

# P8160 - Breast Cancer Data: To lasso or to not lasso

Amy Pitts, Hun Lee, Jimmy Kelliher,  
Tucker Morgan, and Waveley Qiu

2022-03-28

# Motivation

Diagnosing breast cancer is extremely important.

According to NIH there has been an estimated:

- ▶ 281,550 new cases of breast cancer in women in 2021,
- ▶ 43,600 breast cancer in women related deaths in 2021.

American Cancer Society Guideline for Breast Cancer Screening:

- ▶ Women between ages 25-40 should have an annual clinical breast examination.
- ▶ Women between ages 40-44 should begin annual screening via mammogram
- ▶ Women between ages 45-54 should be screened annually via mammogram

# Goal

With using all the collected image data we want to develop an algorithm to predict diagnosis. Since diagnosis is a binary outcome a logistic regression will be utilized.

Methods:

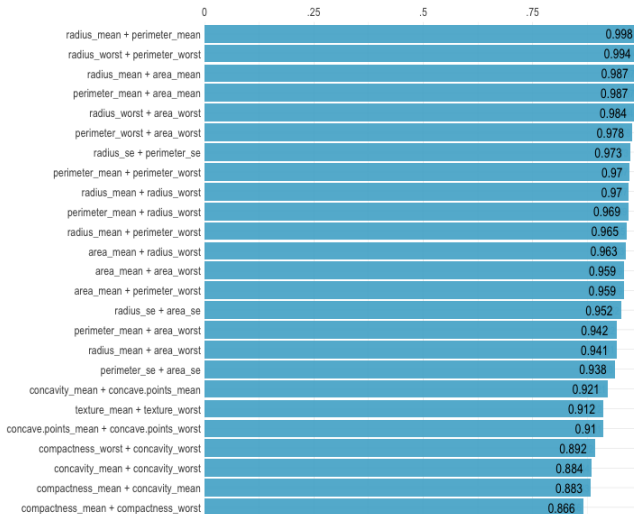
- ▶ Newton-Raphson Algorithm (Full Model)
- ▶ Logistic LASSO Algorithm (Optimal Model)

# Data

- ▶ 569 rows and 31 columns all related to breast tissue images
- ▶ Outcome of interest: Diagnosis (B or M)
  - ▶ 357 benign (B) cases and 212 malignant (M) cases
- ▶ The Covariates include information such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

# Figure 1: Ranked Cross-Correlations

25 most relevant



# Figure 1: Ranked Cross-Correlations

25 most relevant



Best Representative radius\_worst

# Figure 1: Ranked Cross-Correlations

25 most relevant



Best Representative radius\_se

# Remaining Variables

Variable	Diagnosis Received		
	B, N = 357 <sup>†</sup>	M, N = 212 <sup>†</sup>	p-value <sup>‡</sup>
texture_mean	17.91 (4.00)	21.60 (3.78)	<0.001
smoothness_mean	0.09 (0.01)	0.10 (0.01)	<0.001
compactness_mean	0.08 (0.03)	0.15 (0.05)	<0.001
concave points_mean	0.03 (0.02)	0.09 (0.03)	<0.001
symmetry_mean	0.17 (0.02)	0.19 (0.03)	<0.001
fractal_dimension_mean	0.06 (0.01)	0.06 (0.01)	0.5
radius_se	0.28 (0.11)	0.61 (0.35)	<0.001
texture_se	1.22 (0.59)	1.21 (0.48)	0.6
smoothness_se	0.01 (0.00)	0.01 (0.00)	0.2
compactness_se	0.02 (0.02)	0.03 (0.02)	<0.001
concavity_se	0.03 (0.03)	0.04 (0.02)	<0.001
concave points_se	0.01 (0.01)	0.02 (0.01)	<0.001
symmetry_se	0.02 (0.01)	0.02 (0.01)	0.028
fractal_dimension_se	0.00 (0.00)	0.00 (0.00)	<0.001
radius_worst	13.38 (1.98)	21.13 (4.28)	<0.001
smoothness_worst	0.12 (0.02)	0.14 (0.02)	<0.001
compactness_worst	0.18 (0.09)	0.37 (0.17)	<0.001
concavity_worst	0.17 (0.14)	0.45 (0.18)	<0.001
symmetry_worst	0.27 (0.04)	0.32 (0.07)	<0.001
fractal_dimension_worst	0.08 (0.01)	0.09 (0.02)	<0.001

<sup>†</sup> Statistics presented: Mean (SD)

<sup>‡</sup> Statistical tests performed: Wilcoxon rank-sum test



## Full Model (Newton-Raphson)

To implement the Newton-Raphson Method we need the likelihood, gradient, and hessian matrix. Let

$$\pi_i = P(Y_i = 1 | x_{i,1}, \dots, x_{i,p}) = \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j x_{i,j}}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j x_{i,j}}}.$$

**The log-likelihood:**

$$l(\mathbf{X} | \vec{\beta}) = \sum_{i=1}^n \left[ y_i \left( \beta_0 + \sum_{j=1}^p \beta_j x_{i,j} \right) - \log \left( 1 + \exp \left( \beta_0 + \sum_{j=1}^p \beta_j x_{i,j} \right) \right) \right]$$

**The gradient:**

$$\nabla l(\mathbf{X} | \vec{\beta}) = \left[ \sum^n y_i - \pi_i \quad \sum^n x_{i,1}(y_i - \pi_i) \quad \dots \quad \sum^n x_{i,p}(y_i - \pi_i) \right]_{1 \times (p+1)}^T$$

**The hessian:** produces a matrix  $(p + 1 \times p + 1)$

$$\nabla^2 l(\mathbf{X} | \vec{\beta}) = - \sum_{i=1}^n \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{x} \end{pmatrix} \pi_i (1 - \pi_i)$$

# Optimal Model (Logistic LASSO)

also going to be some math

# Optimal Model (Logistic LASSO)

more math

## 5-fold Cross Validation

## Cross Validation Results

Best  $\lambda$  using AUC

# LASSO Coefficients

Best  $\lambda$  using beta plot

# Coefficients Comparison

AUC



# Discussion

## Resources

Cancer Stat Facts: Female Breast Cancer. *National Cancer Institute*  
- *NIH* <https://seer.cancer.gov/statfacts/html/breast.html>

American Cancer Society. (2019). Breast cancer facts & figures 2019–2020. Am Cancer Soc, 1-44.