Application of modern maximum-likelihood variant for high-dimensional logistic regression

Adam Schaffroth

Whiting School of Engineering

Johns Hopkins University

August 22, 2022

Abstract

Logistic regression is a useful statistical model for prediction of variables that fall into categories. Classical statistics has produced theorems that guarantee the usefulness of logistic regression results with the assumption that certain criteria are met. Some of these classical results do not always hold with high dimensional datasets, That is, when the number of predictors p is large relative to the number of observations n. In fact, as the n and p increase with the ratio p/n staying fixed, 3 results can be observed: 1) The maximum likelihood estimates (MLE) is biased, 2) The variability of the MLE is higher than classical estimates, and 3) The likelihood ratio test (LRT) does not have an approximate χ^2 distribution. In addition to exploring these results, a method to correct for dimensionality will be introduced and shown to produce more accurate parameter estimates.

Introduction

Logistic regression is an excellent statistical linear model for classification. Given a binary response variable Y and random variable X, logistic regression models $Pr(Y = 1|X) =$

$p(X)$, where $Y \in \{0, 1\}$. To ensure that the model for $p(X)$ outputs a probability between 0 and 1, a *sigmoid* or *logistic function* is used. In the univariate case, $p(X)$ is modeled as

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

Once a threshold is defined (0.5 is the typical defualt choice), any $p(X)$ greater than or equal to the threshold is predicted to be 1 and 0 otherwise. Some algebraic manipulation reveals

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}[1].$$

This is this called the *odds*, and naturally by taking the natural logarithm of both sides we get $\beta_0 + \beta_1 X$ also known as the *log odds* or *logit*. By changing X by 1 increases the log-odds by β_1 or alternatively, multiplying the odds by e_1^{β}. The optimal estimates for parameters $\hat{\beta}_0, \hat{\beta}_1$ are found via maximum likelihood estimation. In this univariate case, first the log-likelihood function is found,

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} 1 - p(x_i).$$

Under the logistic model's assumptions of observational independence, no multicollinearity, linearity of independent variables and log-odds, and large sample size[2], the MLE should produce β parameter estimates that are asymptotically unbiased. Later in the paper, it will be demonstrated that this does not always hold with higher dimensional datasets. The standard error can be calculated to measure the accuracy of each parameter. Z-statistics are also computed to evaluate the null-hypothesis, where the z-statistic associated with $\beta_i = \hat{\beta}_i / SE(\hat{\beta}_i)$ which in this case is $H_0 : \beta_1 = 0 \implies p(X) = \frac{e_0^{\beta}}{1 + e_0^{\beta}}$. P-values are also computed based on these z-statistics.

To generalize to the logistic regression model to fitting data with p predictors for K classes,

$$\begin{aligned} log\frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} &= \beta_{10} + \beta_1^T x \\ log\frac{Pr(G=2|X=x)}{Pr(G=K|X=x)} &= \beta_{20} + \beta_2^T x \\ &\vdots \\ log\frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} &= \beta_{(K-1)0} + \beta_{K-1}^T x \end{aligned} \tag{1}$$

[3]

It is worth noting that the choice of class K as denominator is arbitrary and any replacement for K in the denominator would work just as well. The $K-1$ log odds ratios are calculated as

$$\begin{aligned} Pr(G=k|X=x) &= \frac{e^{\beta_{k0}+\beta_k^T x}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_l^T x}},\ k=1,\ldots,k=K-1 \\ Pr(G=K|X=x) &= \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_l^T x}} \end{aligned} \tag{2}$$

For the often occurring case when $K=2$, i.e. a binary response variable, the algorithm for maximizing the likelihood become much less complicated and thus this will be the case delineated below. When the binary response variable $y_i = 0$, let $g_i = 2$ and when $y_i = 1$ let $g_i = 1$. The log-likelihood function then becomes

$$\begin{aligned} \mathcal{L}(\beta) &= \sum_{i=1}^{N} \{y_i log p(x_i;\beta) + (1-y_i) log(1-p(x_i;\beta))\} \\ &= \sum_{i=1}^{N} \{y_i \beta^T x_i - log(1+e^{\beta^T x_i})\} \end{aligned} \tag{3}$$

To include the intercept, vector x_i has a 1 in it. The likelihood is then maximized by setting the *score equation* to zero and solving.

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i(y_i - p(x_i;\beta)) = 0 \tag{4}$$

which will generate a $p+1$ system of equations which is popularly solved via the Newton–Raphson algorithm, of which the details are beyond the scope of this paper. [4].

1 Logistic Regression Applied to Housing Data

To demonstrate how simple logistic regression can be applied, a *housing* dataset will be used. In this case, $p = 3$ and $n = 10536$ so it can reasonably be assumed that the conclusions reached by classical theory will apply and results can be trusted (depending on the quality of analysis of course!). This data contains pricing and sale information for various types of single family homes. In this demonstration, some of the data has been filtered and transformed. The *date* variable has been transformed into a binary variable $b4\_1990$ which will be 1 or 2 if the observation was made before or after 1990, respectively. The variable cat_idx refers to the category of observation $\in \{1, 2, 3\}$. Lastly, the variable val indicates the value at which the house was prices at or sold (depending on cat_idx). Due to the approximate gamma distribution of this variable, including all values resulted in unstable results. The value of 50 was chosen as an arbitrary threshold, above which "outliers" were discarded. Figure 1 shows the summarized output of the model produced in R. The model shows that, based on the Wald z-statistic, that all variables included could be statistically significant. The "Estimate" provides the coefficients, which can be interpreted as the MLE for the change in log-odds for a 1 unit increase for the associated predictor. For each \$1 increase in val, the log-odds of the observation being located in the midwest increase be 0.006792. If the sale was before 1990, the log-odds of $in_midwest$ increase by about 0.217. With regards to the cat_idx variable, an increase in the log-odds of around 0.734 is observed if cat_idx is 2 versus 1, and so one for higher categories.

```r
fit <- glm(in_midwest~cat_idx+b4_1990+val, data = df, family = "binomial")
summary(fit)
```

```
##
## Call:
## glm(formula = in_midwest ~ cat_idx + b4_1990 + val, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8149  -0.5483  -0.5123  -0.4657   2.1503
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.221103   0.056840 -39.076  < 2e-16 ***
## cat_idx2     0.733724   0.085900   8.542  < 2e-16 ***
## cat_idx3     0.354567   0.067488   5.254 1.49e-07 ***
## b4_1990TRUE  0.215836   0.075365   2.864  0.00419 **
## val          0.006792   0.002464   2.757  0.00583 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8338.6  on 10535  degrees of freedom
## Residual deviance: 8245.0  on 10531  degrees of freedom
## AIC: 8255
##
## Number of Fisher Scoring iterations: 4
```

Figure 1: logistic regression model summary on housing data

2 High Dimensions: Where Problems Arise

With the advent of new technology, it is becoming easier to collect and store data with thousands or hundreds of thousands of variables, sometimes relatively dwarfing the number of observations. One can observe this change when looking at RNA-seq data[5], MRI imaging data[6], or even manufacturing data [7]. Let us first take a look at what happens to the unbiasedness of coefficients. Here we can take a look at some data generated from covariates following a normal distribution. True coefficients for the logistic model will be randomly drawn integers $\in [0, 100]$. It is important that the ratio p/n be kept constant as dimensions

are scaled. Here, a ratio of 1/5 is chosen so as not to be too extreme. One of the main results from Sur et. al.[8] was estimation of a "signal strength" parameter γ set such that

$$\gamma^2 := \mathbf{Var}(X_i'\beta) = 5. \tag{5}$$

It is extremely important for the log-odds ratio (i.e. $(X_i'\beta)$) does not increase with either n or p. If the log-odds does increase with dimension, the probabilities output by the sigmoid function $f(X') = \dfrac{e^{(X_i'\beta)}}{1 + e^{(X_i'\beta)}}$ will be close to 0 or 1, a trivial result. In R, this could take the form of a lovely warning: "glm.fit: fitted probabilities numerically 0 or 1 occurred".

2.1 Biased Coefficient Estimates

Let's take a look at figure 2.1. For these data, $p = 40$ and $n = 200$ and the aforementioned scaling is implemented. The scattered data points represent MLE estimates for $\hat{\beta}$ and the horizontal lines represent the actual coefficients β. It becomes clear upon initial glance that even under these mild dimensions the MLE estimates for $\hat{\beta}$ are biased, contrary to what classical statistics would dictate. The effect sizes are overestimated when $\beta = 10$ and underestimated when $\beta = -10$. Now to analyze figure 2.1 below. Now the dimensions have been increase to $p = 1000$ and $n = 5000$. Again we can see the bias in overestimation of the effect magnitudes. This implies that if $f(X') \geq .5$ then the probability output by f will overestimate the probability and if $f(X') < .5$, then f will produce a probability lower than the actual. Obviously this is a huge problem for the predictive power of a logistic regression model.

2.2 Standard Errors

From classical theory [9], we can calculate the standard error of our estimates $\hat{beta}$ by first calculating the Fisher Information Matrix

$$I_Y(\beta) = -\mathbf{E}_\beta[\nabla^2 l(\beta)] \tag{6}$$

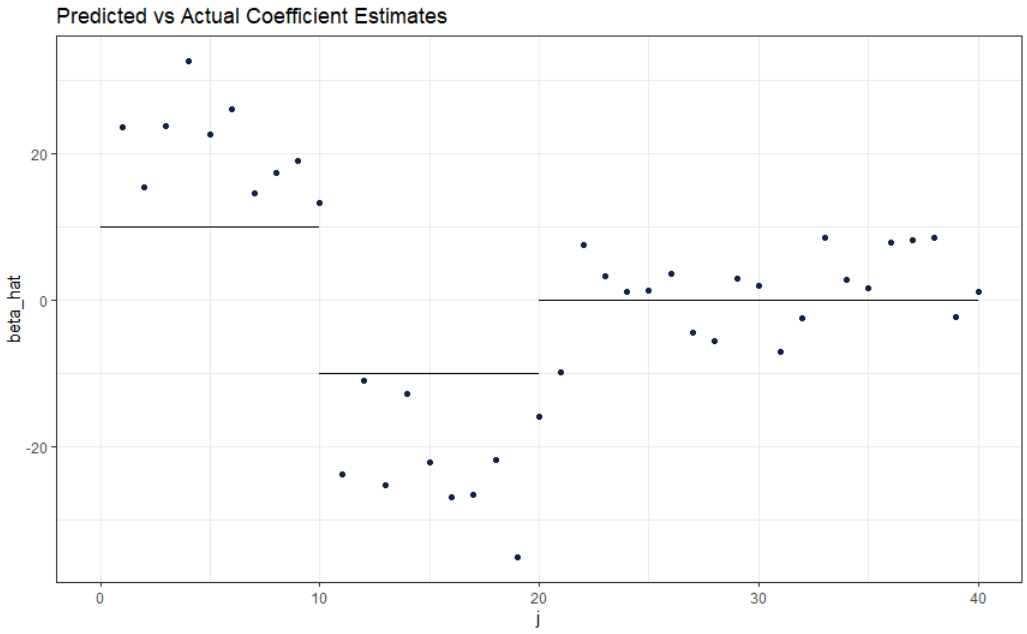


Figure 2: logistic regression coefficient estimate bias p,n = 40,200

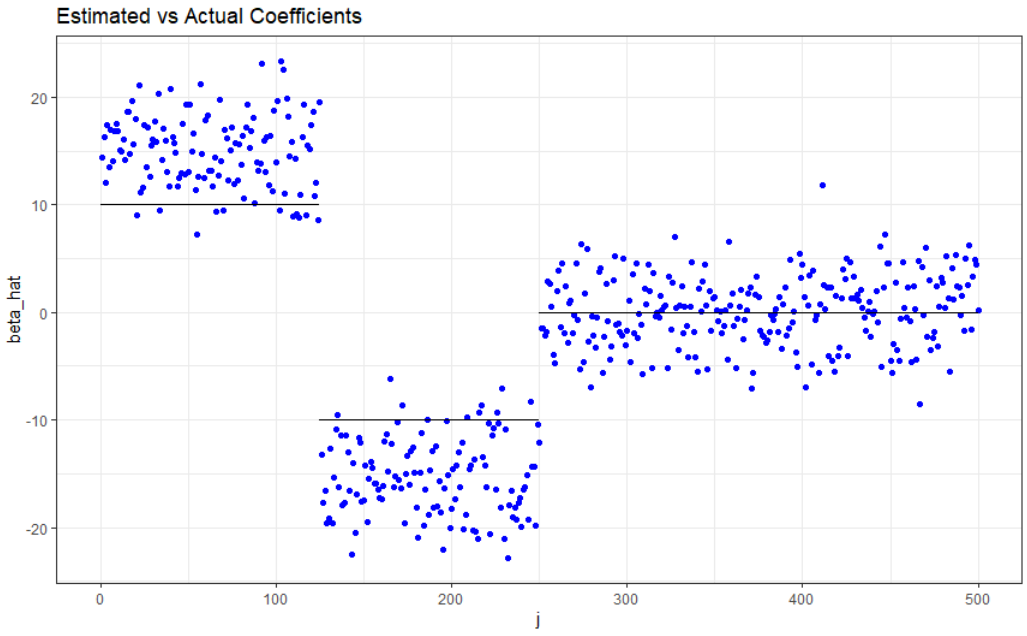


Figure 3: logistic regression coefficient estimate bias p,n = 500,2500

and to get the standard error for $\hat{\beta}_j$

$$\hat{se}_j = \sqrt{I_Y(\beta)^{-}1_{jj}} \tag{7}$$

To save myself the tedious calculation, I borrowed the classical standard error estimate for when half the β values are randomly drawn from $\sim \mathbf{N}(7,1)$[8]. Classical theory suggest that the standard error estimate should be about 2.66. Figure 2.2 shows a relative frequency histogram with a mean far beyond the classical 2.66 (indicated by the vertical red bar).

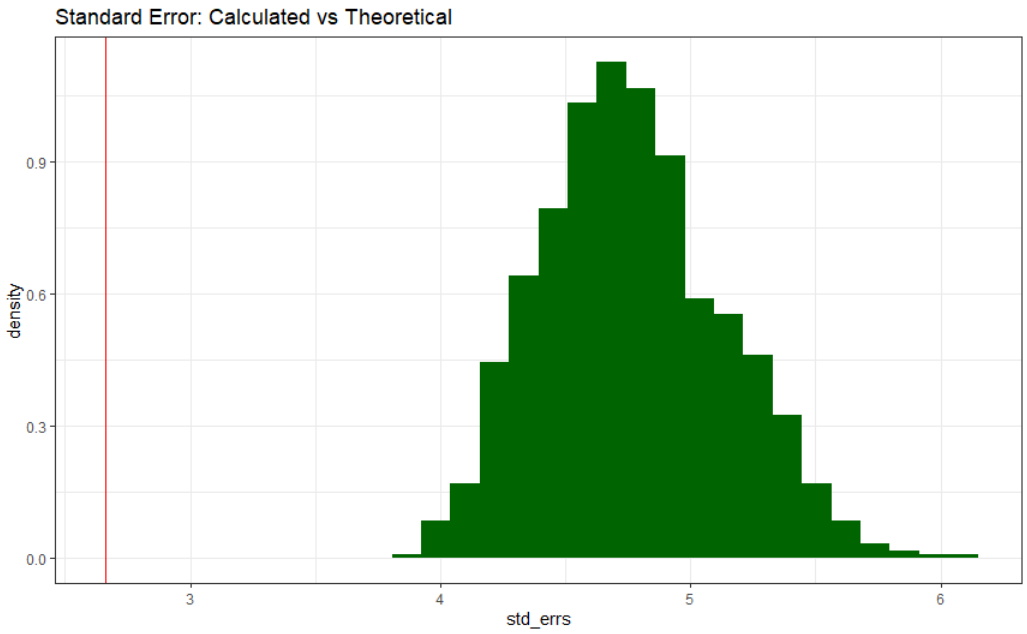


Figure 4: logistic regression coefficient estimate bias p,n = 500,2500

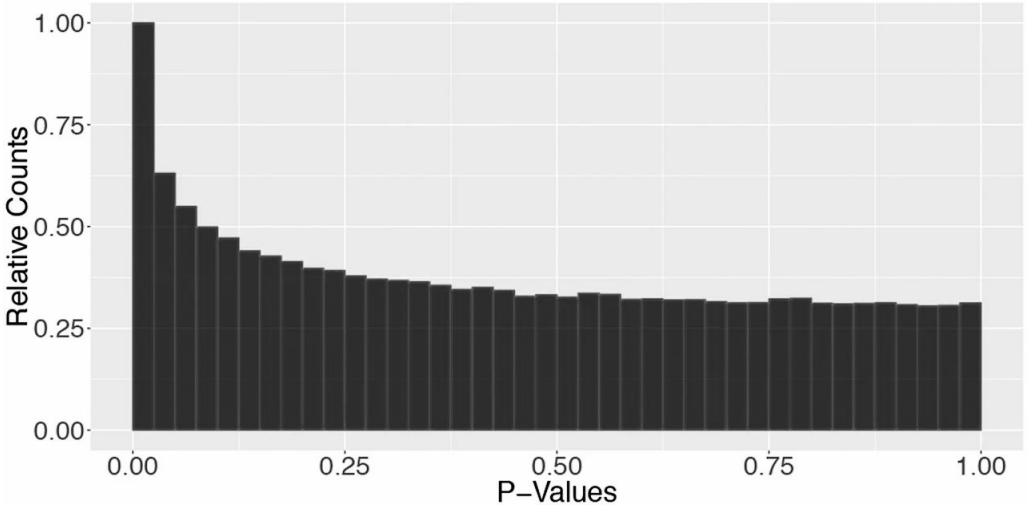
Standard errors that are off by this wide of a margin would most definitely lead to misleading p-values and therefore a potentially disastrous analysis.

2.3 Likelihood Ratio Test

Wilks' Theorem states that as $n \to \infty$, the test statistic $-2ln(\Lambda)$, where $\Lambda = \frac{likelihoodofH_0}{likelihoodofH_A}$, asymptotically approaches $\sim \chi^2_1$ a chi-square distribution with 1 degree of freedom. Alternatively, dropping k predictors from the model would approach $\sim \chi^2_k$. Earlier work from

[10] has already shown that as p and n grow larger, where $p/n \to \kappa \in (1, 1/2)$, 2Λ converges (in distribution) to $\sim \alpha(\kappa)\chi^2_k$, where $\alpha(\kappa)$ is a scaling factor > 1 when $\kappa > 0$.

Work from Sur et. al.[8] shows a distribution of p-values under moderate dimensionality (n=4000, p=8000). These p-values are for testing a $\hat{\beta}$ under a non-global null hypothesis (i.e. response does depend on predictors) based on the chi square distribution from Wilks' Theorem. Unsurprisingly, the distribution is not $\tilde{U}niform$, as it should be. The high counts of p-values near 0 are obviously concerning and will surely result in inaccurate assessments about the predictor-response relationship. The high counts of p-values near 0 are obviously



[8]

concerning and will surely result in inaccurate assessments about the predictor-response relationship.

3 Conclusion

The classical theorems relied upon by statistical programming languages such as R, SAS, etc. can break down when the number of observations increases with the number of variables. The implications of these findings are undoubtedly significant. Recall that many high dimensional datasets come from medical technology, where incorrect inference cause be disastrous for human life. To prevent erroneous inferences, one should be on the lookout

to implement proper parameter scaling, and statistical testing where applicable, especially as the shape of datasets stray from traditional dimensions, where p is insignificantly small relative to n.

References

[1] G. James, D. Witten, T. J. Hastie, and R. J. Tibshirani, *An introduction to statistical learning*. 2 ed.

[2] D. Schreiber-Gregory and K. Bader, "Logistic and linear regression assumptions: Violation recognition and control," 01 2018.

[3] T. Hastie, J. Friedman, and R. Tisbshirani, *The Elements of statistical learning*. Springer, 2017.

[4] S. E. Fienberg and A. Rinaldo, "Maximum likelihood estimation in log-linear models," *The Annals of Statistics*, vol. 40, no. 2, 2012.

[5] A. Lachmann, D. Torre, A. B. Keenan, K. M. Jagodnik, H. J. Lee, L. Wang, M. C. Silverstein, and A. Ma'ayan, "Massive mining of publicly available rna-seq data from human and mouse," Apr 2018.

[6] 2022.

[7] 2022.

[8] P. Sur and E. J. Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," *Proceedings of the National Academy of Sciences*, vol. 116, no. 29, p. 14516–14525, 2019.

[9] "Lecture 26 — logistic regression."

[10] P. Sur, Y. Chen, and E. J. Candès, "The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square," 2017.