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:

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

- The true AUC for the binormal model is:

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

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:

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