Considering Non-Linear Terms
  + Fitting an “Augmented” Model Z and Plotting Effects

## Today’s Agenda, II

* Summarizing/Presenting a Final Model
  + In-Sample and Validated Model Summaries
  + Selecting Model Y or Model Z
  + Describing a Meaningful Effect (see Notes, Chapter 22)
  + ROC curve for the Final Model
  + Nomogram for the Final Model

## Today’s R Setup

knitr::opts\_chunk$set(comment = NA)  
  
library(nhanesA)  
library(broom)  
library(caret)  
library(janitor)  
library(gt)  
library(mosaic)  
library(naniar)  
library(pROC)  
library(rms)  
library(simputation)  
library(tidyverse)  
  
theme\_set(theme\_bw())

# The Data

## NHANES 2011-12 Data

We’ll use data from the 2011-2012 administration of NHANES here, specifically including variables from these data bases.

* **DEMO\_G** for Demographic Variables
* **CDQ\_G** for Cardiovascular Health Questionnaire
* **HSQ\_G** for Current Health Status Questionnaire
* **BPX\_G** for Blood Pressure Examination Results
* **BMX\_G** for Body Measures Examination Results
* **MCQ\_G** for Medical Conditions Questionnaire

## Pulling the Data

demo\_g <- nhanes("DEMO\_G", translated = FALSE) |> tibble(); dim(demo\_g)

[1] 9756 48

cdq\_g <- nhanes("CDQ\_G", translated = FALSE) |> tibble(); dim(cdq\_g)

[1] 3603 17

hsq\_g <- nhanes("HSQ\_G", translated = FALSE) |> tibble(); dim(hsq\_g)

[1] 8956 14

bpx\_g <- nhanes("BPX\_G", translated = FALSE) |> tibble(); dim(bpx\_g)

[1] 9338 27

bmx\_g <- nhanes("BMX\_G", translated = FALSE) |> tibble(); dim(bmx\_g)

[1] 9338 26

mcq\_g <- nhanes("MCQ\_G", translated = FALSE) |> tibble(); dim(mcq\_g)

[1] 9364 92

## Merging the Tibbles

df\_mlist <- list(demo\_g, cdq\_g, hsq\_g, bpx\_g, bmx\_g, mcq\_g)  
  
nh\_merge <- df\_mlist |>  
 reduce(left\_join, by = "SEQN") # reduce is from purrr  
  
dim(nh\_merge)

[1] 9756 219

* We had 224 variables in our original six tibbles, but that counts the SEQN variable six times, and we only have it once in our nh\_merge tibble.

Which of these 219 variables are we actually going to use?

## Seven Variables Used Today

| NHANES | Description | Source |
| --- | --- | --- |
| SEQN | Identifying code | All 6 files |
| CDQ010 | Short of breath on stairs/inclines? | CDQ\_G |
| RIDAGEYR | Age in years at screening | DEMO\_G |
| HSD010 | General health (E/VG/G/F/P) | HSQ\_G |
| BPXDI1 | Diastolic BP (first reading, in mm Hg) | BPX\_G |
| BMXBMI | Body Mass Index (kg/) | BMX\_G |
| MCQ010 | Ever told you have asthma? | MCQ\_G |

## Selecting Today’s Variables

We’re filtering to people with a 1 or 2 on CDQ010, and with RIDAGEYR < 80, and we’ll also rename some of the variables.

nh\_today <- nh\_merge |>  
 select(SEQN, CDQ010, age = RIDAGEYR, sroh = HSD010,   
 dbp = BPXDI1, bmi = BMXBMI,   
 asthma = MCQ010) |>  
 filter(CDQ010 < 3) |>  
 filter(age < 80)  
  
dim(nh\_today)

[1] 3234 7

## Data Checking and Cleaning (1)

1. SEQN should be unique for each row in the data.

identical(nrow(nh\_today), n\_distinct(nh\_today$SEQN))

[1] TRUE

## Data Checking and Cleaning (2)

1. CDQ010 was 1 for Yes, 2 for No. We want 1 for Yes, 0 for No.

* We’ll also name this new variable sbreath (for shortness of breath) and place it immediately after the subject ID SEQN.

nh\_today <- nh\_today |> mutate(sbreath = 2 - CDQ010)  
  
nh\_today |> count(CDQ010, sbreath) # sanity check

# A tibble: 2 × 3  
 CDQ010 sbreath n  
 <dbl> <dbl> <int>  
1 1 1 1015  
2 2 0 2219

nh\_today <- nh\_today |> select(-CDQ010) |>   
 relocate(sbreath, .after = "SEQN")

## Data Checking and Cleaning (3-5)

1. Age should be between 40 and 79 years
2. Body mass index should be between 12.4 and 82.1 kg/
3. Diastolic BP should be between 30 and 120 mm Hg (treat values below 30 as NA)

df\_stats(~ age + bmi + dbp, data = nh\_today) |>   
 rename(var = response) |> gt() |> tab\_options(table.font.size = 24)

| var | min | Q1 | median | Q3 | max | mean | sd | n | missing |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 40.0 | 48.0 | 57.0 | 65.0 | 79.0 | 57.34818 | 10.699621 | 3234 | 0 |
| bmi | 13.6 | 24.7 | 28.3 | 32.9 | 82.1 | 29.52410 | 6.874603 | 3062 | 172 |
| dbp | 0.0 | 66.0 | 74.0 | 80.0 | 120.0 | 73.11493 | 12.438088 | 2854 | 380 |

## Replace dbp values below 30 with NA

nh\_today |> count(dbp < 30)

# A tibble: 3 × 2  
 `dbp < 30` n  
 <lgl> <int>  
1 FALSE 2842  
2 TRUE 12  
3 NA 380

nh\_today <- replace\_with\_na\_at(nh\_today, "dbp", ~ .x < 30)  
  
favstats(~ dbp, data = nh\_today) |> gt() |> tab\_options(table.font.size = 24)

| min | Q1 | median | Q3 | max | mean | sd | n | missing |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 30 | 66 | 74 | 80 | 120 | 73.40957 | 11.59503 | 2842 | 392 |

## Data Checking and Cleaning (6)

1. asthma should be recoded into a two-level factor (it is currently 1 = Yes, 2 = No, and 9 = Don’t Know, which we’ll treat as missing)

nh\_today <- nh\_today |>  
 mutate(asthma = fct\_recode( factor(asthma),   
 "Yes" = "1", "No" = "2", NULL = "9"),  
 asthma = fct\_relevel(asthma, "No"))  
  
nh\_today |> tabyl(asthma) |> adorn\_pct\_formatting()

asthma n percent valid\_percent  
 No 2779 85.9% 86.0%  
 Yes 452 14.0% 14.0%  
 <NA> 3 0.1% -

## Data Checking and Cleaning (7)

1. Self-reported overall health should be a five-level factor

nh\_today <- nh\_today |>  
 mutate(sroh = fct\_recode(factor(sroh), "E" = "1", "VG" = "2",   
 "G" = "3", "F" = "4", "P" = "5"))  
nh\_today |> tabyl(sroh) |> adorn\_pct\_formatting() |>   
 gt() |> tab\_options(table.font.size = 20)

| sroh | n | percent | valid\_percent |
| --- | --- | --- | --- |
| E | 238 | 7.4% | 8.5% |
| VG | 684 | 21.2% | 24.5% |
| G | 1119 | 34.6% | 40.1% |
| F | 612 | 18.9% | 21.9% |
| P | 136 | 4.2% | 4.9% |
| NA | 445 | 13.8% | - |

## So, what is missing?

miss\_var\_summary(nh\_today)

# A tibble: 7 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <dbl>  
1 sroh 445 13.8   
2 dbp 392 12.1   
3 bmi 172 5.32   
4 asthma 3 0.0928  
5 SEQN 0 0   
6 sbreath 0 0   
7 age 0 0

miss\_case\_table(nh\_today)

# A tibble: 4 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 2555 79.0   
2 1 475 14.7   
3 2 75 2.32  
4 3 129 3.99

## Where are we?

We have 3234 rows and 7 columns in the nh\_today data now.

* 2555, or 79%, of rows are complete on these 7 variables.
* Inclusions/Exclusions: Valid (1 or 0) response to sbreath, age between 40 and 79 years, inclusive.

## Using Hmisc::describe()

nh\_today |> describe()

nh\_today   
  
 7 Variables 3234 Observations  
--------------------------------------------------------------------------------  
SEQN   
 n missing distinct Info Mean Gmd .05 .10   
 3234 0 3234 1 67069 3246 62681 63175   
 .25 .50 .75 .90 .95   
 64606 67139 69479 70953 71410   
  
lowest : 62164 62172 62177 62179 62182, highest: 71901 71902 71908 71912 71915  
--------------------------------------------------------------------------------  
sbreath   
 n missing distinct Info Sum Mean Gmd   
 3234 0 2 0.646 1015 0.3139 0.4308   
  
--------------------------------------------------------------------------------  
age   
 n missing distinct Info Mean Gmd .05 .10   
 3234 0 40 0.999 57.35 12.31 41 43   
 .25 .50 .75 .90 .95   
 48 57 65 73 76   
  
lowest : 40 41 42 43 44, highest: 75 76 77 78 79  
--------------------------------------------------------------------------------  
sroh   
 n missing distinct   
 2789 445 5   
   
Value E VG G F P  
Frequency 238 684 1119 612 136  
Proportion 0.085 0.245 0.401 0.219 0.049  
--------------------------------------------------------------------------------  
dbp   
 n missing distinct Info Mean Gmd .05 .10   
 2842 392 45 0.997 73.41 12.88 54 58   
 .25 .50 .75 .90 .95   
 66 74 80 88 92   
  
lowest : 30 32 34 36 38, highest: 110 112 114 118 120  
--------------------------------------------------------------------------------  
bmi   
 n missing distinct Info Mean Gmd .05 .10   
 3062 172 345 1 29.52 7.303 20.90 22.30   
 .25 .50 .75 .90 .95   
 24.70 28.30 32.90 38.39 41.90   
  
lowest : 13.6 14.3 15.4 16.2 16.5, highest: 65.1 66.2 67.3 68.7 82.1  
--------------------------------------------------------------------------------  
asthma   
 n missing distinct   
 3231 3 2   
   
Value No Yes  
Frequency 2779 452  
Proportion 0.86 0.14  
--------------------------------------------------------------------------------

## Revised Codebook

| Name | Description |
| --- | --- |
| SEQN | Identifying code |
| sbreath | Short of breath on stairs/inclines? (1 = Yes, 0 = No) |
| age | Age in years at screening |
| sroh | Self-reported health (E/VG/G/F/P) |
| dbp | Diastolic BP (1st reading, in mm Hg) |
| bmi | Body Mass Index (kg/) |
| asthma | Ever told you have asthma? (Yes or No) |

# Project A Tasks

## Establishing our Research Question

How effectively can we predict whether or not an adult subject has experienced “shortness of breath when hurrying on the level or walking up a slight hill” on the basis of their age, self-reported overall health, diastolic blood pressure, body mass index and whether or not they have been told they have asthma?

* Our data come from NHANES 2011-12, and describe a total of 3234 (unweighted) adult (ages 40-79) subjects.
* We will not use survey weights in this work.

## Specifying / Tidying the Outcome

Our outcome (sbreath) is the subject’s response to:

**Have you had shortness of breath either when hurrying on the level or walking up a slight hill?**

This was asked of adults ages 40 years and up as CDQ010 on the CDQ\_G questionnaire in NHANES 2011-12, and we’re studying those who responded 1 for Yes, or 0 for No.

nh\_today |> tabyl(sbreath) |> adorn\_totals() |>   
 adorn\_pct\_formatting()

sbreath n percent  
 0 2219 68.6%  
 1 1015 31.4%  
 Total 3234 100.0%

## Our Five Candidate Predictors

The five predictors we will consider for this outcome are age, sroh, dbp, bmi and asthma. Our sample has 3234 subjects.

| Name | Description | Missing? |
| --- | --- | --- |
| age | Age in years at screening | None |
| sroh | Self-reported health (E/VG/G/F/P) | 445 |
| dbp | Diastolic BP (1st reading, in mm Hg) | 392 |
| bmi | Body Mass Index (kg/) | 172 |
| asthma | Ever been told you have asthma? | 3 |

## Dealing with Missing Data

We have excluded all cases with missing sbreath so our outcome is complete.

We will assume missing at random (MAR) for the missing values in our candidate predictors, and then use single imputation to complete the rest of our work, including…

* building a Spearman plot, and
* fitting and evaluating our models Y (only linear terms) and Z (include non-linear terms)

## Using Multiple Imputation instead?

If you wanted to use multiple imputation in the project, that’s OK, but I would do that at the end, by refitting the “winning” model and summarizing those results (after imputation) only as part of your **Final Model** materials.

* In this case, given that I have complete data on 79% of my rows (so missing data in 21%) I would probably perform either 25 or 30 imputations.
* I’ll demonstrate multiple imputation in a logistic regression setting (using aregImpute()) next time (Class 10.)

## Single Imputation via simputation

* Variables I must impute: sroh, dbp, bmi, asthma
* Variables with complete data: SEQN (useless identifier), sbreath (my outcome), and age.

set.seed(43212345)  
nh\_today\_i <- nh\_today |> data.frame() |>  
 impute\_rhd(asthma ~ age) |>  
 impute\_rlm(dbp ~ age + asthma) |>  
 impute\_rlm(bmi ~ dbp + age + asthma) |>  
 impute\_cart(sroh ~ age + bmi) |>  
 as\_tibble()  
  
n\_miss(nh\_today\_i) # should now show no missing data

[1] 0

# Model Y: “Main Effects”

## Building our “Main Effects” Model

We assume MAR and analyze the (singly) imputed data nh\_today\_i

d <- datadist(nh\_today\_i)  
options(datadist = "d")  
  
modY\_si <- lrm(sbreath ~ age + sroh + dbp + bmi + asthma,  
 data = nh\_today\_i, x = TRUE, y = TRUE)

## modY\_si results (from lrm fit)

modY\_si

Logistic Regression Model  
  
lrm(formula = sbreath ~ age + sroh + dbp + bmi + asthma, data = nh\_today\_i,   
 x = TRUE, y = TRUE)  
  
 Model Likelihood Discrimination Rank Discrim.   
 Ratio Test Indexes Indexes   
Obs 3234 LR chi2 426.39 R2 0.174 C 0.718   
 0 2219 d.f. 8 R2(8,3234)0.121 Dxy 0.436   
 1 1015 Pr(> chi2) <0.0001 R2(8,2089.3)0.181 gamma 0.436   
max |deriv| 2e-10 Brier 0.187 tau-a 0.188   
  
 Coef S.E. Wald Z Pr(>|Z|)  
Intercept -4.1132 0.4878 -8.43 <0.0001   
age 0.0149 0.0039 3.79 0.0002   
sroh=VG 0.4849 0.2337 2.07 0.0380   
sroh=G 1.1641 0.2179 5.34 <0.0001   
sroh=F 1.8890 0.2269 8.33 <0.0001   
sroh=P 1.9771 0.2780 7.11 <0.0001   
dbp -0.0062 0.0038 -1.63 0.1033   
bmi 0.0544 0.0062 8.79 <0.0001   
asthma=Yes 0.7981 0.1111 7.18 <0.0001

## Two Summaries from modY\_si

Here, the Nagelkerke is 0.174 and the C statistic is 0.718.

* The Nagelkerke measures goodness of fit of our logistic regression model. It’s best interpreted as an improvement from a null (intercept only) model to our fitted one.
  + It is an adjusted Cox-Snell that ranges from 0 to 1, with higher indicating better fit.
  + A Nagelkerke doesn’t describe “a percentage of explained variation” nor is it the square of a correlation.
* The C statistic is the area under the ROC curve.

## glm version of this same fit

modY\_si\_g <- glm(sbreath ~ age + sroh + dbp + bmi + asthma,  
 data = nh\_today\_i,   
 family = binomial(link = logit))  
  
modY\_si\_g

Call: glm(formula = sbreath ~ age + sroh + dbp + bmi + asthma, family = binomial(link = logit),   
 data = nh\_today\_i)  
  
Coefficients:  
(Intercept) age srohVG srohG srohF srohP   
 -4.113185 0.014921 0.484919 1.164107 1.888952 1.977082   
 dbp bmi asthmaYes   
 -0.006171 0.054416 0.798134   
  
Degrees of Freedom: 3233 Total (i.e. Null); 3225 Residual  
Null Deviance: 4024   
Residual Deviance: 3598 AIC: 3616

## modY\_si\_g (Exponentiated) Coefficients

tidy(modY\_si\_g, exponentiate = TRUE,   
 conf.int = TRUE, conf.level = 0.90) |>  
 select(term, estimate, std.error,   
 low90 = conf.low, high90 = conf.high,   
 p = p.value) |>   
 gt() |>   
 fmt\_number(decimals = 3) |>  
 tab\_options(table.font.size = 24)

| term | estimate | std.error | low90 | high90 | p |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 0.016 | 0.488 | 0.007 | 0.036 | 0.000 |
| age | 1.015 | 0.004 | 1.008 | 1.022 | 0.000 |
| srohVG | 1.624 | 0.234 | 1.117 | 2.415 | 0.038 |
| srohG | 3.203 | 0.218 | 2.267 | 4.652 | 0.000 |
| srohF | 6.612 | 0.227 | 4.607 | 9.737 | 0.000 |
| srohP | 7.222 | 0.278 | 4.609 | 11.519 | 0.000 |
| dbp | 0.994 | 0.004 | 0.988 | 1.000 | 0.103 |
| bmi | 1.056 | 0.006 | 1.045 | 1.067 | 0.000 |
| asthmaYes | 2.221 | 0.111 | 1.850 | 2.667 | 0.000 |

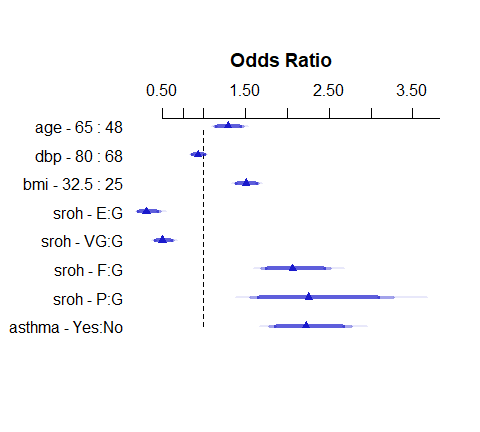
## glance() for modY\_si\_g

glance(modY\_si\_g) |>   
 gt() |> tab\_options(table.font.size = 24)

| null.deviance | df.null | logLik | AIC | BIC | deviance | df.residual | nobs |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 4024.058 | 3233 | -1798.832 | 3615.664 | 3670.398 | 3597.664 | 3225 | 3234 |

## modY\_si Effects (Odds Ratio Scale)

plot(summary(modY\_si))



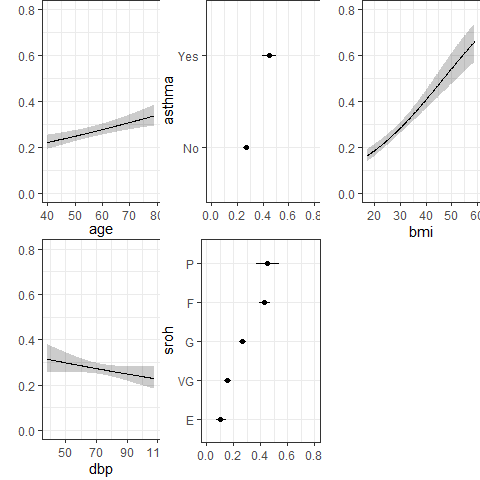
## Details of Effects Plot (modY\_si)

summary(modY\_si)

Effects Response : sbreath   
  
 Factor Low High Diff. Effect S.E. Lower 0.95 Upper 0.95  
 age 48 65.0 17.0 0.253660 0.066962 0.12242 0.384900   
 Odds Ratio 48 65.0 17.0 1.288700 NA 1.13020 1.469500   
 dbp 68 80.0 12.0 -0.074048 0.045455 -0.16314 0.015043   
 Odds Ratio 68 80.0 12.0 0.928630 NA 0.84947 1.015200   
 bmi 25 32.5 7.5 0.408120 0.046411 0.31716 0.499080   
 Odds Ratio 25 32.5 7.5 1.504000 NA 1.37320 1.647200   
 sroh - E:G 3 1.0 NA -1.164100 0.217900 -1.59120 -0.737040   
 Odds Ratio 3 1.0 NA 0.312200 NA 0.20369 0.478530   
 sroh - VG:G 3 2.0 NA -0.679190 0.116900 -0.90831 -0.450060   
 Odds Ratio 3 2.0 NA 0.507030 NA 0.40320 0.637590   
 sroh - F:G 3 4.0 NA 0.724850 0.101520 0.52588 0.923810   
 Odds Ratio 3 4.0 NA 2.064400 NA 1.69190 2.518900   
 sroh - P:G 3 5.0 NA 0.812980 0.190200 0.44019 1.185800   
 Odds Ratio 3 5.0 NA 2.254600 NA 1.55300 3.273200   
 asthma - Yes:No 1 2.0 NA 0.798130 0.111120 0.58035 1.015900   
 Odds Ratio 1 2.0 NA 2.221400 NA 1.78670 2.761900

## Prediction Plot for modY\_si

ggplot(Predict(modY\_si, fun = plogis))



## Confusion Matrix for modY\_si

How well does modY\_si classify subjects using a decision rule at 0.5?

modY\_aug <- augment(modY\_si\_g, type.predict = "response")  
  
modY\_aug <- modY\_aug |>   
 mutate(pred = ifelse(.fitted >= 0.5, "Predict SB", "Predict No SB"))  
  
modY\_aug |> tabyl(pred, sbreath) |>   
 adorn\_totals(where = c("row", "col")) |> adorn\_title()

sbreath   
 pred 0 1 Total  
 Predict No SB 2040 726 2766  
 Predict SB 179 289 468  
 Total 2219 1015 3234

* What fraction of our predictions are correct with this decision rule?

## modY\_si Classification, 1

modY\_aug |> tabyl(pred, sbreath) |>   
 adorn\_totals(where = c("row", "col")) |> adorn\_title()

sbreath   
 pred 0 1 Total  
 Predict No SB 2040 726 2766  
 Predict SB 179 289 468  
 Total 2219 1015 3234

* **Accuracy** is (2040 + 289) / 3234 = 0.720
  + 72.0% of this model’s predictions were accurate.

## modY\_si Classification, 2

modY\_aug |> tabyl(pred, sbreath) |>   
 adorn\_totals(where = c("row", "col")) |> adorn\_title()

sbreath   
 pred 0 1 Total  
 Predict No SB 2040 726 2766  
 Predict SB 179 289 468  
 Total 2219 1015 3234

* **Sensitivity** is 289 / 1015 = 0.285
  + 28.5% of those who actually were short of breath are predicted to be short of breath.
* **Specificity** is 2040 / 2219 = 0.919
  + 91.9% of those who actually weren’t short of breath were predicted not to be short of breath.

## modY\_si Classification, 3

modY\_aug |> tabyl(pred, sbreath) |>   
 adorn\_totals(where = c("row", "col")) |> adorn\_title()

sbreath   
 pred 0 1 Total  
 Predict No SB 2040 726 2766  
 Predict SB 179 289 468  
 Total 2219 1015 3234

* **Positive Predictive Value (PPV)** is 289 / 468 = 0.618
  + 61.8% of those predicted to be short of breath actually were short of breath.
* **Negative Predictive Value (NPV)** is 2040 / 2766 = 0.738
  + 73.8% of those predicted to not be short of breath actually were not short of breath.

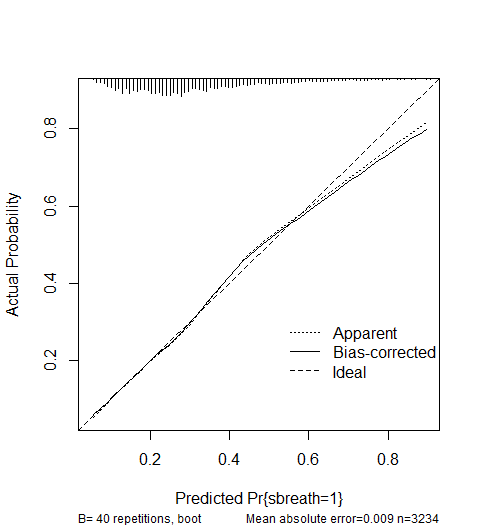
## A confusion matrix from caret

cmY <- confusionMatrix(  
 data = factor(modY\_aug$.fitted >= 0.5),  
 reference = factor(modY\_aug$sbreath == 1),  
 positive = "TRUE")  
  
cmY

Confusion Matrix and Statistics  
  
 Reference  
Prediction FALSE TRUE  
 FALSE 2040 726  
 TRUE 179 289  
   
 Accuracy : 0.7202   
 95% CI : (0.7043, 0.7356)  
 No Information Rate : 0.6861   
 P-Value [Acc > NIR] : 1.393e-05   
   
 Kappa : 0.239   
   
 Mcnemar's Test P-Value : < 2.2e-16   
   
 Sensitivity : 0.28473   
 Specificity : 0.91933   
 Pos Pred Value : 0.61752   
 Neg Pred Value : 0.73753   
 Prevalence : 0.31385   
 Detection Rate : 0.08936   
 Detection Prevalence : 0.14471   
 Balanced Accuracy : 0.60203   
   
 'Positive' Class : TRUE

## Calibration Plot for Model Y

plot(calibrate(modY\_si))



n=3234 Mean absolute error=0.009 Mean squared error=0.00021  
0.9 Quantile of absolute error=0.022

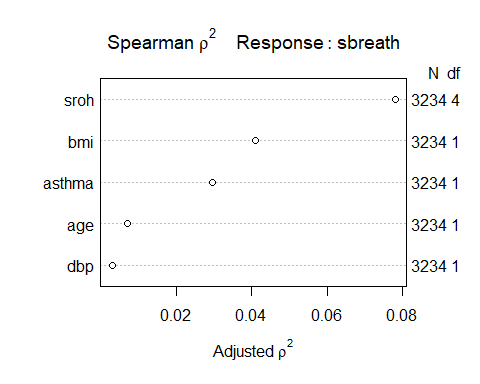
# Model Z: The “Augmented” model

## Considering Non-Linear Terms

* Use 3-6 additional degrees of freedom to account for non-linearity, and add 1-3 non-linear terms.
* We’ll start with the Spearman plot…

## Spearman plot

plot(spearman2(sbreath ~ age + sroh + dbp + bmi + asthma,  
 data = nh\_today\_i))



## Fitting an “Augmented” Model

We’ll add two terms, and stop there.

* the interaction of sroh and bmi, which will add 4 df, and
* a restricted cubic spline with three knots in bmi, which will add one more df.

In all, this should be an additional 5 df.

## note: datadist has already been set up  
  
modZ\_si <- lrm(sbreath ~ age + sroh + rcs(bmi,3) +   
 sroh %ia% bmi + dbp + asthma,  
 data = nh\_today\_i, x = TRUE, y = TRUE)

## Checking our df

As the ANOVA table below shows, we have indeed added 5 degrees of freedom with our non-linear and interaction terms.

anova(modZ\_si)

Wald Statistics Response: sbreath   
  
 Factor Chi-Square d.f. P   
 age 14.08 1 0.0002  
 sroh (Factor+Higher Order Factors) 165.26 8 <.0001  
 All Interactions 1.79 4 0.7745  
 bmi (Factor+Higher Order Factors) 79.65 6 <.0001  
 All Interactions 1.79 4 0.7745  
 Nonlinear 0.21 1 0.6461  
 sroh \* bmi (Factor+Higher Order Factors) 1.79 4 0.7745  
 dbp 2.63 1 0.1047  
 asthma 52.03 1 <.0001  
 TOTAL NONLINEAR + INTERACTION 1.89 5 0.8642  
 TOTAL 352.90 13 <.0001

## modZ\_si results (from lrm fit)

modZ\_si

Logistic Regression Model  
  
lrm(formula = sbreath ~ age + sroh + rcs(bmi, 3) + sroh %ia%   
 bmi + dbp + asthma, data = nh\_today\_i, x = TRUE, y = TRUE)  
  
 Model Likelihood Discrimination Rank Discrim.   
 Ratio Test Indexes Indexes   
Obs 3234 LR chi2 428.25 R2 0.174 C 0.719   
 0 2219 d.f. 13 R2(13,3234)0.120 Dxy 0.438   
 1 1015 Pr(> chi2) <0.0001 R2(13,2089.3)0.180 gamma 0.438   
max |deriv| 4e-09 Brier 0.187 tau-a 0.189   
  
 Coef S.E. Wald Z Pr(>|Z|)  
Intercept -4.2765 1.1558 -3.70 0.0002   
age 0.0148 0.0039 3.75 0.0002   
sroh=VG 0.4052 1.1722 0.35 0.7295   
sroh=G 1.5974 1.0686 1.49 0.1350   
sroh=F 2.1630 1.0985 1.97 0.0489   
sroh=P 2.9370 1.2318 2.38 0.0171   
bmi 0.0596 0.0387 1.54 0.1235   
bmi' 0.0095 0.0207 0.46 0.6461   
sroh=VG \* bmi 0.0023 0.0402 0.06 0.9539   
sroh=G \* bmi -0.0151 0.0368 -0.41 0.6811   
sroh=F \* bmi -0.0100 0.0376 -0.27 0.7907   
sroh=P \* bmi -0.0323 0.0414 -0.78 0.4352   
dbp -0.0061 0.0038 -1.62 0.1047   
asthma=Yes 0.8017 0.1111 7.21 <0.0001

## glm version of Model Z

modZ\_si\_g <- glm(sbreath ~ age + sroh + rcs(bmi, 3) +   
 sroh %ia% bmi + dbp + asthma,  
 data =nh\_today\_i,   
 family =binomial(link =logit))  
  
modZ\_si\_g

Call: glm(formula = sbreath ~ age + sroh + rcs(bmi, 3) + sroh %ia%   
 bmi + dbp + asthma, family = binomial(link = logit), data = nh\_today\_i)  
  
Coefficients:  
 (Intercept) age   
 -4.276453 0.014819   
 srohVG srohG   
 0.405248 1.597374   
 srohF srohP   
 2.163030 2.937031   
 rcs(bmi, 3)bmi rcs(bmi, 3)bmi'   
 0.059582 0.009525   
sroh %ia% bmisroh=VG \* bmi sroh %ia% bmisroh=G \* bmi   
 0.002326 -0.015125   
 sroh %ia% bmisroh=F \* bmi sroh %ia% bmisroh=P \* bmi   
 -0.009987 -0.032268   
 dbp asthmaYes   
 -0.006146 0.801692   
  
Degrees of Freedom: 3233 Total (i.e. Null); 3220 Residual  
Null Deviance: 4024   
Residual Deviance: 3596 AIC: 3624

## Tidied Model Z (Exponentiated)

Here are the tidied coefficients (as odds ratios, after exponentiation) from model modZ\_si\_g.

tidy(modZ\_si\_g, exponentiate = TRUE,   
 conf.int = TRUE, conf.level = 0.90) |>  
 select(term, estimate, std.error,   
 low90 = conf.low, high90 = conf.high) |>   
 gt() |>   
 fmt\_number(decimals = 3) |>   
 tab\_options(table.font.size = 14)

| term | estimate | std.error | low90 | high90 |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.014 | 1.156 | 0.002 | 0.092 |
| age | 1.015 | 0.004 | 1.008 | 1.022 |
| srohVG | 1.500 | 1.172 | 0.215 | 10.758 |
| srohG | 4.940 | 1.068 | 0.836 | 30.246 |
| srohF | 8.697 | 1.098 | 1.402 | 55.705 |
| srohP | 18.860 | 1.232 | 2.436 | 147.704 |
| rcs(bmi, 3)bmi | 1.061 | 0.039 | 0.995 | 1.132 |
| rcs(bmi, 3)bmi' | 1.010 | 0.021 | 0.976 | 1.045 |
| sroh %ia% bmisroh=VG \* bmi | 1.002 | 0.040 | 0.937 | 1.072 |
| sroh %ia% bmisroh=G \* bmi | 0.985 | 0.037 | 0.926 | 1.049 |
| sroh %ia% bmisroh=F \* bmi | 0.990 | 0.038 | 0.930 | 1.055 |
| sroh %ia% bmisroh=P \* bmi | 0.968 | 0.041 | 0.904 | 1.038 |
| dbp | 0.994 | 0.004 | 0.988 | 1.000 |
| asthmaYes | 2.229 | 0.111 | 1.857 | 2.677 |

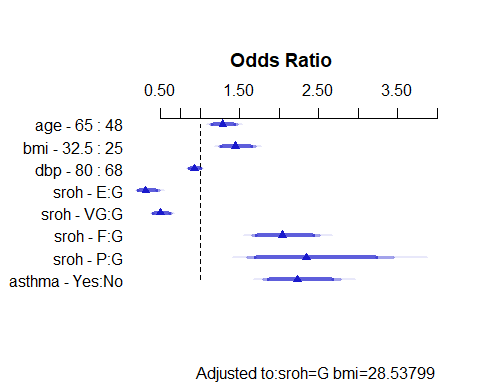
## glance() for modZ\_si\_g

glance(modZ\_si\_g) |>   
 gt() |> tab\_options(table.font.size = 24)

| null.deviance | df.null | logLik | AIC | BIC | deviance | df.residual | nobs |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 4024.058 | 3233 | -1797.904 | 3623.808 | 3708.949 | 3595.808 | 3220 | 3234 |

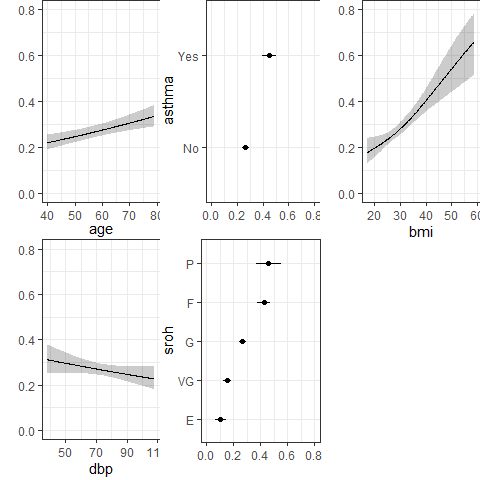
## Model Z Effects (Odds Ratios)

plot(summary(modZ\_si))



## Prediction Plot for Model Z

ggplot(Predict(modZ\_si, fun = plogis))



## Confusion Matrix for Model Z

How well does our Model Z classify subjects using a decision rule at 0.5?

modZ\_aug <- augment(modZ\_si\_g, type.predict = "response")  
  
modZ\_aug <- modZ\_aug |>   
 mutate(pred = ifelse(.fitted >= 0.5,   
 "Predict SB", "Predict No SB"))  
  
modZ\_aug |> tabyl(pred, sbreath) |>   
 adorn\_totals(where = c("row", "col")) |> adorn\_title()

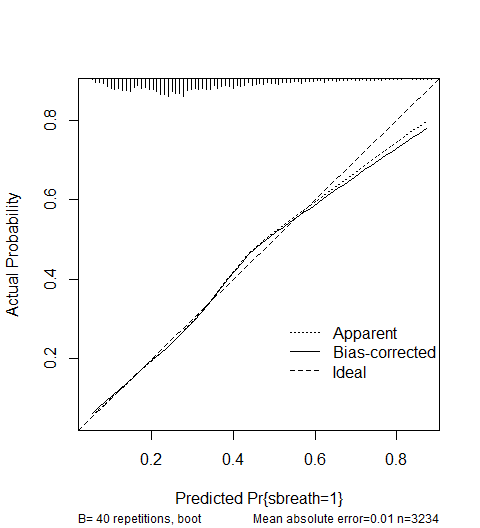
sbreath   
 pred 0 1 Total  
 Predict No SB 2040 725 2765  
 Predict SB 179 290 469  
 Total 2219 1015 3234

## Classification Accuracy

* **Sensitivity** is only slightly changed, to 290 / 1015 = 0.286
  + 28.6% of those who actually were short of breath are predicted to be short of breath.
* **Specificity** is still 2040 / 2219 = 0.919
  + 91.9% of those who actually weren’t short of breath were predicted not to be short of breath.
* **Positive Predictive Value (PPV)** is 290 / 469 = 0.618
  + Again, 61.8% of those predicted to be short of breath actually were short of breath.

## Calibration Plot for Model Z

plot(calibrate(modZ\_si))



n=3234 Mean absolute error=0.01 Mean squared error=0.00024  
0.9 Quantile of absolute error=0.023

# Selecting and Presenting Your Final Model

## Model Y and Z Summaries

temp1 <- bind\_rows(glance(modY\_si\_g), glance(modZ\_si\_g)) |>  
 mutate(model = c("Y", "Z")) |>  
 select(model, AIC, BIC)   
  
temp2 <- tibble(model = c("Y", "Z"),  
 auc = c(modY\_si$stats["C"], modZ\_si$stats["C"]),  
 r2\_nag = c(modY\_si$stats["R2"], modZ\_si$stats["R2"]))  
  
left\_join(temp1, temp2, by = "model") |>   
 gt() |> fmt\_number(columns = AIC:BIC, decimals = 1) |>  
 fmt\_number(columns = auc:r2\_nag, decimals = 4) |>   
 tab\_options(table.font.size = 24)

| model | AIC | BIC | auc | r2\_nag |
| --- | --- | --- | --- | --- |
| Y | 3,615.7 | 3,670.4 | 0.7181 | 0.1735 |
| Z | 3,623.8 | 3,708.9 | 0.7188 | 0.1742 |

## ANOVA comparing Model Y to Z

anova(modZ\_si)

Wald Statistics Response: sbreath   
  
 Factor Chi-Square d.f. P   
 age 14.08 1 0.0002  
 sroh (Factor+Higher Order Factors) 165.26 8 <.0001  
 All Interactions 1.79 4 0.7745  
 bmi (Factor+Higher Order Factors) 79.65 6 <.0001  
 All Interactions 1.79 4 0.7745  
 Nonlinear 0.21 1 0.6461  
 sroh \* bmi (Factor+Higher Order Factors) 1.79 4 0.7745  
 dbp 2.63 1 0.1047  
 asthma 52.03 1 <.0001  
 TOTAL NONLINEAR + INTERACTION 1.89 5 0.8642  
 TOTAL 352.90 13 <.0001

## Validating Model Summaries

set.seed(432123)  
valY <- validate(modY\_si, B = 40)  
valZ <- validate(modZ\_si, B = 40)  
  
val\_1 <- bind\_rows(valY[1,], valZ[1,]) |>  
 mutate(model = c("Y", "Z"),  
 AUC\_nominal = 0.5 + (index.orig/2),   
 AUC\_validated = 0.5 + (index.corrected/2)) |>  
 select(model, AUC\_nominal, AUC\_validated)  
  
val\_2 <- bind\_rows(valY[2,], valZ[2,]) |>  
 mutate(model = c("Y", "Z"),  
 R2\_nominal = index.orig,  
 R2\_validated = index.corrected) |>  
 select(model, R2\_nominal, R2\_validated)  
  
val <- left\_join(val\_1, val\_2, by = "model")   
  
val |> gt() |> fmt\_number(decimals = 4) |>   
 tab\_options(table.font.size = 24)

| model | AUC\_nominal | AUC\_validated | R2\_nominal | R2\_validated |
| --- | --- | --- | --- | --- |
| Y | 0.7181 | 0.7163 | 0.1735 | 0.1693 |
| Z | 0.7188 | 0.7132 | 0.1742 | 0.1632 |

## Describing a Meaningful Effect

This is for you to do.

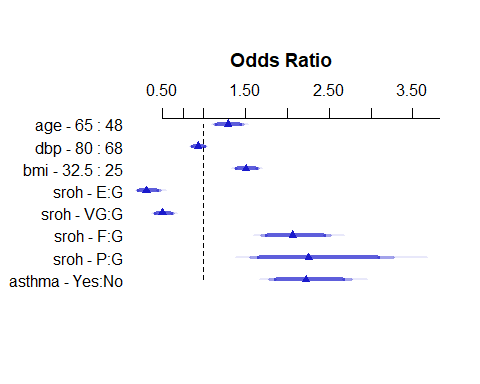
[W]rite a detailed and correct description of the effect of at least one predictor on your outcome for your chosen logistic regression model, providing all necessary elements of such a description, and link this directly to what the (effects) plot is telling you.

See Chapter 22 of the Notes for more details, and this is also the major task in several Lab 4 questions.

* The effects plot for Model Y is repeated in the next slide, and you’ll want the actual summary as well as the plot so you can specify the numbers.
* We prefer you discuss a meaningful effect, should one exist. Pick an effect to describe that is interesting to you.

## Model Y Effects Plot

plot(summary(modY\_si))



## Model Y Effects Summary

summary(modY\_si)

Effects Response : sbreath   
  
 Factor Low High Diff. Effect S.E. Lower 0.95 Upper 0.95  
 age 48 65.0 17.0 0.253660 0.066962 0.12242 0.384900   
 Odds Ratio 48 65.0 17.0 1.288700 NA 1.13020 1.469500   
 dbp 68 80.0 12.0 -0.074048 0.045455 -0.16314 0.015043   
 Odds Ratio 68 80.0 12.0 0.928630 NA 0.84947 1.015200   
 bmi 25 32.5 7.5 0.408120 0.046411 0.31716 0.499080   
 Odds Ratio 25 32.5 7.5 1.504000 NA 1.37320 1.647200   
 sroh - E:G 3 1.0 NA -1.164100 0.217900 -1.59120 -0.737040   
 Odds Ratio 3 1.0 NA 0.312200 NA 0.20369 0.478530   
 sroh - VG:G 3 2.0 NA -0.679190 0.116900 -0.90831 -0.450060   
 Odds Ratio 3 2.0 NA 0.507030 NA 0.40320 0.637590   
 sroh - F:G 3 4.0 NA 0.724850 0.101520 0.52588 0.923810   
 Odds Ratio 3 4.0 NA 2.064400 NA 1.69190 2.518900   
 sroh - P:G 3 5.0 NA 0.812980 0.190200 0.44019 1.185800   
 Odds Ratio 3 5.0 NA 2.254600 NA 1.55300 3.273200   
 asthma - Yes:No 1 2.0 NA 0.798130 0.111120 0.58035 1.015900   
 Odds Ratio 1 2.0 NA 2.221400 NA 1.78670 2.761900

## ROC Calculations for Model Y

roc\_modY <- roc(nh\_today\_i$sbreath ~   
 predict(modY\_si\_g, type="response"), ci = TRUE)

Setting levels: control = 0, case = 1

Setting direction: controls < cases

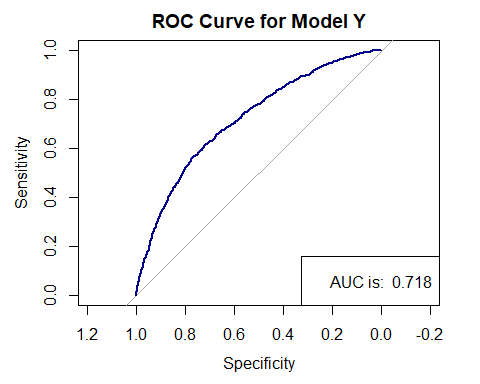
roc\_modY

Call:  
roc.formula(formula = nh\_today\_i$sbreath ~ predict(modY\_si\_g, type = "response"), ci = TRUE)  
  
Data: predict(modY\_si\_g, type = "response") in 2219 controls (nh\_today\_i$sbreath 0) < 1015 cases (nh\_today\_i$sbreath 1).  
Area under the curve: 0.7181  
95% CI: 0.6992-0.7371 (DeLong)

plot(roc\_modY, main = "ROC Curve for Model Y",  
 lwd = 2, col = "salmon")  
legend('bottomright',   
 legend = paste("AUC is: ",round\_half\_up(auc(roc\_modY),3)))

## ROC plot for Model Y

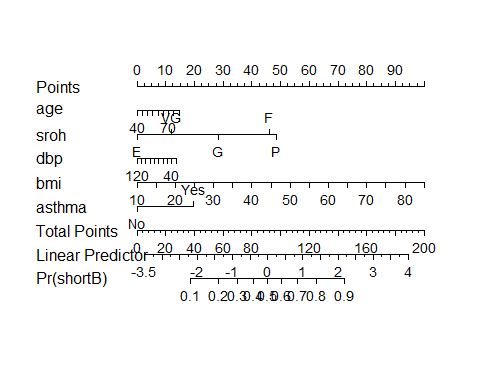
plot(roc\_modY, main = "ROC Curve for Model Y",  
 lwd = 2, col = "blue4")  
legend('bottomright',   
 legend = paste("AUC is: ",round\_half\_up(auc(roc\_modY),3)))



## Nomogram for Model Y

* The final part of your summary of the final model should be a nomogram **with a demonstration of a predicted probability associated with two new subjects of interest** that differ in terms of some of the parameters in your model.
* Your predictions should describe two different subjects. You don’t have to call them Harry and Sally, but it is helpful to give them actual names.

plot(nomogram(modY\_si, fun = plogis,  
 fun.at=c(seq(0.1, 0.9, by = 0.1)),  
 funlabel = "Pr(shortB)"))



## Next Time

More on Logistic Regression

* See Chapter 20 for more on confusion matrices and ROC curves and some material on assessing assumptions through residual plots, all of which are in the context of logistic models fit with glm().
* See Chapter 21 for more on using Spearman’s plot, Nagelkerke , the C statistic, its relationship to Somers’ d, validation and plotting the results, along with some thoughts on identifying influential points, mostly in the context of logistic models fit with lrm().
* See Chapter 22 for some thoughts on estimating and interpreting effect sizes in logistic and in linear models. Some really useful tips to be found here.