# Graphical Diagnostics for Logistic Regression

Hakbae Lee*

**Abstract**

In this paper we discuss graphical and diagnostic methods for logistic regression, in which the response is the number of successes in a fixed number of trials.

Key Words: Logistic regression, Central subspaces, Regression graphics, Sliced average variance estimation.

## 1 Introduction

*Dimension-reduction without loss of information* is a fundamental idea in statistics. Cook and Lee(1999) discussed dimension reduction in binary response regression. In this paper we use a graphical paradigm: Sliced Average Variance Estimation(SAVE)(Cook and Weisberg, 1991).

## 2 The Central Subspace

Let $\mathbf{B}$ denote a fixed $p \times q$, $q \leq p$, matrix so that

$$y \perp\!\!\!\perp \mathbf{x} | \mathbf{B}^T\mathbf{x}. \tag{1}$$

This statement is equivalent to saying that the distribution of $y|\mathbf{x}$ is the same as that of $y|\mathbf{B}^T\mathbf{x}$ for all values of $\mathbf{x}$ in its marginal sample space. It implies that the $p \times 1$ predictor vector $\mathbf{x}$ can be replaced by the $q \times 1$ predictor vector $\mathbf{B}^T\mathbf{x}$ without loss of regression information, and thus represents a potentially useful reduction in the dimension of the predictor vector. If (1) holds then it also holds when $\mathbf{B}$ is replaced with any matrix whose columns form a basis for $\mathcal{S}(\mathbf{B})$. Thus, (1) is appropriately viewed as a statement about $\mathcal{S}(\mathbf{B})$, which is called a *dimension-reduction subspace for the regression of* $y$ *on* $\mathbf{x}$ (Li 1991, Cook 1994a).

Let $\mathcal{S}_{y|\mathbf{x}}$ denote the intersection of all dimension-reduction subspaces. In this article, $\mathcal{S}_{y|\mathbf{x}}$ is assumed to be a dimension-reduction subspace and, following Cook (1994b, 1996, 1998a,b), is called the *central subspace.*

Binary responses cause no conceptual complications for the central subspace, but construction and interpretation of summary plots in practice must recognize the nature of the response. Here we rely on binary response plots as developed by Cook (1996). For example,

*Assistant professor, Department of Applied Statistics, Yonsei University.

if it was inferred that $\dim(\mathcal{S}_{y|\mathbf{x}}) = 3$ then the summary plot would be a three-dimensional binary response plot with the coordinates of $\hat{\eta}^T x$ assigned to the axes of the plot, and the points colored to indicate the states of $y$.

Let $\mathbf{\Sigma_x} = \text{Var}(\mathbf{x})$, which is assumed to be non-singular. Without loss of generality, discussion in the rest of this article will mostly be in terms of the standardized predictor

$$\mathbf{z} = \mathbf{\Sigma}_{\mathbf{x}}^{-1/2}(\mathbf{x} - \text{E}(\mathbf{x})).$$

The corresponding sample version $\hat{\mathbf{z}}$ is obtained by replacing $\mathbf{\Sigma_x}$ and $\text{E}(\mathbf{x})$ with their usual moment estimates, $\hat{\mathbf{\Sigma}}_{\mathbf{x}}$ and $\bar{\mathbf{x}}$. The columns of the matrix $\gamma = \mathbf{\Sigma}_{\mathbf{x}}^{1/2}\eta$ form a basis for $\mathcal{S}_{y|\mathbf{z}}$, the central subspace for the regression of $y$ on $\mathbf{z}$. Thus, there is no loss of generality when working on the $\mathbf{z}$-scale because any basis $\gamma$ for $\mathcal{S}_{y|\mathbf{z}}$ can be back-transformed to a basis $\eta$ for $\mathcal{S}_{y|\mathbf{x}}$.

# 3 SAVE and Logistic Regression

For notational convenience let $\boldsymbol{\mu}_j = \text{E}(\mathbf{z}|y = j)$, $\mathbf{\Sigma}_j = \text{Var}(\mathbf{z}|y = j)$, $j = 0, 1$, and $f = \Pr(y = 1)$. We assume $0 < f < 1$. Finally, let $\boldsymbol{\nu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ and $\mathbf{\Delta} = \mathbf{\Sigma}_1 - \mathbf{\Sigma}_0$.

**Lemma 1**

$$\mathcal{S}_{SAVE} = \mathcal{S}(\mathbf{\Delta}, \boldsymbol{\nu}) \tag{2}$$

This lemma establishes two useful properties of SAVE. First, like the other procedures, it gains information from $(\mathbf{\Delta}, \boldsymbol{\nu})$, allowing use of the equivalent kernel matrix $\mathbf{M} = (\mathbf{\Delta}, \boldsymbol{\nu})$. Second, it is the most comprehensive procedure without requiring the linearity or constant covariance conditions. Those conditions are needed to connect the various method-specific subspaces to the central subspace, but are not needed for Lemma 1.

Since the response is binary the distribution of $y|\mathbf{z}$ can be characterized by the conditional probability of "success", $\Pr(y = 1|\mathbf{z})$. Assuming that $\mathbf{z}|(y = j)$ has a density $g_j$,

$$\log \frac{\Pr(y = 1|\mathbf{z})}{\Pr(y = 0|\mathbf{z})} = \log \frac{g_1(\mathbf{z})}{g_0(\mathbf{z})} + \log \frac{f}{1 - f}.$$

This means that $\Pr(y = 1|\mathbf{z})$ can be expressed via its logit in terms of the log density ratio. Now assuming that $\mathbf{z}|(y = j)$ is normally distributed with mean $\boldsymbol{\mu}_j$ and variance $\mathbf{\Sigma}_j$, $j = 0, 1$, it is known that

$$2 \log \frac{g_1(\mathbf{z})}{g_0(\mathbf{z})} = C + \mathbf{z}^T(\mathbf{\Sigma}_0^{-1} - \mathbf{\Sigma}_1^{-1})\mathbf{z} + 2\mathbf{z}^T(\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0) \tag{3}$$

where $C$ is a constant not depending on $\mathbf{z}$. It follows immediately from this characterizing expression that

$$\mathcal{S}_{y|\mathbf{z}} = \mathcal{S}(\mathbf{\Sigma}_0^{-1} - \mathbf{\Sigma}_1^{-1}, \mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0). \tag{4}$$

**Lemma 2** *Assume that $\mathbf{z}|y$ follows a non-singular normal distribution. Then $\mathcal{S}_{y|\mathbf{z}} = \mathcal{S}_{SAVE}$.*

If the linearity and constant covariance conditions hold, but $\mathbf{z}|y$ is not normally distributed, we will still have $\mathcal{S}_{SAVE} \subset \mathcal{S}_{y|\mathbf{x}}$. However, there is no guarantee of equality because moments higher than the second may be involved. With just the linearity condition there is the added complication that $\mathcal{S}_{SAVE}$ may "overestimate" $\mathcal{S}_{y|\mathbf{x}}$ because of the presence of extraneous directions.

# 4 Example: Diabetes Data

For this first example we consider a data set on 724 patients with complete records from the National Institute of Diabetes and Digestive and Kidney Disease. Smith, Everhart, Dickson, Knowler and Johannes (1988) use this data set to forecast the onset of diabetes mellitus.

The binary response variable $y$ equals 1 if a patient tested positive for diabetes and equals 0 otherwise. An examination of the data on the 6 predictors indicated that power transformations might be used to achieve approximately joint normality. Based on the standardized predictors $z_i$, the two SAVE predictors resulting from this analysis are

$$\begin{aligned} \text{SAVE}_1 &= -0.162\, z_1 + 0.621\, z_2 + 0.194\, z_3 - 0.431\, z_4 + 0.193\, z_5 + 0.572\, z_6 \\ \text{SAVE}_2 &= 0.099\, z_1 - 0.070\, z_2 + 0.181\, z_3 + 0.530\, z_4 + 0.816\, z_5 - 0.079\, z_6. \end{aligned}$$

The span of the two vectors of coefficients in these predictors is the estimate of $\mathcal{S}_{SAVE}$ which, because of the approximate normality of the transformed predictors, we expect is the same as $\mathcal{S}_{y|\mathbf{z}}$.

Shown in Figure 1 is the 2D binary response plot (Cook 1996) based on $\text{SAVE}_1$ and $\text{SAVE}_2$. This plot shows clear separation between the states of $y$, and could be used to guide the remaining analysis. Several options are available, depending on the precise goals of the study. For example, we could fit a logistic model in the predictors $\text{SAVE}_1$ and $\text{SAVE}_2$. Using results by Kay and Little (1987), see from Figure 1 that a logistic model for the regression of $y$ on $\text{SAVE}_1$ and $\text{SAVE}_2$ will likely need a linear term in $\text{SAVE}_1$ and quadratic terms in $\text{SAVE}_1$ and $\text{SAVE}_2$. The linear term in $\text{SAVE}_1$ is needed because the two point clouds have different locations along the $\text{SAVE}_1$ axis. Quadratic terms in $\text{SAVE}_1$ and $\text{SAVE}_2$ would be needed because Figure 2 indicates that $\text{Var}(\text{SAVE}_j|y=0) \neq \text{Var}(\text{SAVE}_j|y=1)$ for $j = 0, 1$. The different variances for $\text{SAVE}_2$ are a little difficult to see in the plot, but are quite apparent when comparing marginal kernel density estimates (Figure 2). Figure 3 shows a plot of chi-residuals versus $-1.03 + 1.89\text{SAVE}_1 - 0.29\text{SAVE}_1^2 + 0.13\text{SAVE}_2^2$. The lowess smooth on Figure 3 is nearly constant, suggesting no evidence against the fitted mean function.

Summary plots such as those in Figure 1 is often useful in the development of first models for the regression. Let $g_j(\mathbf{x})$ denote the conditional density of $\mathbf{x}|(y=j)$, $j = 0, 1$, and assume that $g_0(\mathbf{x})$ is multivariate normal with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}$. Assume further that $g_1(\mathbf{x})$ is a mixture of normal densities,

$$g_1(\mathbf{x}) = \alpha g_{11}(\mathbf{x}) + (1-\alpha) g_{12}(\mathbf{x})$$

where $g_{1k}$ is the multivariate normal density with mean $\boldsymbol{\mu}_{1k}$ and covariance matrix $\boldsymbol{\Sigma}$, $k = 1, 2$. After a little algebra, the regression odds ratio can be expressed as

$$\frac{\Pr(y=0|\mathbf{x})}{\Pr(y=1|\mathbf{x})} = \frac{\exp\{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_{11})^T \boldsymbol{\Sigma}^{-1}\mathbf{x}\}}{\omega_0 + \omega_1 \exp\{(\boldsymbol{\mu}_{12} - \boldsymbol{\mu}_{11})^T \boldsymbol{\Sigma}^{-1}\mathbf{x}\}}$$

where $\omega_0$ and $\omega_1$ are unknown constants not depending on $\mathbf{x}$. It follows that $\dim(\mathcal{S}_{y|\mathbf{x}}) = 2$ with

$$\mathcal{S}_{y|\mathbf{x}} = \boldsymbol{\Sigma}^{-1}\mathcal{S}\{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_{11}), (\boldsymbol{\mu}_{12} - \boldsymbol{\mu}_{11})\}$$

The two vectors defining this subspace can be estimated directly from the subsets of the data corresponding the sub-populations.

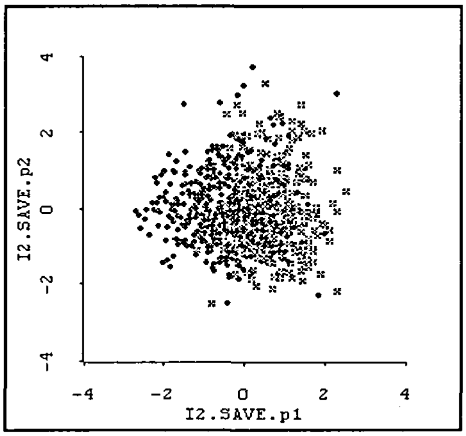


Figure 1. Two_dimensional Binary Response Plot of Direction 1 from SAVE versus Direction 2 from SAVE. y=0 is marked with an open circle; y=1 with a star

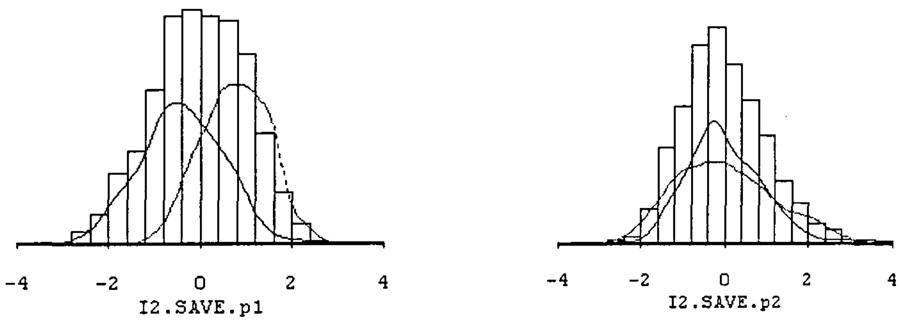


Figure 2. Histogram for two SAVE variables, with separate density estimates for the two values of y.

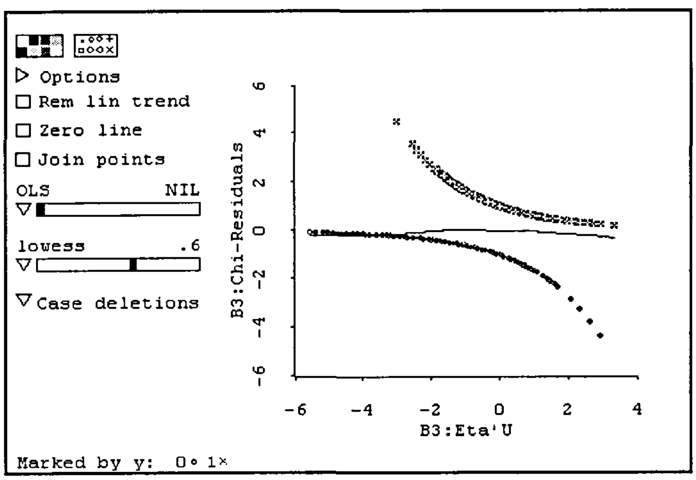


Figure 3. Chi-square residuals versus fitted value using SAVE variables. A lowess smooth is shown on the plot.

Finally, we can re-express the odds ratio as

$$\log \frac{\Pr(y=0|\mathbf{x})}{\Pr(y=1|\mathbf{x})} = \eta_{01} + \eta_1^T \mathbf{x} - \log(1 + \exp(\eta_{02} + \eta_2^T \mathbf{x}))$$

which provides a first (nonlinear) logistic model for the regression. It seems unlikely that we would have arrived at a model of this form without the guidance available from the summary plot.

# 5 Discussion

Approaching a regression through its central subspace is intended to allow construction of a low dimensional summary plot that contains or is inferred to contain all of the regression information available from the sample. Since a parametric model is not required, such plots can be particularly useful at the beginning of an analysis.

# References


Cook, R. D. (1994a). On the interpretation of regression plots. *Journal of the American Statistical Association* **89**, 177–189.

Cook, R. D. (1994b). Using dimension-reduction subspaces to identify important inputs in models of physical systems. *Proceedings of the Section on Physical and Engineering Sciences*, Alexandria, VA: American Statistical Association, pp. 18–25.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association* **91**, 983–992.

Cook, R. D. (1998a). Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association* **93**, 84–100.

Cook, R. D. (1998b). *Regression Graphics, Ideas for studying regressions through graphics.* New York: Wiley.

Cook, R. D. and Lee, H. (1999). Dimension Reduction in Binary Response Regression. *Journal of the American Statistical Association* **94**, 1187–1200.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression" by K. C. Li. *Journal of the American Statistical Association* **86** 328–332.

Kay, R. and Little, S. (1987). Transforming the explanatory variables in the logistic regression model for binary data. *Biometrika* **74** 495–501.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**, 316–342.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C. and Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Symposium on Computer Applications and Medical Care*, 261-265. IEEE Computer Society Press.