Explanatory and Predictive Power for Logistic Regression Models

Andrew Zieffler
Department of Educational Psychology

Explanatory/Predictive Power of a Model

In ordinary regression, the measures *R* and R^2 provide measure of explanatory/predictive power of a model.

- R^2 provides measure of the reduction in variation (via the sum of squares) between the fitted model and the marginal model.
- Values closer to 1 would indicate better explanation/prediction

The comparable analogs used in logistic regression are not as useful.

In this unit we will examine some of the alternative ways to summarize the explanatory/predictive power of a logistic regression model

- Pseudo R^2 measures
- Classification tables
- ROC curves

Pseudo R^2 Measures

In ordinary regression, R^2 is

$$R^2 = 1 - \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{\sum\left(Y_i - \bar{Y}\right)^2}$$

There are several advantages and interpretations of R^2

- **Explained variation:** Since the denominator (of the fraction) represents the variation in the data and the numerator the variation *not* explained by the model, *R2* can be thought of as what proportion of variation the model does explain
- **Improvement from the null model to the fitted model:** The denominator (of the fraction) also represents the proportion of variation in the data that is explained by the simplest (null) model, $Y \sim 1$.
- **Square of the multiple correlation coefficient:** This is the interpretation that leads to the name, "*R-squared*". $R_{Y,\hat{Y}}$

The Y-hat values from logistic regression are based on likelihood estimates. These are not based on minimizing the sum of squared error, so the R^2 measures do not apply as fit measures.

Pseudo *R²* values

- They resemble R^2, in that they are on the same scale, namely [0, 1]
- Bigger values indicate better fit (higher predictive/explanatory power)
- Not interpreted the same way as ordinary R^2
- Different methods of producing the Pseudo R^2 often give very different values

X_1

```
# Fit Model K
> glm.k = glm(racture ~ momfrac + prior + age + age:prior, data = glow,
     family = binomial(link = "logit"))
```

```
> summary(glm.k)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.89631    1.09870  -5.367 8.02e-08 ***
momfrac      0.70771    0.29674   2.385  0.01708 *  
prior        5.34763    1.83323   2.917  0.00353 ** 
age          0.06400    0.01561   4.101 4.12e-05 ***
prior:age   -0.06275    0.02533  -2.477  0.01326 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 517.80  on 495  degrees of freedom
AIC: 527.8

Number of Fisher Scoring iterations: 4
```

Efron's Pseudo R-Squared

$$R^2 = 1 - \frac{\sum \left(Y_i - \hat{\pi}_i\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2}$$

where $\hat{\pi}_i$ is the predicted probability from the logistic model

- This measure is also equal to the squared correlation between the predicted and observed values.

- The scales between the observed values of Y (discrete; 0 or 1) is not on the same scale as the predicted probabilities (continuous). Thus differences (expressed in the numerator) are not easily interpretable

```
# Observed values
> y = glow$fracture

# Fitted/predicted values
> pi_hat = fitted(glm.k)

# Efron's pseudo R-squared
> 1 - sum((y - pi_hat) ^ 2) / sum((y - mean(y)) ^ 2)
[1] 0.08670543
```

McFadden's Pseudo R-Squared

$$R^2 = 1 - \frac{\ell\left(\mathrm{M}_{\mathrm{Fitted}}\right)}{\ell\left(\mathrm{M}_{\mathrm{Null}}\right)}$$

where $\ell\left(\mathrm{M}_{\mathrm{Fitted}}\right)$ is the estimated log-likelihood of the fitted model and $\ell\left(\mathrm{M}_{\mathrm{Null}}\right)$ is the estimated log-likelihood of the model with no predictors.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.89631    1.09870  -5.367 8.02e-08 ***
momfrac      0.70771    0.29674   2.385  0.01708 *  
prior        5.34763    1.83323   2.917  0.00353 ** 
age          0.06400    0.01561   4.101 4.12e-05 ***
prior:age   -0.06275    0.02533  -2.477  0.01326 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 517.80  on 495  degrees of freedom
AIC: 527.8

Number of Fisher Scoring iterations: 4
```

$$\text{Deviance} = -2 \times \ell(\mathrm{M})$$

$$\ell(\mathrm{M}_{\mathrm{Fitted}}) = \frac{517.80}{-2} = -258.9$$

$$\ell(\mathrm{M}_{\mathrm{Null}}) = \frac{562.34}{-2} = -281.17$$

$$R^2 = 1 - \frac{-258.9}{-281.17} = 0.079$$

There is also an Adjusted McFadden's Pseudo R-squared. This adjusts the numerator by subtracting the number of predictors

Cox and Snell's Pseudo R-Squared

$$R^2 = 1 - \left[\frac{\mathcal{L}\left(\mathrm{M}_{\mathrm{Null}}\right)}{\mathcal{L}\left(\mathrm{M}_{\mathrm{Fitted}}\right)} \right]^{2/N}$$

where $\mathcal{L}$ is the estimated likelihood of the given model.

- The estimated likelihood of a model is the *conditional probability* of Y given the predictors.

- The model likelihood is the product of N conditional probabilities. Taking the Nth root gives an estimate of the likelihood for each observation.

- The maximum value of this pseudo R-squared measure is *not* 1. If the fitted model predicts perfectly then $\mathcal{L}\left(\mathrm{M}_{\mathrm{Fitted}}\right) = 1$ and the pseudo R-squared will be < 1.

$$\mathcal{L}(\mathrm{M}) = exp \left[\frac{\mathrm{Deviance}}{-2} \right]$$

```
# Cox and Snell’s pseudo R-squared
> 1 - ( exp(562.34 / -2) / exp(517.8 / -2) ) ^ (2/500)
[1] 0.08522761
```

Nagelkerke / Cragg and Uhler's Pseudo R-Squared

$$R^2 = \frac{1 - \left[\frac{\mathcal{L}(\mathrm{M}_{\mathrm{Null}})}{\mathcal{L}(\mathrm{M}_{\mathrm{Fitted}})}\right]^{2/N}}{1 - \mathcal{L}\left(\mathrm{M}_{\mathrm{Null}}\right)^{2/N}}$$

- This measure adjusts Cox and Snell's pseudo R-squared so that the maximum is 1

```
# Nagelkerke’s pseudo R-squared
> (1 - ( ( exp(562.34 / -2) / exp(517.8 / -2) ) ^ (2/500) )) / ( 1 - exp(562.34 / -2) ^ (2/500) )
[1] 0.1262176
```

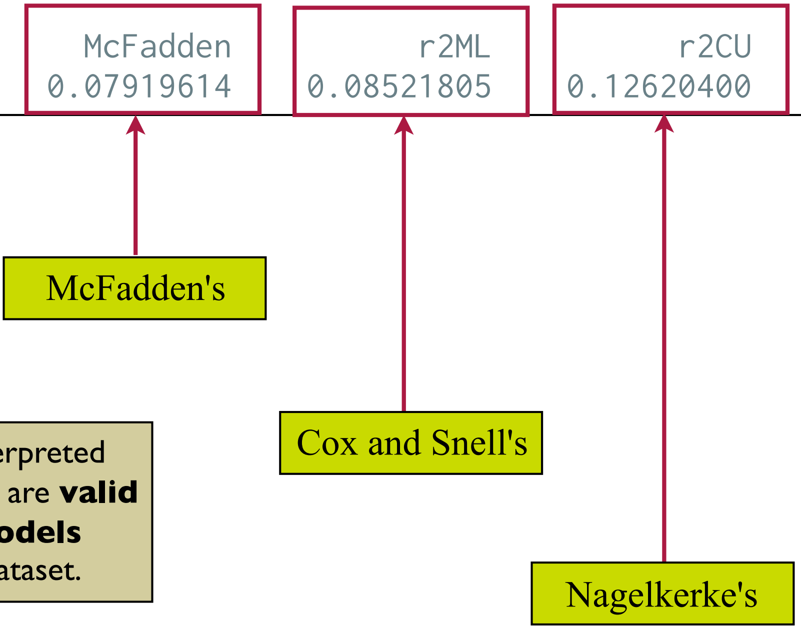
Using R to Obtain Pseudo R² Measures

```
# Load pscl library
> library(pscl)

# Get pseudo R-squared measures
> pR2(glm.k)

          llh       llhNull            G2      McFadden          r2ML          r2CU
-258.90018597 -281.16757231   44.53477267    0.07919614    0.08521805    0.12620400
```

Note that the **pseudo R-squared values vary greatly** from each other within the same model.



McFadden's

Cox and Snell's

Nagelkerke's

While pseudo R-squared values cannot be interpreted independently or compared across datasets, they are **valid and useful in evaluating multiple models** predicting the same outcome on the same dataset.

```
# Alternative library
> library(BaylorEdPsych)
> PseudoR2(glm.k)
```

	Model K	
Predictor	B	SE
Mother hip fracture	0.71*	0.30
Age	0.06***	0.02
Prior fracture	5.35**	1.83
Age x Prior fracture	–0.06*	0.03
(Intercept)	–5.90***	1.10
Model evaluation		
AIC	527.8	
BIC	548.9	
Pseudo R-squared		
Cox and Snell's	0.085	
Efron's	0.087	
McFadden's	0.079	
Nagelkerke	0.126	

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Classification Accuracy

One way of thinking about predictive power is to **cross-classify** the outcome with the a dichotomous variable whose values are derived from the predicted probabilities from the model

To derive the dichotomous variable we compare each of the predicted values from the model with some cutpoint, *c*.

- If $\hat{\pi}_i > c$ then = 1
- else = 0

```
# First 10 fitted values
> fitted(glm.k)

0.1269606
0.1498098
0.5670090
0.3434319
0.1200341
0.3859294
0.3728450
0.3904078
0.3916051
0.1011859
```

Classified as 1 if *c* = 0.5

```
# Classification using c = 0.5
> table(fitted(glm.k) > 0.5)

FALSE  TRUE
  480    20
```

Classification is predicting *group membership*. In this case, the model is predicting whether or not a subject would have a fracture.

- The model predicts 480 subjects would not have a fracture
- The model predicts 20 subjects would have a fracture.

Classification Accuracy

```
# Cross-classify with fitted values and outcome (using c = 0.5)
> table(fitted(glm.k) > 0.5, glow$fracture)

          0   1
  FALSE 364 116
  TRUE   11   9
```

The cross-classification shows how well the model predicts by showing whether the predicted classification matches the observed outcome.

- The model predicts correctly for 364 + 9 = 373 subjects

- The model predicts incorrectly for 11 + 116 = 127 subjects

The overall rate of correct classification is 373/500 = 74.6%.

- ✓ For subjects without a fracture the correct classification rate is 364/375 = 97% (**specificity**).

- ✓ For subjects with a fracture the correct classification rate is 9/125 = 7.2% (**sensitivity**)

Classification Table Based on Logistic Regression Model Using a Cutpoint of 0.5.

	Observed		
Classified	Fracture = 0	Fracture = 1	Total
Fracture = 0	364	116	480
Fracture = 1	11	9	20
Total	375	125	500

Classification is sensitive to the relative sizes of the two groups of the outcome.

- ✓ Generally favors classification into the larger group (Fracture = 0).
- ✓ For subjects with a fracture the correct classification rate is 9/125 = 7.2% (**sensitivity**)

The measures used from the cross-classification (e.g., sensitivity, specificity) should **not be used as measures of model fit**.

- They are heavily dependent on the distribution of the estimated probabilities in the sample
  - ✓ Differences in the sensitivity or specificity between models may be a result of "patient mix" rather than superiority of one model versus another

Classification Table Based on Logistic Regression Model Using a Cutpoint of 0.5.

	Observed		
Classified	Fracture = 0	Fracture = 1	Total
Fracture = 0	364	116	480
Fracture = 1	11	9	20
Total	375	125	500

Sensitivity = 9/125 = **7.2%**

Specificity = 364/375 = **97.0%**

Overall = (9 + 364)/500 = **74.6%**

Classification Table Based on Logistic Regression Model Using a Cutpoint of 0.40.

	Observed		
Classified	Fracture = 0	Fracture = 1	Total
Fracture = 0	356	108	464
Fracture = 1	19	17	36
Total	375	125	500

Sensitivity = 17/125 = **13.6%**

Specificity = 356/375 = **94.9%**

Overall = (17 + 356)/500 = **74.6%**

Classification Table Based on Logistic Regression Model Using a Cutpoint of 0.30.

	Observed		
Classified	Fracture = 0	Fracture = 1	Total
Fracture = 0	266	52	318
Fracture = 1	109	73	182
Total	375	125	500

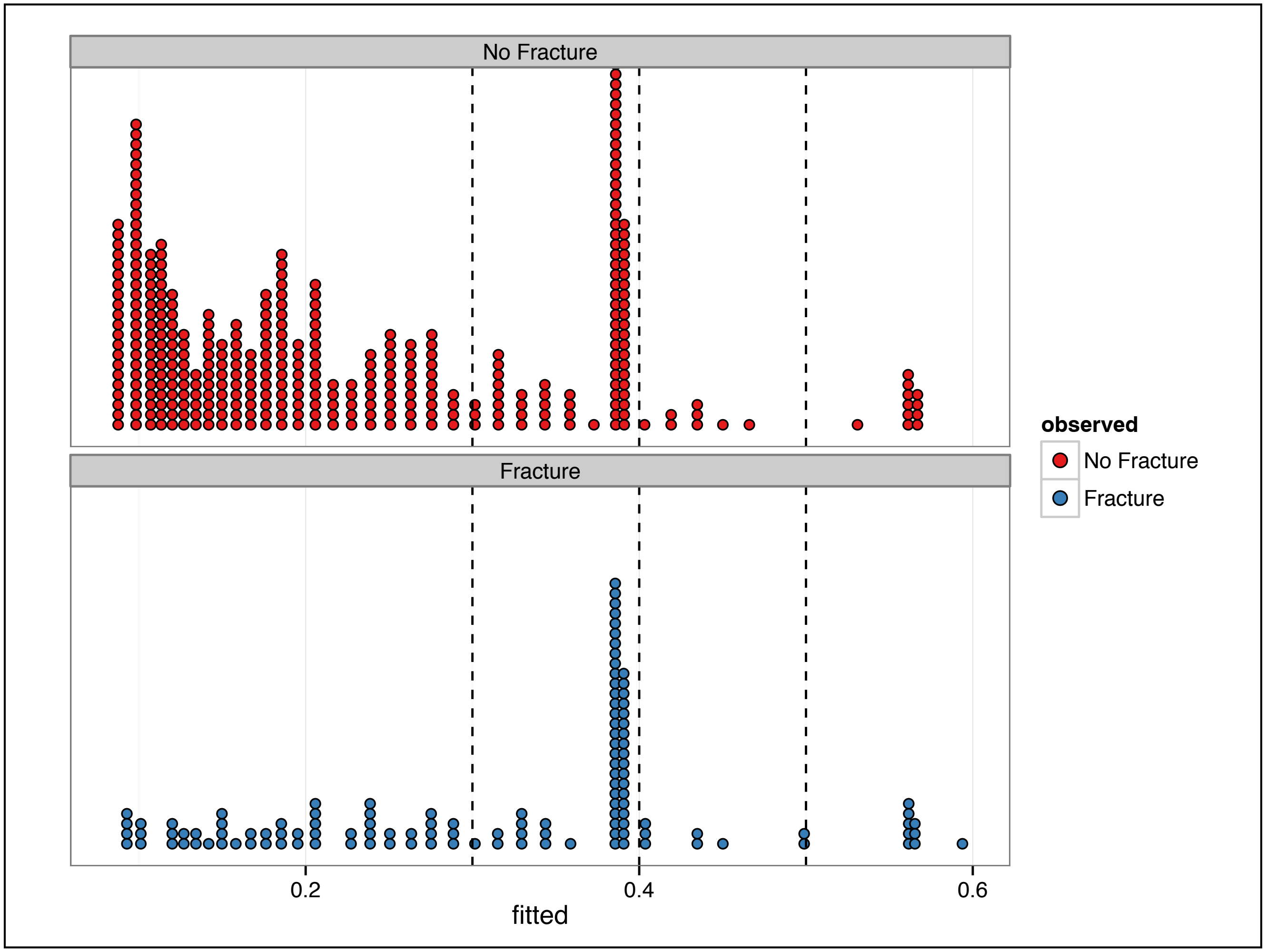
Sensitivity = 73/125 = **58.4%**

Specificity = 266/375 = **70.9%**

Overall = (73 + 266)/500 = **67.8%**

- Sensitivity and specificity are inversely related (as one gets larger, the other gets smaller)
- Sensitivity, specificity, and overall correct classification rate are dependent on the cutpoint. (Different cutpoints give different values)

Cutpoint	Sensitivity	Specificity	Overall
0.30	58.4%	70.9%	67.8%
0.40	13.6%	94.9%	74.6%
0.50	7.2%	97.0%	74.6%



The fitted values for both observed groups seems to be distributed across the entire range [0.0, 0.6]

✓ No matter where we put the cutpoint there will be a great deal of misclassification.

Pseudo R² to Accompany Classification Table

Count Pseudo R-Squared

$$R^2 = \frac{\#\ \text{Correct}}{\text{Total}}$$

where # *Correct* is the number of correct predictions based on converting the predicted probabilities from the fitted model into either 0 or 1.

- Any case with a predicted probability of 0.5 or greater gets a predicted outcome of 1 and any record with a predicted probability less than 0.5 gets a predicted outcome of 0.

- This is the overall correct classification rate

w/a cutpoint of 0.5

$$R^2 = \frac{364 + 9}{500} = 0.746$$

Classification Table Based on Logistic Regression Model Using a Cutpoint of 0.5.

	Observed		
Classified	Fracture = 0	Fracture = 1	Total
Fracture = 0	364	116	480
Fracture = 1	11	9	20
Total	375	125	500

One thing to note is that you can always get an okay classification rate by just predicting every observation into the largest category.

- For example, if you predicted every patient would not have a fracture you would have a correct classification rate of 375/500 = 75%
  - ✓ This is actually better than the model classification rate of 74.6%!

Adjusted Count Pseudo R-Squared

$$R^2 = \frac{\#\ \text{Correct} - n}{\text{Total} - n}$$

where n is the count of the most frequent outcome

- One prediction model would be to predict every outcome as having the same value as the most frequently occurring outcome
- The adjustment here controls for that by computing the improvement over such a model

w/a cutpoint of 0.5

$$R^2 = \frac{364 + 9 - 375}{500 - 375} = -0.016$$

Plot of Sensitivity and Specificity

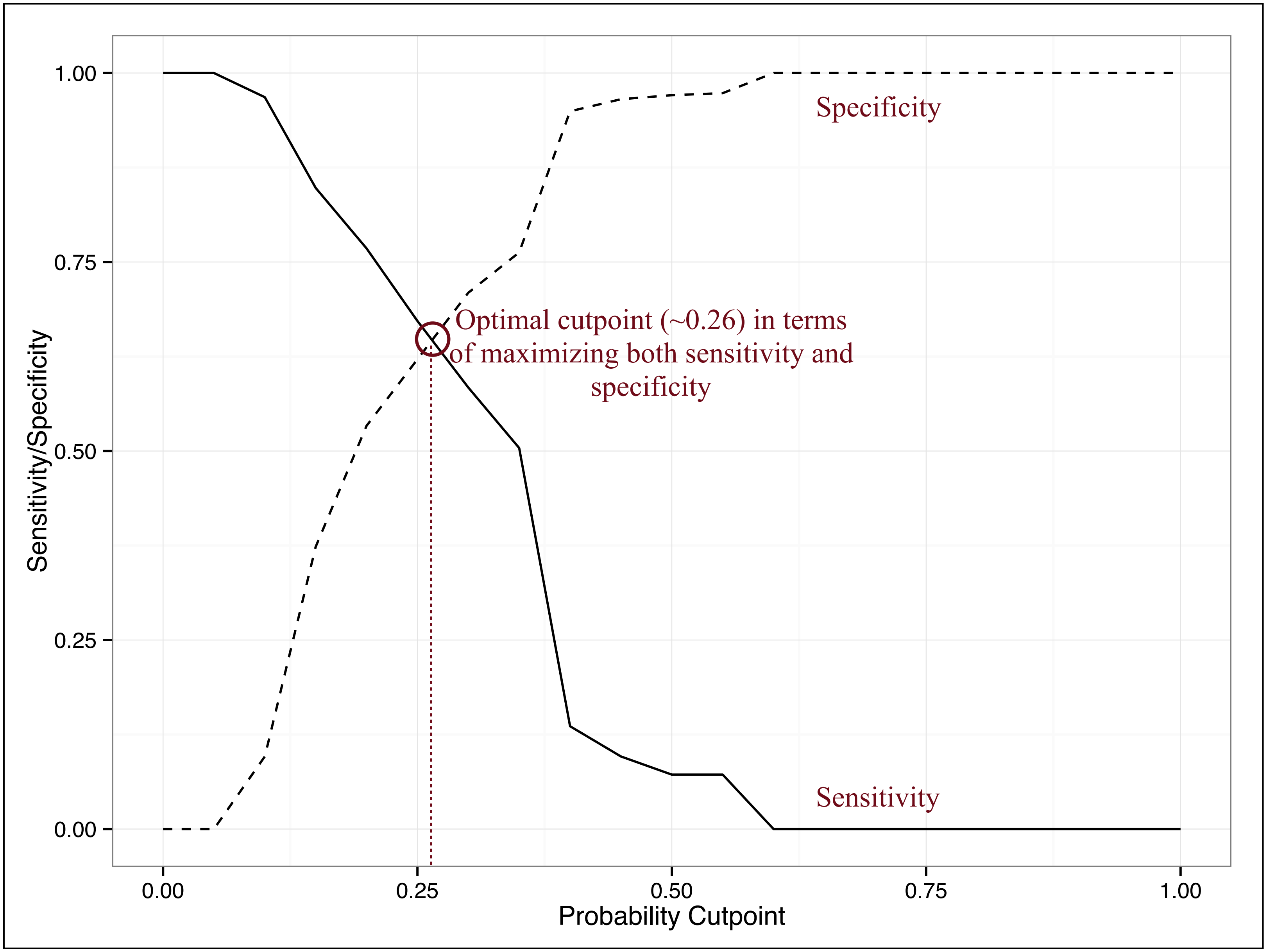
As a Function of the Cutpoint

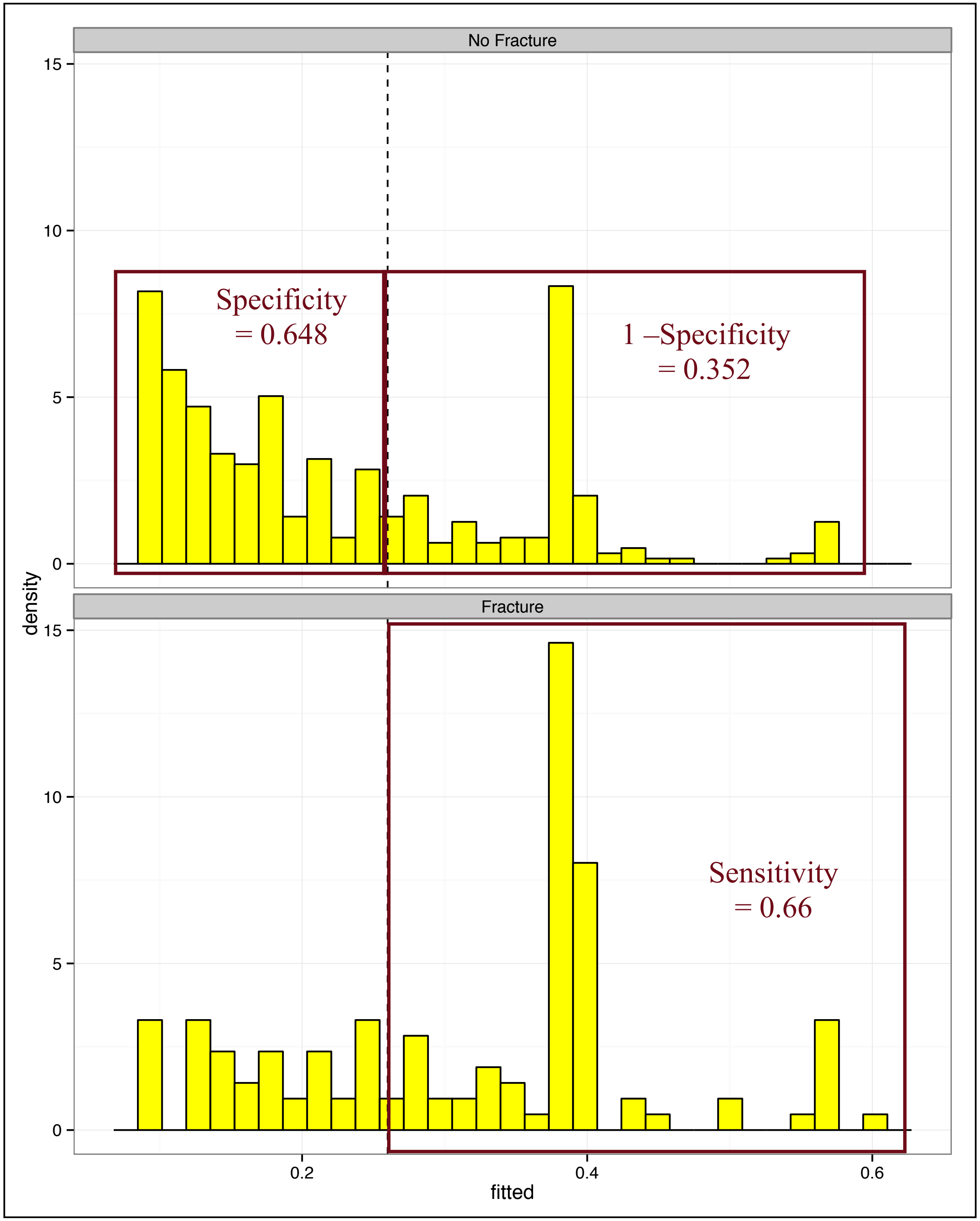
```
# Classification using c = 0.5
> new = data.frame(
     cutpoint = seq(0, 1, by = 0.05)
     )

> for(i in 1:nrow(new)){
     new$sensitivity[i] = sum(fitted(glm.k) > new$cutpoint[i] & glow$fracture == 1) / 125
     new$specificity[i] = sum(fitted(glm.k) <= new$cutpoint[i] & glow$fracture == 0) / 375
     }

> head(new)

   cutpoint sensitivity specificity
1      0.00       1.000   0.0000000
2      0.05       1.000   0.0000000
3      0.10       0.968   0.0960000
4      0.15       0.848   0.3733333
5      0.20       0.768   0.5333333
6      0.25       0.672   0.6213333
```

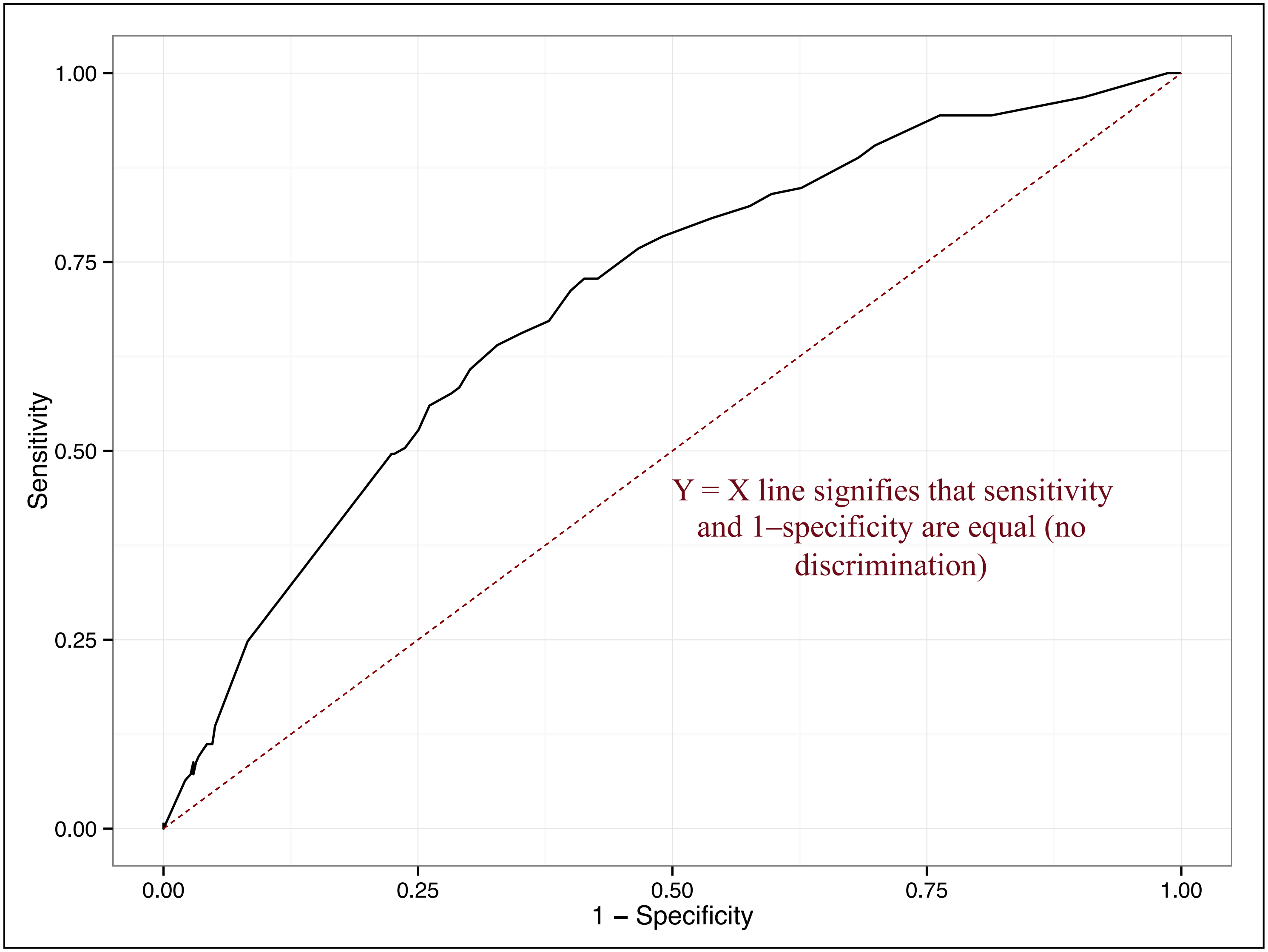

Sensitivity/Specificity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Probability Cutpoint
Specificity
Sensitivity
Optimal cutpoint (~0.26) in terms of maximizing both sensitivity and specificity



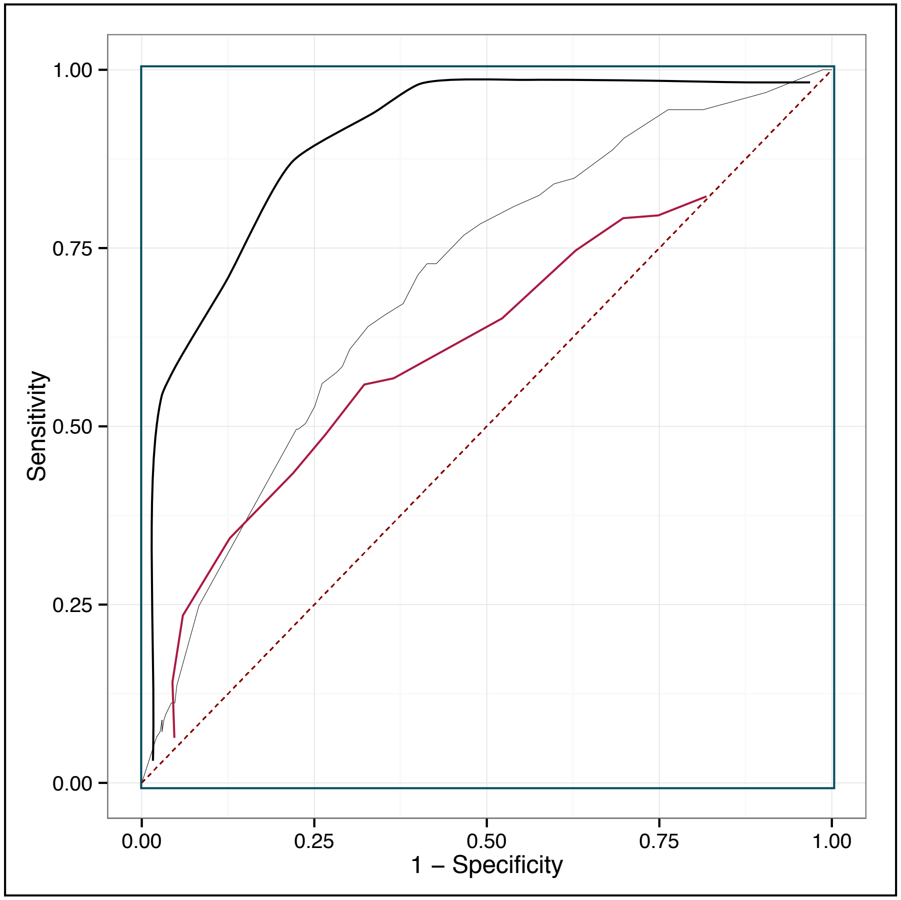
- Cutpoint is at 0.26.
- If the two distributions were identical then the area to the right of the cutpoint would be the same in both histograms.
  - ✓ Sensitivity = 1–Specificity
  - ✓ This would indicate the model could not really discriminate between subjects with and without fractures
- Perfect discrimination is when the histograms do not overlap at all.
  - ✓ Sensitivity = 100%
  - ✓ 1–Specificity = 0%
- To examine the discrimination ability of the model we could plot the sensitivity versus 1–Specificity for all cutpoints

Receiver Operating Characteristic (ROC) Curve

Function of the Sensitivity versus 1–Specificity



Y = X line signifies that sensitivity and 1–specificity are equal (no discrimination)

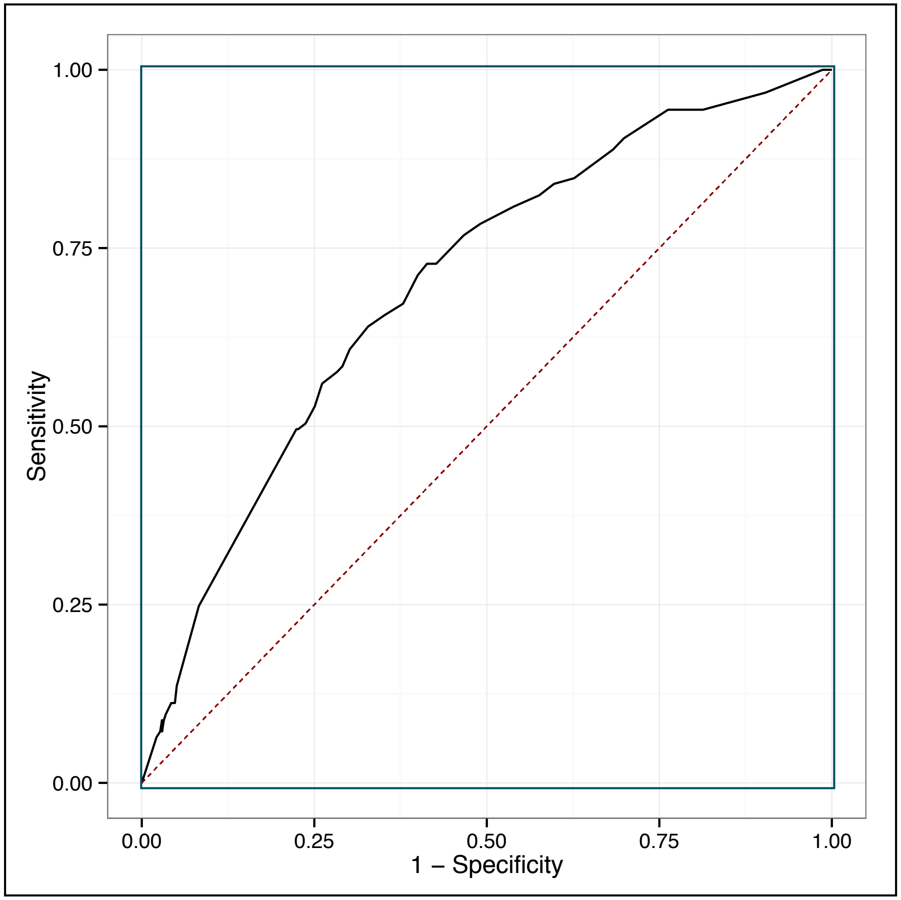


The pink ROC curve shows less discrimination than the black curve

The closer the AUC gets to 1 the better the discrimination.

Interpreting ROC Curve

Examine are under curve



The estimated probability that a subject who has a fracture will have a higher predicted probability than a subject who does not have a fracture using the fitted model is 0.696.

Imagine the 1x1 square (in blue)

The area under the Y = X line is 0.5. This is indicative of the discrimination for a 50/50 random model (i.e., you flip a coin to categorize patients).

The area under the ROC line represents the discrimination of the model.
Area Under Curve (AUC) = 0.696

The area under the ROC line represents the estimated probability that a subject who has an outcome of 1 will have a higher predicted probability than a subject who has an outcome value of 0

```
# Get ROC curve and AUC
> library(epicalc)

> lroc(glm.k)

$model.description
[1] ”fracture ~ momfrac + prior + age + age:prior”

$auc
[1] 0.696096

$predicted.table
 predicted.prob Non-diseased Diseased
         0.0850            5        0
         0.0901           16        2
         0.0955           15        2
         0.1011           16        3
         0.1071           18        0
            ....             ....      ....

$diagnostic.table
   1-Specificity Sensitivity
     1.000000000       1.000
>    0.986666667       1.000
>    0.944000000       0.984
>    0.904000000       0.968
>    0.861333333       0.944
>    0.813333333       0.944
           ....            ....
```

- What value for the AUC do we need for "good" discrimination?
  - ✓ No magic number...only rule of thumb

  - ✓ ROC = 0.5 (No discrimination)
  - ✓ 0.5 < ROC < 0.7 (Poor discrimination)
  - ✓ 0.7 < ROC < 0.8 (Acceptable discrmination)
  - ✓ 0.8 < ROC < 0.9 (Excellent discrimination)
  - ✓ ROC ≥ 0.9 (Outstanding discrimination)

For our fitted model, the AUC is 0.696. This is indicative of poor discrimination.

"...the classification table is most appropriate when classification is a stated goal of the analysis; otherwise, it should only supplement more rigorous methods of assessment of fit."

– Hosmer, Lemeshow & Sturdivant (2013)