PM 592
Regression Analysis for Public Health Data Science

Week 10

Predictive Modeling

Predictive Modeling

Introduction to Prediction Models

Predictive Model Building

Predictive Power

Optimizing Classification

Lecture Objectives

- Implement a complete prediction model-building method
- Explain how to diagnose the predictive ability of these models
- Describe the ROC curve and its metrics
- Determine the best cut point for a prediction model

1. Review

✓ Assumptions in logistic regression – similarities and differences from OLS regression

✓ Goodness-of-fit measures

✓ Diagnosing outliers and influential values

✓ Automated selection procedures

There are two ways to approach the development of regression models:

1. **Testing a hypothesized association**

   "Does a violence prevention program in high schools successfully reduce the chance that students will experience bullying?"

   Must consider potential confounders and effect modifiers.

2. **Developing a prediction model**

   "What are the factors that contribute to developing coronary heart disease?" (Framingham coronary risk model)

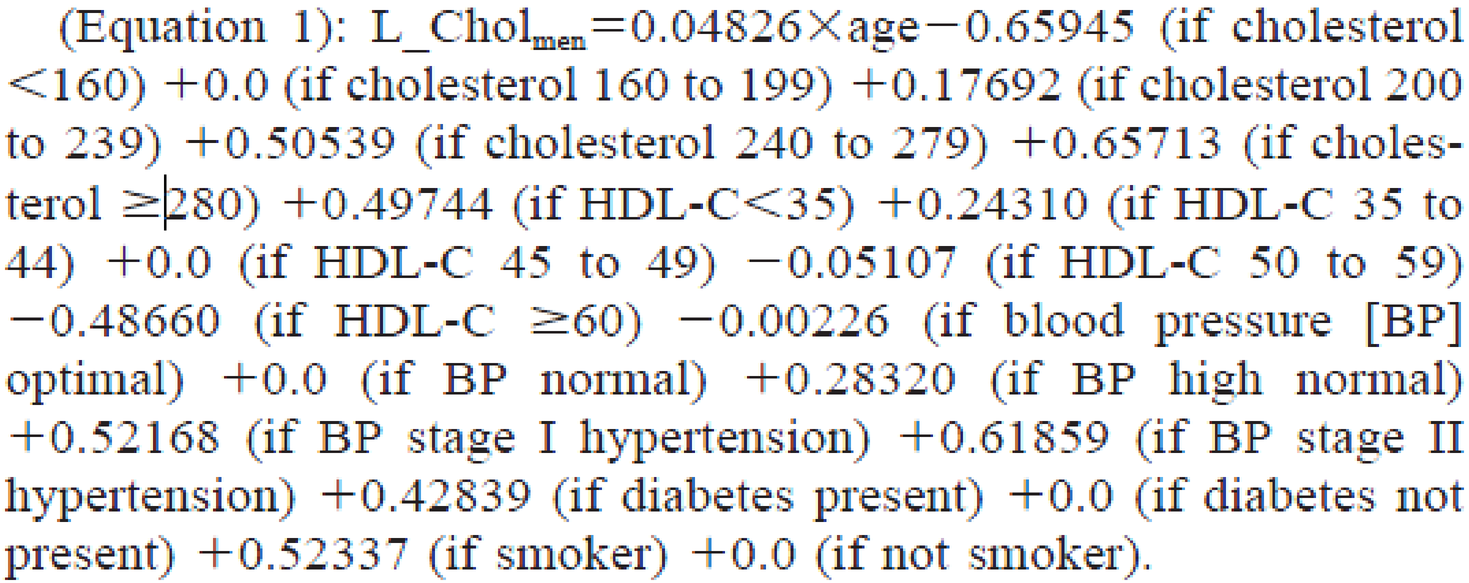
   Confounders aren't of interest, as there is no specific association we are examining.

   We just want a model that has a successful prediction rate.

Regression is one tool used for prediction models
(We discuss only logistic regression in this course)

- Linear Regression
- Logistic Regression
- Time Series
- Decision Tree
- Neural Networks
- Random Forests

The model determines how variables, **as a set**, predict the outcome.



(Equation 1): L_Chol$_{\text{men}}$=0.04826×age−0.65945 (if cholesterol <160) +0.0 (if cholesterol 160 to 199) +0.17692 (if cholesterol 200 to 239) +0.50539 (if cholesterol 240 to 279) +0.65713 (if cholesterol ≥280) +0.49744 (if HDL-C<35) +0.24310 (if HDL-C 35 to 44) +0.0 (if HDL-C 45 to 49) −0.05107 (if HDL-C 50 to 59) −0.48660 (if HDL-C ≥60) −0.00226 (if blood pressure [BP] optimal) +0.0 (if BP normal) +0.28320 (if BP high normal) +0.52168 (if BP stage I hypertension) +0.61859 (if BP stage II hypertension) +0.42839 (if diabetes present) +0.0 (if diabetes not present) +0.52337 (if smoker) +0.0 (if not smoker).

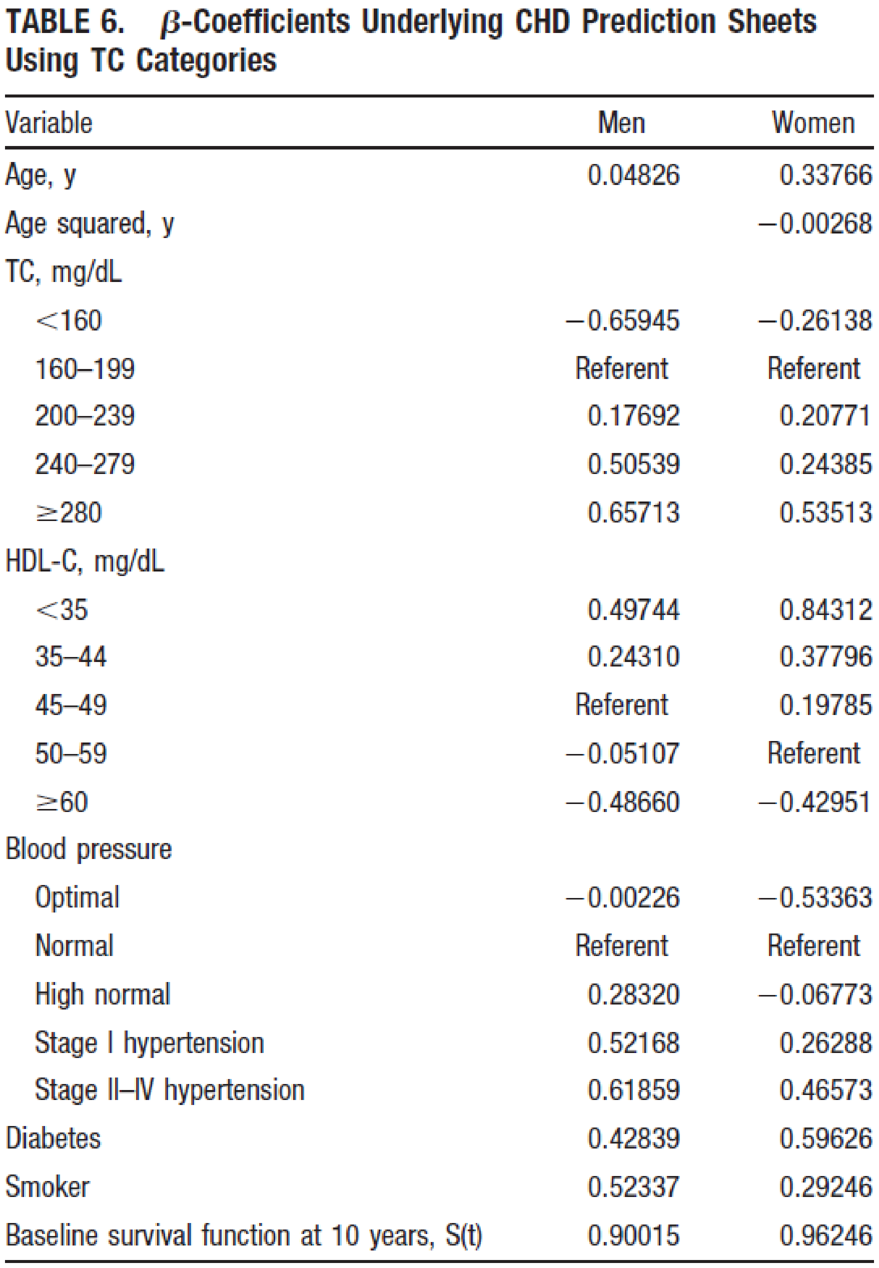


TABLE 6. β-Coefficients Underlying CHD Prediction Sheets Using TC Categories

Variable	Men	Women
Age, y	0.04826	0.33766
Age squared, y		−0.00268
TC, mg/dL		
<160	−0.65945	−0.26138
160–199	Referent	Referent
200–239	0.17692	0.20771
240–279	0.50539	0.24385
≥280	0.65713	0.53513
HDL-C, mg/dL		
<35	0.49744	0.84312
35–44	0.24310	0.37796
45–49	Referent	0.19785
50–59	−0.05107	Referent
≥60	−0.48660	−0.42951
Blood pressure		
Optimal	−0.00226	−0.53363
Normal	Referent	Referent
High normal	0.28320	−0.06773
Stage I hypertension	0.52168	0.26288
Stage II–IV hypertension	0.61859	0.46573
Diabetes	0.42839	0.59626
Smoker	0.52337	0.29246
Baseline survival function at 10 years, S(t)	0.90015	0.96246

Which independent variables should be considered?

- Anything that may help predict the outcome.
- Many independent variables can be included.
- The variables don’t have to be etiologically relevant.
- It is not necessary to consider confounding.

How many variables should be in the final model?

- Enough to predict the outcome well, but not overfit the model.
- Rule of thumb: there should be at least 10 of each outcome (Y=0 & Y=1) for each predictor in the model (so beta and SE estimates are well-powered and not biased)

Parsimony

- Every model is a simplification of reality (parsimony).
- If two models fit the data equally well, the more parsimonious model is the one with fewer predictors.
- Models that are very complex may “over-fit” the data; providing very good prediction for our current sample but may not be **generalizable** to other samples (lacks **external validity**).
- Models that are very complex are difficult to interpret.

Validation

- To ensure the model isn’t over-fit to the data, development of a prediction model is usually done in two steps
    1. Develop a model with a training data set
    2. Validate the model with a testing data set
- The training and testing sets should be independent (i.e., don’t validate on the training data)
- The validation component will provide evidence for external validity.
- If you split a larger dataset into two sets for this purpose, generally 70% (up to 80%) of the data is used for training.

Recap

- Prediction models develop the best *prediction* of outcome and are not concerned with any particular association of interest
- Good prediction models are as simple as possible while having predictive ability
- We will additionally need to validate a prediction model against an independent data set

Recap

- Explain the differences in the approach for prediction modeling vs. models of association

Test Yourself

For each of the following, state whether it should be treated as an association or prediction model:

- M. Arduto collected blood lab values for ER patients who were diagnosed with Covid-19. Given these variables, she wanted to know whether patients would die during their hospital stay.
- Dr. Hah wondered whether a new billboard ad could reduce smoking prevalence in Koreatown, Los Angeles.
- Dr. Peiji wanted to determine whether artificial sweeteners made people more hungry after consumption, but thought this effect might be different depending on gender and BMI.
- M. Melvae works for the MTA and wants to be able to determine how many passengers will ride the bus system in a given day, based on variables such as temperature, humidity, and day of week.

Test Yourself

For each of the following, state whether it should be treated as an association or prediction model:

- M. Arduto collected blood lab values for ER patients who were diagnosed with Covid-19. Given these variables, she wanted to know whether patients would die during their hospital stay. Prediction.
- Dr. Hah wondered whether a new billboard ad could reduce smoking prevalence in Koreatown, Los Angeles. Association.
- Dr. Peiji wanted to determine whether artificial sweeteners made people more hungry after consumption, but thought this effect might be different depending on gender and BMI. Association.
- M. Melvae works for the MTA and wants to be able to determine how many passengers will ride the bus system in a given day, based on variables such as temperature, humidity, and day of week. Prediction.

Overview

1. Univariate analysis
2. Variable selection for multivariate model
3. Preliminary main effects model (does each variable retain significance?)
4. Main effects model (check linearity, scale of variables)
5. Preliminary final model (check for interactions)
6. Final model (check model fit and adequacy)

3. Steps in Predictive Model Building

Just like in cooking, a recipe is only as good as the ingredients you put into it.

Prediction models are only as good as the variables you put into it.

Check your variables and data before beginning regression modeling.

Example

Schools in rural areas face increased risk of mental health issues. Researchers want to determine whether characteristics of students' friendship networks, in addition to demographic variables, can be used to identify students at risk of suicide attempt.

```
> with(sos_dat,
+      gmodels::CrossTable(s_att))

   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  11042

          |         0 |         1 |
          |-----------|-----------|
          |      9958 |      1084 |
          |     0.902 |     0.098 |
          |-----------|-----------|
```

Example

We will predict suicide attempt using several variables:

male – male gender
age – student age
grade – student grade in school (9-12)
odg – out-degree (# of friends student named)
dens – density of the ego's friendship network (range from 0 to 1)
recip – reciprocity of friendship nominations (range from 0 to 1)
tatot – number of trusted adults named by the student (0-7)
bullied – student was bullied at school
bully – student bullied others at school

1. Determine the Form of the Outcome

- For a binary outcome, you may use:
  - Logistic regression
  - Probit regression (not covered in this course)
- For a non-binary outcome, you may use:
  - Linear regression
  - Poisson regression
  - Others

- For continuous/non-binary outcomes, it helps to evaluate any potential transformations with the entire set of possible X variables.

2. Univariate Analyses

- Categorical Predictors
  - Examine the distribution of X for Y=0 and Y=1
  - Get univariate odds ratios
  - Pay attention to empty (zero) cells – we will need to deal with these somehow

The χ^2 test from the contingency table is asymptotically equivalent to the LR χ^2 test from the logistic regression model.

- Continuous Predictors:
  - Examine the distribution of X for Y=0 and Y=1
  - Get univariate odds ratios
  - Assess the linearity assumption (grouped smooth, LOESS, fractional polynomials)

The 2-sample t-test is is asymptotically equivalent to the χ^2 tests from the logistic regression model.

Example

Relationship between grade and attempt: contingency table.

```
> with(sos_dat, gmodels::CrossTable(grade, s_att, prop.chisq=F, prop.t=F, chisq=T))

   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|-------------------------|

             | s_att 
       grade |         0 |         1 | Row Total | 
-------------|-----------|-----------|-----------|
           9 |      2475 |       286 |      2761 | 
             |     0.896 |     0.104 |     0.250 | 
             |     0.249 |     0.264 |           | 
-------------|-----------|-----------|-----------|
          10 |      2502 |       259 |      2761 | 
             |     0.906 |     0.094 |     0.250 | 
             |     0.251 |     0.239 |           | 
-------------|-----------|-----------|-----------|
          11 |      2450 |       310 |      2760 | 
             |     0.888 |     0.112 |     0.250 | 
             |     0.246 |     0.286 |           | 
-------------|-----------|-----------|-----------|
          12 |      2531 |       229 |      2760 | 
             |     0.917 |     0.083 |     0.250 | 
             |     0.254 |     0.211 |           | 
-------------|-----------|-----------|-----------|
Column Total |      9958 |      1084 |     11042 | 
             |     0.902 |     0.098 |           | 
-------------|-----------|-----------|-----------|

Pearson's Chi-squared test 
------------------------------------------------------------
Chi^2 =  14.95095     d.f. =  3     p =  0.001859049
```

3. Steps in Predictive Model Building

Example

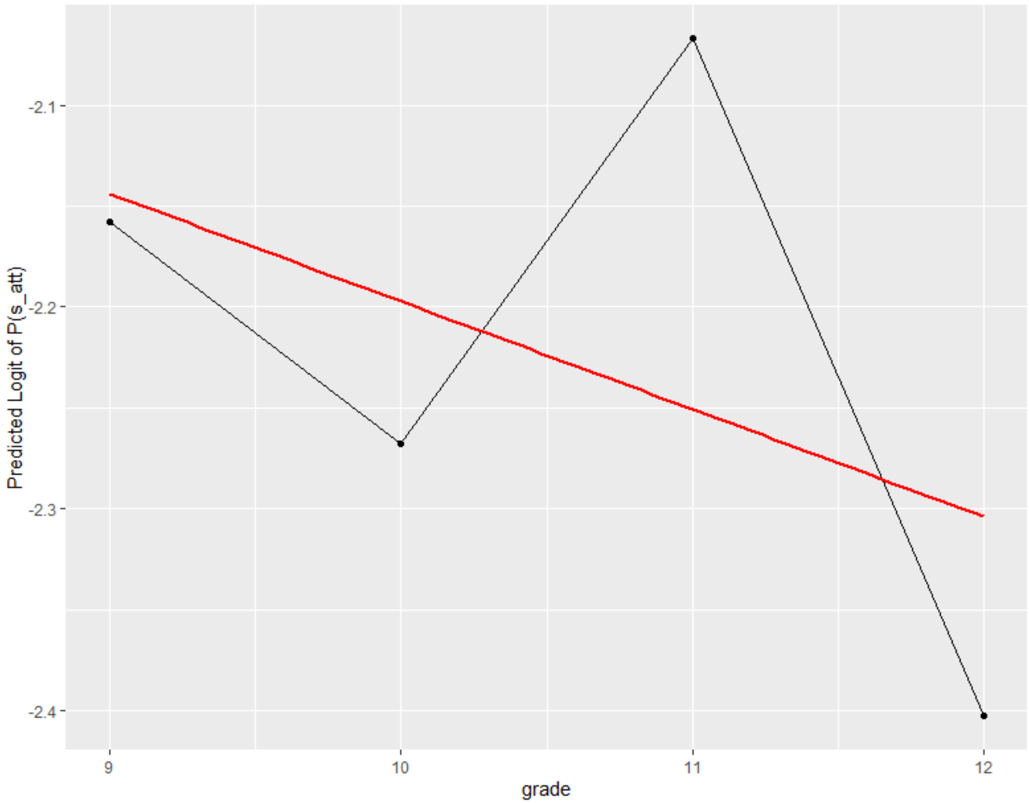
Relationship between grade and attempt: logistic regression.

```
> summary(univ_grade.m)

Call:
glm(formula = s_att ~ factor(grade), family = binomial, data = sos_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4881  -0.4677  -0.4439  -0.4162   2.2313

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.15800    0.06245 -34.553  < 2e-16 ***
factor(grade)10 -0.11001    0.09034  -1.218  0.22331
factor(grade)11  0.09073    0.08680   1.045  0.29589
factor(grade)12 -0.24464    0.09307  -2.629  0.00858 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(univ_grade.m, test="LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: s_att

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                          11041     7090.0
factor(grade)  3   15.055     11038     7074.9  0.00177 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. Steps in Predictive Model Building

We may want to keep this variable categorical.

```
> group_smooth("grade", "s_att", sos_dat)
`summarise()` ungrouping output (override with `.groups` argument)
Analysis of Deviance Table

Model 1: y ~ meanx
Model 2: y ~ factor(meanx)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     11040     7087.0
2     11038     7074.9  2   12.127 0.002326 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

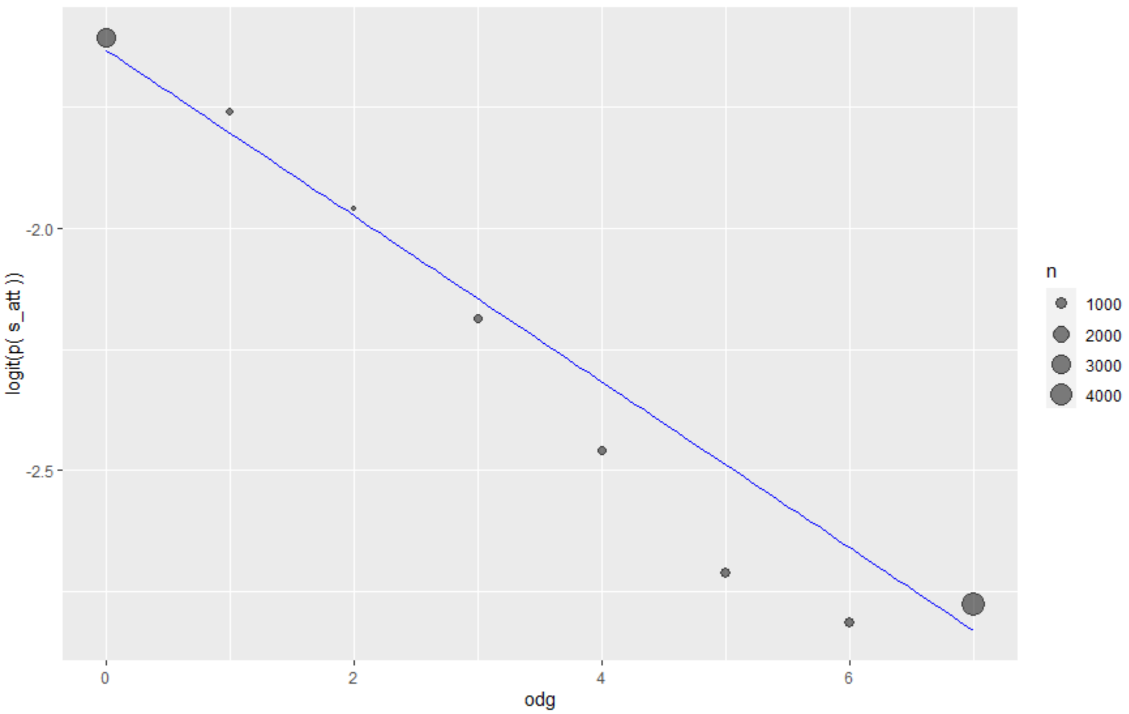


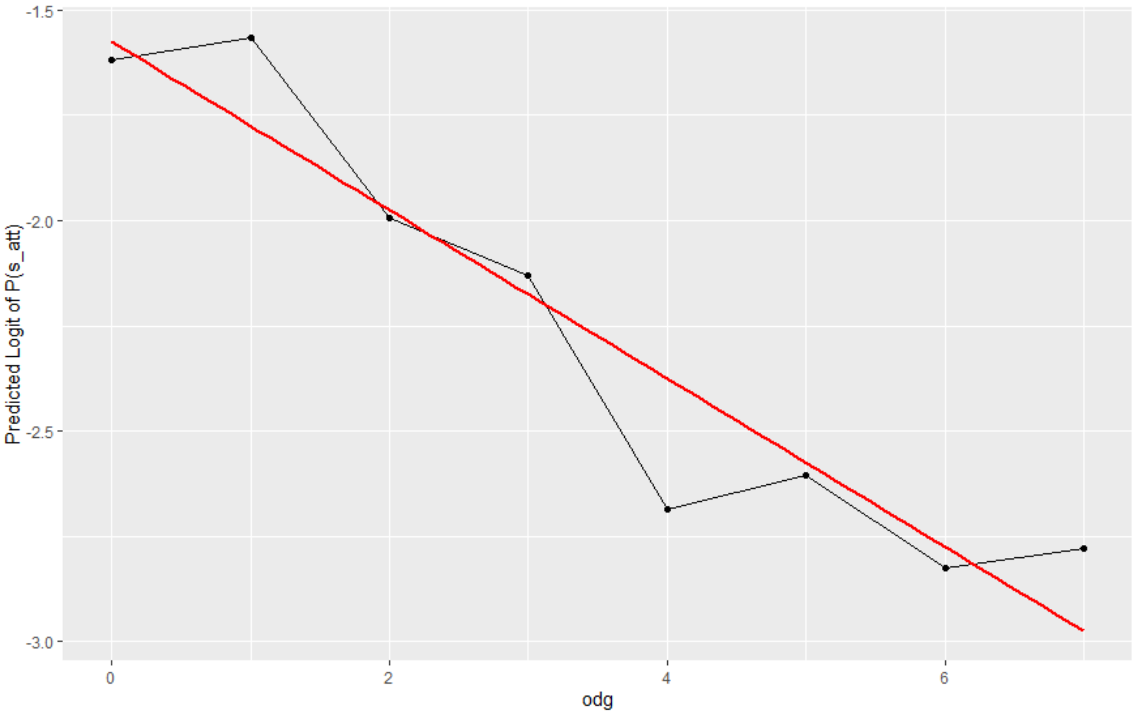
Example

Distribution of out-degree by attempt status

```
> sos_dat %>%
+   group_by(s_att) %>%
+   select(s_att, odg) %>%
+   skimr::skim()
-- Data Summary ------------------------
                           Values
Name                       Piped data
Number of rows             11043
Number of columns          2
_______________________
Column type frequency:
  numeric                  1
________________________
Group variables            s_att

-- Variable type: numeric --------------------------------------------------------------------------------------------
# A tibble: 3 x 12
  skim_variable s_att n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
* <chr>         <dbl>     <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 odg               0         0             1  4.26  2.98     0     0     6  7        7 ▆▁▃▁▇
2 odg               1         0             1  2.64  3.01     0     0     1  6.25     7 ▇▁▁▁▆
3 odg              NA         0             1  5    NA         5     5     5  5        5 ▁▁▇▁▁
```

Relationship between out-degree and attempt: LOESS and Grouped Smooth.





Relationship between out-degree and attempt: MFP.

```
> mfp(s_att ~ fp(odg), family = binomial, data = sos_dat)
Call:
mfp(formula = s_att ~ fp(odg), data = sos_dat, family = binomial)


Deviance table:
                       Resid. Dev
Null model  7089.952
Linear model           6812.877
Final model            6812.877

Fractional polynomials:
    df.initial select alpha df.final power1 power2
odg          4      1  0.05        1      1      .


Transformations of covariates:
              formula
odg I(((odg+1)/10)^1)

Re-Scaling:
Non-positive values in some of the covariates. No re-scaling was performed.

Coefficients:
Intercept      odg.1
   -1.438     -1.754

Degrees of Freedom: 11041 Total (i.e. Null);  11040 Residual
Null Deviance:            7090
Residual Deviance: 6813          AIC: 6817
```

3. Variable Selection

After all variables have been examined univariately, include

- All variables with “clinical importance”
- If this produces too many variables, include all with a univariate p<.25
- Constrain the total number of variables to follow the sample size rule of thumb, either by redefining the definition of “clinically important” or by choosing a lower p-value threshold for inclusion.

We use a less strict p-value to include variables as they may become significant later (in combination with others).

4. Preliminary Modeling

Use purposeful model building to examine several different models.

- Be careful of automated selection procedures (stepwise regression)
- Forward selection produces more “noise” variables
- Some good models cannot be found with automated selection
- In general automated procedures do not do well with correlated predictors
- Automated selection discourages thinking about the actual problem
- Automated procedures can be helpful for hypothesis-generating analyses (after the primary analysis is done)

4. Preliminary Modeling

Be skeptical of p-values.

- P-values are only valid for testing pre-specified hypotheses
- Since we are screening variables, p-values only indicate relative importance among all variables included
- The larger candidate pool of variables, the more variables will appear significant when they in fact aren’t related to outcome

Scrutinize the models for biological plausibility.

Choose the “best” multivariate model

Verify Each Variable

Verify, delete, refit, etc. until you’re satisfied that:

- All important variables are included in the model
- All excluded variables are clinically or statistically unimportant

Then you will have the **preliminary main effects model**.

Preliminary main effects model.

```
> summary(best_subset_att)$bestmodel %>% glm(., data = sos_dat, family = binomial) %>% summary(.)

Call:
glm(formula = ., family = binomial, data = sos_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-1.4184  -0.4834  -0.3466  -0.2384   3.0518  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.61553    0.11612 -22.524  < 2e-16 ***
odg         -0.16279    0.01189 -13.690  < 2e-16 ***
dens        -0.64609    0.12164  -5.311 1.09e-07 ***
recip       -0.69673    0.13829  -5.038 4.70e-07 ***
male        -0.73708    0.06830 -10.791  < 2e-16 ***
bully        0.90381    0.05021  17.999  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7090.0  on 11041  degrees of freedom
Residual deviance: 6292.7  on 11036  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 6304.7

Number of Fisher Scoring iterations: 6
```

Some questions:

- Why did age and grade drop out of the model?
- Why is only “bully” in the model (and not “bullied”)?

5. Refine your model

Re-assess linearity for continuous variables.

- This assumption is usually not critical in the variable selection stages; the model-building process is quite forgiving for modest violations of linearity (except for U-shaped relationships)
- Scatterplots (e.g., LOESS) are not easily extended to multivariable models
- Grouped-smooth and fractional polynomials are good approaches to assess linearity in multivariable models

Make sure your model makes sense clinically and scientifically.

You now have the **main effects model**.

6. Check for interactions

List all pairs of variables that have scientific plausibility for interaction and add them to the model.

- May need to discuss “plausibility” with your co-investigators with content expertise
- Add one at a time to the main effects model
- Include interactions significant at p<.05 (we’re stricter here because non-significant interactions tend to inflate the standard errors of beta coefficients).

You now have the **preliminary final model**.

Preliminary final model.

```
> summary(best_subset2_att)$bestmodel %>% glm(., data = sos_dat, family = binomial) %>% summary(.)

Call:
glm(formula = ., family = binomial, data = sos_dat)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.4696  -0.4941  -0.3465  -0.1946   3.2520  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -2.95479    0.12359 -23.907  < 2e-16 ***
odg           -0.05241    0.01547  -3.388 0.000704 ***
dens           0.36294    0.14785   2.455 0.014096 *  
recip         -0.30085    0.14169  -2.123 0.033724 *  
male          -0.74965    0.06848 -10.947  < 2e-16 ***
bully          0.90464    0.05063  17.866  < 2e-16 ***
I(odg * dens) -0.48294    0.04761 -10.144  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7090.0  on 11041  degrees of freedom
Residual deviance: 6177.5  on 11035  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 6191.5

Number of Fisher Scoring iterations: 6
```

Some questions:

- What does it mean for an out-degree x density interaction?

7. Check Model Fit

Fit Statistics

- Pearson’s GOF
- Hosmer-Lemeshow

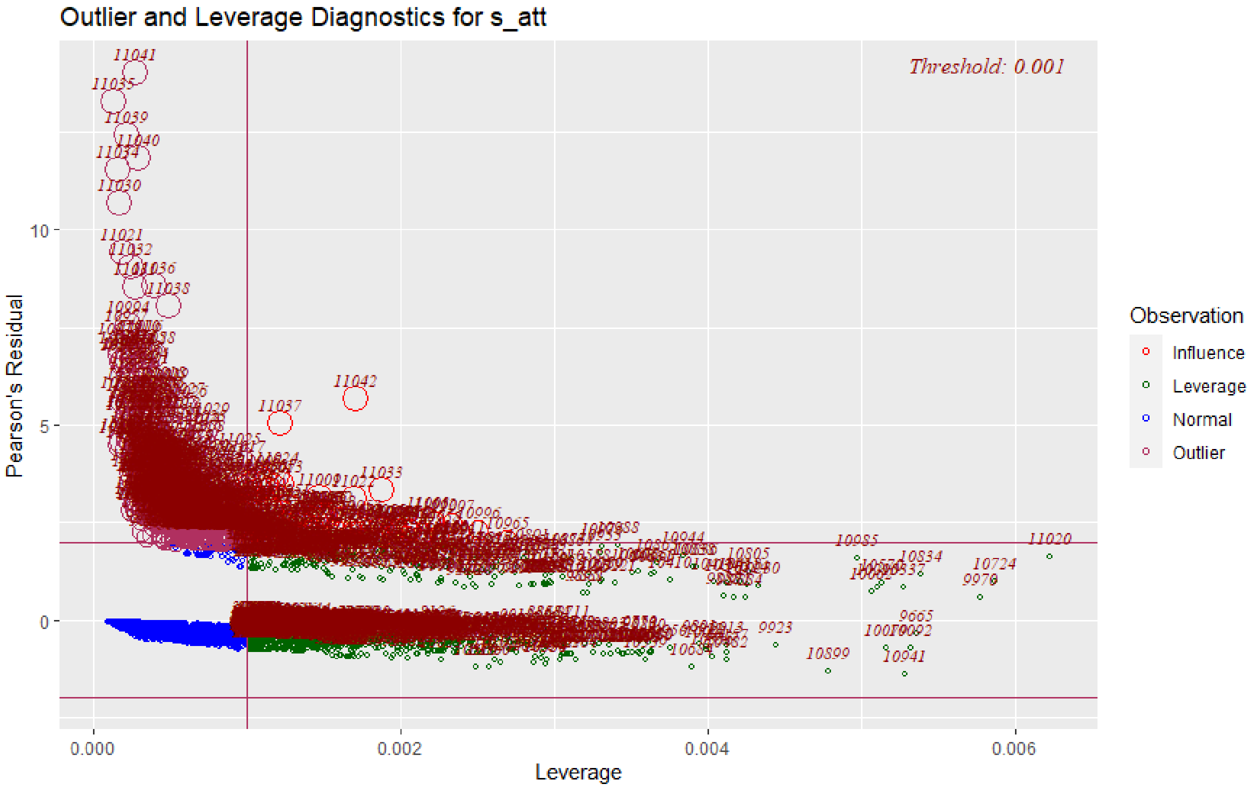
Model Diagnostics

- Closely examine influential points
- Do NOT exclude influential points simply to get better fit
- Consult with investigators and content experts to see if there is a reason why points might be excluded

You now have the **final model**.

Fit Statistics

```
> ResourceSelection::hoslem.test(best_model$y, fitted(best_model), g=10)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  best_model$y, fitted(best_model)
X-squared = 3.745, df = 8, p-value = 0.8794

> ResourceSelection::hoslem.test(best_model$y, fitted(best_model), g=20)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  best_model$y, fitted(best_model)
X-squared = 13.001, df = 18, p-value = 0.7915
```



Recap

- Model building requires statistical knowledge but is also part art; think of yourself as crafting a model and getting to know the variables along the way
- Good model-building is driven by theory as well; check to make sure your results make sense along the way, and there is theoretical justification for how you handle the variables
- There may be a lot of trial-and-error and refinement in this process

Recap

- Implement the methods in this section to build a prediction model

From Probability to Outcome

Recall that logistic regression provides logit values, which are converted to $\hat{\pi}$, the estimated probability that Y=1.

4. Predictive Power

Generally, we classify people into predicted outcome status by:

i) Fitting the logistic model

ii) Obtaining the predicted probabilities for each subject ($\hat{\pi}$)

iii) Choosing a cutpoint *c* (usually *c*=0.5)

iv) Classifying individuals into an estimated outcome based on their predicted probability and c, such that:

If $\hat{\pi}_i$ > c, $\hat{Y}_i$ = 1

If $\hat{\pi}_i$ < c, $\hat{Y}_i$ = 0

4. Predictive Power

Classification

Let’s classify individuals in this data set.

Our accuracy rate is 0.9036 – that is, we correctly classified about 90% of participants!

Suppose our model instead classified everyone as Y=0. What would have been the correct classification rate?

```
> DescTools::Conf(best_model, pos = 1)

Confusion Matrix and Statistics

          Reference
Prediction    1    0
         1   35   16
         0 1049 9942

               Total n : 11'042
              Accuracy : 0.9036
                95% CI : (0.8979, 0.9089)
   No Information Rate : 0.9018
   P-Value [Acc > NIR] : 0.2780

                 Kappa : 0.0533
 Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.0323
           Specificity : 0.9984
        Pos Pred Value : 0.6863
        Neg Pred Value : 0.9046
            Prevalence : 0.0982
        Detection Rate : 0.0046
  Detection Prevalence : 0.0032
     Balanced Accuracy : 0.5153
      F-val Accuracy : 0.0617
   Matthews Cor.-Coef : 0.1346

      'Positive' Class : 1
```

Classification

Let’s classify individuals in this data set.

```
> DescTools::Conf(best_model, pos = 1)

Confusion Matrix and Statistics

           Reference
Prediction     1    0
         1    35   16
         0  1049 9942

               Total n : 11'042
              Accuracy : 0.9036
                95% CI : (0.8979, 0.9089)
   No Information Rate : 0.9018
   P-Value [Acc > NIR] : 0.2780

                 Kappa : 0.0533
 Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.0323
           Specificity : 0.9984
        Pos Pred Value : 0.6863
        Neg Pred Value : 0.9046
            Prevalence : 0.0982
        Detection Rate : 0.0046
  Detection Prevalence : 0.0032
     Balanced Accuracy : 0.5153
     F-val Accuracy : 0.0617
   Matthews Cor.-Coef : 0.1346

       'Positive' Class : 1
```

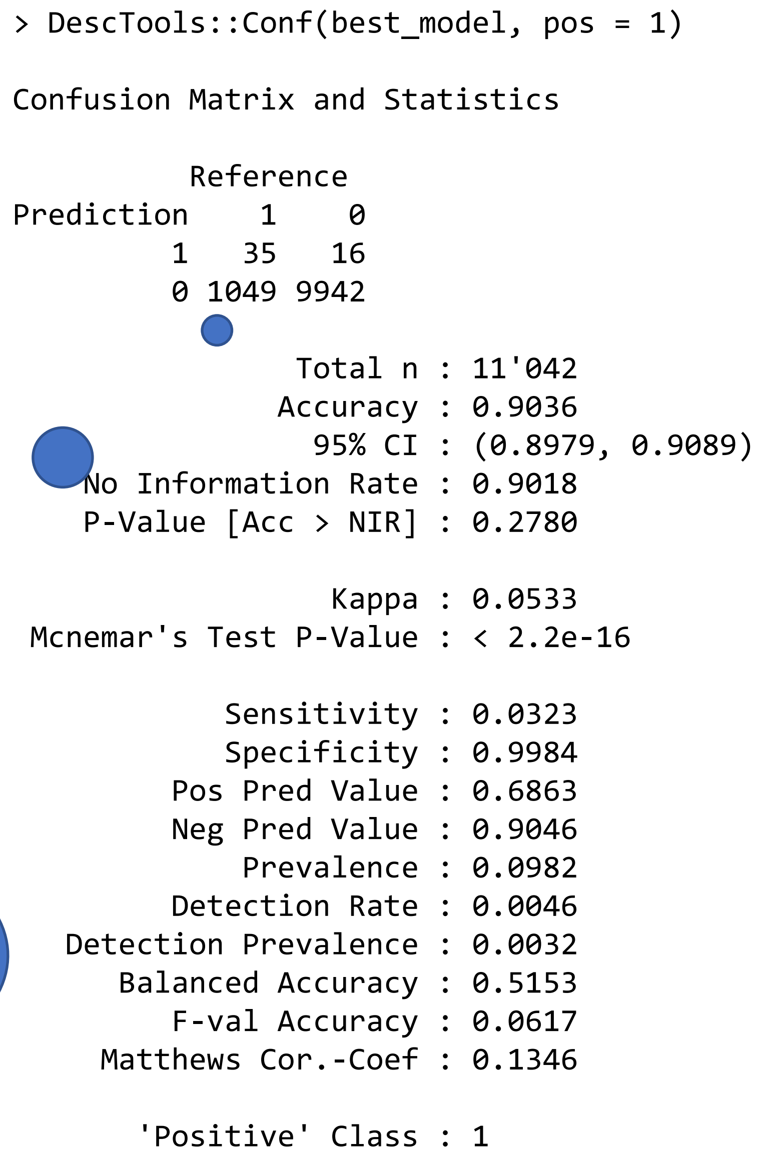
The Accuracy is not better than the No Information Rate

4. Predictive Power

Classification

Let’s classify individuals in this data set.

Our accuracy rate is 0.9036 – that is, we correctly classified about 90% of participants!



```
> DescTools::Conf(best_model, pos = 1)

Confusion Matrix and Statistics

          Reference
Prediction    1    0
         1   35   16
         0 1049 9942

               Total n : 11'042
              Accuracy : 0.9036
                95% CI : (0.8979, 0.9089)
   No Information Rate : 0.9018
   P-Value [Acc > NIR] : 0.2780

                 Kappa : 0.0533
 Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.0323
           Specificity : 0.9984
        Pos Pred Value : 0.6863
        Neg Pred Value : 0.9046
            Prevalence : 0.0982
        Detection Rate : 0.0046
  Detection Prevalence : 0.0032
     Balanced Accuracy : 0.5153
      F-val Accuracy : 0.0617
   Matthews Cor.-Coef : 0.1346

       'Positive' Class : 1
```

(1049 + 35) = 1084 individuals had Y=1. However, our model only predicted (16 + 35) = 51 to have the outcome. Our model seems to be a bit conservative in assigning Y=1.

Classification

Let’s classify individuals in this data set.

Sensitivity: the proportion of those with Y=1 that had $\hat{Y}=1$.

How good is the model at identifying individuals who actually have the outcome?

```
> DescTools::Conf(best_model, pos = 1)

Confusion Matrix and Statistics

          Reference
Prediction    1    0
         1   35   16
         0 1049 9942

               Total n : 11'042
              Accuracy : 0.9036
                95% CI : (0.8979, 0.9089)
   No Information Rate : 0.9018
   P-Value [Acc > NIR] : 0.2780

                 Kappa : 0.0533
 Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.0323
           Specificity : 0.9984
        Pos Pred Value : 0.6863
        Neg Pred Value : 0.9046
            Prevalence : 0.0982
        Detection Rate : 0.0046
  Detection Prevalence : 0.0032
     Balanced Accuracy : 0.5153
      F-val Accuracy : 0.0617
     Matthews Cor.-Coef : 0.1346

      'Positive' Class : 1
```

Classification

Let’s classify individuals in this data set.

Specificity: the proportion of those with Y=0 that had $\hat{Y}$ =0.

9942/(16 + 9942)

How good is the model at discerning which individuals do not have the outcome?

```
> DescTools::Conf(best_model, pos = 1)

Confusion Matrix and Statistics

          Reference
Prediction    1    0
         1   35   16
         0 1049 9942

                Total n : 11'042
               Accuracy : 0.9036
                 95% CI : (0.8979, 0.9089)
    No Information Rate : 0.9018
    P-Value [Acc > NIR] : 0.2780

                  Kappa : 0.0533
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.0323
            Specificity : 0.9984
         Pos Pred Value : 0.6863
         Neg Pred Value : 0.9046
             Prevalence : 0.0982
         Detection Rate : 0.0046
   Detection Prevalence : 0.0032
      Balanced Accuracy : 0.5153
         F-val Accuracy : 0.0617
     Matthews Cor.-Coef : 0.1346

       'Positive' Class : 1
```

Classification

Let’s classify individuals in this data set.

```
> DescTools::Conf(best_model, pos = 1)

Confusion Matrix and Statistics

          Reference
Prediction    1    0
         1   35   16
         0 1049 9942

               Total n : 11'042
              Accuracy : 0.9036
                95% CI : (0.8979, 0.9089)
   No Information Rate : 0.9018
   P-Value [Acc > NIR] : 0.2780

                 Kappa : 0.0533
 Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.0323
           Specificity : 0.9984
        Pos Pred Value : 0.6863
        Neg Pred Value : 0.9046
            Prevalence : 0.0982
        Detection Rate : 0.0046
  Detection Prevalence : 0.0032
     Balanced Accuracy : 0.5153
      F-val Accuracy : 0.0617
    Matthews Cor.-Coef : 0.1346

      'Positive' Class : 1
```

PPV: the proportion of those with $\hat{Y}=1$ that also had Y=1.

35/(35 + 16)

How much can a patient trust they actually have the disease after a positive diagnosis?

Classification

Let’s classify individuals in this data set.

NPV: the proportion of those with $\hat{Y}$ =0 that also had Y=0.

9942/(1049+9942)

How much can a patient trust they don’t have the disease after receiving a negative diagnosis?

```
> DescTools::Conf(best_model, pos = 1)

Confusion Matrix and Statistics

          Reference
Prediction    1    0
         1   35   16
         0 1049 9942

                Total n : 11'042
               Accuracy : 0.9036
                 95% CI : (0.8979, 0.9089)
    No Information Rate : 0.9018
    P-Value [Acc > NIR] : 0.2780

                  Kappa : 0.0533
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.0323
            Specificity : 0.9984
         Pos Pred Value : 0.6863
         Neg Pred Value : 0.9046
             Prevalence : 0.0982
         Detection Rate : 0.0046
   Detection Prevalence : 0.0032
      Balanced Accuracy : 0.5153
       F-val Accuracy : 0.0617
     Matthews Cor.-Coef : 0.1346

       'Positive' Class : 1
```

4. Predictive Power

As you can see, sensitivity and specificity depend highly on the relative sizes of each group (Y=1, Y=0).

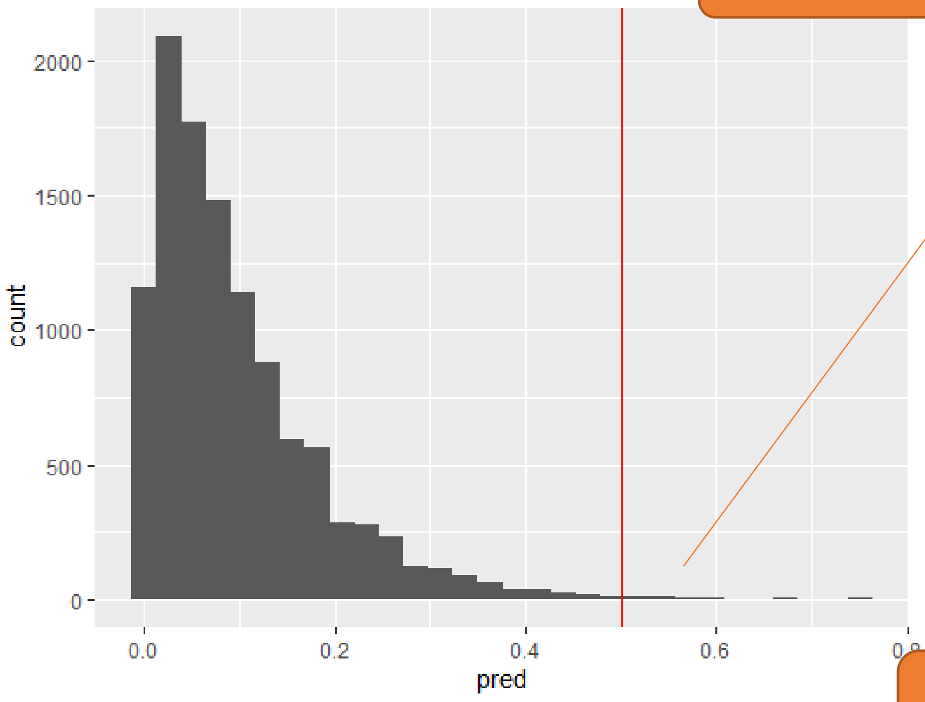
It’s typically more likely that individuals will be classified into the larger group, and this likelihood increases as the relative size of the larger group increases.

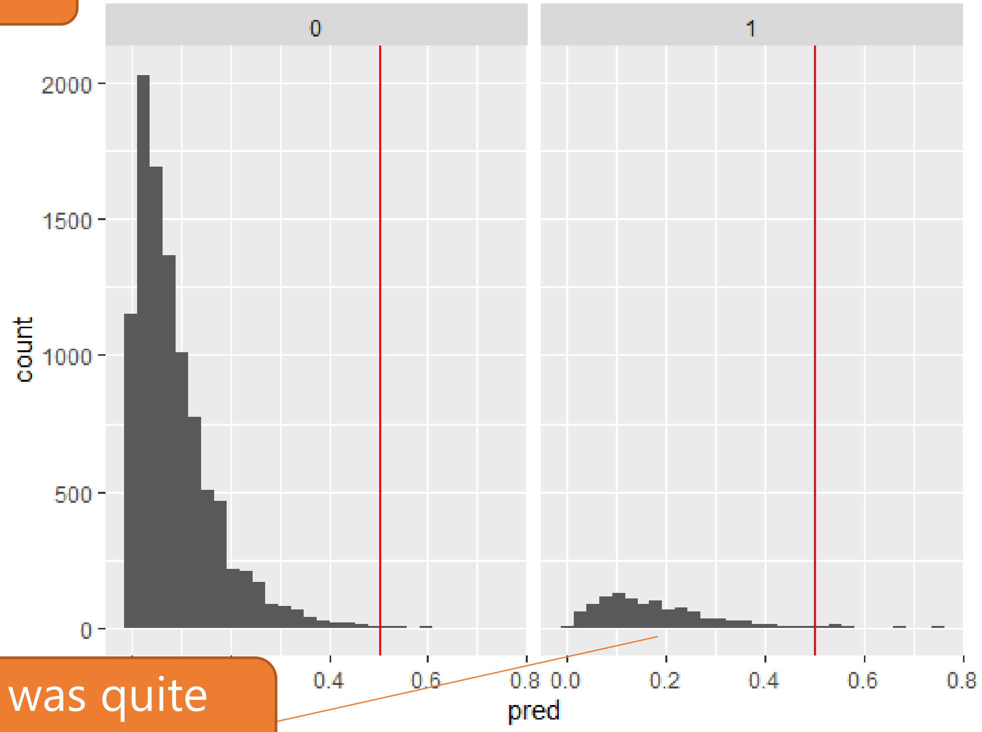
4. Predictive Power

As you can see, sensitivity and specificity depend highly on the relative sizes of each group (Y=1, Y=0).

It's typically more likely that individuals will be classified into the larger group, and this likelihood increases as the relative size of the larger group increases.

Almost nobody had $\hat{\pi}>0.50$.





The distribution of $\hat{\pi}$ was quite low, even for those with attempt.

Classification also depends on how similar individuals in the population are (with respect to $\hat{\pi}_i$):

1) In a **homogenous** population, many individuals will have $\hat{\pi}_i$ close to the classification threshold.

   This may be problematic as observations with similar $\hat{\pi}_i$ are forced into discrete outcome categories.

   E.g., A subject with $\hat{\pi}_i$=0.495 will be classified as no-outcome, and a subject with $\hat{\pi}_i$=0.505 will be classified as having outcome.

2) In a **polarized** population, the $\hat{\pi}_i$ are distributed at the extremes.

In both cases, we **shouldn’t expect perfect fit**!

- If most individuals have $\hat{\pi}_i$ close to 0.5, then we should expect about 50% misclassification.
- If most individuals have $\hat{\pi}_i$ close to 0.05 or 0.95, then we should expect about 5% misclassification.
- Classification measures (e.g. sensitivity, specificity) depend on the distribution of $\hat{\pi}_i$ in the sample and, therefore, are not absolute measures of goodness of classification.

4. Predictive Power

We can also **change the cutpoint** to make it easier or harder to classify Y=1.

This should be done depending on the research question.

Example

We want to use this diagnostic tool in order to identify students at school who may be particularly at-risk for suicide attempt.

We will use the tool to discreetly contact students and refer them to resources at school that can help them.

How would we change the cut point in this case?

Example

We would lower it.

- When you **raise the cutpoint:**

  sensitivity decreases – P$(\hat{Y} = 1|Y = 1)$
  specificity increases – P $(\hat{Y} = 0|Y = 0)$

- When you **lower the cutpoint:**

  sensitivity increases – P$(\hat{Y} = 1|Y = 1)$
  specificity decreases – P $(\hat{Y} = 0|Y = 0)$

Therefore we would be able to detect more students with attempt, at the cost of classifying some without attempt as having it.

Sensitivity vs. Specificity – Advantages

- **Sensitive** models are helpful to identify those who actually have the disease, even at the cost of misdiagnosing some individuals without the disease.
    - Screening tests with the opportunity for further follow-up
    - Examples: mammograms, HIV screening, airport security
- **Specific** models should be used when we want to verify that an individual does not have the disease, even at the cost of misdiagnosing some individuals who actually have the disease.
    - Useful after preliminary screening when being diagnosed “positive” has large risk of physical, emotional, or monetary harm (e.g., biopsies).

Predictive Value

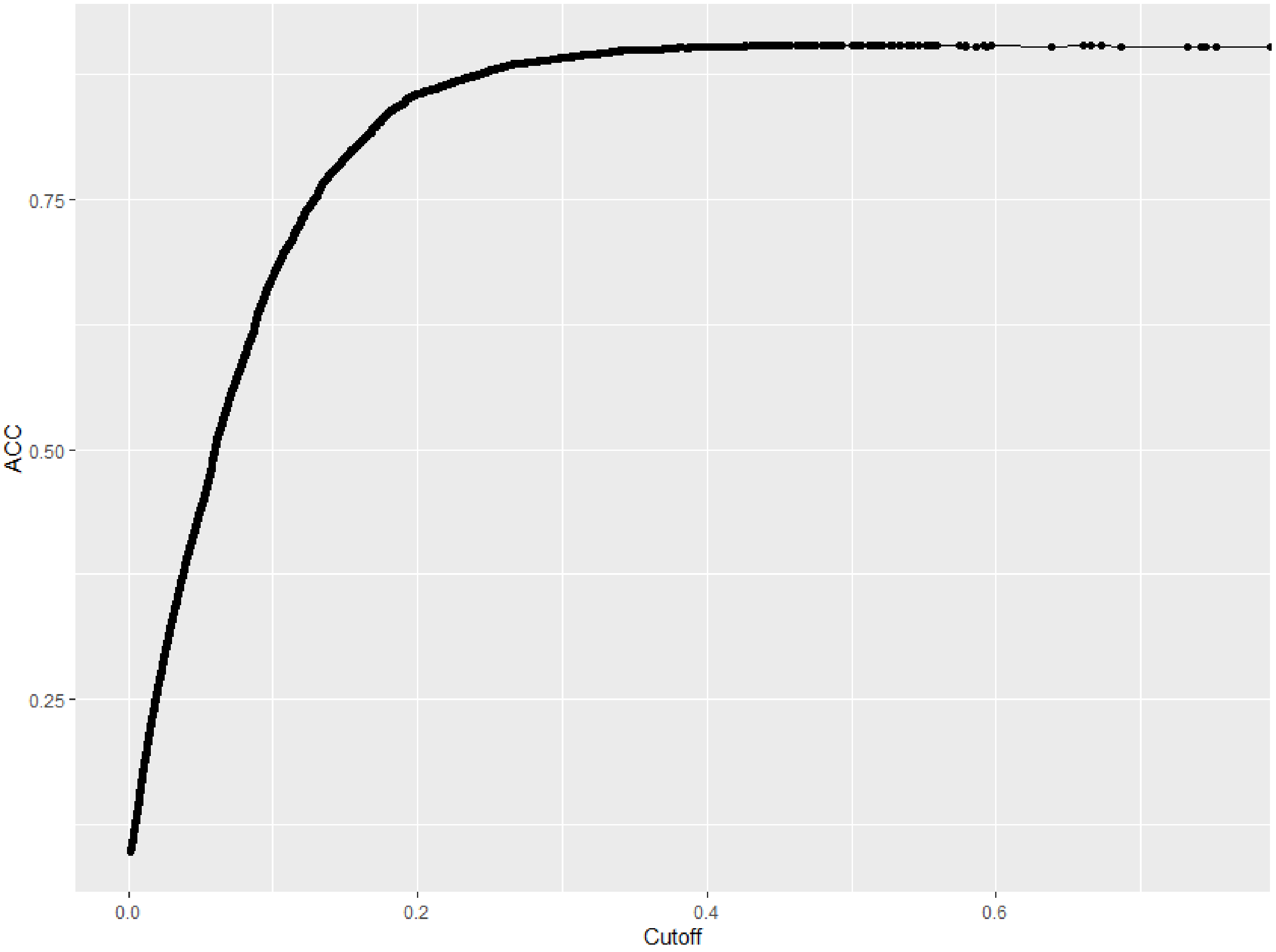
- Sensitivity and specificity are more useful for clinicians & researchers, when considering diagnosing individuals at the population-level.
- PPV and NPV are more useful to the patient

Changing the Cutpoint

Here we change the cutpoint of $\hat{\pi}$ for classifying $\hat{Y}$=1 vs. $\hat{Y}$ =0.

We graph the accuracy over all possible cutpoints.

Note that the accuracy doesn't change much, likely because of the small number of individuals classified with Y=1.



Recap

- Sensitivity, specificity, positive predictive value, and negative predictive value are all ways of evaluating the predictive power of a model.

- The cutpoint for classifying Y=1 can be changed, and will alter the specificity and sensitivity depending on the goal of the prediction model.

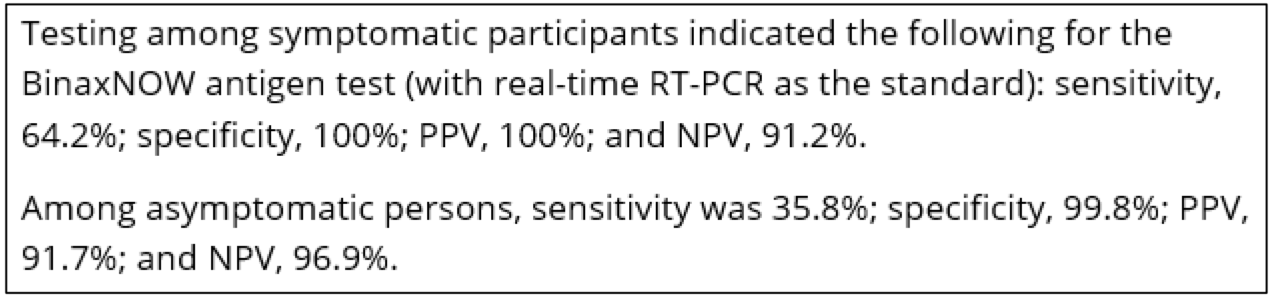
- These metrics are highly confounded with the proportion of individuals in the sample with Y=0, Y=1.

Recap

- ➢ Explain the concepts of sensitivity, specificity, positive predictive value, and negative value
- ➢ Explain how these measures are affected by the overall proportion of individuals with the outcome

Test Yourself

Rita had a stuffy nose and sore throat. Her BinaxNOW rapid test came back positive. What is the probability she has Covid-19?



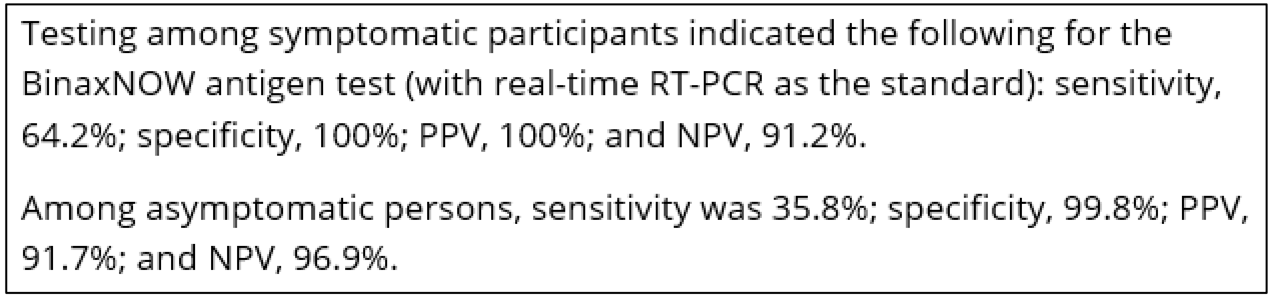
Testing among symptomatic participants indicated the following for the BinaxNOW antigen test (with real-time RT-PCR as the standard): sensitivity, 64.2%; specificity, 100%; PPV, 100%; and NPV, 91.2%.

Among asymptomatic persons, sensitivity was 35.8%; specificity, 99.8%; PPV, 91.7%; and NPV, 96.9%.

Test Yourself

Rita had a stuffy nose and sore throat. Her BinaxNOW rapid test came back positive. What is the probability she has Covid-19?

We want to know the probability $P(Y = 1|\hat{Y} = 1)$. This is the positive predictive value (PPV), which is 100%.

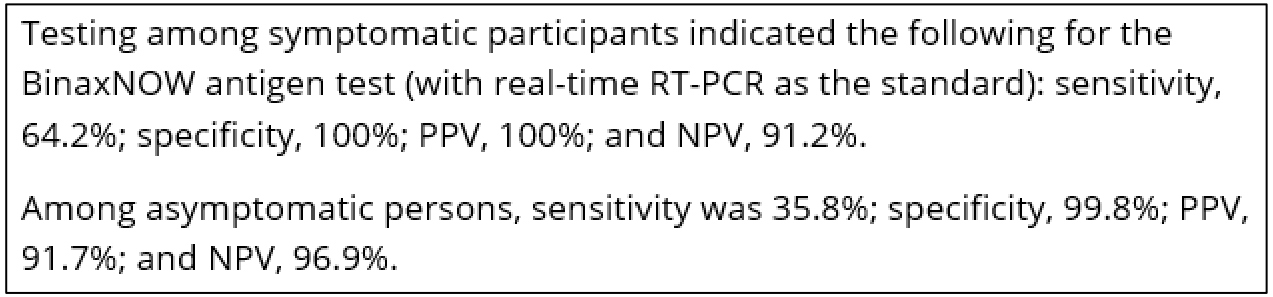


Testing among symptomatic participants indicated the following for the BinaxNOW antigen test (with real-time RT-PCR as the standard): sensitivity, 64.2%; specificity, 100%; PPV, 100%; and NPV, 91.2%.

Among asymptomatic persons, sensitivity was 35.8%; specificity, 99.8%; PPV, 91.7%; and NPV, 96.9%.

Test Yourself

Trevor was exposed to Rita and took this test 5 days after exposure. He did not have any symptoms and received a negative result. What is the probability that Trevor does NOT have Covid-19?



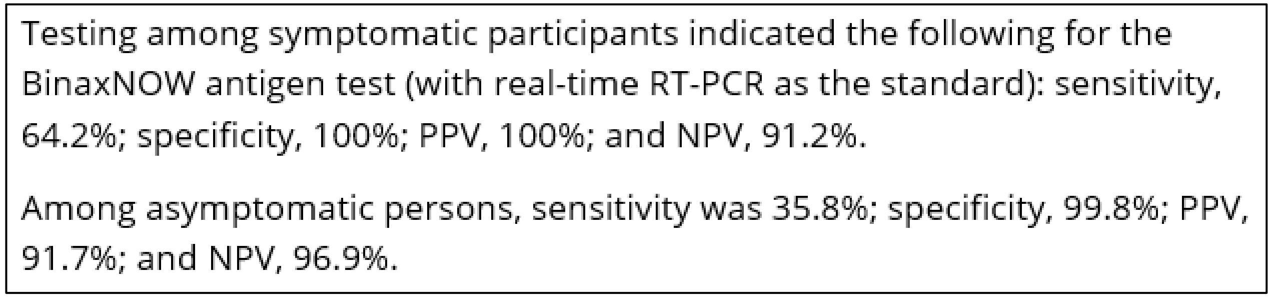
Testing among symptomatic participants indicated the following for the BinaxNOW antigen test (with real-time RT-PCR as the standard): sensitivity, 64.2%; specificity, 100%; PPV, 100%; and NPV, 91.2%.

Among asymptomatic persons, sensitivity was 35.8%; specificity, 99.8%; PPV, 91.7%; and NPV, 96.9%.

Test Yourself

Trevor was exposed to Rita and took this test 5 days after exposure. He did not have any symptoms and received a negative result. What is the probability that Trevor does NOT have Covid-19?

We want to know the probability $P(Y = 0|\hat{Y} = 0)$. This is the negative predictive value (NPV), which is 96.9%.

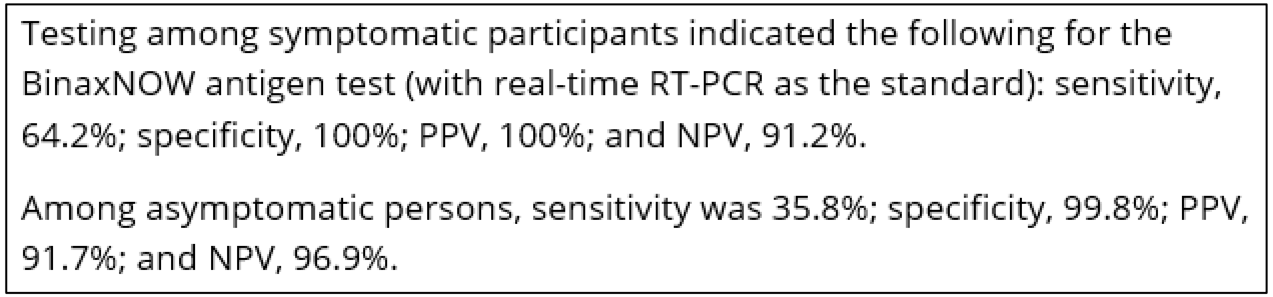


Testing among symptomatic participants indicated the following for the BinaxNOW antigen test (with real-time RT-PCR as the standard): sensitivity, 64.2%; specificity, 100%; PPV, 100%; and NPV, 91.2%.

Among asymptomatic persons, sensitivity was 35.8%; specificity, 99.8%; PPV, 91.7%; and NPV, 96.9%.

Test Yourself

Suppose you were in charge of Covid screening at the Soto building during the pandemic. If you had 50 BinaxNOW tests, and your goal was to identify individuals with Covid, should you be using the tests on symptomatic or asymptomatic individuals?



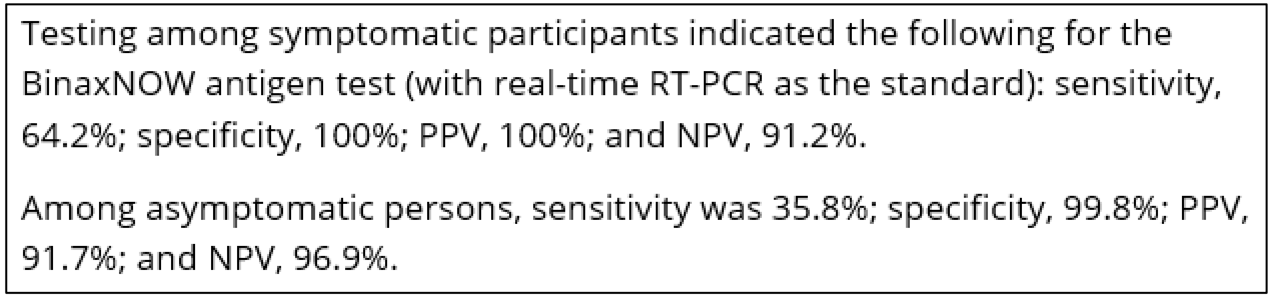
Testing among symptomatic participants indicated the following for the BinaxNOW antigen test (with real-time RT-PCR as the standard): sensitivity, 64.2%; specificity, 100%; PPV, 100%; and NPV, 91.2%.

Among asymptomatic persons, sensitivity was 35.8%; specificity, 99.8%; PPV, 91.7%; and NPV, 96.9%.

Test Yourself

Suppose you were in charge of Covid screening at the Soto building during the pandemic. If you had 50 BinaxNOW tests, and your goal was to identify individuals with Covid, should you be using the tests on symptomatic or asymptomatic individuals?

If our goal is to identify those with Covid, we want to have a high value of $P\left(\hat{Y} = 1 \middle| Y = 1\right)$ - we want the test to predict people have Covid if they actually have it. This is the sensitivity, and it is higher in those who have symptoms (64.2%) vs. those who are asymptomatic (35.8%). Test symptomatic individuals.



Testing among symptomatic participants indicated the following for the BinaxNOW antigen test (with real-time RT-PCR as the standard): sensitivity, 64.2%; specificity, 100%; PPV, 100%; and NPV, 91.2%.

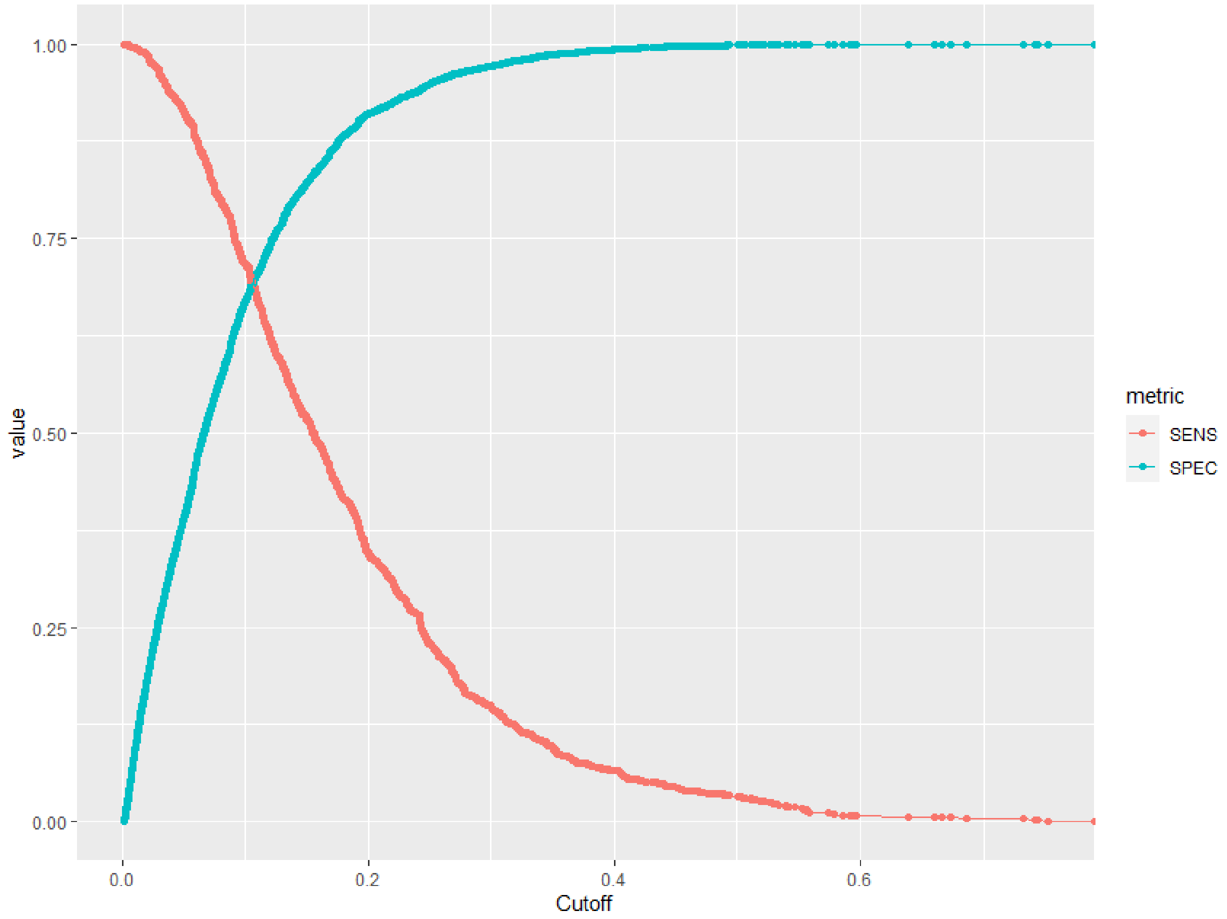
Among asymptomatic persons, sensitivity was 35.8%; specificity, 99.8%; PPV, 91.7%; and NPV, 96.9%.

5. Optimizing Classification

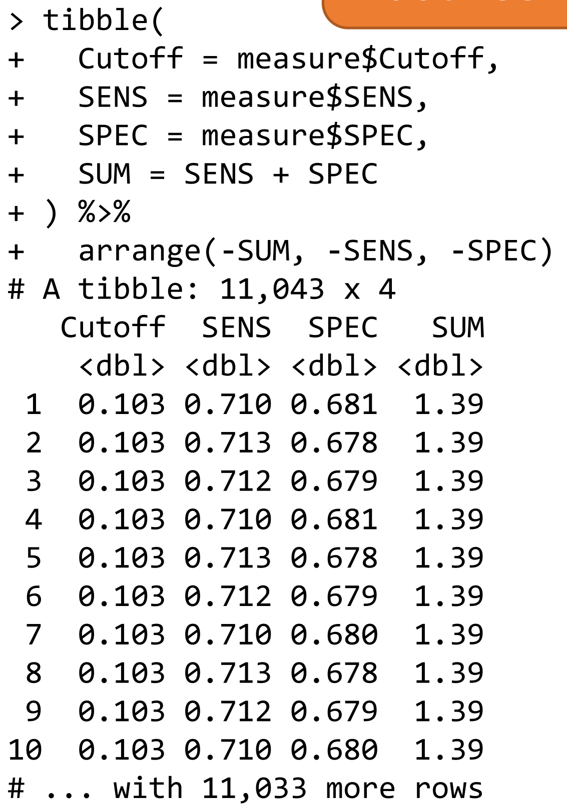
We can see that the cutpoint we choose may depend on the nature of the diagnostic tool you'd like to create.

Is there another way to optimize how individuals are classified?

We can create a graph that shows us the **tradeoff between sensitivity and specificity.**



A cutpoint of 0.103 maximizes both sensitivity and specificity.



```
> tibble(
+   Cutoff = measure$Cutoff,
+   SENS = measure$SENS,
+   SPEC = measure$SPEC,
+   SUM = SENS + SPEC
+ ) %>%
+   arrange(-SUM, -SENS, -SPEC)
# A tibble: 11,043 x 4
   Cutoff  SENS  SPEC   SUM
    <dbl> <dbl> <dbl> <dbl>
 1  0.103 0.710 0.681  1.39
 2  0.103 0.713 0.678  1.39
 3  0.103 0.712 0.679  1.39
 4  0.103 0.710 0.681  1.39
 5  0.103 0.713 0.678  1.39
 6  0.103 0.712 0.679  1.39
 7  0.103 0.710 0.680  1.39
 8  0.103 0.713 0.678  1.39
 9  0.103 0.712 0.679  1.39
10  0.103 0.710 0.680  1.39
# ... with 11,033 more rows
```

ROC (receiver operating characteristic) curves are another tool used to optimize classification, and provide a more complete assessment of classification accuracy.

The area under these curves is indicative of how good a model fits.

> In world war 2, radar operators had to interpret blips on radar screens as either friendly, hostile, or noise. The blip was the “signal”, and an ROC curve was one way of “signal detection,” or a set of ways to measure/study how, and to what extent, the receivers could make sense of the signal.

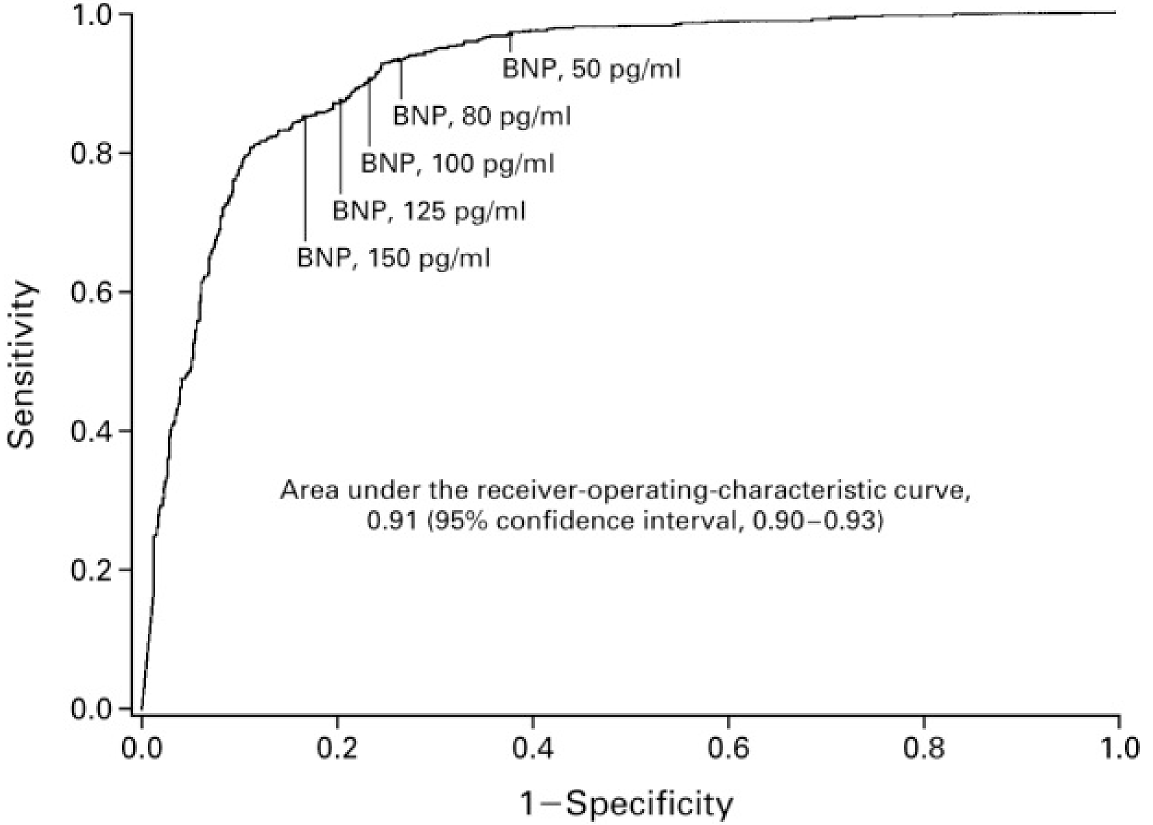
Example

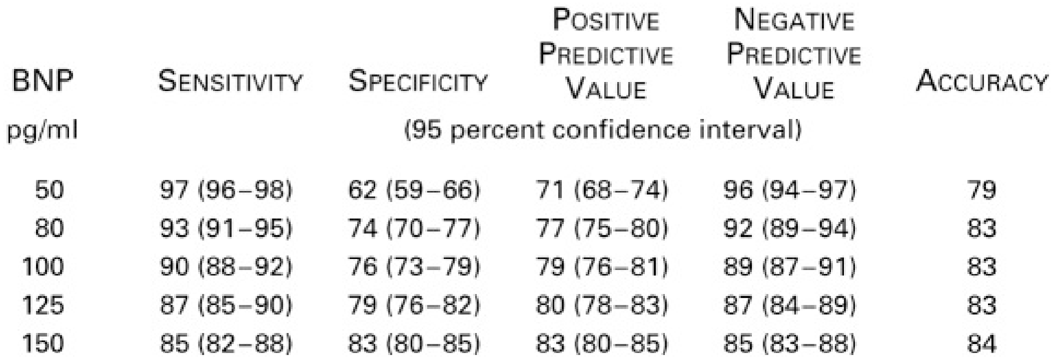
BNP is a cardiac peptide secreted in the heart in response to volume expansion. Therefore, BNP levels can be used to detect cardiac problems. Furthermore, patients presenting with dyspnea (difficult breathing) may be experiencing this symptom due to congestive heart failure.

BNP is a potential diagnostic tool for congestive heart failure.

(Florkowski 2008)

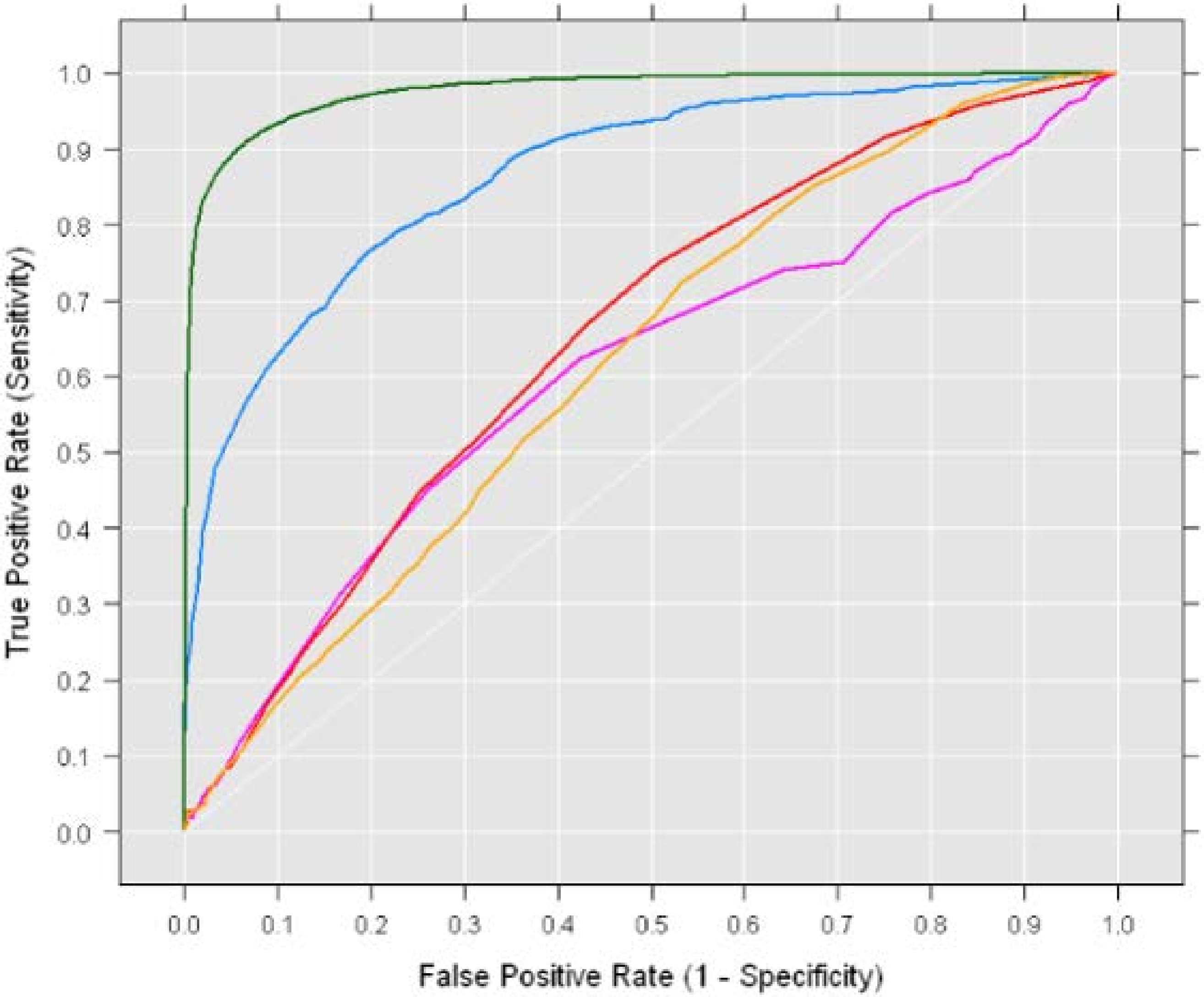
5. Optimizing Classification





BNP	SENSITIVITY	SPECIFICITY	POSITIVE PREDICTIVE VALUE	NEGATIVE PREDICTIVE VALUE	ACCURACY
pg/ml		(95 percent confidence interval)			
50	97 (96−98)	62 (59−66)	71 (68−74)	96 (94−97)	79
80	93 (91−95)	74 (70−77)	77 (75−80)	92 (89−94)	83
100	90 (88−92)	76 (73−79)	79 (76−81)	89 (87−91)	83
125	87 (85−90)	79 (76−82)	80 (78−83)	87 (84−89)	83
150	85 (82−88)	83 (80−85)	83 (80−85)	85 (83−88)	84

A cutoff can be designed here to be highly specific, highly sensitive, or a combination of the two.



For the green curve, a cutpoint can be chosen with high sensitivity and specificity.

For the red curve, there is more of a tradeoff between the two metrics.

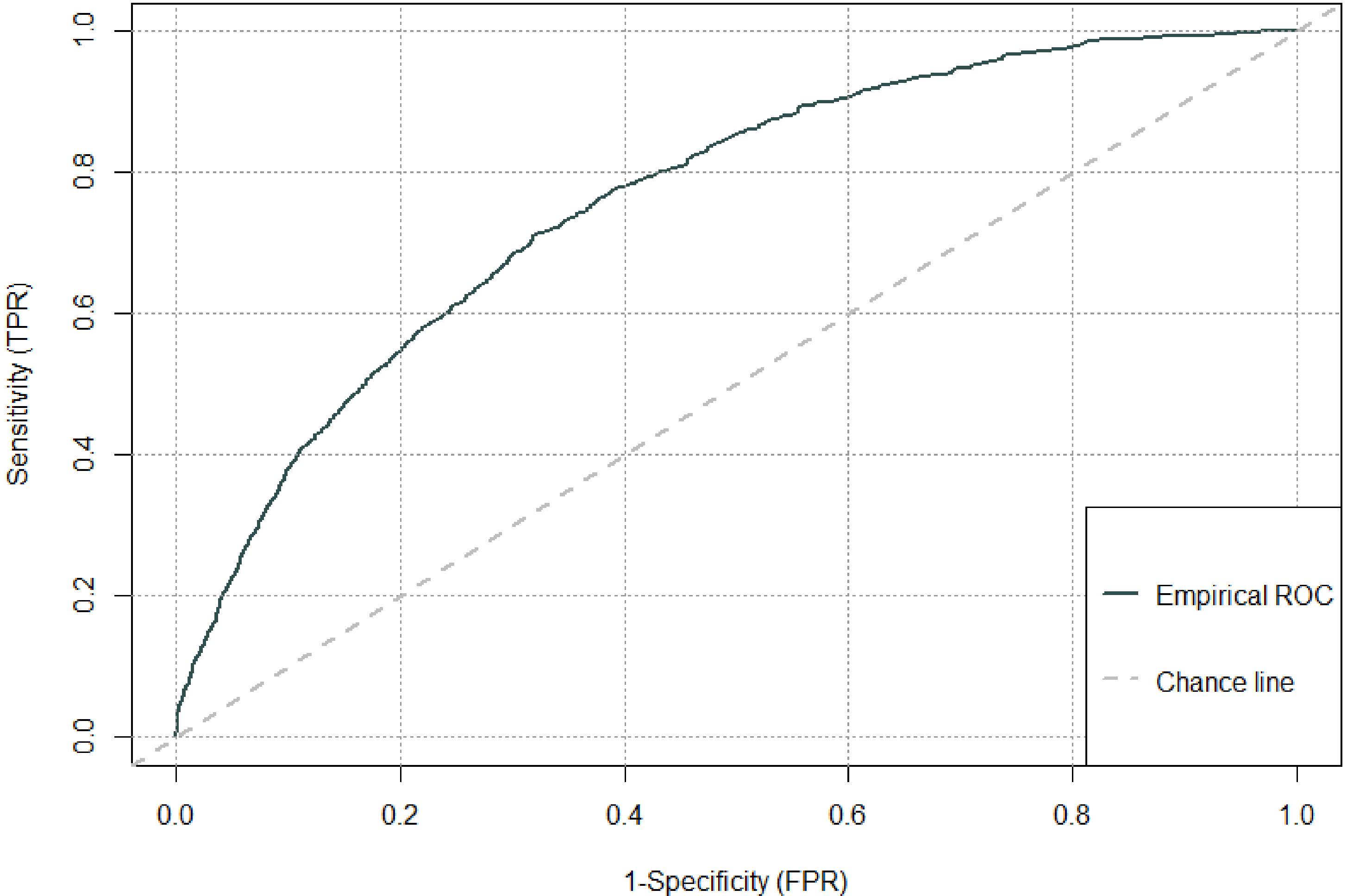
The white curve (y=x), represents a poor instrument.

5. Optimizing Classification

General Guideline for an instrument’s discriminative ability.

AUROC (Area Under ROC)	Classification
0.5	Useless (essentially a coin flip)
0.5-0.7	Poor
0.7-0.8	Acceptable-Good
0.8-0.9	Excellent
0.9-1.0	Nearly perfect

The ROC for our model indicates acceptable ability to discriminate against those with and without outcome.



```
> summary(roc_empirical)

 Method used: empirical
 Number of positive(s): 1084
 Number of negative(s): 9958
 Area under curve: 0.7609

 > ciAUC(roc_empirical)

   estimated AUC : 0.760899653081688
   AUC estimation method : empirical

   CI of AUC
   confidence level = 95%
   lower = 0.743807020194161
   upper = 0.777992285969216
```

We can also examine optimal cutpoints using differing criteria:

```
> optimal.cutpoints(X = "score", status = "class",
+                   data = data.frame(model_output),
+                   methods = c("Youden", "MaxSpSe", "MaxProdSpSe"), tag.healthy = 0)

Call:
optimal.cutpoints.default(X = "score", status = "class", tag.healthy = 0,
    methods = c("Youden", "MaxSpSe", "MaxProdSpSe"), data = data.frame(model_output))

Optimal cutoffs:
  Youden MaxSpSe MaxProdSpSe
1 0.1035  0.1058      0.1035

Area under the ROC curve (AUC):  0.761 (0.747, 0.775
```

Youden = max(Sp + Se – 1)
MaxSpSe = max(min(Sp, Se))
MaxProdSpSe = max(Sp*Se)

Recap

- Discrimination is another tool for assessing prediction models, in addition to correct classification rates.
- Models that best discriminate between those with Y=0 and Y=1 maximize both sensitivity and specificity.
- ROC curves show the tradeoff between sensitivity and specificity for different cutpoints.
- Higher area under the ROC curve (AUROC) indicates a model with better discrimination.

Recap

- Use AUC as a metric to explain a model’s discriminant ability

Additional Reading

- More on the Youden Index: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.200410135

Packages and Functions

- `DescTools::Conf()`
- `ROCit::measureit()`
- `ROCit::rocit()`
- `ROCit::ciAUC()`
- `OptimalCutpoints::optimal.cutpoints()`
- `plotROC`