432 Class 06 Slides

thomaselove.github.io/432

2022-01-27

Moving Forward

- Logistic Regression Models and the `smart3_sh` data

Setup

```
library(here); library(magrittr)
library(janitor); library(knitr)
library(patchwork); library(broom)
library(equatiomatic); library(simputation)
library(naniar)
library(rsample); library(yardstick)
library(tidyverse)

theme_set(theme_bw())
```

smart3 Variables, by Type

Variable	Type	Description
landline	Binary (1/0)	survey conducted by landline? (vs. cell)
healthplan	Binary (1/0)	subject has health insurance?
age_imp	Quantitative	age (imputed from groups - see Notes)
fruit_day	Quantitative	mean servings of fruit / day
drinks_wk	Quantitative	mean alcoholic drinks / week
bmi	Quantitative	body-mass index (in kg/m^2)
physhealth	Count (0-30)	of last 30 days, # in poor physical health
dm_status	Categorical	diabetes status (4 levels, *we'll collapse to 2*)
activity	Categorical	physical activity level (4 levels, *we'll re-level*)
smoker	Categorical	smoking status (4 levels, *we'll collapse to 3*)
genhealth	Categorical	self-reported overall health (5 levels)

The smart3 data (built last time)

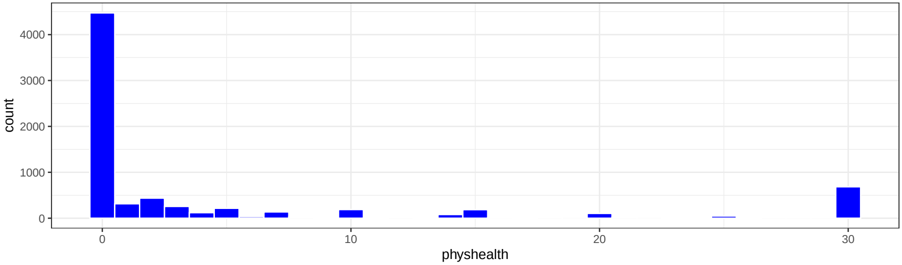
```
smart3_sh <- readRDS(here("data", "smart3_sh.Rds"))

str(smart3_sh)
```

```
tibble [7,412 x 28] (S3: tbl_df/tbl/data.frame)
 $ SEQNO      : chr [1:7412] "2017000001" "2017000002" "2017
 $ mmsa       : chr [1:7412] "Cincinnati" "Cincinnati" "Cinc
 $ mmsa_wt    : num [1:7412] 670 407 356 203 194 ...
 $ landline   : int [1:7412] 1 1 1 1 1 1 1 1 1 1 ...
 $ age_imp    : num [1:7412] 36 41 55 61 57 24 65 53 51 42 
 $ healthplan : chr [1:7412] "1" "1" "1" "1" ...
 $ dm_status  : Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 
 $ fruit_day  : num [1:7412] 1.43 1 3 0.5 0.72 ...
 $ drinks_wk  : num [1:7412] 4.67 0 0 0 0.23 1.87 0 0 0.23 
 $ activity   : Factor w/ 4 levels "Highly_Active",..: 2 2 
 $ smoker     : Factor w/ 3 levels "Current","Former",..: 3 
 $ physhealth : int [1:7412] 0 0 2 0 2 0 0 30 2 30 ...
```

Days (in last 30) of poor physical health

```
ggplot(smart3_sh, aes(x = physhealth)) +
    geom_histogram(binwidth = 1,
                   fill = "blue", col = "white")
```




```
smart3_sh %$% tabyl(physhealth > 0)
```

```
 physhealth > 0    n   percent
          FALSE 4472 0.6033459
           TRUE 2940 0.3966541
```

Create day6 data: predicting Pr(physhealth > 0)?

```
day6 <- smart3_sh %>%
    mutate(sick = as.numeric(physhealth > 0),
           id = as.character(
               as.numeric(SEQNO)-2017000000)) %>%
    select(id, sick, age = age_imp, dm_status, smoker,
           bmi, physhealth)

slice(day6, 17:19) # show rows 17-19
```

```
# A tibble: 3 x 7
  id     sick   age dm_status smoker    bmi physhealth
  <chr> <dbl> <dbl> <fct>     <fct>   <dbl>      <int>
1 17        0    72 No        Former   31.4          0
2 18        1    82 No        Never    27.6          5
3 19        0    62 Yes       Current  27.5          0
```

Before fitting models, let's split our sample

```
set.seed(4322022)

day6_split <- initial_split(day6, prop = 0.7,
                            strata = smoker)

d6_train <- training(day6_split)
d6_test <- testing(day6_split)
```

What does `strata = smoker` do?

Impact of `strata = smoker` in split

```
d6_train %>% tabyl(smoker)
```

```
  smoker    n   percent
 Current  933 0.1798728
  Former 1443 0.2781955
   Never 2811 0.5419318
```

```
d6_test %>% tabyl(smoker)
```

```
  smoker    n   percent
 Current  400 0.1797753
  Former  619 0.2782022
   Never 1206 0.5420225
```

The logistic regression model

$$logit(event) = log\left(\frac{Pr(event)}{1 - Pr(event)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

$$odds(event) = \frac{Pr(event)}{1 - Pr(event)}$$

$$Pr(event) = \frac{odds(event)}{odds(event) + 1}$$

$$Pr(event) = \frac{exp(logit(event))}{1 + exp(logit(event))}$$

Model 1: predict `sick` from `smoker` and `age`

Fit the model with and without an interaction term?

```
mod1 <- glm(sick ~ age + smoker,
                family = binomial(link = "logit"),
                data = d6_train)

mod2 <- glm(sick ~ age * smoker,
               family = binomial(link = "logit"),
               data = d6_train)
```

1. Can we use the models to make predictions?
2. How should we interpret the model coefficients?
3. Can we compare the models based on in-sample performance?
4. How can we assess predictions using our test sample?

Model 1

```
extract_eq(mod1, wrap = TRUE, terms_per_line = 2,
           operator_location = "start", use_coefs = TRUE,
           coef_digits = 3)
```

$$\log\left[\frac{P(\widehat{\text{sick} = 1})}{1 - P(\widehat{\text{sick} = 1})}\right] = -0.322 + 0.005(\text{age}) \\ - 0.318(\text{smoker}_{\text{Former}}) - 0.51(\text{smoker}_{\text{Never}}) \tag{1}$$

Likelihood Ratio Tests: Model 1

```
anova(mod1, test = "LRT")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: sick

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    5186     6964.6
age     1    7.422      5185     6957.2  0.006442 **
smoker  2   45.045      5183     6912.1 1.654e-10 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 1

```
tidy(mod1, conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, std.error, conf.low, conf.high, p.v
    kable(dig = 3)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	-0.322	0.105	-0.494	-0.150	0.002
age	0.005	0.002	0.002	0.007	0.003
smokerFormer	-0.318	0.086	-0.460	-0.176	0.000
smokerNever	-0.510	0.077	-0.636	-0.384	0.000

Model 1 Predictions for subjects A-F

- logit(sick) = -0.322 + 0.005 age - 0.318 Former - 0.510 Never

ID	age	smoker	logit(sick)	odds(sick)	Pr(sick)
A	33	Current	-0.157	0.8547	0.461
B	33	Former	-0.475	0.6219	0.383
C	33	Never	-0.667	0.5132	0.339
D	55	Current	-0.047	0.9541	0.488
E	55	Former	-0.365	0.6942	0.410
F	55	Never	-0.557	0.5729	0.364

Sample Calculation (for E):

- logit(sick) = -0.322 + 0.005 (55) - 0.318 (1) - 0.510 (0) = -0.365
- odds(sick) = exp(-0.365) = 0.6942
- Prob(sick) = 0.6942 / (1 + 0.6942) = 0.410

Model 1 (coefficients exponentiated)

```
tidy(mod1, exponentiate = TRUE, conf.int = TRUE,
     conf.level = 0.90) %>%
    select(term, estimate,
           lo90 = conf.low, hi90 = conf.high) %>%
    kable(dig = 3)
```

term	estimate	lo90	hi90
(Intercept)	0.725	0.610	0.861
age	1.005	1.002	1.007
smokerFormer	0.728	0.631	0.839
smokerNever	0.601	0.529	0.681

- So what can we conclude about, for instance, the effect of Never smoking (as compared to Current smoking)?

Model 1 (coefficients exponentiated)

```
tidy(mod1, exponentiate = TRUE, conf.int = TRUE,
     conf.level = 0.90) %>%
    select(term, estimate,
           lo90 = conf.low, hi90 = conf.high) %>%
    kable(dig = 3)
```

term	estimate	lo90	hi90
(Intercept)	0.725	0.610	0.861
age	1.005	1.002	1.007
smokerFormer	0.728	0.631	0.839
smokerNever	0.601	0.529	0.681

- So what can we conclude about, for instance, the effect of Never smoking (as compared to Current smoking)?
- Suppose Chloe and Nancy are the same age, where Nancy never smoked and Chloe is a current smoker.

Model 1 (Chloe and Nancy)

term	estimate	lo90	hi90
(Intercept)	0.725	0.610	0.861
age	1.005	1.002	1.007
smokerFormer	0.728	0.631	0.839
smokerNever	0.601	0.529	0.681

- Chloe and Nancy are the same age; Nancy never smoked and Chloe smokes currently. What can we conclude about the relative odds for Nancy of a sick day as compared to Chloe?

Model 1 (Chloe and Nancy)

term	estimate	lo90	hi90
(Intercept)	0.725	0.610	0.861
age	1.005	1.002	1.007
smokerFormer	0.728	0.631	0.839
smokerNever	0.601	0.529	0.681

- Chloe and Nancy are the same age; Nancy never smoked and Chloe smokes currently. What can we conclude about the relative odds for Nancy of a sick day as compared to Chloe?
- Nancy's odds of at least one sick day in the past 30 are 60.1% of Chloe's odds.

Model 1 (Chloe and Nancy)

term	estimate	lo90	hi90
(Intercept)	0.725	0.610	0.861
age	1.005	1.002	1.007
smokerFormer	0.728	0.631	0.839
smokerNever	0.601	0.529	0.681

- Chloe and Nancy are the same age; Nancy never smoked and Chloe smokes currently. What can we conclude about the relative odds for Nancy of a sick day as compared to Chloe?
- Nancy's odds of at least one sick day in the past 30 are 60.1% of Chloe's odds.
- 90% CI for this odds ratio is (0.529, 0.681).

Model 1 (Chloe and Nancy)

term	estimate	lo90	hi90
(Intercept)	0.725	0.610	0.861
age	1.005	1.002	1.007
smokerFormer	0.728	0.631	0.839
smokerNever	0.601	0.529	0.681

- Chloe and Nancy are the same age; Nancy never smoked and Chloe smokes currently. What can we conclude about the relative odds for Nancy of a sick day as compared to Chloe?
- Nancy's odds of at least one sick day in the past 30 are 60.1% of Chloe's odds.
- 90% CI for this odds ratio is (0.529, 0.681).
- Chloe's odds of a sick day are $(1/0.601 = 1.664)$ times those of Nancy.

Does this match the predictions we made?

- Suppose both Chloe and Nancy are 33 years old.
- We saw that Nancy's odds(sick) should be 0.601 times Chloe's odds(sick).

ID	age	smoker	logit(sick)	odds(sick)	Pr(sick)
Chloe	33	Current	-0.157	0.8547	0.461
Nancy	33	Never	-0.667	0.5132	0.339

- and we have 0.5132 / 0.8547 = 0.600 for the ratio of Nancy's odds to Chloe's odds
- and Chloe's odds are 0.8547 / 0.5132 = 1.665 times those of Nancy.
- These discrepancies are just due to rounding error in my table.

Model 1 results from `glance`

- We'll have some additional measures of fit quality, in time.
- Deviance = -2(log Likelihood)

```
glance(mod1) %>%
    select(nobs, df.null, null.deviance,
           deviance, df.residual) %>%
    kable(dig = 1)
```

nobs	df.null	null.deviance	deviance	df.residual
5187	5186	6964.6	6912.1	5183

```
glance(mod1) %>%
    select(nobs, logLik, AIC, BIC) %>%
    kable(dig = 1)
```

nobs	logLik	AIC	BIC
5187	3456.1	6920.1	6946.4

Model 2

```
extract_eq(mod2, wrap = TRUE, terms_per_line = 1,
           operator_location = "start", use_coefs = TRUE,
           coef_digits = 3)
```

$$
\begin{aligned}
\log\left[\frac { \widehat{P(\operatorname{sick} = 1)} }{ 1 - \widehat{P(\operatorname{sick} = 1)} } \right] &= -0.188 \\
&\quad + 0.002(\operatorname{age}) \\
&\quad - 0.563(\operatorname{smoker}_{\operatorname{Former}}) \\
&\quad - 0.642(\operatorname{smoker}_{\operatorname{Never}}) \\
&\quad + 0.004(\operatorname{age} \times \operatorname{smoker}_{\operatorname{Former}}) \\
&\quad + 0.003(\operatorname{age} \times \operatorname{smoker}_{\operatorname{Never}})
\end{aligned} \tag{2}
$$

Likelihood Ratio Tests: Model 2

```
anova(mod2, test = "LRT")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: sick

Terms added sequentially (first to last)

           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        5186     6964.6
age         1    7.422      5185     6957.2  0.006442 **
smoker      2   45.045      5183     6912.1 1.654e-10 ***
age:smoker  2    0.733      5181     6911.4  0.693211
---
Signif. codes:
```

Model 2

```
tidy(mod2, conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, std.error, conf.low, conf.high, p.v
    kable(dig = 3)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	-0.188	0.214	-0.542	0.164	0.380
age	0.002	0.004	-0.005	0.009	0.608
smokerFormer	-0.563	0.300	-1.057	-0.070	0.060
smokerNever	-0.642	0.245	-1.046	-0.238	0.009
age:smokerFormer	0.004	0.005	-0.004	0.013	0.392
age:smokerNever	0.003	0.004	-0.005	0.010	0.564

- logit(sick) = `-0.188 + 0.002 age - 0.563 Former - 0.642 Never + 0.004 age*Former + 0.003 age*Never`

Model 2 predictions for subjects A-F

- logit(sick) = `-0.188 + 0.002 age - 0.563 Former - 0.642 Never + 0.004 age*Former + 0.003 age*Never`

ID	age	smoker	logit(sick)	odds(sick)	Pr(sick)
A	33	Current	-0.122	0.8851	0.470
B	33	Former	-0.553	0.5752	0.365
C	33	Never	-0.665	0.5143	0.340
D	55	Current	-0.078	0.9250	0.481
E	55	Former	-0.421	0.6564	0.396
F	55	Never	-0.555	0.5741	0.365

- Subject E: logit(sick) = `-0.188 + 0.002 (55) - 0.563 (1) - 0.642 (0) + 0.004(55)(1) + 0.003(55)(0) = -0.421`
- odds(sick) = exp(-0.421) = 0.6564 so
- Prob(sick) = 0.6564 / (1 + 0.6564) = 0.396 for subject E.

Model 2 (coefficients exponentiated)

```
tidy(mod2, exponentiate = TRUE, conf.int = TRUE,
     conf.level = 0.90) %>%
    select(term, estimate,
           lo90 = conf.low, hi90 = conf.high) %>%
    kable(dig = 3)
```

term	estimate	lo90	hi90
(Intercept)	0.828	0.582	1.178
age	1.002	0.995	1.009
smokerFormer	0.570	0.348	0.932
smokerNever	0.526	0.351	0.788
age:smokerFormer	1.004	0.996	1.013
age:smokerNever	1.003	0.995	1.010

Model 2 (Chloe and Nancy)

term	estimate	lo90	hi90
(Intercept)	0.828	0.582	1.178
age	1.002	0.995	1.009
smokerFormer	0.570	0.348	0.932
smokerNever	0.526	0.351	0.788
age:smokerFormer	1.004	0.996	1.013
age:smokerNever	1.003	0.995	1.010

- Chloe and Nancy are the same age; Nancy never smoked and Chloe smokes currently. What can we conclude about the relative odds for Nancy of a sick day as compared to Chloe?

Model 2 (Chloe and Nancy)

term	estimate	lo90	hi90
(Intercept)	0.828	0.582	1.178
age	1.002	0.995	1.009
smokerFormer	0.570	0.348	0.932
smokerNever	0.526	0.351	0.788
age:smokerFormer	1.004	0.996	1.013
age:smokerNever	1.003	0.995	1.010

- Chloe and Nancy are the same age; Nancy never smoked and Chloe smokes currently. What can we conclude about the relative odds for Nancy of a sick day as compared to Chloe?
- We cannot conclude anything **unless** we know what age Chloe and Nancy are, since the effect of smoking depends on age.

Model 2 (Chloe and Nancy)

If Chloe (current smoker) and Nancy (never smoker) are each 33, then ...

ID	age	smoker	logit(sick)	odds(sick)	Pr(sick)
Chloe	33	Current	-0.122	0.8851	0.470
Nancy	33	Never	-0.665	0.5143	0.340

- Chloe's odds of being sick are $0.8851/0.5143 = 1.72$ times that of Nancy, **if** they are each 33 years old.
- If Chloe and Nancy are each 55, then from the table below, Chloe's odds are 0.9250 / 0.5741 = 1.61 times Nancy's odds of being sick.

ID	age	smoker	logit(sick)	odds(sick)	Pr(sick)
D	55	Current	-0.078	0.9250	0.481
F	55	Never	-0.555	0.5741	0.365

Comparing Model 1 to Model 2 with AIC and BIC

```
bind_rows(glance(mod1) %>% select(nobs, AIC, BIC),
          glance(mod2) %>% select(nobs, AIC, BIC)) %>%
    mutate(mod = c("m1 (no int.)", "m2 (interaction)")) %>%
    kable(digits = 1)
```

nobs	AIC	BIC	mod
5187	6920.1	6946.4	m1 (no int.)
5187	6923.4	6962.7	m2 (interaction)

- Which model looks like it performs better in the training sample?

Comparison with Mallows' C_p statistic?

```
anova(mod1, mod2, test = "Cp")
```

```
Analysis of Deviance Table

Model 1: sick ~ age + smoker
Model 2: sick ~ age * smoker
  Resid. Df Resid. Dev Df Deviance     Cp
1      5183     6912.1              6920.1
2      5181     6911.4  2  0.73284 6923.4
```

- Same as what we got from glance for AIC in this case.

Can we compare the models with a Test?

```
anova(mod1, mod2, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: sick ~ age + smoker
Model 2: sick ~ age * smoker
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      5183     6912.1
2      5181     6911.4  2  0.73284   0.6932
```

Could also consider:

- Rao's efficient score test (`test = "Rao"`)
- Pearson's chi-square test (`test = "Chisq"`)

Let's get predicted probabilities in training sample

```
m1_aug <- augment(mod1, type.predict = "response")
m2_aug <- augment(mod2, type.predict = "response")
```

The predicted probabilities are in the `.fitted` column.

```
m1_aug %>% select(age, smoker, sick, .fitted) %>% slice(1)
```

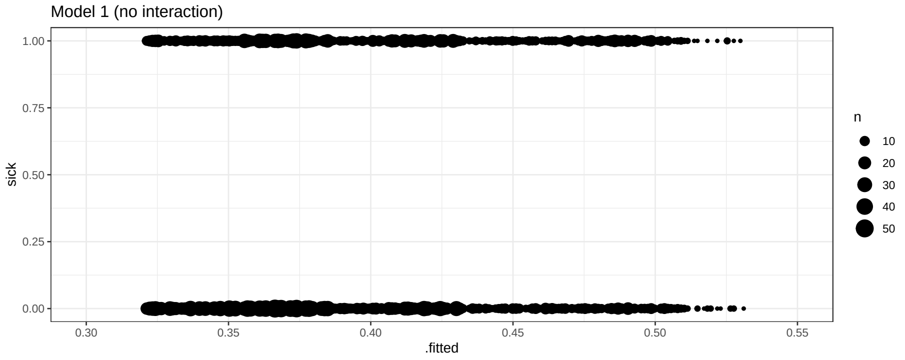
```
# A tibble: 1 x 4
    age smoker   sick .fitted
  <dbl> <fct>   <dbl>   <dbl>
1    57 Current     1   0.486
```

```
m2_aug %>% select(age, smoker, sick, .fitted) %>% slice(1)
```

```
# A tibble: 1 x 4
    age smoker   sick .fitted
  <dbl> <fct>   <dbl>   <dbl>
1    57 Current     1   0.482
```

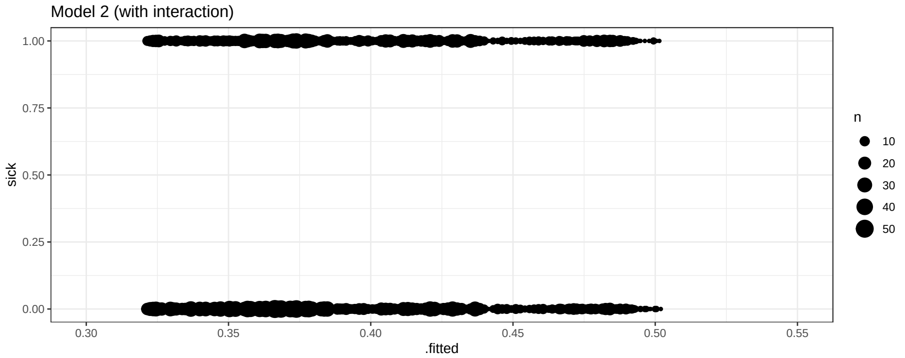
Observed (sick status) vs. Model 1 fitted Pr(sick)

```
ggplot(m1_aug, aes(x = .fitted, y = sick)) +
    geom_count() + xlim(0.30, 0.55) +
    labs(title = "Model 1 (no interaction)",
         sub = "Training Data")
```



Observed (sick status) vs. Model 2 fitted Pr(sick)

```r
ggplot(m2_aug, aes(x = .fitted, y = sick)) +
    geom_count() + xlim(0.30, 0.55) +
    labs(title = "Model 2 (with interaction)",
         sub = "Training Data")
```



Making Classification Decisions

- Our outcome is `sick`, where `sick` $= 1$ if `physact` > 0, otherwise `sick` $= 0$.
- We can establish a classification rule based on our model's predicted probabilities of `sick` $= 1$.
- 0.5 is a natural (but not inevitable) cut point.
  - if `.fitted` is below 0.50, we'll predict `sick` $= 0$
  - if `.fitted` is 0.50 or larger, we'll predict `sick` $= 1$.

```
m1_aug %$% table(.fitted >= 0.50, sick)
```

```
       sick
           0    1
  FALSE 3058 2005
  TRUE    75   49
```

Standard Epidemiological Format

Confusion matrix for Model `mod1` in the training sample.

```
confuse_m1 <- m1_aug %>%
    mutate(sick_obs = factor(sick == "1"),
           sick_pred = factor(.fitted >= 0.50),
           sick_obs = fct_relevel(sick_obs, "TRUE"),
           sick_pred = fct_relevel(sick_pred, "TRUE")) %$%
    table(sick_pred, sick_obs)

confuse_m1
```

```
         sick_obs
sick_pred TRUE FALSE
    TRUE    49    75
    FALSE 2005  3058
```

Terminology associated with the Confusion Matrix

```
confuse_m1
```

```
         sick_obs
sick_pred TRUE FALSE
    TRUE    49    75
    FALSE 2005  3058
```

- Total Observations = 49 + 75 + 2005 + 3058 = 5187
- Correct Predictions = 49 + 3058 = 3107, or 59.9% accuracy
- Incorrect Predictions = 75 + 2005 = 2080 (40.1%)
- Observed TRUE = 49 + 2005 = 2054, or 39.6% prevalence
- Predicted TRUE = 49 + 75 = 124, or 2.4% detection prevalence

Other Summaries from a Confusion Matrix

```
confuse_m1
```

```
         sick_obs
sick_pred TRUE FALSE
    TRUE    49    75
    FALSE 2005  3058
```

- Sensitivity = 49 / (49 + 2005) = 2.4% (also called Recall)
  - if the subject actually was sick, our model predicts that 2.4% of the time
- Specificity = 3058 / (3058 + 75) = 97.6%
  - if the subject was actually not sick, our model predicts that 97.6% of the time
- Positive Predictive Value (or Precision) = 49 / (49 + 75) = 39.5%
  - our predictions of sick were correct 39.5% of the time
- Negative Predictive Value = 3058 / (3058 + 2005) = 60.4%
  - our predictions of "not sick" were correct 60.4% of the time

Confusion Matrix for `mod2` (training sample)

We can obtain a similar confusion matrix for model `mod2` using the same (arbitrary) cutoff of `.fitted >= 0.5` to indicate `sick`.

```
confuse_m1
```

```
         sick_obs
sick_pred TRUE FALSE
    TRUE    49    75
    FALSE 2005  3058
```

```
confuse_m2
```

```
         sick_obs
sick_pred TRUE FALSE
    TRUE     2     3
    FALSE 2052  3130
```

Which of these confusion matrices looks better?

Get confusion matrix more easily?

Switch to a 0.45 cutoff. . .

```
m1_aug <- m1_aug %>%
    mutate(obs = factor(sick),
           pred = factor(ifelse(.fitted >= 0.45, 1, 0)))

conf_mat(data = m1_aug, truth = obs, estimate = pred)
```

```
          Truth
Prediction    0    1
         0 2679 1642
         1  454  412
```

Accuracy and Kappa Results for `mod_1`

```
metrics(data = m1_aug, truth = obs, estimate = pred) %>%
    kable(digits = 6)
```

.metric	.estimator	.estimate
accuracy	binary	0.595913
kap	binary	0.061834

- Kappa = a correlation statistic from -1 to +1, with complete agreement +1 and complete disagreement -1.
- Kappa measures the inter-rater reliability of our predicted and true classifications.

Confusion Matrix for mod_2 with 0.45 cutoff

```
m2_aug <- m2_aug %>%
    mutate(obs = factor(sick),
           pred = factor(ifelse(.fitted >= 0.45, 1, 0)))

conf_mat(data = m2_aug, truth = obs, estimate = pred)
```

```
          Truth
Prediction    0    1
         0 2582 1561
         1  551  493
```

- 493 + 2582 = 3075 accurate predictions (59.3% accuracy)
- Sensitivity = 493 / (493 + 1561) = 24.0%
  - for the people who were actually sick, we made correct predictions 24% of the time
- Specificity = 2582 / (2582 + 551) = 82.4%
  - for the people who weren't actually sick, we made correct predictions 82.4% of the time with this decision rule and model `mod_2`.

Holdout Sample?

```
mod1_aug_test <- augment(mod1, newdata = d6_test,
                         type.predict = "response") %>%
    mutate(obs = factor(sick),
           pred = factor(ifelse(.fitted >= 0.45, 1, 0)))

mod2_aug_test <- augment(mod2, newdata = d6_test,
                         type.predict = "response") %>%
    mutate(obs = factor(sick),
           pred = factor(ifelse(.fitted >= 0.45, 1, 0)))
```

metrics for test sample, models 1 and 2

```
bind_cols(
    metrics(data = mod1_aug_test,
            truth = obs, estimate = pred) %>%
        select(.metric, mod1 = .estimate),
    metrics(data = mod2_aug_test,
            truth = obs, estimate = pred) %>%
        select(mod2 = .estimate)
)
```

```
# A tibble: 2 x 3
  .metric    mod1   mod2
  <chr>     <dbl>  <dbl>
1 accuracy 0.609  0.604
2 kap      0.0938 0.0979
```

What's Next?

Expanding our options with tidymodels and the Harrell-verse...

- Fitting linear and logistic regression models in new ways
- Evaluating the success of our models in new ways
- Incorporating imputation approaches more seamlessly

Please don't forget to submit your Project A proposal by Monday at 9 PM.