Simulation Algorithms and Methodology of Eli Lilly Biomarker Project

Bel and Cathy

1 Summary of the Chen and Ghosal (2021) Paper

This paper discusses advances in the analysis of Receiver Operating Characteristic (ROC) curves using placement values (PVs). The PV of a diseased test score measures the proportion of healthy test scores greater than the diseased score. The authors propose modeling PVs instead of traditional ROC analysis approaches and compare several parametric models using simulation.

The placement value of a diseased test score y is the proportion of healthy test scores with values greater than y. Mathematically, the placement value Z is defined as:

$$Z = \Pr(X > y) = 1 - F_0(y)$$

where X represents the test scores from the healthy population, and F_0 is the cumulative distribution function (CDF) of X. In this way, the placement value Z quantifies how well a test score from the diseased population separates from the scores in the healthy population.

Statistical Models

The paper evaluates four parametric models for placement values $Z \in (0, 1)$:

- **Logit-Normal Model:**

$$\tau(Z) = \log\left(\frac{Z}{1-Z}\right) \sim N(\mu, \sigma^2)$$

where the ROC curve is estimated as:

$$\text{ROC}(t) = \Phi\left(\frac{\tau(t) - \mu}{\sigma}\right), \quad t \in (0, 1)$$

and the area under the curve (AUC) is estimated as:

$$\text{AUC} = 1 - \tau^{-1}(\mu)$$

- **Probit-Normal Model:**

$$\tau(Z) = \Phi^{-1}(Z) \sim N(\mu, \sigma^2)$$

where Φ is the CDF of the standard normal distribution.

- **Log-Normal Model:**

$$Z \sim \text{Log-Normal}(\theta, \lambda^2)$$

where the ROC curve is expressed as:

$$\text{ROC}(t) = F_{\text{LN}}(t; \theta, \lambda)$$

and the AUC is:

$$\text{AUC} = 1 - e^{\theta + \lambda^2/2}$$

- **Beta Model:**

$$Z \sim \text{Beta}(a, b)$$

the ROC curve is derived as:

$$\text{ROC}(t) = F_{\text{Beta}}(t; a, b)$$

where F_{Beta} is the cumulative distribution function (CDF) of the Beta distribution with parameters a and b, evaluated at t.

And the AUC is:

$$\text{AUC} = \frac{b}{a + b}$$

Methodology and Simulation Setup

The authors conduct a simulation study to compare the performance of these models under three data-generating mechanisms:

1. Bi-Normal Mechanism
2. Bi-Gamma Mechanism
3. Bi-Mixture-Normal Mechanism

The performance metrics used are:

- Bias of AUC estimates
- Empirical Mean Squared Error (EMSE) of ROC curve estimates

For each scenario, 1000 datasets are simulated with sample sizes $n = 100, 200, 400$ and varying true AUC levels (0.5 to 0.9). The models are fit to the simulated data and compared using both analytic-based and sampling-based estimators.

Simulation Results

Key findings from the simulation include:

- The *Probit-Normal* and *Logit-Normal* models perform best, especially under the Bi-Normal mechanism.
- The *Log-Normal* model exhibits significant bias, making it impractical for use in most cases.
- The *Beta* model shows reasonable performance but has larger biases at high AUC values compared to the normal models.

Conclusion

The paper concludes that normal-based models (Logit and Probit) are the most reliable for modeling placement values in ROC curve analysis, with Probit-Normal showing slightly better performance. The authors also note that the Beta model, while reasonable, underperforms compared to the normal-based models.

2 Modified model with covariates

2.1. Probit-Normal Model with Covariate

In the Probit-Normal model, the placement value Z is transformed by the inverse normal CDF (probit function). To include the covariate $\mathbf{X}$, we specify the mean as a function of $\mathbf{X}$:

$$\Phi^{-1}(Z) \sim N(\mu(\mathbf{X}), \sigma^2)$$

where:

$$\mu(\mathbf{X}) = \alpha + \beta\mathbf{X}$$

The ROC curve is then given by:

$$\text{ROC}(t|\mathbf{X}) = \Phi\left(\frac{\Phi^{-1}(t) - \mu(\mathbf{X})}{\sigma}\right)$$

and the AUC is:

$$\text{AUC}(\mathbf{X}) = 1 - \Phi(\mu(\mathbf{X}))$$

2.2. Logit-Normal Model with Covariate

In the Logit-Normal model, the placement value Z is transformed using the logit function. With covariate $\mathbf{X}$, we model the linear predictor $\mu(\mathbf{X})$ as follows:

$$\text{logit}(Z) = \log\left(\frac{Z}{1-Z}\right) \sim N(\mu(\mathbf{X}), \sigma^2)$$

where:

$$\mu(X) = \alpha + \beta\mathbf{X}$$

The ROC curve becomes:

$$\text{ROC}(t|\mathbf{X}) = \Phi\left(\frac{\text{logit}(t) - \mu(\mathbf{X})}{\sigma}\right)$$

and the AUC is:

$$\text{AUC}(\mathbf{X}) = 1 - \text{logit}^{-1}(\mu(\mathbf{X}))$$

2.3. Beta Model with Covariate

In the Beta model, the placement value Z follows a Beta distribution. To incorporate the covariate $\mathbf{X}$, we model the mean $\mu(\mathbf{X})$ and the scale parameter $\phi = a + b$ through a regression framework.

The beta regression model for the mean μ_t(using Sarah paper's notation) and variance $\text{Var}(Z)$ is given by:

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t$$

where $g(\cdot)$ is a monotonic link function, and x_{ti} are observations on k covariates. For the logit link function, we model the mean as:

$$\mu_t = \frac{1}{1 + e^{-x_t^\top \beta}}$$

Given the beta distribution for $Z \sim \text{Beta}(a, b)$, the parameters a and b can be expressed in terms of μ and the scale parameter ϕ:

$$a(\mathbf{X}) = \phi \cdot \mu(t), \quad b(\mathbf{X}) = \phi \cdot (1 - \mu(t))$$

The ROC curve for the beta model is then expressed as:

$$\text{ROC}(t|\mathbf{X}) = F_{\text{Beta}}\left(t; a(\mathbf{X}), b(\mathbf{X})\right)$$

And the AUC for the beta model is given by:

$$\text{AUC}(\mathbf{X}) = \frac{b(\mathbf{X})}{a(\mathbf{X}) + b(\mathbf{X})}$$

3 Derivation of TRUE ROC and AUC with Generic Notations

Model Setup:

$$Y_1 = c_1 + d_1 X + \epsilon_1, \quad \epsilon_1 \sim N(0, \sigma_1^2)$$
$$Y_0 = c_0 + d_0 X + \epsilon_0, \quad \epsilon_0 \sim N(0, \sigma_0^2)$$

Difference Between Diseased and Non-Diseased:

$$Y_1 - Y_0 = (c_1 - c_0) + (d_1 - d_0)X + (\epsilon_1 - \epsilon_0)$$

where

$$\epsilon_1 - \epsilon_0 \sim N(0, \sigma_1^2 + \sigma_0^2)$$

3.1. Probit Model

Placement Value:

$$Z = \Phi\left(\frac{(c_1 - c_0) + (d_1 - d_0)X}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right)$$

ROC Curve:

$$\mathrm{ROC}(t) = \Phi\left(\frac{(c_1 - c_0) + (d_1 - d_0)X}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\Phi^{-1}(t)\right)$$

AUC:

$$\mathrm{AUC}(X) = \Phi\left(\frac{(c_1 - c_0) + (d_1 - d_0)X}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right)$$

3.2. Logit Model

Placement Value:

$$Z = \mathrm{logit}^{-1}\left(\frac{(c_1 - c_0) + (d_1 - d_0)X}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right)$$

ROC Curve:

$$\mathrm{ROC}(t) = \mathrm{logit}^{-1}\left(\frac{(c_1 - c_0) + (d_1 - d_0)X}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\mathrm{logit}^{-1}(t)\right)$$

AUC:

$$\mathrm{AUC}(X) = \mathrm{logit}^{-1}\left(\frac{(c_1 - c_0) + (d_1 - d_0)X}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right)$$

4 Theoretical models for PV in ROC (Simulation)

Step 1: Binormal Data Generation

- The covariate $\mathbf{X}$ is uniformly distributed between 0 and 1:

$$\mathbf{X} \sim U(0, 1)$$

- For the diseased population, $\mathbf{Y}_1$ is generated as follows:

$$\mathbf{Y}_1 = 2 + 4\mathbf{X} + \epsilon_1, \quad \epsilon_1 \sim N(0, 1.5^2)$$

- For the non-diseased population, Y_0 is generated as:

$$\mathbf{Y}_0 = 1.5 + 3\mathbf{X} + \epsilon_0, \quad \epsilon_0 \sim N(0, 1.5^2)$$

Step 2: Quantile Regression for the Reference Population

- We fit a quantile regression model on the non-diseased population to estimate the reference survival function:

$$\hat{S}_{0,\mathbf{X}}(t) = \text{QuantileRegression}(\mathbf{Y}_0 \sim \mathbf{X}, \text{quantiles} = t)$$

- This provides covariate-specific survival function estimates for the reference population at each t.

Step 3: Placement Values for the Diseased Population

- The placement values PV_i for each observation in the diseased population are computed by interpolating the survival estimates from the reference population for each observed Y_1:

$$PV_i = 1 - S_{0,\mathbf{X}}(\mathbf{Y}_1|\mathbf{X})$$

Step 4: Beta Regression for Placement Values

- After obtaining the placement values PV, we fit a beta regression model using the logit link function:

$$PV \sim \text{Beta}\left(\frac{1}{1+\exp(-(\beta_0+\beta_1\mathbf{X}))}, \frac{1-\frac{1}{1+\exp(-(\beta_0+\beta_1\mathbf{X}))}}{1+\exp(-(\beta_0+\beta_1\mathbf{X}))}\right)$$

- Using the estimates from beta regression, the beta parameters are computed as:

$$\hat{\alpha} = \left(\frac{1}{1+\exp(-(\hat{\beta}_0+\hat{\beta}_1\mathbf{X}))}\right)\cdot\hat{\phi}$$

$$\hat{\beta} = \left(1-\frac{1}{1+\exp(-(\hat{\beta}_0+\hat{\beta}_1\mathbf{X}))}\right)\cdot\hat{\phi}$$

ROC Curve for Beta Model

- The ROC curve for the beta model is computed by evaluating the Beta CDF at different values of the false positive rate (FPR) t:

$$\text{ROC}_{\text{beta}}(t) = F_{\text{Beta}}(t;\alpha,\beta)$$

True ROC for Binormal Model (Comparison)

- The true ROC curve for the binormal model is calculated as:

$$\text{True ROC}(t) = \Phi\left(\frac{c_1+d_1X-(c_0+d_0X)}{\sigma_1}+\frac{\sigma_0}{\sigma_1}\Phi^{-1}(t)\right)$$

- The true AUC for the binormal model is:

$$\text{True AUC}(X) = \Phi\left(\frac{c_1+d_1X-(c_0+d_0X)}{\sqrt{\sigma_1^2+\sigma_0^2}}\right)$$

Step 5: Probit Model

- The placement values are transformed using the inverse normal CDF to get the probit-transformed placement values:

$$PV_{\text{probit}} = \Phi^{-1}(PV)$$

- A linear model is fit using the probit-transformed placement values:

$$PV_{\text{probit}} \sim \beta_0 + \beta_1 \mathbf{X}$$

- The ROC curve for the probit model is then computed as:

$$\text{ROC}_{\text{probit}}(t) = \Phi\left(\frac{\Phi^{-1}(t) - (\beta_0 + \beta_1 \mathbf{X})}{\sigma_{\hat{\epsilon}}}\right)$$

where $\sigma_{\hat{\epsilon}}$ is the residual standard deviation from the probit regression model.
- The AUC for the probit model is:

$$\text{AUC}_{\text{probit}}(X) = 1 - \Phi(\beta_0 + \beta_1 \mathbf{X})$$

Step 6: Logit Model

- The placement values are transformed using the logit function to get the **logit-transformed placement values**:

$$PV_{\text{logit}} = \log\left(\frac{PV}{1 - PV}\right)$$

- A linear model is fit using the logit-transformed placement values:

$$PV_{\text{logit}} \sim \beta_0 + \beta_1 X$$

- The ROC curve for the logit model is then computed as:

$$\text{ROC}_{\text{logit}}(t) = \Phi\left(\frac{\log\left(\frac{t}{1-t}\right) - (\beta_0 + \beta_1 \mathbf{X})}{\sigma_{\hat{\epsilon}}}\right)$$

where $\sigma_{\hat{\epsilon}}$ is the residual standard deviation from the logit regression model.
- The AUC for the logit model is:

$$\text{AUC}_{\text{logit}}(\mathbf{X}) = 1 - \text{logit}^{-1}(\beta_0 + \beta_1 \mathbf{X})$$

Step 7: Mean Squared Error (MSE) Computation

- The **Mean Squared Error (MSE)** for the ROC curve is calculated by comparing the estimated ROC curves (from the beta, probit, and logit models) to the true binormal ROC curve:

$$\text{MSE}_{\text{model}} = \frac{1}{n_T} \sum_{i=1}^{n_T} \left(\text{ROC}_{\text{model}}(t_i) - \text{True ROC}(t_i)\right)^2$$

Step 8: Comparison of AUC Values

- The estimated AUCs from the beta, probit, and logit models are compared to the true binormal AUC:

$$\text{AUC Comparison} = |\text{AUC}_{\text{model}} - \text{True AUC}|$$

5 Comments regarding questions about modeling covariates

I have been thinking about my response, this may be a little rough but hopefully we can work through the rough patches and decide the best way forward.

- I do not think the beta model approach is meant for the problem at hand. Issues include;
  - The beta approach is used to model probabilities, in this case the survival value for a disease subject using a survival function created for non-diseased subjects where the endpoint(biomarker) for diseased and non-diseased subjects is a continuous rv Y. No distribution assumptions are made for Y. The CDF for the placement values is the ROC. The ROC provides a graphic for the extent of overlap between the diseased and non-diseased groups when using a biomarker or endpoint Y. Once the placement values(Z) are determined, the rv Y is no longer used.
  - The other three approaches are modeling transformations of Z with normal or log-normal distributions. The logit-normal model models the log odds for Z with a normal distribution.[1] They do share a common property in that the ROC and AUC are known functions of the simulation parameters which is very useful in accessing simulations for estimating the AUC and ROC. The beta approach is entirely different. The parameters use to simulate the biomarker for the diseased and non-diseased groups can be used to create cases where the two group's overlap is large or small. Yet,

[1] Since, I have not used any of these I do not have an opinion to share concerning their merit.

this visual overlap doesn't provide the exact values for either the AUC or ROC. Instead, one must rely upon the placement probabilities which are used to form estimates (not exact values) for the beta parameters. Which in turn provide estimates for the AUC and ROC.

– There are questions as to the validity of using quantile regression models for estimating Z. Allow me to give you some historical context. Originally, I and my students where modeling the AUC regression model in the presence of covariates (discrete and continuous). The AUC is the expected value of the Mann-Whitney statistic which is formed using concurrent and disconcurrent pairs. At that time, Alonzo and Pepe were considering a similar problem for which they used quantile regression to compute the needed Mann-Whitney statistics. They placed their code on STATA and we used it. My students modified it for use in R and Sarah extended it to the ROC regression model. Katie placed Sarah's code into RStudio. I have access to her code and ran parts (no lengthy simulations).

  So what did I do? I compute the survival curve for the non-diseased subjects using the endpoint Y as the event of interest. SAS PROC LIFEREG tables for Y, $F_Y(y)$, and $Z = 1 - F_Y(y)$. I now add Y for the diseased data and sort the combined data from smallest to largest. The values for $F_Y(\cdot)$ and $Z = 1 - F_Y(\cdot)$ are missing for the diseased data. I fill in the missingness by using the order of Y and the first non-missing $F_Y \leq Y$. This part of the procedure is basically just a table look-up method. The placement values for the diseased data are then used as the dependent variable in the Beta regression. The non-diseased group are never used again since the information that I want comes from the beta output on the diseased data.

  This is an important property that the other models do not have. The two group problem has been reduced to one group (diseased subjects) and the adjustment for the non-diseased subjects is reflected in the placement values Z_i. In which case, one does not need to estimate the survival curve for the non-diseased group using the covariates under consideration in the diseased model.

Turning our attention to section 2.3 we have

- The covariate $\mathbf{X}$ is uniformly distributed between 0 and 1:

$$\mathbf{X} \sim U(0,1)$$

- For the diseased population, $\mathbf{Y}_1$ is generated as follows:

$$\mathbf{Y}_1 = 2 + 4\mathbf{X} + \epsilon_1, \quad \epsilon_1 \sim N(0, 1.5^2)$$

- For the non-diseased population, Y_0 is generated as:

$$\mathbf{Y}_0 = 2 + 3\mathbf{X} + \epsilon_0, \quad \epsilon_0 \sim N(0, 1.5^2)$$

Recall that the placement value method is distribution free. The above models may be useful for comparison purposes (with the other models) but these values are meaningless in light of what the survival function for the non-diseased observations will be. I propose changing the model for the non-diseased population to

$$\mathbf{Y}_0 = \mu_0 + \epsilon_0, \quad \epsilon_0 \sim N(0, 1.5^2)$$

and for diseased population to

$$\mathbf{Y}_1 = \mu_0 + 4\mathbf{X} + \epsilon_1, \quad \epsilon_1 \sim N(0, 1.5^2)$$

for μ_0 is any specified constant. Under these changes, we have that the values for Y_1 are 2 units greater than Y_0 on average. Yet, nothing is known about either groups' survival curve nor the placement values for any subject.

– The beta method is very intuitive and easy to compute. Its appeal is seen in its ability to characterize the extent of overlap between the two groups when using any continuous biomarker, Y. This overlap is probably best characterized by the Youden index.

I threw some shade on using quantile regression for estimating the placement values. The data set has to be large before the output of the quantile regression has any real merit. For example, what is the .99 quantile when the sample size is 50? Furthermore, one must compute the regression at every value of $t \in (0,1)$. The output of the quantile tends to be very similar when t is small,

say $t \leq .05$ and when t is large, say $t \geq .9$. This lack of precision for placement values close to 0 or 1 has a huge impact on the accuracy of the estimated beta parameters.

A similar argument could be made when using the estimated survival curve. Indeed, I find little value in the precision of either the ROC or the AUC as apposed to being able to say the AUC is "big enough" when compared to .5 and that the ROC is not near the diagonal line when 1 - specificity is less than .5. I say this because I doubt that the new drug regulators will ever use methods based upon the diagnostic testing literature. Instead, these methods have potential value as in-house investigative tools for problems concerning "borrowing", biomarker identification and potential "cutpoints".[2]

[2]Furthermore, I believe (this is my opinion) that activities of this type is what industry statisticians should be doing as compared to writing methodological papers.