BIOS 7345: Advanced Regression Analysis I

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 19: Receiver operating characteristic regression

Version: 09/17/2023

Table of Contents

Sensitivity and specificity

Diagnostic tools:

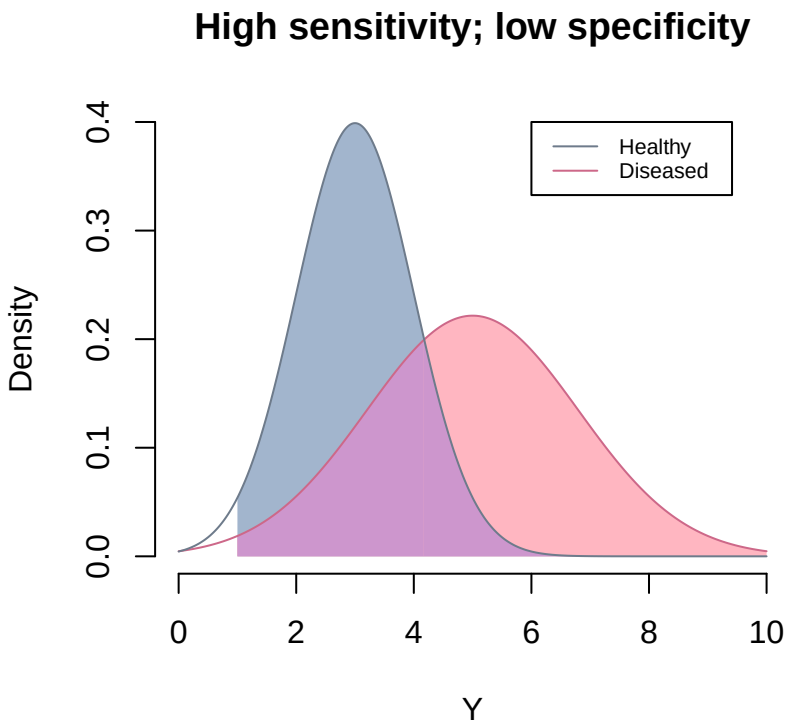
- It is often useful to use continuous measures to classify people into groups (e.g., disease diagnosis via biomarkers).
- Suppose $Y_0 \sim F_0$ (healthy) and $Y_1 \sim F_1$ (diseased).
- The survivor functions are marked by $S_0(c) = 1 - F_0(c) = \mathrm{P}(Y_0 \geq c)$ and $S_1(c) = 1 - F_1(c) = \mathrm{P}(Y_1 \geq c)$.
- One way to quantify how well a test can discriminate between groups is to choose a threshold c and determine the proportion of healthy and diseased individuals with values meeting or exceeding that threshold (assume without loss of generality that greater values test positively).
  - In the ideal scenario, we can find a c so that the former proportion is low and the latter is high.

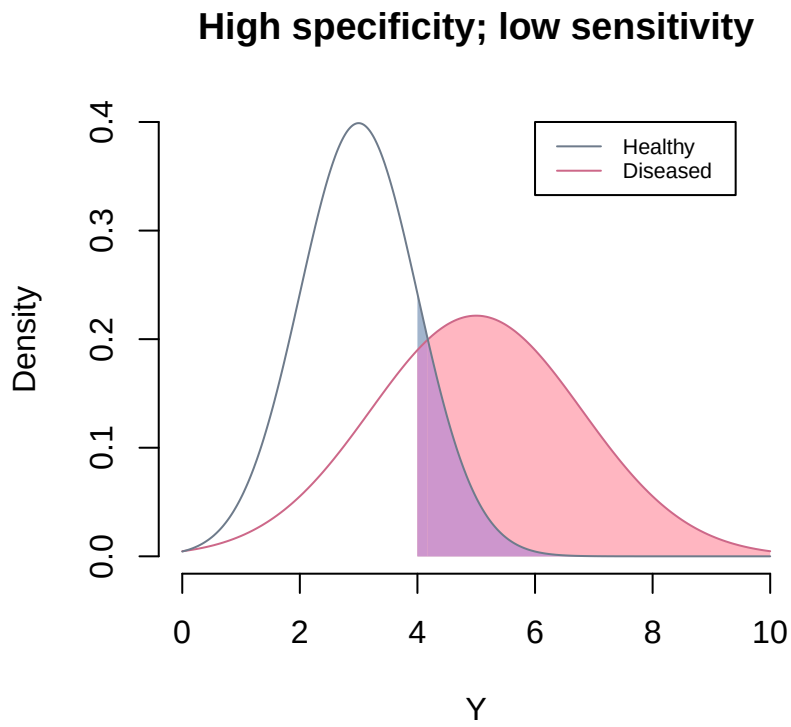
Sensitivity and specificity

Measures of diagnostic accuracy:

- The **sensitivity** of a test based on threshold c is the *true positive rate*, $\text{TPR}(c) = S_1(c)$.
  - What proportion of the diseased individuals are being correctly classified as such on the basis of the test?
- The **specificity** of a test based on threshold c is *one minus the false positive rate*, $1 - \text{FPR}(c) = 1 - S_0(c)$.
  - What proportion of the healthy individuals are being correctly classified as such on the basis of the test?
- As you can imagine, these two quantities generally come with an inherent trade-off, illustrated on the following slide.

Illustration: Trade-off

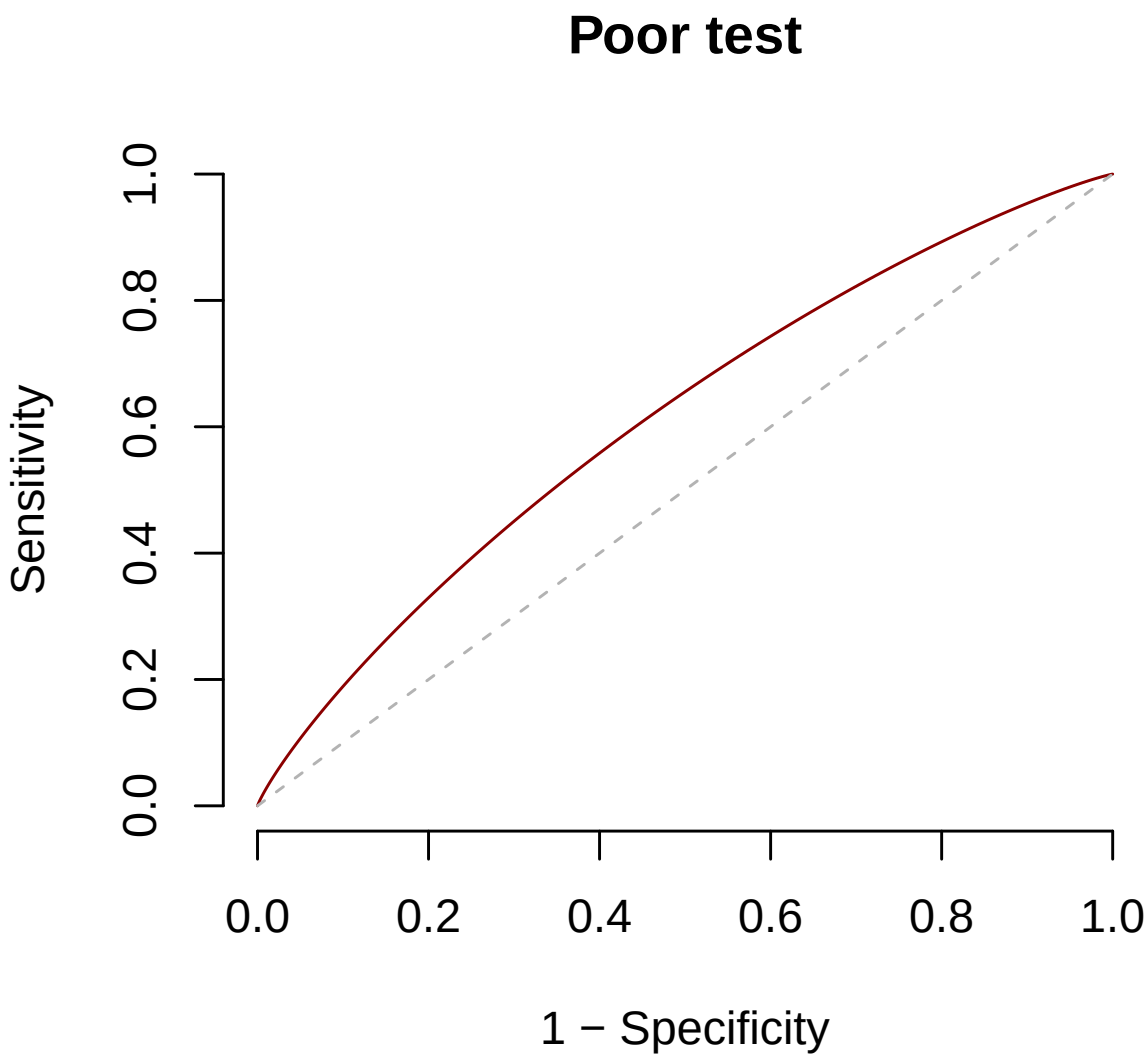




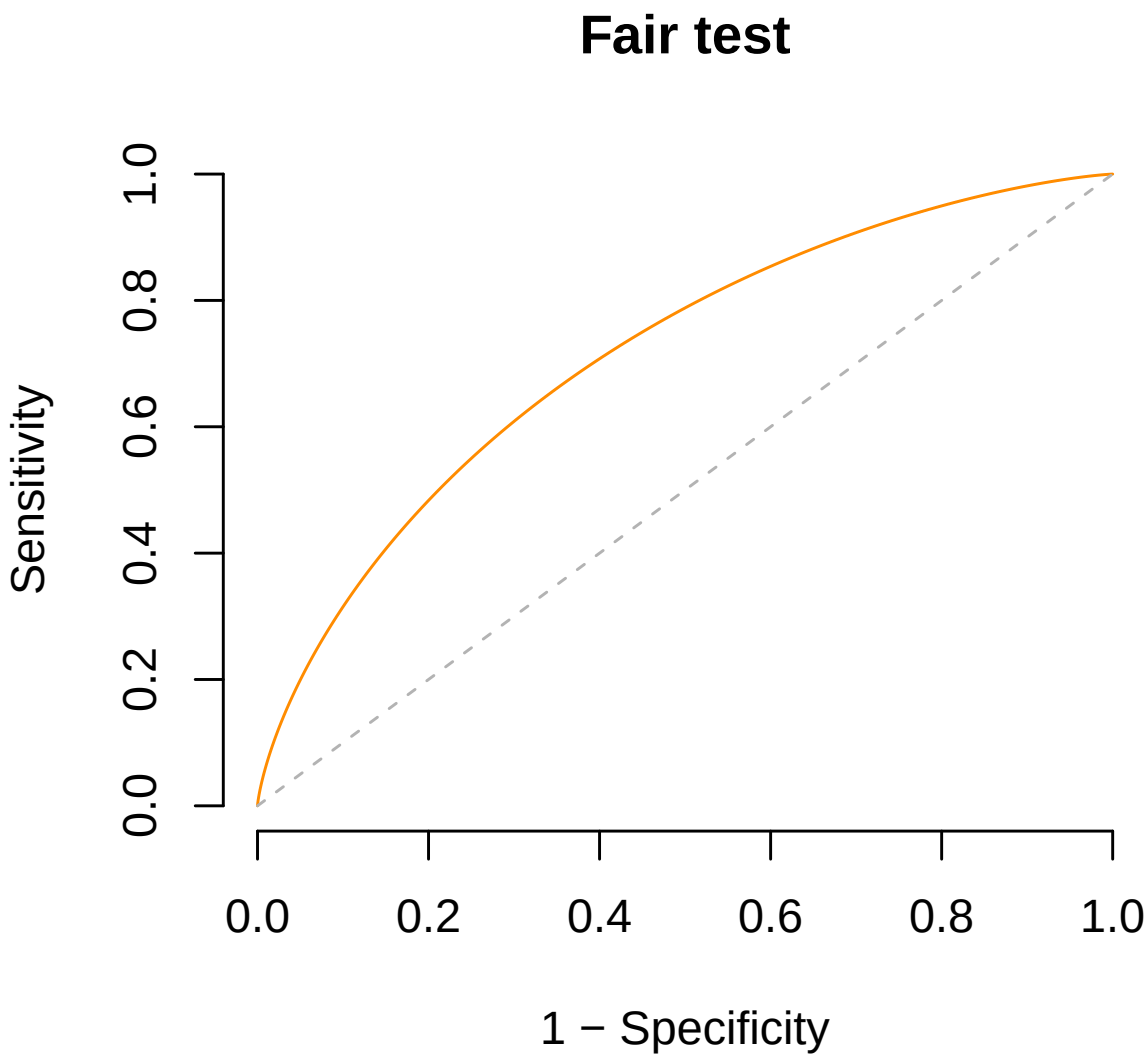
The ROC curve:

- Rather than restricting attention to a single cut-point, you can gain a global understanding of a diagnostic test by considering the set of sensitivity-specificity pairs induced by a range of cut-points.
- The set of ordered pairs $(1 - S_0(c), S_1(c))$ is referred to as the receiver operating characteristic (ROC) curve.
- **Key observation**: $\text{ROC}(p) = S_1(S_0^{-1}(p))$ for $0 \leq p \leq 1$.
  - Suppose you want to know the sensitivity of a test that has 80% specificity.
  - Step 1: Figure out the quantile for the healthy individuals, q, that corresponds to 20% of healthy individuals meeting testing positive.
  - Step 2: Figure out the proportion of diseased individuals who exceed threshold q. This is the sensitivity of that test.

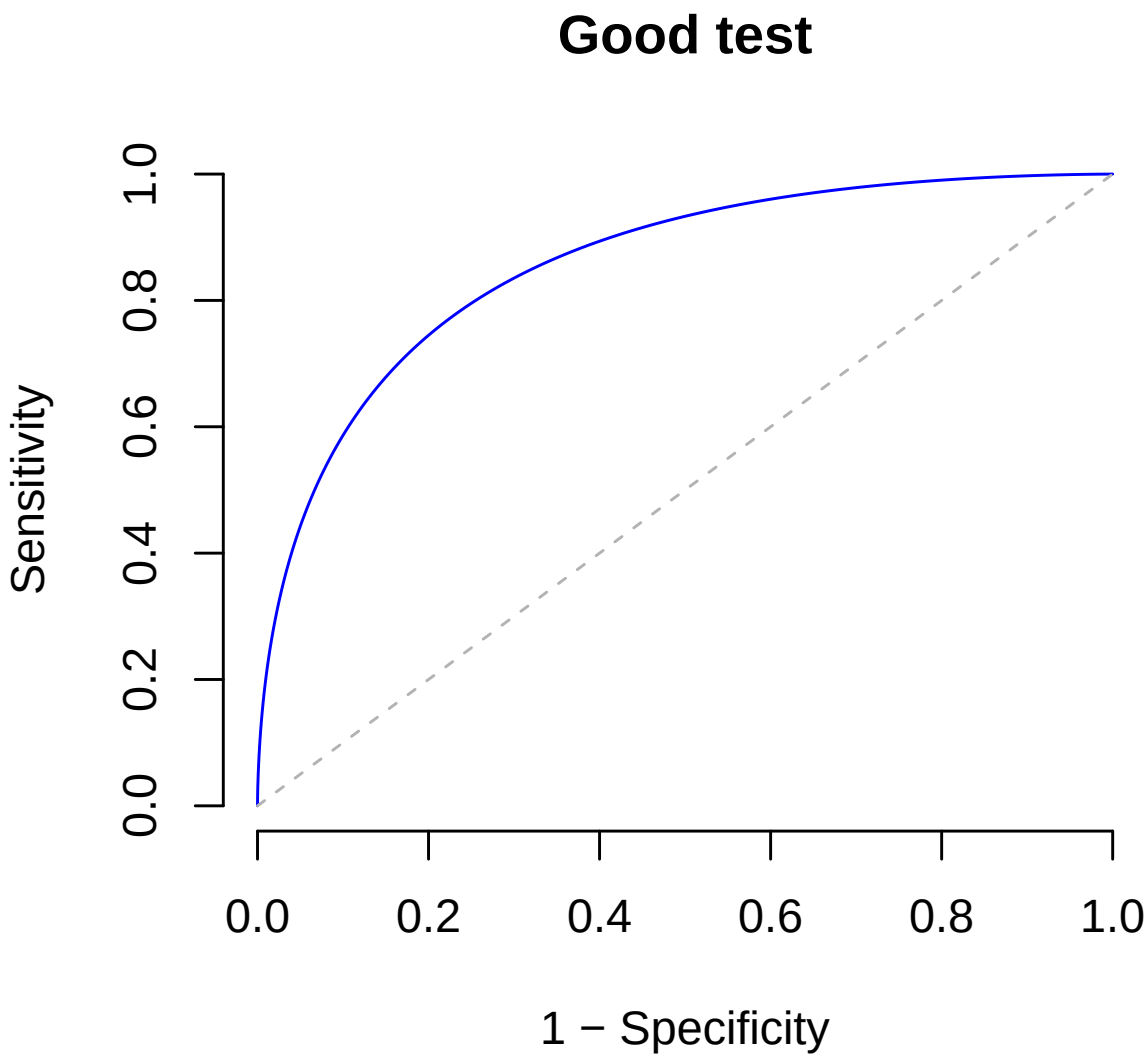
Sensitivity and specificity



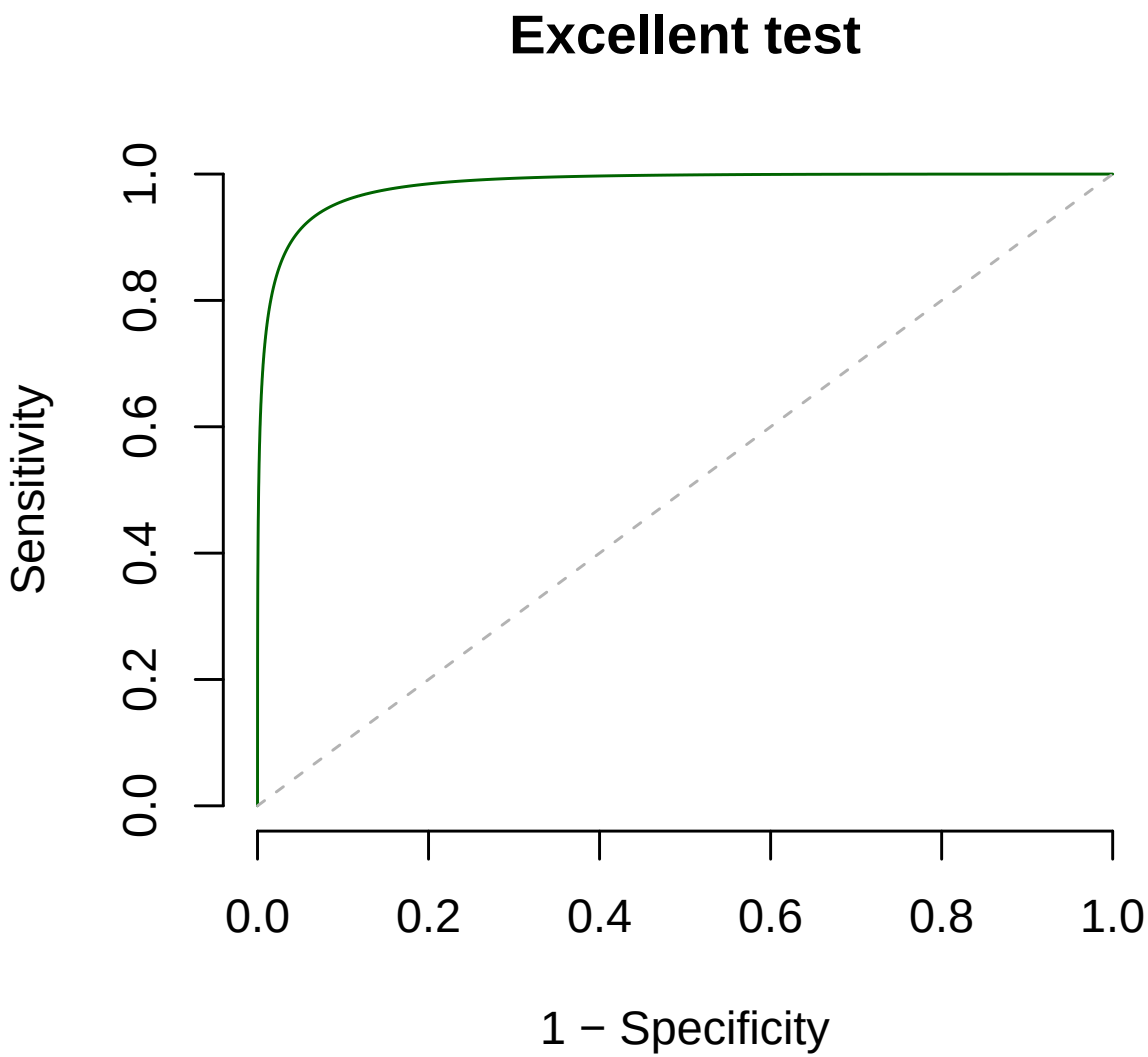
Sensitivity and specificity



Sensitivity and specificity



Sensitivity and specificity

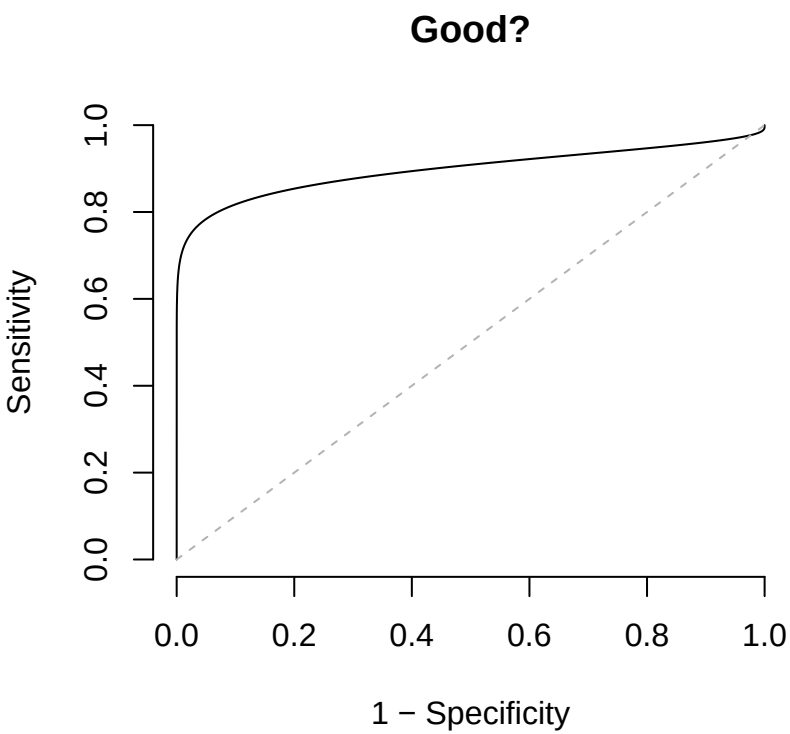


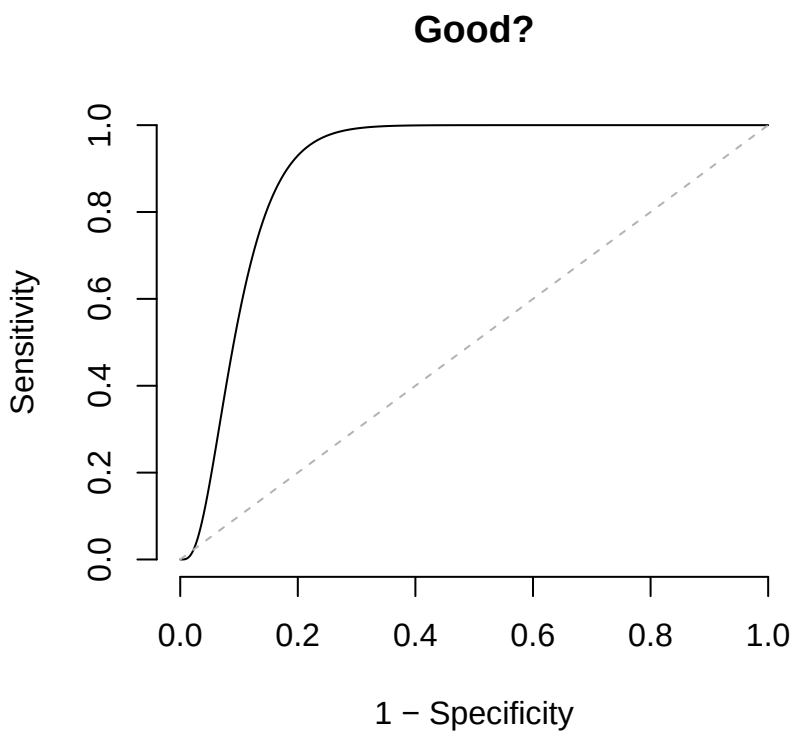
Sensitivity and specificity

The ROC curve:

- ROC curves are by no means always symmetric about the line $y = 1 - x$ (though they were in the examples I just showed).
- Sometimes the only way to achieve very high sensitivity is to sacrifice a meaningful amount of specificity.
  - Can you draw of a pair of distributions that would have this property?
- Sometimes the only way to achieve very high specificity is to sacrifice a meaningful amount of sensitivity.
  - Can you draw of a pair of distributions that would have this property?
- How you choose to optimize these trade-offs depends upon what sort of disease you're seeking to test, how costly/invasive/time-consuming the test is, etc.

Sensitivity and specificity





The ROC curve: Empirical

- The ROC curve is often estimated empirically (nonparametric).
- Does not presume that either the healthy or the diseased groups conform to a specific set of distributions.
  - Distribution-free.
  - Non-parametric.

The ROC curve: AUC

- One way to characterize the quality of the diagnostic test is through the area under the ROC curve (AUC, for "area under the curve").

$$\text{AUC} = \int_0^1 \text{ROC}(p)dp = \text{P}(Y_1 > Y_0).$$

- The area under the curve is a measure of *stochastic ordering* and is closely related to the Mann-Whitney U test
  - $\widehat{\text{AUC}}_{\text{Empirical}} = U/(N_0 N_1)$.
- As a rank-based measure, it suffers from lack of transitivity (it is possible to create a data set where group A does better than B, B better than C, and C better than A).
  - Who's up for a game of rock, paper, scissors?
  - Best to compare two groups.

The ROC curve: Empirical

```
rocplot <- function(y0, y1, switch=FALSE, AUC=TRUE)
{
  if (switch == TRUE) {
    y0old <- y0; y1old <- y1
    y0 <- y1old; y1 <- y0old
  }
  n0 <- length(y0); n1 <- length(y1)
  spec <- c(0:n0)/n0
  q <- quantile(y0, spec)
  sens <- matrix(0, nrow = length(spec), ncol = 1)
  for (j in 1:length(spec)){
    sens[j,1] <- mean(as.numeric(y1 >= q[j]))
  }
  spec <- c(0,spec,1); sens <- c(1,sens,0)
  plot(1 - spec, sens, frame.plot = FALSE, xlab = "1 - Specificity",
       ylab = "Sensitivity", pch = 20, cex = 0.6, xlim = c(0,1), ylim = c(0,1),
       main = "ROC curve")
  for (j in 1:(length(spec) - 1))
  {
    segments(1 - spec[j], sens[j], 1 - spec[j], sens[j + 1])
    segments(1 - spec[j], sens[j + 1], 1 - spec[j+1], sens[j + 1])
  }
  segments(0,0,1,1, col = "gray80", lty = 2)
  if (AUC == TRUE){
    AUC <- wilcox.test(y0,y1)$statistic/(n0*n1)
    text(0.8, 0.4, paste("AUC =",round(AUC,2)))
  }
}
```

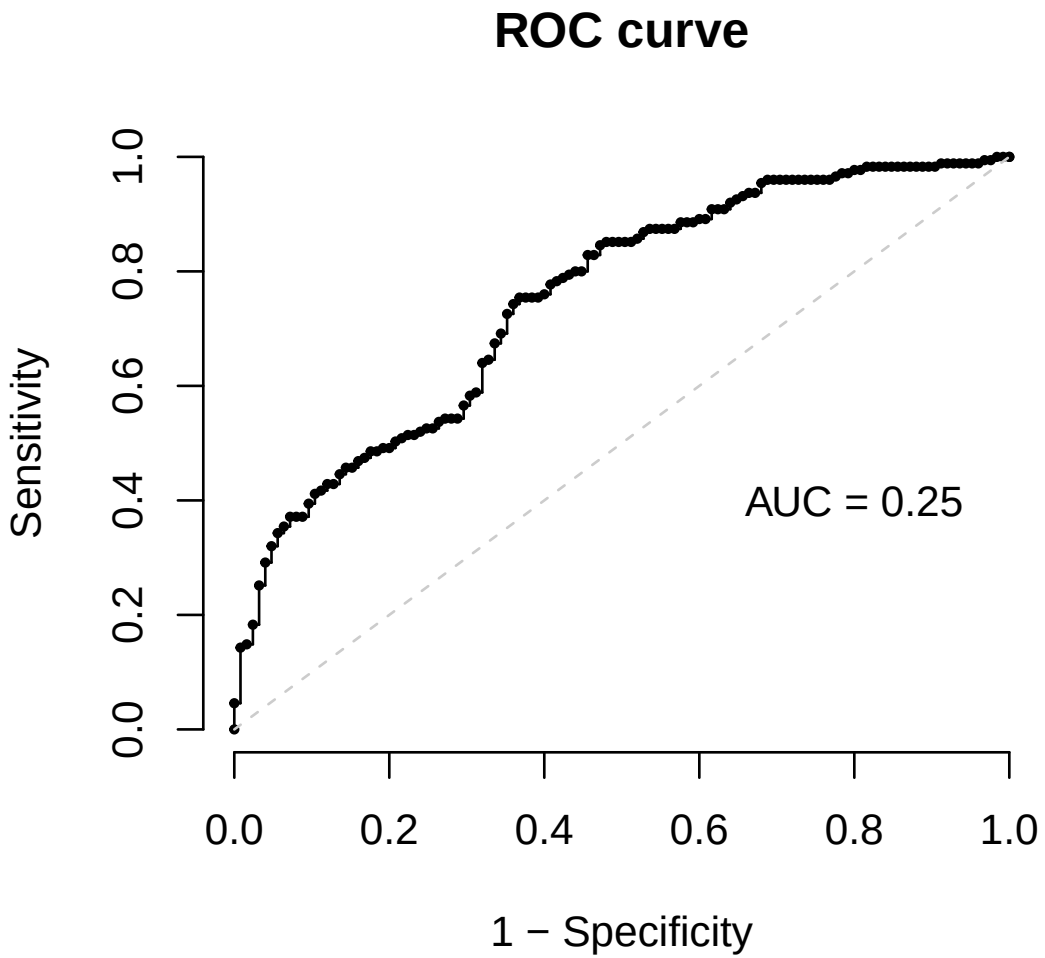
The ROC curve: Empirical

```
## Set seed for reproducibility
set.seed(7345)

## Generate data
n0 <- 125
n1 <- 175
y0 <- rnorm(n0, 3, 1)
y1 <- rnorm(n1, 4, 1.)

## Plot empirical ROC curve
rocplot(y0,y1)
```

Sensitivity and specificity



The ROC curve:

- Might we benefit from estimating the ROC curve parametrically?
  - It is likely we will (if our parametric assumptions are correct).
- Can we learn about how the diagnostic utility of Y varies across subgroups defined by their value of X?
  - Regression of the ROC curve.

Table of Contents

Parametric estimation of the ROC curve

The binormal curve:

- Suppose $Y_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$.
- The corresponding ROC curve can be readily derived, since $S_0^{-1}(p) = \mu_0 - \sigma_0 \Phi^{-1}(p)$ and $S_1(c) : \Phi((c - \mu_1)/\sigma_1)$:

$$\begin{aligned} \text{ROC}(p) &= S_1(S_0^{-1}(p)) \\ &= 1 - \Phi((\mu_0 + \sigma_0 \Phi^{-1}(1-p) - \mu_1)/\sigma_1) \\ &= \vdots \text{ (math)} \\ &= \Phi\left(\frac{\mu_1 - \mu_0}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\Phi^{-1}(p)\right). \end{aligned}$$

- The maximum likelihood estimator for the ROC curve (point-wise) is based on the plug-in estimator, though we can use the sample variance rather than the MLE.
- Try this for $Y_0 \sim \text{Exponential}(\lambda_0)$ and $Y_1 \sim \text{Exponential}(\lambda_1)$!

Parametric regression of the ROC curve

The binormal curve:

- Suppose we seek to estimate how the diagnostic utility of a continuous outcome (Y) varies across:
  - Predictors, $\mathbf{X}$, that apply both to the diseased and the healthy groups.
    - ★ Age.
    - ★ Sex.
    - ★ Other baseline characteristics.
  - Predictors, $\mathbf{X}_1$, that apply to the diseased group only.
    - ★ Cancer stage/morphology.
    - ★ What transplant donor related or unrelated?

Parametric regression of the ROC curve

The binormal curve:

- Suppose now that:
  - $Y_0 \sim \mathcal{N}(\alpha_0 + \mathbf{X}^T\boldsymbol{\gamma}_0, \sigma_0^2)$.
  - $Y_1 \sim \mathcal{N}(\alpha_0 + \alpha_1 + \mathbf{X}^T(\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1) + \mathbf{X}_1^T\boldsymbol{\zeta}_1, \sigma_1^2)$.
- The corresponding *conditional* ROC curve can be readily derived:

$$\text{ROC}(p|\mathbf{X}, \mathbf{X}_1) = \Phi\left(\frac{1}{\sigma_1}\alpha_1 + \frac{1}{\sigma_1}\mathbf{X}^T\boldsymbol{\gamma}_1 + \frac{1}{\sigma_1}\mathbf{X}_1^T\boldsymbol{\zeta}_1 + \frac{\sigma_0}{\sigma_1}\Phi^{-1}(p)\right).$$

- We can estimate the parameters via maximum likelihood (with primary emphasis on $\boldsymbol{\gamma}_1$ and $\boldsymbol{\zeta}_1$ for inference purposes).

The binormal curve: Example

- $X \sim \mathcal{N}(\mu_X = 2, \sigma_X^2 = 1)$.
- $Y_0 \sim \mathcal{N}(\mu_0 = X, \sigma_0^2 = 1)$.
- $Y_1 \sim \mathcal{N}(\mu_1 = 2X, \sigma_1^2 = 1)$.

$$\text{ROC}(p|X = x) \quad = \quad \Phi\left(x + \Phi^{-1}(p)\right).$$

- The diagnostic test based on Y has more ability to discriminate between the healthy and diseased groups among those with a higher value of X.

Parametric estimation of the ROC curve

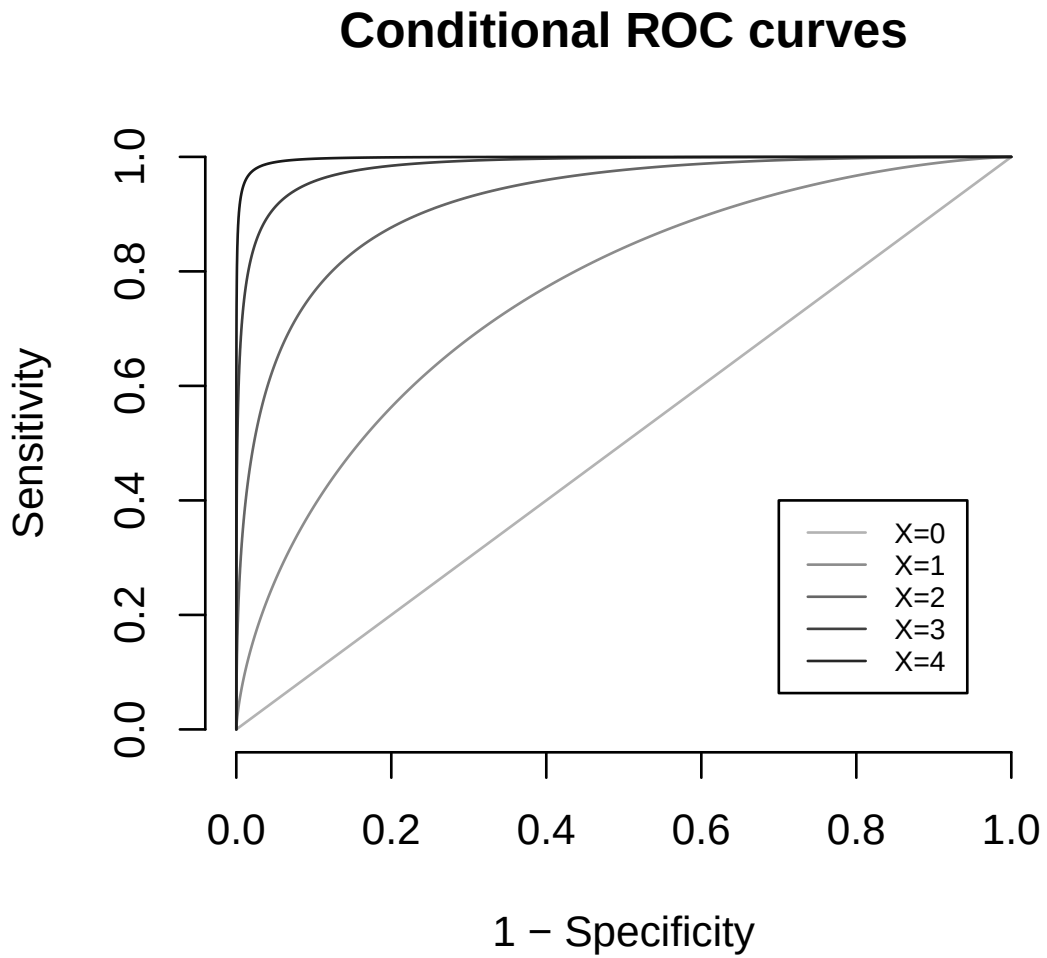


Table of Contents

Semi-parametric estimation of the ROC curve

The binormal curve: A more general class of ROC curves

- Notice that the binormal ROC curve takes the following form:

$$\begin{aligned} \text{ROC}(p) &= \Phi\left(\frac{\mu_1 - \mu_0}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\Phi^{-1}(p)\right) \\ &= \Phi\left(\beta_0 + \beta_1\Phi^{-1}(p)\right) \end{aligned}$$

- Consider the hypothesis tests:
  - $H_0 : \beta_0 = 0$ (AUC is 0.5).
  - $H_0 : \beta_0 = 0$ and $\beta_1 = 1$ (ROC curve tracks line $y = x$).
- Since the actual values of μ_0, μ_1, σ_0, and σ_1 are not the relevant parameters, is there a way to skip over this step and estimate β_0 and β_1 directly?

Semi-parametric estimation of the ROC curve

The binormal curve: A more general class of ROC curves

- Define $U_{ij} = 1(Y_{1j} \geq Y_{0i})$. Amazingly, we have the following result:

$$\mathrm{E}[U_{ij}|S_0(Y_{0i}) = p] = \mathrm{ROC}(p).$$

- Now, if we believe the ROC curve follows the binormal family, we can use a probit regression model on the U_{ij}'s:

$$\mathrm{E}[U_{ij}] = \Phi(\beta_0 + \beta_1 \Phi^{-1}(p)).$$

- The estimating equations are therefore given by:

$$\sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathbf{x}_{ij} \frac{\phi(z_{ij})}{\Phi(z_{ij})(1 - \Phi(z_{ij}))} (U_{ij} - \Phi(z_{ij})) = \mathbf{0},$$

where $\mathbf{x}_{ij} = (1, \Phi^{-1}(\widehat{S}_0(Y_{0i})))^T$ and $z_{ij} = \beta_0 + \beta_1 \Phi^{-1}(\widehat{S}_0(Y_{0i}))$.

Semi-parametric estimation of the ROC curve

The binormal curve: A more general class of ROC curves

- Importantly, the "binormal" terminology originates from the setting of normally distributed outcomes, but a binormal curve itself can originate from distributions that are not themselves normal.
- Semi-parametric, much like the proportional hazards model.
  - Not modeling group-specific hazard functions, but the Cox model is only valid when the log-hazard hazard functions in subgroups defined by the right-hand side of the model differ by a constant.
- If the sample size is given by $N = N_0 + N_1$, how many U_{ij}'s are there?
  - Do not trust the standard errors that come from the `glm()` function.

The binormal curve: Example 1

```
## Set seed for reproducibility
set.seed(7345)

## Set sample size
n0 <- n1 <- 30

## Set number of simulations (ideally higher)
nsim <- 300

## Set number of bootstrap replicates (ideally higher)
B <- 300

## Create a space to store results of hypothesis test
hypoth <- matrix(0, nrow = nsim, ncol = 1)

## Create a space to store estimates
est <- matrix(0, nrow = nsim, ncol = 2)

## Create a space to store standard error estimates
ster <- matrix(0, nrow = nsim, ncol = 2)
```

The binormal curve: Example 1

```
## Begin simulation (EXAMPLE 1)
for (k in 1:nsim) {
  ## Generate data
  Y0 <- rnorm(n0, 4, 0.9)
  Y1 <- rnorm(n1, 4, 1.2)
  ## Create space to store bootstrap results
  bres <- matrix(0, nrow = B, ncol = 2)
  for (b in 1:B)
  {
    ## Re-sample with replacement
    y0 <- Y0[sample(1:n0, replace = TRUE)]
    y1 <- Y1[sample(1:n1, replace = TRUE)]
    ## Create U outcome and estimate S0(c)
    U <- as.numeric(t(outer(y1, y0, ">=")))
    Pij <- matrix(colMeans(outer(y0, y0, ">=")), nrow = n0, ncol = n1)
    P <- matrix(Pij, ncol = 1)
    Q <- qnorm(P)
    Q[P == 0] <- min(Q[P != 0])
    Q[P == 1] <- max(Q[P != 1])
    ## Run ROC-GLM
    zz <- glm(U ~ Q, family = binomial(link = "probit"))
    bres[b,] <- as.numeric(coef(zz))
  }
  ## Conduct hypothesis test and extract results
  if (quantile(bres[,1], 0.025) > 0 | quantile(bres[,1], 0.975) < 0) {
     hypoth[k,1] <- 1
  }
  est[k,] <- colMeans(bres)
  ster[k,] <- apply(bres, 2, sd)
}
```

The binormal curve: Example 1

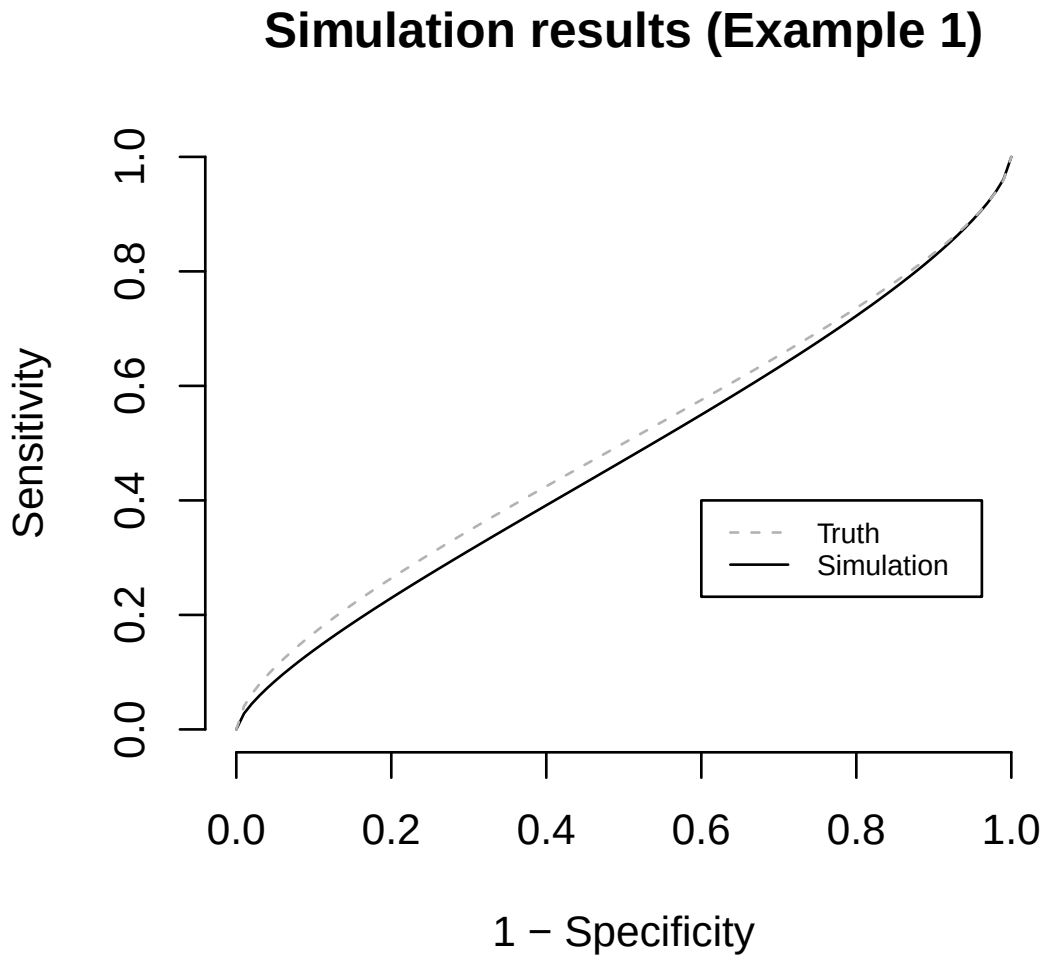
```
## Type 1 error
> mean(hypoth)
[1] 0.06333333

## Average coefficient estimates
> colMeans(est)
[1] -0.07538844  0.79019014

## Average standard errors
> colMeans(ster)
[1] 0.2534285 0.1848871

## Empirical standard errors
> apply(est, 2, sd)
[1] 0.2527257 0.1698172
```

Semi-parametric estimation of the ROC curve



The binormal curve: Example 2

```
## Set seed for reproducibility
set.seed(7345)

## Set sample size
n0 <- n1 <- 50

## Set number of simulations (ideally higher)
nsim <- 300

## Set number of bootstrap replicates (ideally higher)
B <- 300

## Create a space to store results of hypothesis test
hypoth <- matrix(0, nrow = nsim, ncol = 1)

## Create a space to store estimates
est <- matrix(0, nrow = nsim, ncol = 2)

## Create a space to store standard error estimates
ster <- matrix(0, nrow = nsim, ncol = 2)
```

The binormal curve: Example 2

```
## Begin simulation (EXAMPLE 2)
for (k in 1:nsim) {
  ## Generate data
  Y0 <- rnorm(n0, 4.0, 1)
  Y1 <- rnorm(n1, 5.5, 1)
  ## Create space to store bootstrap results
  bres <- matrix(0, nrow = B, ncol = 2)
  for (b in 1:B)
  {
    ## Re-sample with replacement
    y0 <- Y0[sample(1:n0, replace = TRUE)]
    y1 <- Y1[sample(1:n1, replace = TRUE)]
    ## Create U outcome and estimate S0(c)
    U <- as.numeric(t(outer(y1, y0, ">=")))
    Pij <- matrix(colMeans(outer(y0, y0, ">=")), nrow = n0, ncol = n1)
    P <- matrix(Pij, ncol = 1)
    Q <- qnorm(P)
    Q[P == 0] <- min(Q[P != 0])
    Q[P == 1] <- max(Q[P != 1])
    ## Run ROC-GLM
    zz <- glm(U ~ Q, family = binomial(link = "probit"))
    bres[b,] <- as.numeric(coef(zz))
  }
  ## Conduct hypothesis test and extract results
  if (quantile(bres[,1], 0.025) > 0 | quantile(bres[,1], 0.975) < 0) {
     hypoth[k,1] <- 1
  }
  est[k,] <- colMeans(bres)
  ster[k,] <- apply(bres, 2, sd)
}
```

The binormal curve: Example 2

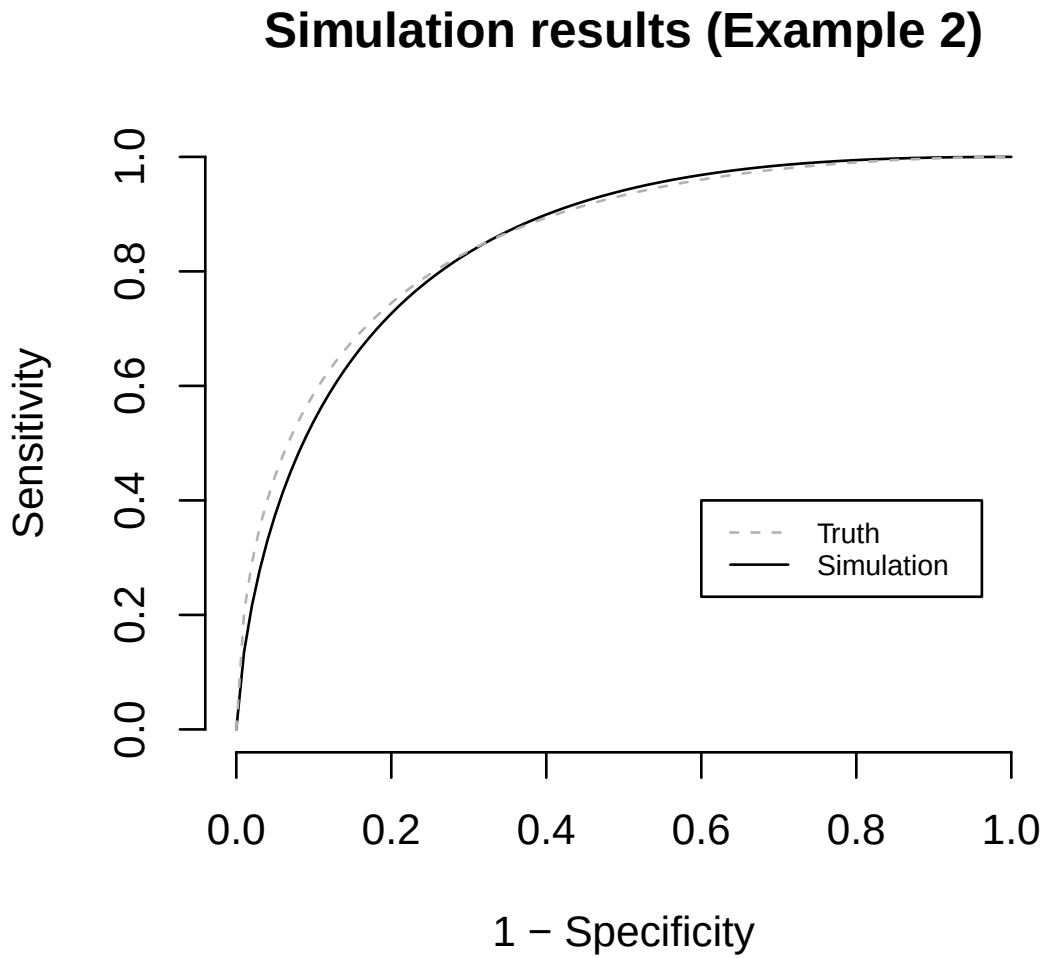
```
## Power
> mean(hypoth)
[1] 1

## Average coefficient estimates
> colMeans(est)
[1] 1.584796 1.154461

## Average standard errors
> colMeans(ster)
[1] 0.3451116 0.2614128

## Empirical standard errors
> apply(est, 2, sd)
[1] 0.3321131 0.2675124
```

Semi-parametric estimation of the ROC curve



Semi-parametric estimation of the ROC curve

The binormal curve:

- The simulation examples are a little cheap in that I'm using normally distributed data to generate the binormal curve.
- Nevertheless, you should be able to convince yourself that if Y_0 follows a reasonable distribution, then you will often be able to find a distribution function for Y_1 such that the resulting ROC curve is binormal (and vice versa).
- The methodology only relies on the family of curves to which ROC(p) can belong.
- If the data are normally distributed, estimation based on the MLEs for μ_0, μ_1, σ_0^2, and σ_1^2 will be more efficient, asymptotically. We're familiar with this trade-off.

Semi-parametric regression of the ROC curve

ROC-GLM regression:

- We're now in a position where we can easily add covariates to the ROC-GLM framework.
- Assume the conditional ROC curve takes the following form:

$$\text{ROC}(p|\mathbf{X}, \mathbf{X}_1) = \Phi\left(\beta_0 + \mathbf{X}^T\boldsymbol{\beta}_1 + \mathbf{X}_1^T\boldsymbol{\beta}_2 + \beta_3\Phi^{-1}(p)\right).$$

- The estimating equations are the same as those shown previously, except now:
  - $\mathbf{x}_{ij} = (1, \mathbf{x}_j, \mathbf{x}_{1j}, \Phi^{-1}(\widehat{S}_0(Y_{0i}|\mathbf{x}_j)))^T$.
  - $z_{ij} = \beta_0 + \beta_1 X_j + \beta_2 X_{1j} + \beta_3\Phi^{-1}(\widehat{S}_0(Y_{0i}|\mathbf{x}_j))$.
- Here, we need to estimate the conditional survivor function of interest using, e.g., quantile regression techniques.

Summary

So far:

- Regression of the ROC curve.

Summary

Up next:

- Categorical and ordinal outcomes.