STAT021 Statistical Methods II

Lecture 23 Model Inference and Assessment

Lu Chen
Swarthmore College

12/6/2018

Review

- **Binary response variable** $Y = 1$ or 0.
- **Bernoulli distribution** for binary data $Y \sim Bernoulli(\pi)$
  - $\pi = P(Y = 1)$; mean of Y is π and SD of Y is $\sqrt{\pi(1-\pi)}$.
- **Logistic regression model**

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K \text{ or } \pi = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_KX_K}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_KX_K}}$$

  where $\pi = P(Y = 1|X_1, X_2, \cdots, X_K)$.

- **Probability** π, **odds** $\frac{\pi}{1-\pi}$, **log-odds** $\log\frac{\pi}{1-\pi}$ and **odds ratio** (ratio of two odds).
- **Empirical probability** (from data) and **estimated probability** (from the logistic regression model).

Outline

- Examples
  - Teenager sleep $Sleep \sim Age$
  - Medical school $Acceptance \sim GPA$
- Inference for slope(s): z (Wald) test and confidence interval
- Inference for the model: likelihood ratio test (LRT)
- Model assessment: AIC
- Predictive accuracy
  - Sensitivity, specificity, ROC curve and AUC

Teenager sleep *Sleep ~ Age*

```r
teenagersleep <- glm(Sleep ~ Age, family="binomial", data=TeenSleep)
summary(teenagersleep)$coefficients
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.1186     1.3337    2.34    0.019 *
## Age          -0.1514     0.0823   -1.84    0.066 .
```

- $b_0 = 3.12$, $b_1 = -0.15$. **Estimated odds ratio:** $e^{b_1} = e^{-0.15} = 0.86$.
- The odds that teenagers at age $x + 1$ sleep at least 7 hours a night is 0.86 of the odds that teenagers at age x sleep at least 7 hours a night.
  - The odds that teenagers sleep at least 7 hours a night is 14% lower as age increases 1 year.

Medical school *Acceptance* ~ *GPA*

```r
medaccept <- glm(Acceptance ~ GPA, family="binomial", data=Med)
summary(medaccept)$coefficients
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -19.21       5.63   -3.41  0.00064 ***
## GPA             5.45       1.58    3.45  0.00055 ***
```

- $b_0 = -19.21$, $b_1 = 5.45$. **Estimated odds ratio**: $e^{b_1} = e^{5.45} = 233.76$.
- The odds of being accepted by medical schools is 233.76 times higher for every 1 unit increase in *GPA*.
- The odds of being accepted by medical schools is $e^{b_1 \times 0.1} = e^{0.545} = 1.72$ times higher for every 0.1 unit increase in *GPA*.
- Sometimes, interpretation of the odds ratio should be adapted to the actual meaning the variables.

z (Wald) test and confidence interval

To test whether the slope for the predictor X_k $(k = 1, 2, \cdots, K)$ in a logistic regression model is significantly different from zero, the hypotheses are $H_0 : \beta_k = 0, H_a : \beta_k \neq 0$ and the test statistic is

$$z = \frac{b_k}{SE_{b_k}} \sim N(0, 1)$$

Assuming we have a reasonably large sample (with independent, random outcomes), the P-value is determined from a Normal distribution. This z-statistic is also called the **Wald statistic**.

The confidence interval for the slope is

$$b_k \pm z^* SE_{b_k}$$

where z^* is found using the Normal distribution and the desired level of confidence.

z (Wald) test

Teenagers sleep example

```
summary(teenagersleep)$coefficients
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.1186     1.3337    2.34    0.019 *
## Age          -0.1514     0.0823   -1.84    0.066 .
```

- $H_0 : \beta_1 = 0; H_a : \beta_1 \neq 0$
- $b_1 = -0.15$
- $z = -1.84$, $P = 0.066 > 0.05$, P-value is close to 0.05.
- Age is marginally significantly associated with whether a teenager sleeps at least 7 hours a night.

Confidence interval

Teenagers sleep example

```r
confint(teenagersleep) # 95% confidence interval for the slope
```

```
##                  2.5 %      97.5 %
## (Intercept)  0.5219296 5.758495250
## Age         -0.3139621 0.009352902
```

```r
exp(confint(teenagersleep)) # 95% confidence interval for odds ratio
```

```
##                 2.5 %     97.5 %
## (Intercept) 1.6852764 316.871158
## Age         0.7305467   1.009397
```

- 95% CI for β_1: $[-0.31, 0.01]$ (contains 0 $\Leftrightarrow P > 0.05$)
- 95% CI for e^{β_1}: $[e^{-0.31}, e^{0.01}] = [0.73, 1.01]$ (contains 1 $\Leftrightarrow P > 0.05$)
- The odds that teenagers at age $x + 1$ sleep at least 7 hours a night is 0.86 (with 95% CI [0.73, 1.01]) of the odds that teenagers at age x sleep at least 7 hours a night.

z (Wald) test and confidence interval

The medical school acceptance example

```
summary(medaccept)$coefficients
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -19.21       5.63   -3.41  0.00064 ***
## GPA             5.45       1.58    3.45  0.00055 ***
```

```
exp(confint(medaccept))
```

```
##                    2.5 %       97.5 %
## (Intercept) 1.686955e-14 8.472476e-05
## GPA         1.482501e+01 7.829246e+03
```

- $z = 3.45$, $P = 0.00055 < 0.05$; An applicant's *GPA* is significantly associated with the acceptance by medical schools.
- $e^{b_1} = 233.76$ with 95% CI [14.83, 7829.25]. The odds of being accepted by medical schools is 233.76 times higher (significant with 95% CI [14.83, 7829.25]) for every 1 unit increase in *GPA*.

Likelihood ratio test (LRT)

The method of **maximum likelihood** chooses parameter values to maximize L, or, equivalently, to minimize $-2\log L$, which is called the **deviance**. To test the overall effectiveness of a logistic regression model with predictors $X_1, \cdots, X_K$,

$H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$ versus

H_a : at least one $\beta_k \neq 0$, where $k = 1, 2, \cdots, K$,

we use the test statistic

$$G = -2\log \hat{L}_0 - (-2\log \hat{L}) \sim \chi^2(K)$$

where $\hat{L}_0$ is the likelihood for a model without predictors and $\hat{L}$ is the likelihood using the logistic model. We compare this improvement in $-2\log L$ to a chi-square distribution with K degrees of freedom.

Likelihood ratio test (LRT)

```r
summary(teenagersleep)
```

```
##
## Call:
## glm(formula = Sleep ~ Age, family = "binomial", data = TeenSleep)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6205  -1.4161   0.8443   0.8991   1.0152
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.11864    1.33375   2.338   0.0194 *
## Age         -0.15136    0.08235  -1.838   0.0661 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 569.60  on 445  degrees of freedom
## Residual deviance: 566.19  on 444  degrees of freedom
## AIC: 570.19
##
```

- $-2\log\hat{L}_0 = 569.60$
- $-2\log\hat{L} = 566.19$
- $G = 569.60 - 566.19 = 3.41 \sim \chi^{2}$

Likelihood ratio test (LRT)

```r
library(lmtest)
lrtest(teenagersleep)
```

```
## Likelihood ratio test
##
## Model 1: Sleep ~ Age
## Model 2: Sleep ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   2 -283.1
## 2   1 -284.8 -1 3.4062    0.06495 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $G = (-2) \times (-283.1) - (-2) \times (-284.8) = 3.41 \sim \chi^2(1)\, P = 0.065 > 0.05$. The likelihood ratio test is marginally significant.
- The model with the predictor *Age* is marignally significant in explaining *Sleep*.
- Note: the likelihood ratio test for the model is usually close to but **not exactly the same** as the z (Wald) test for the slope ($P = 0.066$).

Likelihood ratio test (LRT)

```r
lrtest(medaccept)
```

```
## Likelihood ratio test
##
## Model 1: Acceptance ~ GPA
## Model 2: Acceptance ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -28.420
## 2   1 -37.896 -1 18.952   1.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $G = (-2) \times (-37.90) - (-2) \times (-28.42) = 18.95 \sim \chi^2(1)$, $P = 1.34 \times 10^{-5} < 0.05$.
- The model with the predictor *GPA* is highly significant in explaining the *Acceptance* to medical schools.
- In the z (Wald) test for the slope, $P = 0.00055$.

Logistic regression AIC

> Suppose that we have a statistical model with p parameters to be estimated ($p = K + 1$ in logistic regression, where K is the number of predictors). Let $\hat{L}$ be the maximized value of the likelihood function for the model. Then the AIC value of the model is computed by
>
> $$AIC = 2p - 2\log\hat{L} = 2(K+1) - 2\log\hat{L}$$
>
> Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

```
AIC(teenagersleep, medaccept)
```

```
##                df       AIC
## teenagersleep   2 570.19488
## medaccept       2  60.83901
```

Note: these 2 models are not comparable because they are built on different datasets. AIC can be compared between models built on the same data and with the same response variable.

Predictive accuracy

```r
logit_pi <- predict(medaccept)
est_pi <- exp(logit_pi)/(1+exp(logit_pi))
# Use 0.5 as the cutoff for predicting Y
pred_y <- as.numeric(est_pi >= 0.5)
head(data.frame(Med, est_pi, pred_y), 12)
```

```
##    Acceptance  GPA    est_pi pred_y
## 1           0 3.62 0.6312488      1
## 2           1 3.84 0.8503685      1
## 3           1 3.23 0.1694476      0
## 4           1 3.69 0.7149136      1
## 5           1 3.38 0.3161716      0
## 6           1 3.72 0.7470602      1
## 7           1 3.89 0.8818641      1
## 8           0 3.34 0.2709933      0
## 9           1 3.71 0.7366158      1
## 10          1 3.89 0.8818641      1
## 11          1 3.97 0.9203078      1
## 12          1 3.49 0.4572388      0
```

- Suppose we use cutoff $c = 0.5$ for predicting *Acceptance*.
  - For any $\hat{\pi} \geq 0.5, \hat{y} = 1$
  - For any $\hat{\pi} < 0.5, \hat{y} = 0$
- For the observed *Acceptance* y and predicted *Acceptance* $\hat{y}$ (`pred_y`), there are 4 scenarios:

1. $y = 1$ and $\hat{y} = 1 \Rightarrow$ True positive
2. $y = 0$ and $\hat{y} = 1 \Rightarrow$ False positive
3. $y = 1$ and $\hat{y} = 0 \Rightarrow$ False negative
4. $y = 0$ and $\hat{y} = 0 \Rightarrow$ True negative

Predictive accuracy

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	True positive	False negative
$y = 0$	False positive	True negative

$c = 0.5$	$\widehat{Acceptance} = 1$	$\widehat{Acceptance} = 0$
$Acceptance = 1$	24	6
$Acceptance = 0$	9	16

Sensitivity (**true positive rate**) measures the proportion of positives that are correctly identified as such.

$$\text{Sensitivity} = \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false negatives}}$$

Specificity (**true negative rate**) measures the proportion of negatives that are correctly identified as such.

$$\text{Specificity} = \frac{\#\text{ true negatives}}{\#\text{ true negatives} + \#\text{ false positives}}$$

Predictive accuracy

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	True positive	False negative
$y = 0$	False positive	True negative

$c = 0.5$	$\widehat{Acceptance} = 1$	$\widehat{Acceptance} = 0$
$Acceptance = 1$	24	6
$Acceptance = 0$	9	16

$$\text{Sensitivity} = \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false negatives}} = \frac{24}{24+6} = 0.8$$

$$\text{Specificity} = \frac{\#\text{ true negatives}}{\#\text{ true negatives} + \#\text{ false positives}} = \frac{16}{16+9} = 0.64$$

- This model using *GPA* to predict *Acceptance* to medical schools (cutoff $= 0.5$) has quite high sensitivity (80%) and relatively high specificity (64%).
- If we change the cutoff value, the values of sensitivity and specificity will also change.

Predictive accuracy

$c = 0.4$	$\widehat{Acceptance} = 1$	$\widehat{Acceptance} = 0$
$Acceptance = 1$	26	4
$Acceptance = 0$	9	16

$c = 0.5$	$\widehat{Acceptance} = 1$	$\widehat{Acceptance} = 0$
$Acceptance = 1$	24	6
$Acceptance = 0$	9	16

$c = 0.6$	$\widehat{Acceptance} = 1$	$\widehat{Acceptance} = 0$
$Acceptance = 1$	20	10
$Acceptance = 0$	7	18

- $c = 0.4$

  $\text{Sensitivity} = \frac{26}{26+4} = 0.87$

  $\text{Specificity} = \frac{16}{16+9} = 0.64$

- $c = 0.5$

  $\text{Sensitivity} = \frac{24}{24+6} = 0.80$

  $\text{Specificity} = \frac{16}{16+9} = 0.64$

- $c = 0.6$

  $\text{Sensitivity} = \frac{20}{20+10} = 0.67$

  $\text{Specificity} = \frac{18}{18+7} = 0.72$

Predictive accuracy

```r
# Medical school Acceptance ~ GPA
```

```
##       Cutoff Sensitivity Specificity
##  [1,]    0.0        1.00        0.00
##  [2,]    0.1        1.00        0.16
##  [3,]    0.2        0.93        0.24
##  [4,]    0.3        0.93        0.40
##  [5,]    0.4        0.87        0.64
##  [6,]    0.5        0.80        0.64
##  [7,]    0.6        0.67        0.72
##  [8,]    0.7        0.60        0.88
##  [9,]    0.8        0.43        1.00
## [10,]    0.9        0.10        1.00
## [11,]    1.0        0.00        1.00
```

- As cutoff increases, sensitivity decreases since less and less predicted probabilities will be categorized as positive.
- As cutoff increases, specificity increases since more and more predicted probabilities will be categorized as negative.
- We need a comprehensive measure for predictive accuracy that takes into account all sensitivity and specificity values given all cutoff values.

Predictive accuracy

```r
# Medical school Acceptance ~ GPA
```

```
##        Cutoff Sensitivity Specificity
##  [1,]     0.0        1.00        0.00
##  [2,]     0.1        1.00        0.16
##  [3,]     0.2        0.93        0.24
##  [4,]     0.3        0.93        0.40
##  [5,]     0.4        0.87        0.64
##  [6,]     0.5        0.80        0.64
##  [7,]     0.6        0.67        0.72
##  [8,]     0.7        0.60        0.88
##  [9,]     0.8        0.43        1.00
## [10,]     0.9        0.10        1.00
## [11,]     1.0        0.00        1.00
```

- In some studies, we prefer lower cutoff and higher sensitivity. A model with 100% sensitivity will identify $y = 1$ as positive for sure. A negative result will definitely indicate $y = 0$. It is useful in ruling out disease among patients.
- In other cases, we prefer higher cutoff and higher specificity. A model with 100% specificity will identify $y = 0$ as negative for sure. A positive result will definitely indicate $y = 1$. It is useful in ruling in disease among healthy people.

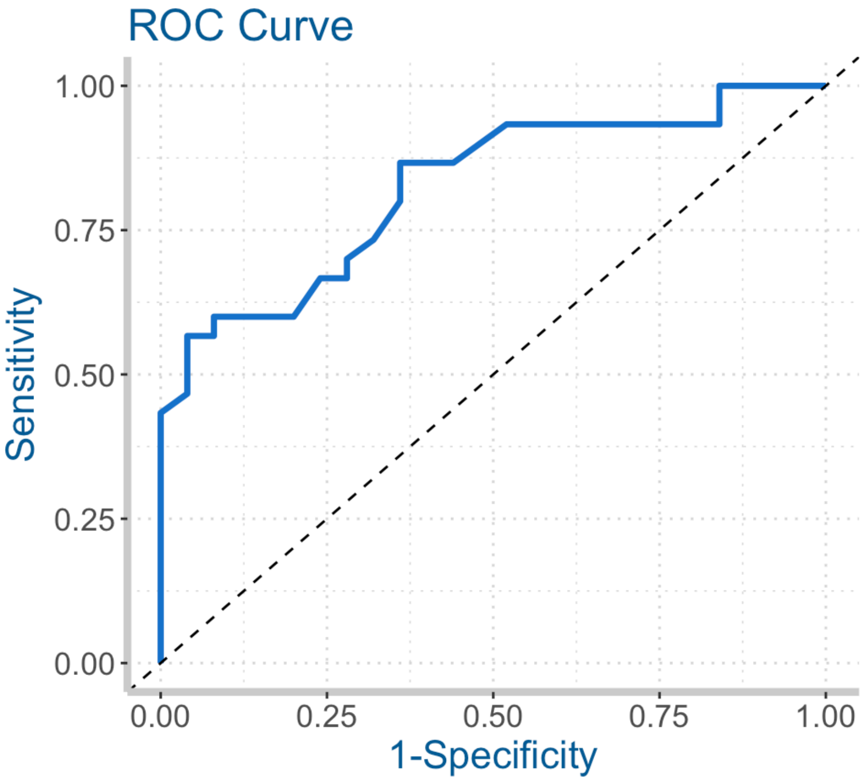
Receiver operating characteristic (ROC) curve

> The **receiver operating characteristic curve**, i.e. **ROC curve**, is created by plotting **sensitivity** against **1− specificity** at various cutoff settings. Sensitivity is also known as the true positive rate (**TPR**) and the 1− specificity is also known as false positive rate (**FPR**).

- ROC curve is a curve of sensitivity (TPR) versus 1− specificity (FPR).
- Since sensitivity is a decreasing function of specificity, sensitivity is an increasing function of 1− specificity.

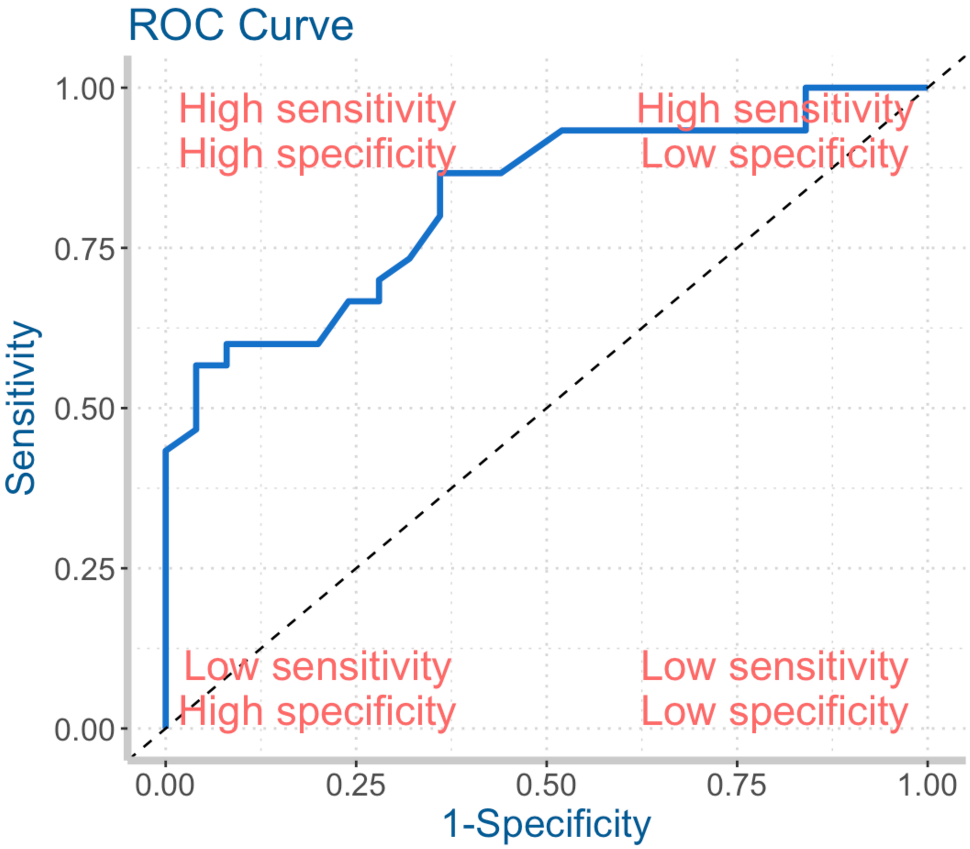
Receiver operating characteristic (ROC) curve

```r
library(ROCR) # install and library package ROCR
library(ggplot2)
ROC(medaccept, color="dodgerblue3", title="ROC Curve")
```



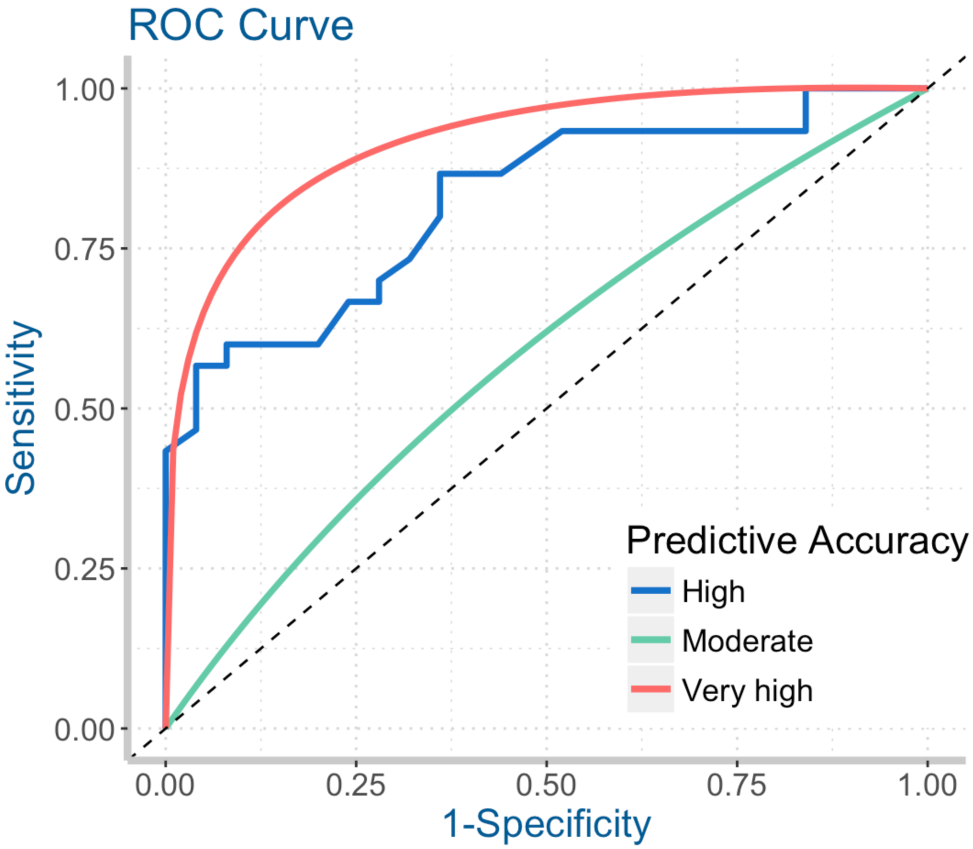
- As $1 - Specificity$ increases, $Sensitivity$ increases. The curve starts from $1 - Specificity = 0$ ($Specificity = 1$) and $Sensitivity = 0$, and ends at $1 - Specificity = 1$ ($Specificity = 0$) and $Sensitivity = 1$.
- We prefer large $Specificity$ (small $1 - Specificity$) and large $Sensitivity$.

Receiver operating characteristic (ROC) curve



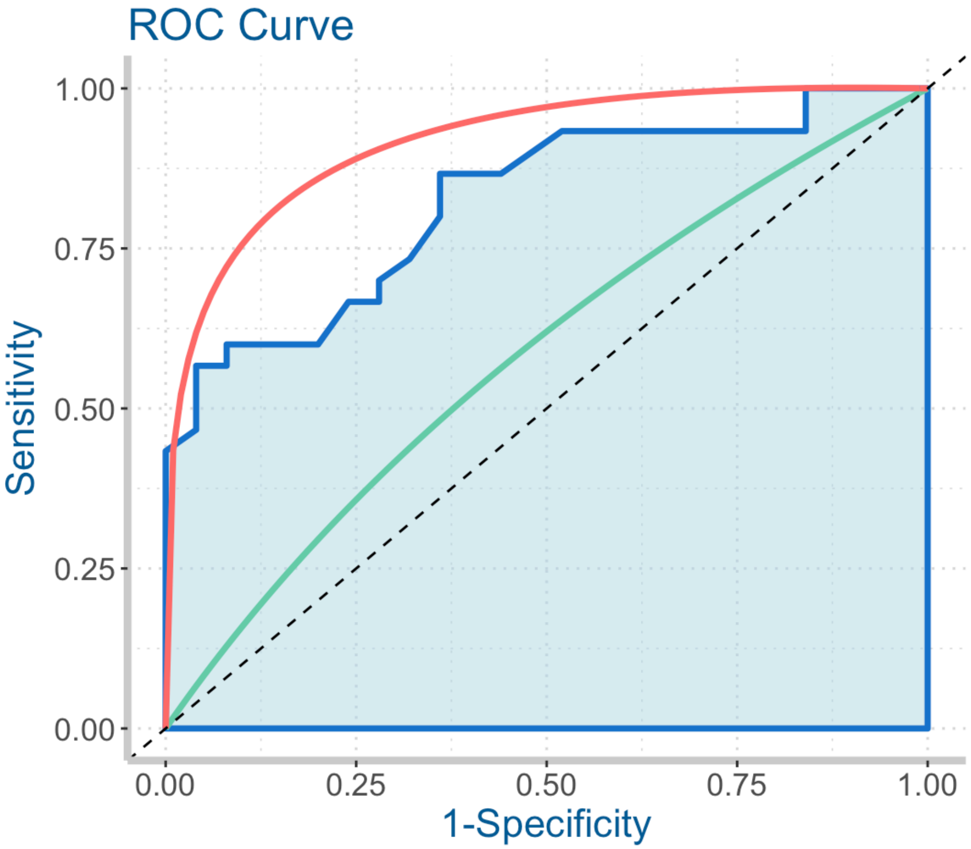
- The upper left corner is associated with high sensitivity and high specificity values.
- The ROC curve of a model with high predictive accuracy should be pulled very close to the upper left corner.

Receiver operating characteristic (ROC) curve



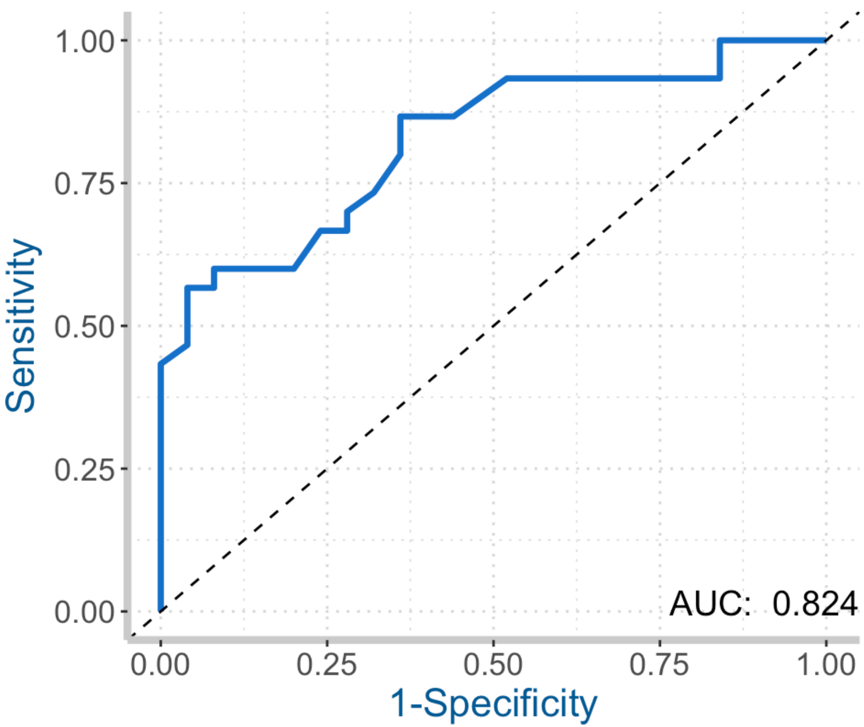
- The red ROC curve indicates a model with very high predictive accuracy.
- The green ROC curve indicates a model with moderate predictive accuracy.
- What the curve looks like in the worst case?
- A model with no predictive accuracy will perform the same as random guess, which results in $Sensitivity + Specificity = 1$and thus $Sensitivity = 1 - Specificity$, i.e. the $y = x$ line in the graph.

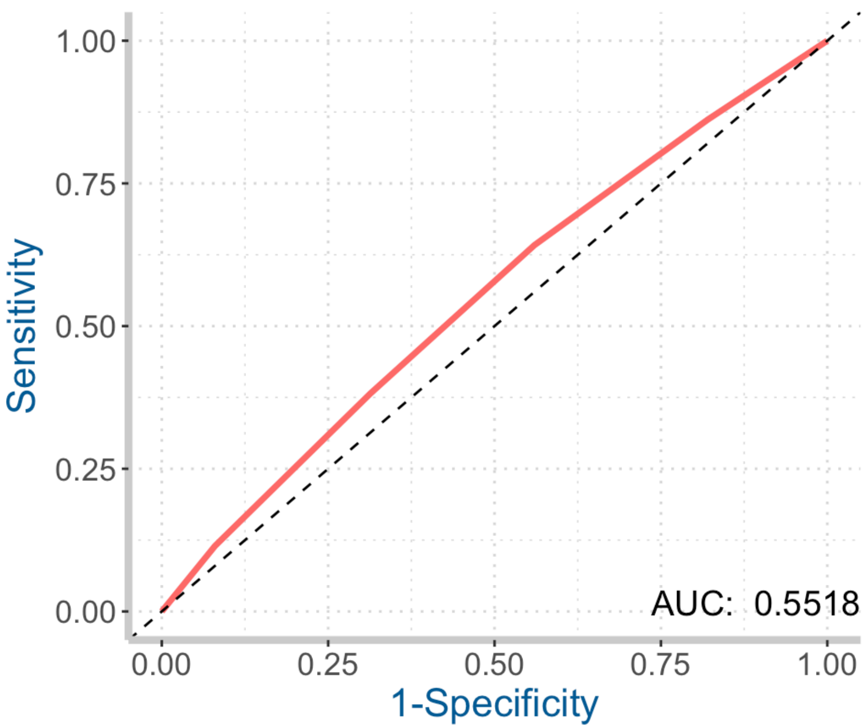
Area under the curve (AUC)



- We use the **area under the ROC curve**, i.e. **AUC** to quantify the predictive accuracy.
- Larger AUC is associated with higher predictive accuracy. Smaller AUC is associated with lower predictive accuracy.
- For the *Acceptance* ~ *GPA* example, $AUC = 0.824$
- The red curve has $AUC = 0.917$ and the green one has $AUC = 0.582$.
- $0.5 \leq AUC \leq 1$.

Area under the curve (AUC)


Medical School Acceptance


Teenagers Sleep

Function for ROC curve and AUC

```
# Function for ROC curve with AUC
ROC <- function(model, color, title){
  logit_pi <- predict(model)
  est_pi <- exp(logit_pi)/(1+exp(logit_pi))
  roc_pred <- prediction(est_pi, model$y)
  roc_ss <- performance(roc_pred, "sens", "spec")
  auc <- round(performance(roc_pred, "auc")@y.values[[1]], 4)
  roc_curve <- data.frame(Sensitivity = roc_ss@y.values[[1]],
                          One_Specificity = 1-roc_ss@x.values[[1]])
  ggplot(data=roc_curve, aes(x=One_Specificity, y=Sensitivity))+
    geom_path(color=color, size=1.2)+
    geom_abline(intercept=0, slope=1, linetype=2)+
    annotate("text", Inf, -Inf, label=paste("AUC: ", auc),
             hjust=1, vjust=-1, size=5)+
    xlab("1-Specificity")+
    ggtitle(title)
}
```

Summary

- Inference for slope(s)
  - z (Wald) test for the slope
  - Confidence interval for the odds ratio
- Inference for the model
  - Likelihood ratio test (LRT)
- Model assessment: AIC
- Predictive accuracy
  - Sensitivity and specificity at a certain cutoff
  - ROC curve
  - Area under the ROC curve (AUC)