Welcome!

Distribution-free ROC Analysis Using Binary Regression Techniques

Todd A. Alonzo and Margaret S. Pepe

As presented by: Andrew J. Spieker

University of Washington

Dept. of Biostatistics

Goals

Overarching Goals:

- Identify diagnostic tests with ability to discriminate between states of health
- Identify factors which influence diagnostic accuracy

Methodological Goal:

- Devise a method for ROC regression which:
  - Provides valid estimates
  - Is simple to implement
  - Is computationally efficient

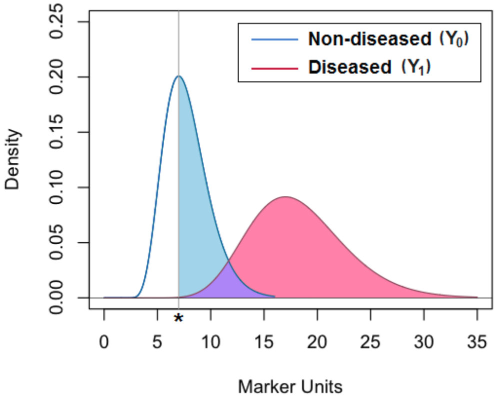
Notation

- Y_0 and Y_1: test values for *healthy* and *diseased* participants
- X: covariate(s) for all subjects
- X_1 : variables specific to diseased group
- S_0 and S_1: survivor functions for Y_0 and Y_1, respectively:
- ...meaning, $S_0(c) = P(Y_0 \geq c)$, and $S_1(c) = P(Y_1 \geq c)$

Key Observation

$$ROC_{Y_0,Y_1|X,X_1}(p) = S_1\left(S_0^{-1}(p|X)\middle| X, X_1\right)$$

ROC


Non-diseased (Y_0)
Diseased (Y_1)
Density
0.00
0.05
0.10
0.15
0.20
0.25
*
0
5
10
15
20
25
30
35
Marker Units

A Suitable Starting Point

Inherently parametric methods

- Parametrically model the test results
- And determine the induced ROC curve

Model ROC curve directly rather than presume a distribution for the data

- Generalized linear model framework (2000)
  - Much easier to program, somewhat intuitive
  - Computationally less efficient than desirable

A Suitable Starting Point

Inherently parametric methods

- Parametrically model the test results
- And determine the induced ROC curve

Model ROC curve directly rather than presume a distribution for the data

- Generalized linear model framework (2000)
  - Much easier to program, somewhat intuitive
  - Computationally less efficient than desirable

Estimation (Pepe, 2000)

We observe Y- and X-values on n_0 healthy controls and n_1 diseased participants ...

Key Observation

- *If* $U_{ij} = \mathbf{1}(Y_{1j} \geq Y_{0i})$, *then ...*

$$\begin{aligned}
\mathbb{E}[U_{ij}|S_0(Y_{0i}|X_i) = p, X_i, X_j, X_{1j}] \\
&= \mathbb{P}(Y_{1j} \geq Y_{0i}|S_0(Y_{0i}|X_i) = p, X_i, X_j, X_{1j}) \\
&= \mathbb{P}(Y_{1j} \geq S_0^{-1}(p|X_i)|X_j, X_{1j}) \\
&= S_1(S_0^{-1}(p|X_i)|X_j, X_{1j}) \\
&= ROC_{Y_{0i}, Y_{1j}|X_i, X_j, X_{1j}}(p)
\end{aligned}$$

Estimation (Pepe, 2000)

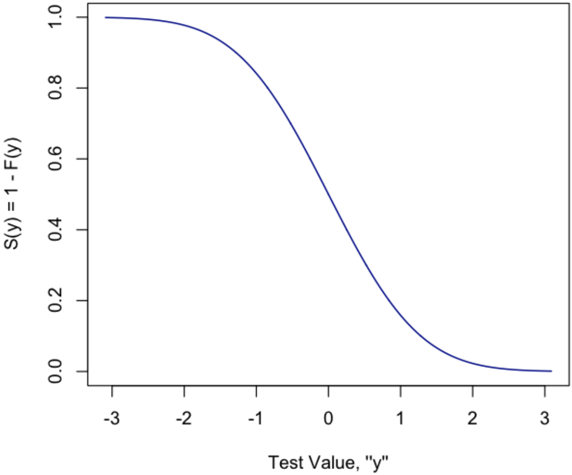
We observe Y- and X-values on n_0 healthy controls and n_1 diseased participants ...

Key Observation

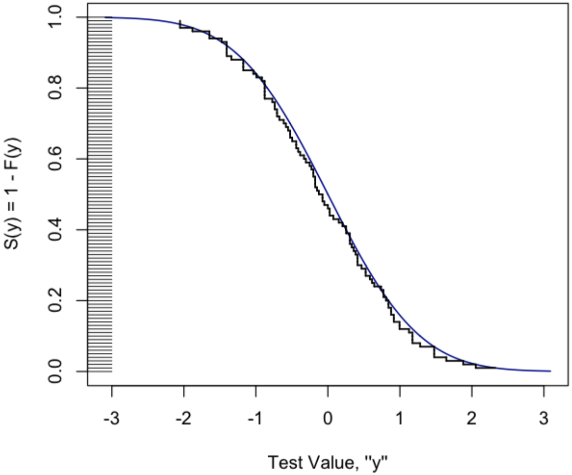
- *If* $U_{ij} = \mathbf{1}(Y_{1j} \geq Y_{0i})$, *then* ...

$$\begin{aligned}
&\mathbb{E}[U_{ij}|S_0(Y_{0i}|X_i) = p, X_i, X_j, X_{1j}] \\
&\quad = \mathbb{P}(Y_{1j} \geq Y_{0i}|S_0(Y_{0i}|X_i) = p, X_i, X_j, X_{1j}) \\
&\quad = \mathbb{P}(Y_{1j} \geq S_0^{-1}(p|X_i)|X_j, X_{1j}) \\
&\quad = S_1(S_0^{-1}(p|X_i)|X_j, X_{1j}) \\
&\quad = ROC_{Y_{0i}, Y_{1j}|X_i, X_j, X_{1j}}(p)
\end{aligned}$$

Estimation of S_0



Estimation of S_0



Transitioning: Pepe (2000) $\longrightarrow$ Alonzo & Pepe (2002)

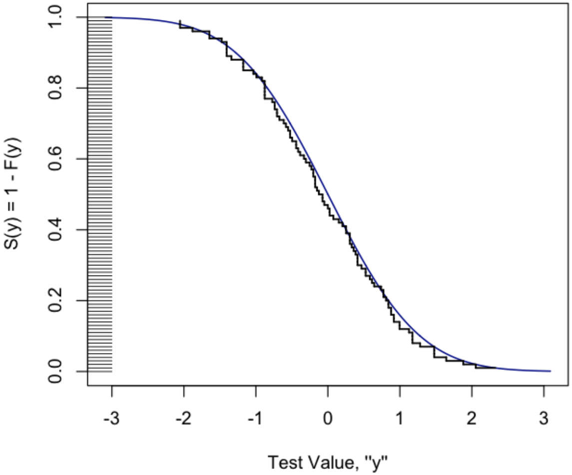
Pepe (2000):

- If $U_{ij} = \mathbf{1}(Y_{1j} \geq Y_{0i})$, then $\mathbb{E}[U_{ij} | G_0(Y_{0i}|X_i) = p, X_i, X_j, X_{1i}] = \text{ROC}_{Y_{0i}, Y_{1j}|X_i, X_j, X_{1j}}(p)$

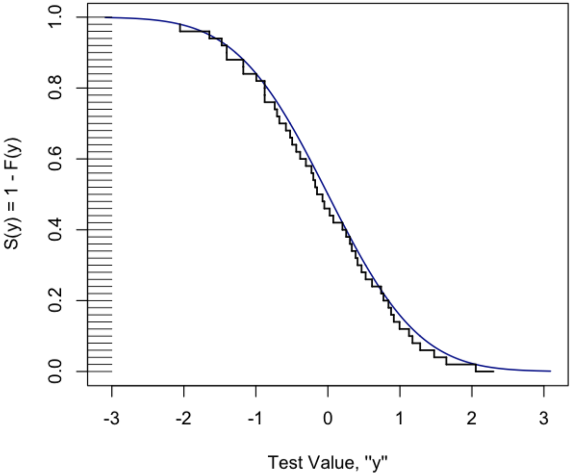
Alonzo & Pepe (2002):

- Estimate S_0 on a user-determined grid, G_ℓ
- If $U_{pj} = \mathbf{1}(Y_{1j} \geq \hat{S}_0^{-1}(p|X_i))$, then $\mathbb{E}[U_{pj} | X_j, X_{1j}] = \text{ROC}_{Y_{0i}, Y_{1j}|X_i, X_j, X_{1j}}(p)$

Estimation of S_0



Estimation of S_0



Transitioning: Pepe (2000) $\longrightarrow$ Alonzo & Pepe (2002)

Pepe (2000):

- If $U_{ij} = \mathbf{1}(Y_{1j} \geq Y_{0i})$, then $\mathbb{E}[U_{ij}|G_0(Y_{0i}|X_i) = p, X_i, X_j, X_{1i}] = \mathrm{ROC}_{Y_{0i},Y_{1j}|X_i,X_j,X_{1j}}(p)$

Alonzo & Pepe (2002):

- Estimate S_0 on a user-determined grid, G_ℓ
- If $U_{pj} = \mathbf{1}(Y_{1j} \geq \hat{S}_0^{-1}(p|X_i))$, then $\mathbb{E}[U_{pj}|X_j, X_{1j}] = \mathrm{ROC}_{Y_{0i},Y_{1j}|X_i,X_j,X_{1j}}(p)$

If it looks like a GLM and "links" like a GLM...

Focusing attention on "binormal" setup (GLM with probit link):

$$\mathrm{ROC}_{Y_0,Y_1|X,X_1}(p) = \Phi(\alpha_0 + \alpha_1\Phi^{-1}(p) + \beta_0 X + \beta_1 X_1).$$

$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})^T$ solve the following estimating equations:

$$\sum_{p \in G_\ell} \sum_{j=1}^{n_1} \mathbf{X}_{pj} \frac{\phi(z_{pj})}{\Phi(z_{pj})\,(1 - \Phi(z_{pj}))}\,(U_{pj} - \Phi(z_{pj})) = 0,$$

where $\mathbf{X}_{pj} = (1, \Phi^{-1}(p), X_j, X_{1j})^T$, and $z_{pj} = \mathbf{X}_{pj}^T \boldsymbol{\theta}$.

We refer to this approach as ROC-GLM.

Compare with Likelihood Approach

Assume:

- $Y_{0i}|X_i \overset{\text{iid}}{\sim} \mathcal{N}(\gamma_0 + \zeta_0 X_i, \sigma_0^2) =_d P_{\tau_0}$
- $Y_{1j}|(X_j, X_{1j}) \overset{\text{iid}}{\sim} \mathcal{N}(\gamma_0 + \gamma_1 + (\zeta_0 + \zeta_1)X_j + \zeta_2 X_{1j}, \sigma_1^2) =_d P_{\tau_1}$

Then, defining $\boldsymbol{\tau} = (\gamma_0, \gamma_1, \zeta_0, \zeta_1, \sigma_0, \sigma_1)$:

- $\hat{\boldsymbol{\tau}} = \arg\max_{\tau} \sum_{i=1}^{n_0} \log p_{\tau_0}(Y_{0i}; X_{0i}) + \sum_{j=1}^{n_1} \log p_{\tau_1}(Y_{1j}; X_j, X_{1j})$
- $\mathrm{ROC}_{Y_0,Y_1|X,X_1}(s) = \Phi\left(\frac{\gamma_1}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\Phi^{-1}(p) + \frac{\zeta_1}{\sigma_1}X + \frac{\zeta_2}{\sigma_1}X_1\right).$

Simulation Studies

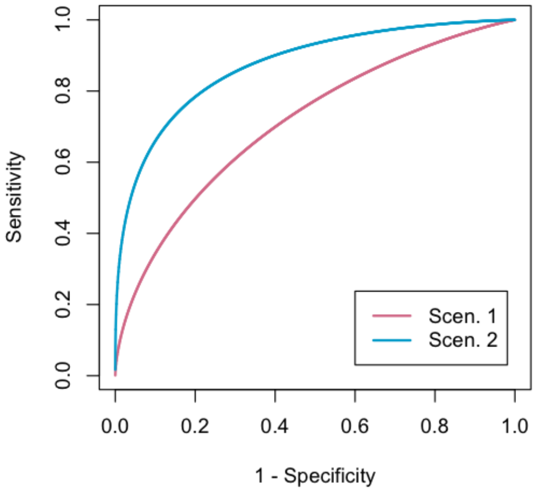
- Reminder: Trade-off between computational and statistical efficiency based on G_ℓ
- Reminder: Efficiency loss of ROC-GLM when likelihood is correctly specified
- ROC-GLM robustness to misspecified ROC curve

Simulation Setup

Consider case of *no* covariates, for simplicity.

- $Y_{0i} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$; $Y_{1j} \overset{\text{iid}}{\sim} \mathcal{N}\left(\alpha_0/\alpha_1, 1/\alpha_1^2\right)$
- ... this is done so that $\text{ROC}_{Y_0,Y_1}(p)$ is binormal with parameters α_0 and α_1
  - Scenario 1: $\boldsymbol{\alpha} = (0.75, 0.90)$
  - Scenario 2: $\boldsymbol{\alpha} = (1.50, 0.85)$

Simulation Setup



Efficiency with Choice of G_ℓ

G_ℓ divides the interval $[0, 1]$ into ℓ equally spaced subdivisions:

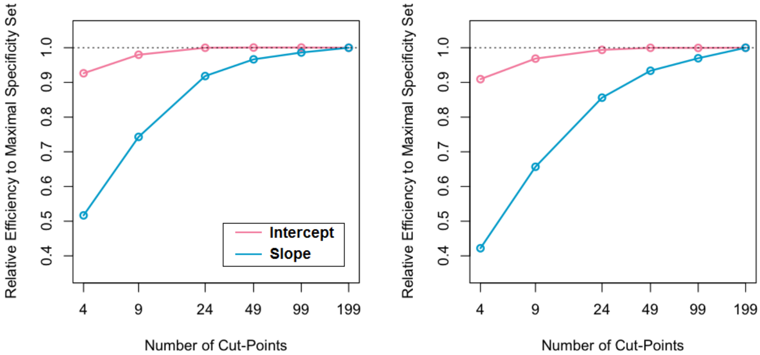
$$G_\ell = \left\{ \frac{i}{\ell} : i = 1, \ldots, \ell \right\}$$

Goal: determine the effect of ℓ on statistical and computational efficiency.

Recall: Statistical Efficiency Loss

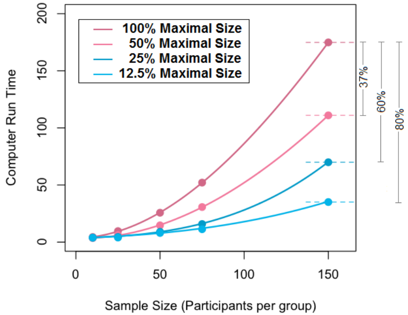
Consider $n_0 = n_1 = 200$:

$\boldsymbol{\alpha} = (0.75, 0.90)$ $\boldsymbol{\alpha} = (1.50, 0.85)$

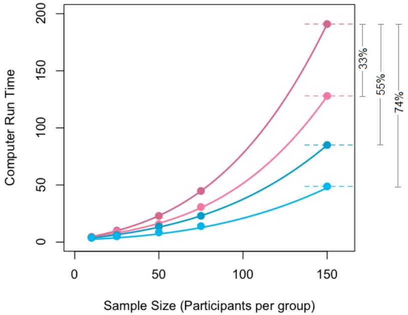


Recall: Computational Efficiency Gain

$\boldsymbol{\alpha} = (0.75, 0.90)$



$\boldsymbol{\alpha} = (1.50, 0.85)$



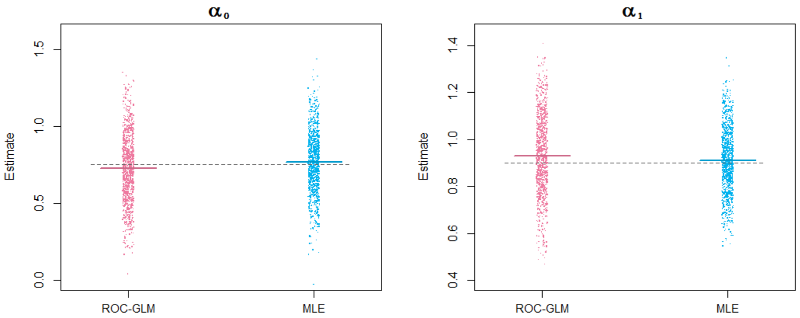
Simulation Setup

Consider case of *no* covariates, for simplicity.

- $Y_{0i} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$; $Y_{1j} \overset{\text{iid}}{\sim} \mathcal{N}\left(\alpha_0/\alpha_1, 1/\alpha_1^2\right)$
- ...this is done so that $\text{ROC}_{Y_0,Y_1}(p)$ is binormal with parameters α_0 and α_1
  - Scenario 1: $\boldsymbol{\alpha} = (0.75, 0.90)$
  - Scenario 2: $\boldsymbol{\alpha} = (1.50, 0.85)$
- $n_0 = n_1 = 50$
- G_ℓ is maximal ($\ell = 50$)

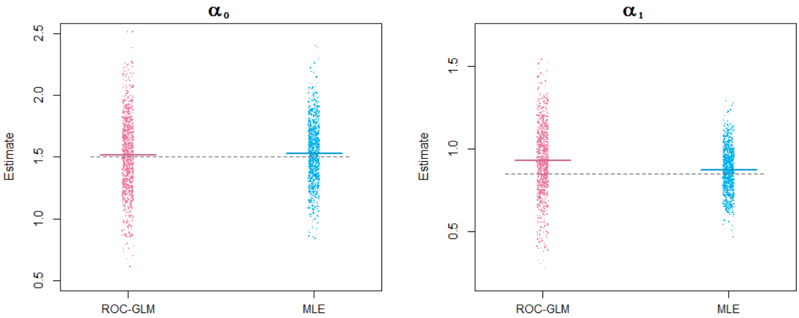
Comparison to MLE

Scenario 1: $\alpha = (0.75, 0.90)$



Comparison to MLE

Scenario 2: $\alpha = (1.50, 0.85)$



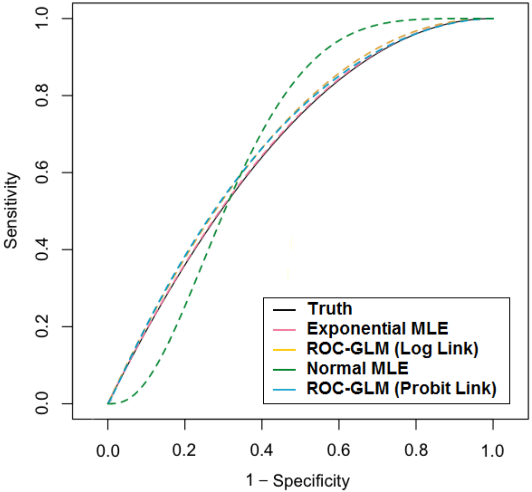
Model Misspecification

Suppose $Y_{0i} \overset{\text{iid}}{\sim}$ **Exponential**(2) and $Y_{1j} \overset{\text{iid}}{\sim}$ **Exponential**(4), so that

$$\mathrm{ROC}_{Y_0,Y_1}(p) = \exp\left(\frac{4}{2}\log(p)\right) = p^2.$$

Some Options:	**Correct?**	**Hope works?**
Exponential MLE	✓	✓
ROC-GLM (log-link)	✓	✓
Normal MLE	✗	✗
ROC-GLM (probit-link)	✗	✓

Model Misspecification



CPAO Data

- Childhood Predictors of Adult Obesity Study (CPAO)
- 823 adults (133 obese and 690 non-obese)
- Determine whether childhood BMI can predict adult obesity
- Adjusted model: include age, sex (for everyone), and adult BMI (for the obese participants) in the model

CPAO Example

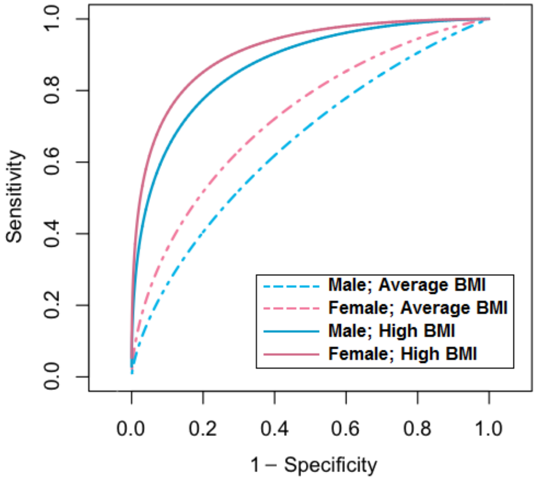
ROC-GLM Models for CPAO Data

Method	Estimate	Bootstrap S.E.	p-value
Adjusted:			
Intercept	0.150	0.376	0.69
Age	0.0669	0.0276	0.012
Gender	-0.284	0.238	0.23
Adult BMI	0.236	0.136	0.083
$\Phi^{-1}(p)$	0.923	0.0926	< 0.001

Table 4. *Results of ROC regression analysis applied to the CPAO study*

Variable	Coefficient	Standard error	*p*-value
Intercept	0.210	0.225	0.348
AGE (years)	0.080	0.014	<0.0001
GENDER (female = 0, male = 1)	−0.313	0.185	0.090
aBMIz (*z*-score)	0.285	0.084	0.001
$\Phi^{-1}(t)$	1.140	0.069	<0.0001

CPAO Example



In Summary

- Computational efficiency gains with fewer cut-points ✓
- . . . at a cost of "statistical" efficiency ✗/✓
- Loss might be acceptable when working with an absolutely massive data set ✗/✓
- Robustness to model misspecification ✓✓✓✓✓
- We want a more compelling reason to adopt this method over full-size G_ℓ
  - Consider that part of the critique
  - See "extra" slides on bootstrap estimation

Extra Slides

- CPAO example interpretation
- Bootstrap standard errors

CPAO Example Interpretation

ROC-GLM Models for CPAO Data

Method	Estimate	Bootstrap S.E.	p-value
Age	0.0669	0.0276	**0.012**

"Two obese adults of the same gender and adult BMI, but differing in age by one year, differ in estimated probability of having a BMI exceed the healthy quantiles of the same respective covariates by 0.0669 on the probit scale (with the older participant having the higher probability)"

Bootstrap

Variance Estimation

Scenario:	**1**	**2**
$\boldsymbol{\alpha} =$	(0.75, 0.90)	(1.50, 0.85)
$n_0 = n_1 =$	(50, 50)	(50, 50)
Simulated Var$[\hat{\alpha}]$		
$\ell = 50$	(0.047, 0.026)	(0.094, 0.045)
$\ell = 10$	(0.061, 0.043)	(0.11, 0.072)
$\hat{\mathrm{Var}}[\hat{\alpha}]$		
$\ell = 50$	(0.030, 0.015)	(0.069, 0.027)
$\ell = 10$	(0.065, 0.049)	(0.13, 0.077)