High Dimensional Binary Regression and Classification

David J. Olive and Mohanned S. Alsaudi *
Southern Illinois University

December 12, 2024

Abstract

Consider a binary regression model with binary response variable Y, that takes on values 0 and 1, with predictors $\boldsymbol{x} = (x_1, ..., x_p)$. Let n be the number of cases. For a high dimensional binary regression, n/p is small. We consider some high dimensional binary regression models that have some simple large sample theory. As is well known, binary regression estimators can be used for classification.

KEY WORDS: Dimension reduction, lasso, data spitting, marginal maximum likelihood estimator.

1 INTRODUCTION

This section reviews binary regression models, including variable selection and data splitting. Consider a binary regression model with binary response variable $Y \in \{0, 1\}$ and predictors $\boldsymbol{x} = (x_1, ..., x_p)$. Then there are n cases $(Y_i, \boldsymbol{x}_i^T)^T$, and the sufficient predictor $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$. For the binary regression models, the conditioning and subscripts, such as i, will often be suppressed. A binary regression model is $Y = Y|SP \sim$ binomial$(1, \rho(\text{SP}))$ where $\rho(SP) = P(Y = 1|SP)$. There are many binary regression models, including binary logistic regression, binary probit regression, and support vector machines (with $Z_i = 2Y_i - 1$). See Hosmer and Lemeshow (2000) and James et al. (2021). The binary logistic regression model has

$$\rho(SP) = \frac{e^{SP}}{1 + e^{SP}}.$$

Variable selection estimators include forward selection or backward elimination when $n \geq 10p$. When n/p is not large, sparse regression methods such as forward selection, lasso, and the elastic net can be useful: the binary logistic regression submodel uses the predictors that had nonzero sparse regression estimated coefficients. See Friedman et al. (2007), Friedman, Hastie, and Tibshirani (2010), and Zou and Hastie (2005).

*David J. Olive is Professor, School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901, USA.

Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \quad (1)$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (1) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I.$$

Thus $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. The model using $\boldsymbol{x}^T\boldsymbol{\beta}$ is the full model.

To clarify notation, suppose $p = 3$, a constant α is always in the model, and $\boldsymbol{\beta} = (\beta_1, 0, 0)^T$. Then the $J = 2^p = 8$ possible subsets of $\{1, 2, ..., p\}$ are $I_1 = \emptyset$, $S = I_2 = \{1\}$, $I_3 = \{2\}$, $I_4 = \{3\}$, $I_5 = \{1, 2\}$, $I_6 = \{1, 3\}$, $I_7 = \{2, 3\}$, and $I_8 = \{1, 2, 3\}$. There are $2^{p-a_S} = 4$ subsets I_2, I_5, I_6, and I_8 such that $S \subseteq I_j$. Let $\hat{\boldsymbol{\beta}}_{I_7} = (\hat{\beta}_2, \hat{\beta}_3)^T$ and $\boldsymbol{x}_{I_7} = (x_2, x_3)^T$.

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, use zero padding to form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets, e.g. $J = 2^p$.

Theory for the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$ is complicated. See Pelawa Watagoda and Olive (2021) for multiple linear regression, and Rathnayake and Olive (2023) for models such as GLMs and Cox (1972) proportional hazards regression. For fixed p, these two papers showed that $\hat{\boldsymbol{\beta}}_{VS}$ is $\sqrt{n}$ consistent with a complicated nonnormal limiting distribution.

The marginal maximum likelihood estimator (MMLE) is due to Fan and Lv (2008) and Fan and Song (2010). This estimator computes the marginal regression, such as the binary logistic regression, of Y on x_i resulting in the estimator $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M})$ for $i = 1, ..., p$. Then $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_{1,M}, ..., \hat{\beta}_{p,M})^T$.

For estimation with ordinary least squares (OLS) and the discriminant function, let the covariance matrix of $\boldsymbol{x}$ be $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T = E(\boldsymbol{x}\boldsymbol{x}^T) - E(\boldsymbol{x})E(\boldsymbol{x}^T)$ and $\boldsymbol{\eta} = \text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = E[(\boldsymbol{x} - E(\boldsymbol{X})(Y - E(Y))] = E(\boldsymbol{x}Y) - E(\boldsymbol{x})E(Y) = E[(\boldsymbol{x} - E(\boldsymbol{x}))Y] = E[\boldsymbol{x}(Y - E(Y))]$. Let

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_n = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \boldsymbol{S}_{\boldsymbol{x}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$$

and

$$\tilde{\boldsymbol{\eta}} = \tilde{\boldsymbol{\eta}}_n = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}).$$

Then the OLS estimators are $\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T \overline{\boldsymbol{x}}$ and

$$\hat{\boldsymbol{\beta}}_{OLS} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\eta}}.$$

For a multiple linear regression model with independent, identically distributed (iid) cases, $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ under mild regularity conditions, while $\hat{\alpha}_{OLS}$ is a consistent estimator of $E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x})$.

Another binary regression model is the discriminant function model. See Hosmer and Lemeshow (2000, pp. 43–44). Assume that $\pi_j = P(Y = j)$ and that $\boldsymbol{x}|Y = j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{pool})$ for $j = 0, 1$. That is, the conditional distribution of $\boldsymbol{x}$ given $Y = j$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_{pool}$ which does not depend on j. Notice that $\boldsymbol{\Sigma}_{pool} = \mathrm{Cov}(\boldsymbol{x}|Y) \neq \mathrm{Cov}(\boldsymbol{x})$. Then as for the binary logistic regression model,

$$P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})}.$$

Under the conditions above, the *discriminant function* parameters are given by

$$\boldsymbol{\beta} = \boldsymbol{\beta}_{DF} = \boldsymbol{\Sigma}_{pool}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{2}$$

$$\text{and} \quad \alpha = \log\left(\frac{\pi_1}{\pi_0}\right) - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_{pool}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

Under the above conditions (multivariate normality with the same covariance matrix but possibly different means), the population quantity estimated by the discriminant function model is the same as that estimated by logistic regression: $\boldsymbol{\beta} = \boldsymbol{\beta}_{DF} = \boldsymbol{\beta}_{LR}$. In general, the above conditions fail to hold, and $\boldsymbol{\beta} = \boldsymbol{\beta}_{DF} \neq \boldsymbol{\beta}_{LR}$.

To compare the OLS estimator with binary regression estimators such as binary logistic regression, Olive (2017a, pp. 396-397) gave the following derivation. Let $\pi_j = P(Y = j)$ for $j = 0, 1$. Let $\boldsymbol{\mu}_j = E(\boldsymbol{x}|Y = j)$ for $j = 0, 1$. Let N_i be the number of Ys that are equal to i for $i = 0, 1$. Then

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{N_i} \sum_{j:Y_j = i} \boldsymbol{x}_j$$

for $i = 0, 1$ while $\hat{\pi}_i = N_i/n$ and $\hat{\pi}_1 = 1 - \hat{\pi}_0$. Hence $\hat{\boldsymbol{\mu}}_i = \overline{\boldsymbol{x}}_i$ is the sample mean of the $\boldsymbol{x}_k$ corresponding to $Y_k = j$ for $j = 0, 1$. Then

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i Y_i - \overline{\boldsymbol{x}}\, \overline{Y}.$$

$$\text{Thus} \quad \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}} = \frac{1}{\mathrm{n}} \left[\sum_{\mathrm{j:Y_j=1}} \boldsymbol{x}_{\mathrm{j}}(1) + \sum_{\mathrm{j:Y_j=0}} \boldsymbol{x}_{\mathrm{j}}(0) \right] - \overline{\boldsymbol{x}}\, \hat{\pi}_1 =$$

$$\frac{1}{n}(N_1 \hat{\boldsymbol{\mu}}_1) - \frac{1}{n}(N_1 \hat{\boldsymbol{\mu}}_1 + N_0 \hat{\boldsymbol{\mu}}_0)\hat{\pi}_1 = \hat{\pi}_1 \hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1^2 \hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1 \hat{\pi}_0 \hat{\boldsymbol{\mu}}_0 =$$

$$\hat{\pi}_1(1-\hat{\pi}_1)\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 = \hat{\pi}_1\hat{\pi}_0(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0).$$

This result means

$$\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \pi_1\pi_0(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \tag{3}$$

and $\boldsymbol{\phi} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ are quantities of interest for binary regression. Note that

$$\boldsymbol{\beta}_{DF} = \frac{1}{\pi_1\pi_0}\boldsymbol{\Sigma}_{pool}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \frac{1}{\pi_1\pi_0}\boldsymbol{\Sigma}_{pool}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \frac{1}{\pi_1\pi_0}\boldsymbol{\Sigma}_{pool}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}.$$

Let $\boldsymbol{\beta} = \lambda\boldsymbol{\eta} = \gamma\boldsymbol{\phi}$. To compute $\hat{\lambda}$ or $\hat{\boldsymbol{\phi}}$, plug in $\hat{\boldsymbol{\eta}}^T\boldsymbol{x}$ or $\hat{\boldsymbol{\phi}}^T\boldsymbol{x}$ into a binary regression program such as logistic regression, probit regression, support vector machines (with $Z_i = 2Y_i - 1$), et cetera. Then $\hat{\boldsymbol{\beta}} = \hat{\lambda}\hat{\boldsymbol{\eta}}$ or $\hat{\boldsymbol{\beta}} = \hat{\gamma}\hat{\boldsymbol{\phi}}$ This procedure is very similar to the one component partial least squares estimator for multiple linear regression. See Olive and Zhang (2024).

Data splitting divides the training data set of n cases into two sets: H and the validation set V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. An application of data splitting is to use a variable selection method, such as forward selection or lasso, on H to get submodel I_{min} with a predictors, then fit the selected model to the cases in the validation set V using standard inference. See, for example, Rinaldo et al. (2019).

High dimensional regression has n/p small. A fitted or population regression model is sparse if a of the predictors are active (have nonzero $\hat{\beta}_i$ or β_i) where $n \geq Ja$ with $J \geq 10$. Otherwise the model is nonsparse. A high dimensional population regression model is abundant or dense if the regression information is spread out among the p predictors (nearly all of the predictors are active). Hence an abundant model is a nonsparse model.

Olive and Zhang (2024) proved that there are often many valid population models for binary regression, gave theory for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$, gave theory for data splitting estimators, and gave some theory for the MMLE for multiple linear regression.

Section 2 gives some large sample theory, including tests of hypotheses.

2 Large Sample Theory and Testing

The MMLE is interesting since if each predictor satisfies a marginal model, then the marginal model theory can be used to find a confidence interval for β_i for $i = 1, ..., p$ where β_i is the ith component of $\boldsymbol{\beta}_{MMLE}$. This regularity condition is quite strong for high dimensional binary regression.

Suppose $\boldsymbol{\beta}_{BR} = \lambda\boldsymbol{\eta}$ is found by plugging $\hat{\boldsymbol{\eta}}^T\boldsymbol{x}$ into a binary regression program to get $\hat{\lambda}$. Then $\hat{\boldsymbol{\beta}}_{BR} = \hat{\lambda}\hat{\boldsymbol{\eta}}$. Testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{BR} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\beta}_{BR} \neq \boldsymbol{0}$ is equivalent to testing $H_0 : \boldsymbol{A}\boldsymbol{\eta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\eta} \neq \boldsymbol{0}$ where $\boldsymbol{A}$ is a $k \times p$ constant matrix with $k \leq p$. If the cases $(\boldsymbol{x}_i, Y_i)$ are iid and $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, then the testing is exactly as in Olive and Zhang (2024).

Now suppose $\boldsymbol{\eta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\hat{\boldsymbol{\eta}} = \overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_0$ where $\overline{\boldsymbol{x}}_i$ is the sample mean of the predictors corresponding to $Y = i$ for $i = 0, 1$. Assume the cases in each group $Y = i$ are iid and that the groups are independent. Then the large sample theory for $\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_0$ is given by Rupasinghe Arachchige Don and Pelawa Watagoda (2018) and Rupasinghe

Arachchige Don and Olive (2019). To simplify the large sample theory, assume $n_i = \pi_i n$ where $0 < \pi_i < 1$ and $\sum_{i=1}^{2} \pi_i = 1$. Assume H_0 is true, and let $\boldsymbol{\mu}_i = \boldsymbol{\mu}$ for $i = 0, 1$. Suppose $\sqrt{n_i}(T_i - \boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_i)$, and $\sqrt{n}(T_i - \boldsymbol{\mu}) \xrightarrow{D} N_p\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_i}{\pi_i}\right)$. Then

$$\sqrt{n}[(T_1 - T_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)] \xrightarrow{D} N_p\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_1}{\pi_1} + \frac{\boldsymbol{\Sigma}_0}{\pi_0}\right).$$

Thus

$$\sqrt{n}[(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)] \xrightarrow{D} N_p\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_1}{\pi_1} + \frac{\boldsymbol{\Sigma}_0}{\pi_0}\right) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}}), \tag{4}$$

with

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{W}} = \frac{n\hat{\boldsymbol{\Sigma}}_1}{n_1} + \frac{n\hat{\boldsymbol{\Sigma}}_0}{n_0}.$$

High Dimensional Tests

Some tests when n/p is not large are simple. Testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{BR} = \mathbf{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\beta}_{BR} \neq \mathbf{0}$ is equivalent to testing $H_0 : \boldsymbol{A}\boldsymbol{\eta} = \mathbf{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\eta} \neq \mathbf{0}$ where $\boldsymbol{A}$ is a $k \times p$ constant matrix. Let $\mathrm{Cov}(\hat{\boldsymbol{\eta}}) = \boldsymbol{\Sigma}_{\boldsymbol{W}}$ be the asymptotic covariance matrix of $\hat{\boldsymbol{\eta}}$. In high dimensions where $n < 5p$, we can't get a good nonsingular estimator of $\mathrm{Cov}(\hat{\boldsymbol{\eta}})$, but we can get good nonsingular estimators of $\mathrm{Cov}((\hat{\eta}_{i1}, ..., \hat{\eta}_{ik})^T)$ with $\boldsymbol{u} = (x_{i1}, ..., x_{ik})^T$ where $n \geq Jk$ with $J \geq 10$. (Values of J much larger than 10 may be needed if some of the k predictors are skewed or if a π_i in near 0 or 1.) Simply use the sample covariance matrix with $\boldsymbol{u}$ replacing $\boldsymbol{x}$. Hence we can test hypotheses like $H_0 : \beta_i - \beta_j = 0$. In particular, testing $H_0 : \beta_i = 0$ is equivalent to testing $H_0 : \eta_i = 0$.

Data splitting uses model selection (variable selection is a special case) to reduce the high dimensional problem to a low dimensional problem. The above procedure also reduces the high dimensional problem to a low dimensional problem.

3 CONCLUSIONS

Binary regression is closely related to two sample tests. Note that $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$ can use other multivariate location estimators than sample means. For example, sample coordinatewise medians, sample coordinatewise trimmed means, and the Olive (2017b) T_{RMVN} estimator have large sample theory given by Rupasinghe Arachchige Don and Olive (2019) and Rupasinghe Arachchige Don and Pelawa Watagoda (2018).

Some papers on binary regression include Cai, Guo, and Ma (2023), Candès and Sur (2020), Mukherjee, Pillai, and Lin (2015), Sur and Candès (2019), Sur, Chen, and Candès (2019), and Tang and Ye (2020). Empirically, often $\boldsymbol{\beta}_{LR} \approx d\ \boldsymbol{\beta}_{OLS}$. Haggstrom (1983) suggests that d is not far from 1/MSE for logistic regression.

These binary regression estimators also give new ways to compare multivariate location estimators from two groups. The tests using k predictors can be performed. High dimensional tests for means from two groups can also be used. The tests that make very strong assumptions, such as multivariate normality or equal covariance matrices for the two groups, should be avoided. See Feng and Sun (2015), Gregory et al. (2015), Hu and Bai (2015), Rajapaksha and Olive (2024), and Xue and Yao (2020).

Software

The *R* software was used in the simulations. See R Core Team (2024). Programs will be added to the Olive (2025) collections of *R* functions *slpack.txt*, available from (http://parker.ad.siu.edu/Olive/slpack.txt).

References

Cai, T.T., Guo, Z., and Ma, R. (2023), "Statistical Inference for High-Dimensional Generalized Linear Models with Binary Outcomes," *Journal of the American Statistical*, 118, 1319-1332.

Candès, E.J., and Sur, P. (2020), "The Phase Transition for the Existence of the Maximum Likelihood Estimate in High-Dimensional Logistic Regression," *The Annals of Statistics*, 48, 27-42.

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society, B*, 70, 849-911.

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with np-Dimensionality," *The Annals of Statistics*, 38, 3217-3841.

Feng, L., and Sun, F. (2015), "A Note on High–Dimensional Two–Sample Test," *Statistics & Probability Letters,* 105, 29-36.

Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *Annals of Applied Statistics*, 1, 302-332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1-22.

Gregory, K.B., Carroll, R.J., Baladandayuthapani, V., and Lahari, S.N. (2015), "A Two–Sample Test for Equality of Means in High Dimension," *Journal of the American Statistical Association*, 110, 837-849.

Haggstrom, G.W. (1983), "Logistic Regression and Discriminant Analysis by Ordinary Least Squares," *Journal of Business & Economic Statistics*, 1, 229-238.

Hosmer, D.W., and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd ed., Wiley, New York, NY.

Hu, J., and Bai, Z. (2015), "A Review of 20 Years of Naive Tests of Significance for High-Dimensional Mean Vectors and Covariance Matrices," *Science China Mathematics*, 55, online.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021), *An Introduction to Statistical Learning with Applications in R*, 2nd ed., Springer, New York, NY.

Mukherjee, R., Pillai, N.S., and Lin, X. (2015), "Hypothesis Testing for High-Dimensional Sparse Binary Regression," *The Annals of Statistics,* 43, 352-381.

Olive, D.J. (2013), "Plots for Generalized Additive Models," *Communications in Statistics: Theory and Methods*, 42, 2610-2628.

Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2025), *Prediction and Statistical Learning*, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).

Olive, D.J., and Zhang, L. (2025), "One Component Partial Least Squares, High Dimensional Regression, Data Splitting, and the Multitude of Models," *Communications*

in Statistics: Theory and Methods, 54, 130-145.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021), "Comparing Six Shrinkage Estimators with Large Sample Theory and Asymptotically Optimal Prediction Intervals," *Statistical Papers*, 62, 2407-2431.

R Core Team (2024), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rajapaksha, K.W.G.D.H., and Olive, D.J. (2024), "Wald Type Tests with the Wrong Dispersion Matrix," *Communications in Statistics: Theory and Methods*, 53, 2236-2251.

Rathnayake, R.C., and Olive, D.J. (2023), "Bootstrapping Some GLM and Survival Regression Variable Selection Estimators," *Communications in Statistics: Theory and Methods*, 52, 2625-2645.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019), "Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference," *The Annals of Statistics*, 47, 3438-3469.

Rupasinghe Arachchige Don, H.S., and Olive, D.J. (2019), "Bootstrapping Analogs of the One Way MANOVA Test," *Communications in Statistics: Theory and Methods*, 48, 5546-5558.

Rupasinghe Arachchige Don, H.S., and Pelawa Watagoda, L.C.R. (2018), "Bootstrapping Analogs of the Two Sample Hotelling's T^2 Test," *Communications in Statistics: Theory and Methods*, 47, 2172-2182.

Sur, P., and Candès, E.J. (2019), "A Modern Maximum-Likelihood Theory for High-Dimensional Logistic Regression," *Proceedings of the National Academy of Sciences*, 116, 14516-14525.

Sur, P., Chen, Y., and Candès, E. (2019), "The Likelihood Ratio Test in High-Dimensional Logistic Regression is Asymptotically a Rescaled Chi-Square," *Probability Theory and Related Fields*, 175, 487-558.

Tang, W., and Ye, Y. (2020), "The Existence of Maximum Likelihood Estimate in High-Dimensional Binary Response Generalized Linear Models," *Electronic Journal of Statistics*, 14, 4028-4053.

Xue, K., and Yao, F. (2020), "Distribution and Correlation-Free Two-Sample Test of High-Dimensional Mean," *The Annals of Statistics*, 48, 1304-1328.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society Series, B,* 67, 301-320.