Regression Diagnostics & Influential Observations in Linear Models

Won I. Lee

1 Introduction

An oft-overlooked but important aspect of regression analysis is the detection of outliers and influential observations in the data. Idiosyncrasies in the data can lead to distorted or misleading conclusions when analyzing only overall summary statistics, such as $R^2, \hat{\beta}$. In order to combat this issue, a number of diagnostic tools and methods have been developed for both linear and generalized linear models. This paper outlines the development of these diagnostics and detection methods and their foundations in the theory of linear models.

2 Cook's Distance for Linear Models

2.1 Matrix Inversion Lemma

We first take a brief technical aside to introduce a widely-used lemma and technique for simplifying formulas after the deletion of an observation, adopting the approach of Hager (1989). The *matrix inversion lemma*, also known as the Sherman-Morrison-Woodbury formula, is a method for computing the inverse of a modified matrix based on the inverses of the original matrix and the modification matrices. In other words, given a matrix $\mathbf{A}$, we would like to compute the inverse $\mathbf{B}^{-1}$, where $\mathbf{B}$ is a modified version of $\mathbf{A}$. For example, the deletion of an observation leads to a modified matrix that has one row removed, and other cases may arise in the solution of systems of linear equations for which the coefficient matrix $\mathbf{B}$ is a perturbation of a more convenient matrix $\mathbf{A}$.

The primary result is that when $\mathbf{B} = \mathbf{A} - \mathbf{U}\mathbf{V}$ for some matrices $\mathbf{U}, \mathbf{V}$, and both $\mathbf{A}$ and $\mathbf{A} - \mathbf{U}\mathbf{V}$ are invertible, then we have:

$$\mathbf{B}^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{U}(\mathbf{I} - \mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{A}^{-1}$$

In particular, note that the computation involves only inverting the original matrix $\mathbf{A}$ and products of $\mathbf{A}, \mathbf{U}, \mathbf{V}$. Moreover, in deletion diagnostics, we only consider cases in which $\mathbf{U}$ is a column vector, denoted $\mathbf{u}$, and $\mathbf{V}$ is a row vector, denoted $\mathbf{v}$. This yields a much simplified version of the formula:

$$\mathbf{B}^{-1} = \mathbf{A}^{-1} + \alpha\mathbf{A}^{-1}\mathbf{u}\mathbf{v}\mathbf{A}^{-1}$$

in which $\alpha = 1/(1 - \mathbf{v}\mathbf{A}^{-1}\mathbf{u})$. While Hager (1989) cites the utility of this formula in its ability to update the inverse matrix when a new observation enters the model matrix, it has found much applicability in simplifying formulas for the method of deletion diagnostics, as outlined below.

2.2 Motivation for Cook's Distance

We follow the standard convention for linear models, where the response variable $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ and $\mathbf{e}$ denotes the residual, with $E(\mathbf{e}) = 0$ and $\mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{I}$. We have $\mathbf{y}$ as an $n \times 1$ vector, $\mathbf{X}$ as an $n \times p$ full-rank model matrix, and β as a $p \times 1$ vector of parameters. Correspondingly, the least squares estimates are $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, while the hat matrix is $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

While not necessary for the development of the metric, we make the normal assumption on $\mathbf{y}$ in order to motivate the formula for Cook's distance. Thus, assume that $\mathbf{y} \sim N(\mu, \sigma^2\mathbf{I})$; then, we have:

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

since $\hat{\beta}$ is a linear function of $\mathbf{y}$. Consequently, we can form a quadratic form using the inverse of the variance matrix to yield a chi-squared distribution:

$$\frac{(\beta - \hat{\beta})^T\mathbf{X}^T\mathbf{X}(\beta - \hat{\beta})}{\sigma^2} \sim \chi^2_p$$

Similarly, we have $\frac{1}{\sigma^2}(y-\hat{\mu})^T(y-\hat{\mu}) = \frac{n-p}{\sigma^2}s^2 \sim \chi^2_{n-p}$. Thus, forming the ratio of two independent chi-squared distributions (due to the orthogonality of the model and error spaces) over their degrees of freedom:

$$\frac{(\beta - \hat{\beta})^T\mathbf{X}^T\mathbf{X}(\beta - \hat{\beta})}{ps^2} \sim F_{p,n-p}$$

which measures the number of residuals that a given $\mu = \mathbf{X}\beta$ is from the estimate $\hat{\mu} = \mathbf{X}\hat{\beta}$. Thus, it provides a measure of distance of a modified estimate from the original when the analysis is altered in some fashion, as in the case when an observation is deleted.

This idea suggests a measure for the degree of influence of an observation; letting $\hat{\beta}_{(-i)}$ denote the least squares estimate of β with the i^{th} observation deleted, we define *Cook's distance*:

$$D_i \equiv \frac{(\hat{\beta}_{(-i)} - \hat{\beta})^T\mathbf{X}^T\mathbf{X}(\hat{\beta}_{(-i)} - \hat{\beta})}{ps^2}$$

Thus, we see that based on the F-statistic for the deviation of the estimated $\hat{\mu}$ from the true value μ under the normal assumption, Cook's distance is a natural measure of the distance between the estimated $\hat{\mu}_{(-i)}$ with the i^{th} observation deleted and $\hat{\mu}$ with the full data. Moreover, aside from the scale factor ps^2, D_i is simply the Euclidean distance that the fitted values $\hat{\mu}$ "move" when the i^{th} observation is removed.

2.3 Simplification via Leverage Values

As Cook (1977) notes, the utility of the above measure is diminished by the fact that evaluating the diagnostic for all n observations involves computing $n+1$ regressions for $\hat{\beta}, \hat{\beta}_{(-1)}, \ldots, \hat{\beta}_{(-n)}$, which can incur significant computational costs and time. Fortunately, by application of the matrix inversion lemma noted above, we can provide a substantial simplification of Cook's distance into a more computationally tractable as well as intuitive form.

To do so, following the proof of the main result in Beckman and Trussell (1974), we first note that:

$$\mathbf{X}^T\mathbf{X} = \mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)} + \mathbf{x}_i\mathbf{x}_i^T$$

since we have $[\mathbf{X}^T\mathbf{X}]_{jk} = ([\mathbf{X}]_j)^T[\mathbf{X}]_k = ([\mathbf{X}_{(-i)}]_j)^T[\mathbf{X}_{(-i)}]_k + [\mathbf{x}\mathbf{x}^T]_{jk}$, where $[\mathbf{A}]_{jk}$ denotes the jk element of the matrix $\mathbf{A}$ and $[\mathbf{A}]_j$ denotes the j^{th} column of the matrix $\mathbf{A}$. Moreover, $\mathbf{X}^T = [\mathbf{X}_{(-i)}^T, \mathbf{x}_i]$, where for simplicity of exposition we assume that i is the last observation (but can be placed among any of the rows). Thus, using the formula given by the matrix inversion lemma for the case of modification by a single row and column vector:

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \\
&= [\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)} + \mathbf{x}_i\mathbf{x}_i^T]^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left[(\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1} - \frac{1}{1+c}(\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1}\right]\mathbf{X}^T\mathbf{y} \\
&= \left[\mathbf{I} - \frac{1}{1+c}(\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1}\mathbf{x}_i\mathbf{x}_i^T\right]\hat{\beta}_{(-i)} + \frac{1}{1+c}(\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1}\mathbf{x}_i y_i
\end{aligned}$$

where $c = \mathbf{x}_i^T(\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1}\mathbf{x}_i$ is a scalar. Thus, multiplying by $\mathbf{x}_i^T$ and some simplification yields:

$$\mathbf{x}_i^T\hat{\beta} = \frac{1}{1+c}\mathbf{x}_i^T\hat{\beta}_{(-i)} + \frac{c}{1+c}y_i$$

which, after rearrangement, can be written as:

$$y_i - \mathbf{x}_i^T\hat{\beta} = \frac{1}{1+c}\left(y_i - \mathbf{x}_i^T\hat{\beta}_{(-i)}\right)$$

Finally, rearranging terms and using a generalized left inverse yields the desired difference in estimates term:

$$\hat{\beta} - \hat{\beta}_{(-i)} = (\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1}\mathbf{x}_i(y_i - \mathbf{x}_i^T\hat{\beta})$$

The matrix inversion lemma again demonstrates its utility in this case, as we can now compute the inverse $(\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1}$ from the inverse of the original model matrix as follows:

$$\begin{aligned}
(\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1} &= [\mathbf{X}^T\mathbf{X} - \mathbf{x}_i\mathbf{x}_i^T]^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1} + \frac{1}{1-h_{ii}}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}$$

where $h_{ii} = [\mathbf{H}]_{ii}$ is the i^{th} diagonal term in the hat matrix $\mathbf{H}$, or the *leverage* of the i^{th} observation.

Multiplying on the right by $\mathbf{x}_i$ yields:

$$(\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1}\mathbf{x}_i = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}}$$

Using this expression in our difference in estimates term results in:

$$\hat{\beta} - \hat{\beta}_{(-i)} = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}}(y_i - \mathbf{x}_i^T\hat{\beta})$$

Thus, we can now substitute this expression in the definition of Cook's distance given above, yielding:

$$D_i = \left(\frac{y_i - \mathbf{x}_i^T\hat{\beta}}{s\sqrt{1-h_{ii}}}\right)^2 \frac{h_{ii}}{p(1-h_{ii})}$$

But noting that the squared term is precisely the i^{th} *standardized residual* (or sometimes referred to as the studentized residual), we thus have:

$$\boxed{D_i = r_i^2 \frac{h_{ii}}{p(1-h_{ii})}}$$

This version of Cook's distance, in addition to being more computationally tractable, captures the intuition of the magnitude of values that can result. Prior to Cook's approach, the standard recommendation was to use a combination of residual plots (of the e_i), standardized residual values (r_i), and leverages (h_{ii}) to determine whether an observation was influential for the fit and potentially distorting the estimate. Cook's distance evidently combines these separate analyses into a single metric, which can be compared uniformly across all observations as a scalar.

2.4 Extensions of Cook's Distance

When conducting tests such as the general linear hypothesis, it is often the case that the quantity of interest is not β itself, but rather some linear combination(s) of the β. If we are interested in q linear combinations of β, for example, we can denote the quantities of interest as $\psi = \mathbf{\Lambda}\beta$ where $\mathbf{\Lambda}$ is a $q \times p$ rank q matrix.

We can arrive at a generalization of Cook's distance for the given matrix $\mathbf{\Lambda}$ by following the same approach as in the motivation of the original formula. Under the normal assumption, we have $\psi \sim N(\mathbf{\Lambda}\hat{\beta}, \sigma^2\mathbf{\Lambda}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda}^T)$. Thus, the quadratic form

$$\frac{1}{\sigma^2}(\psi - \hat{\psi})^T[\mathbf{\Lambda}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda}]^{-1}(\psi - \hat{\psi}) \sim \chi_q^2$$

since $\mathbf{\Lambda}$ is of rank q; meanwhile, $\frac{n-p}{\sigma^2}s^2 \sim \chi_{n-p}^2$ as before. Thus, forming the ratio over the respective degrees of freedom:

$$\frac{(\psi - \hat{\psi})^T[\mathbf{\Lambda}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda}]^{-1}(\psi - \hat{\psi})}{qs^2} \sim F_{q,n-p}$$

We therefore define the *generalized Cook's distance* to be that ratio with $\hat{\psi}_{(-i)} = \mathbf{\Lambda}\hat{\beta}_{(-i)}$ denoting the estimate of the desired quantity under deletion of observation i:

$$D_i(\mathbf{\Lambda}) \equiv \frac{(\hat{\psi}_{(-i)} - \hat{\psi})^T[\mathbf{\Lambda}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda}]^{-1}(\hat{\psi}_{(-i)} - \hat{\psi})}{qs^2}$$

While the simplification for the general case is not particularly illuminating, it is useful to consider the special case in which $q = 1$; that is, when we are considering a single linear combination of β. This case is useful because it allows for the consideration of the effect of an observation on a single parameter β_i, as well as the impact on a *contrast* $\beta_i - \beta_j$. In this case, we have $\mathbf{\Lambda} = \lambda$ where λ is a row vector. In this case, $\mathbf{\Lambda}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda}^T = \lambda(\mathbf{X}^T\mathbf{X})^{-1}\lambda^T \in \mathbb{R}$ is a scalar.

Using the fact that:

$$\hat{\psi} - \hat{\psi}_{(-i)} = \lambda(\hat{\beta} - \hat{\beta}_{(-i)}) = \frac{y_i - \mathbf{x}_i^T\hat{\beta}}{1 - h_{ii}}\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$$

as derived in the simplification of the original Cook's distance formula, we have:

$$D_i(\lambda) = \frac{r_i^2}{1 - h_{ii}}\frac{\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\lambda^T\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{\lambda(\mathbf{X}^T\mathbf{X})^{-1}\lambda^T}$$

We now note that the correlation between $\mathbf{x}_i^T\hat{\beta}$ and $\lambda\hat{\beta}$ is given by the following:

$$\begin{aligned}
\rho(\mathbf{x}_i^T\hat{\beta}, \lambda\hat{\beta}) &= \frac{\text{cov}(\mathbf{x}_i^T\hat{\beta}, \lambda\hat{\beta})}{\sqrt{\text{var}(\mathbf{x}_i^T\hat{\beta})\text{var}(\lambda\hat{\beta})}} \\
&= \frac{\mathbf{x}_i^T[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]\lambda^T}{\sqrt{[\sigma^2\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i]\cdot[\sigma^2\lambda(\mathbf{X}^T\mathbf{X})^{-1}\lambda^T]}} \\
&= \frac{\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\lambda^T}{\sqrt{h_{ii}\lambda(\mathbf{X}^T\mathbf{X})^{-1}\lambda^T}}
\end{aligned}$$

Thus, substituting into the expression for $D_i(\lambda)$ and noting that $r_i^2\frac{h_{ii}}{1-h_{ii}} = pD_i$, we finally obtain:

$$\boxed{D_i(\lambda) = pD_i \cdot \rho(\mathbf{x}_i^T\hat{\beta}, \lambda\hat{\beta})^2}$$

3 Generalization: Empirical Influence Function

One may consider two potential approaches to further generalize the Cook's distance metric. The first is to consider alternative quadratic forms to define the distance measure, rather than the standard covariance matrix $\mathbf{X}^T\mathbf{X}$. While this matrix yielded an intuitive interpretation of Cook's distance as the Euclidean distance that the fitted values moved when the i^{th} observation was deleted, it may be the case that other choices for the matrix result in simpler or more readily comparable quadratic forms. The second approach is to consider deleting more than one observation at a time and formulating metrics for the deletion of *subsets* of observations.

Both of these potential generalizations can be formulated in terms of the *empirical influence function*, a special case of the influence function by Hampel (1974), defined for an arbitrary subset of the data A as:

$$IF_A = \hat{\beta}_A - \hat{\beta}$$

We can then define a generalized notion of a distance, or location/scale-invariant norm, for the subset A as:

$$D_A(\mathbf{M}, c) \equiv \frac{(IF_A)^T \mathbf{M} (IF_A)}{c}$$

for a given matrix $\mathbf{M}$ and scale factor c. These parameters, in a sense, may be chosen to reflect the quantities or changes of interest in the particular analysis. We note that the original Cook's distance is a special case of this formula with $\mathbf{M} = \mathbf{X}^T\mathbf{X}$ and $c = ps^2$. Moreover, when a single observation i is deleted, we denote the distance as $D_i(\mathbf{M}, c)$.

3.1 Generalized Single-Deletion Metrics

Cook and Weisberg (1980) discuss a number of potential alternatives for the choices of $\mathbf{M}$ and c, but note that all choices that yield location- and scale-invariant $D_i(\mathbf{M}, c)$ provide "approximately the same information." However, both the geometric interpretation of the measures as well as their reduced forms vary from the original metric.

The first alternatives are to consider $\mathbf{M} = \mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)}$ and $c = ps_{(-i)}^2$. In a sense, the latter scale factor corresponds to a 'Studentized' version of the metric, in the sense of Studentized residuals, in which the sample variance is estimated without the i^{th} observation. **Table 1** provides a number of possible alternative metrics and their reduced forms, based on the assumption of normality regarding the response variable. Given this assumption, we note that:

$$F_i \equiv r_i^2 \frac{n-p-1}{n-p-r_i^2} \sim F_{1,n-p-1}$$

assuming that the correct model is given by $\mathbf{y} = \mathbf{X}\beta + \epsilon$.

$\mathbf{M}$	c	Reduced Form
$\mathbf{X}^T\mathbf{X}$	$ps_{(-i)}^2$	$\frac{n-p}{p} F_i \frac{h_{ii}}{1-h_{ii}}$
$\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)}$	ps^2	$r_i^2 \frac{h_{ii}}{p}$
$\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)}$	$ps_{(-i)}^2$	$F_i \frac{h_{ii}}{p}$
$[\text{diag}(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}$	$ps_{(-i)}^2$	$\frac{n-p}{p} F_i \frac{\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}M(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}}$

Table 1: Alternative quadratic forms $D_i(\mathbf{M}, c)$ based on different choices for $\mathbf{M}$ and c, and their simplification to reduced form.

3.2 General Metrics for Subsets & Linear Combinations

While the single-deletion case is often the most important and sufficient, some data sets contains points that are *jointly influential*, but individually uninfluential. An example is given in **Figure 1**, which contains data regarding average attitudes towards inequality based on the country's the Gini coefficient.[1]

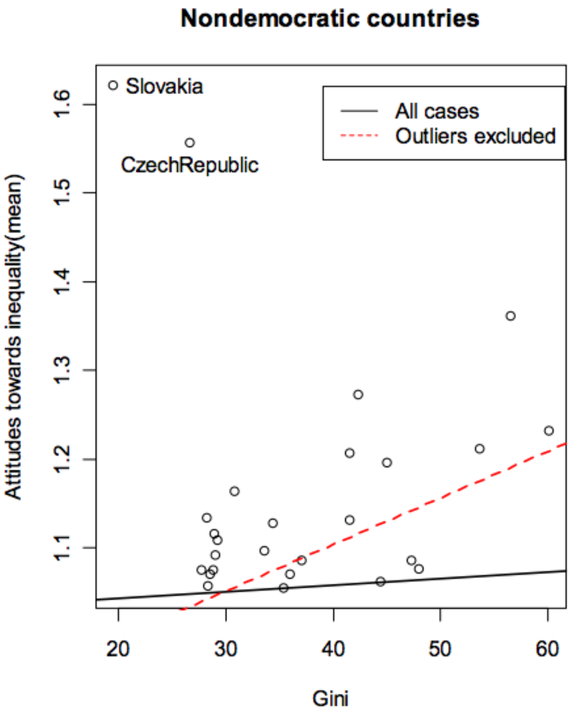

Figure 1: Data regarding attitudes towards inequality, given a country's Gini coefficient (Jacoby, 2005).

Evidently, removing either of the outliers individually (Slovakia or the Czech Republic) will only minimally impact the parameter estimates, and least squares will still yield a fitted line similar to the "all cases" line indicated in Figure 1. On the other hand, removing both Slovakia and the Czech Republic yields the "outliers excluded" fit indicated by the dotted red line, which appears to be a better fit to the majority of the data. Thus, it is at times important to be able to characterize distance between estimated parameter values when a subset of the observations are deleted.

To obtain such a metric, we return to the general formulation using the empirical distance function. We consider I to be the index set $\{i_1, \ldots, i_m\}$ indicating the cases that are deleted; then, $\hat{\beta}_{(-I)}$ represents, by analogy to the single-deletion case, the estimated parameter values without the observations index by I. Then, we define:

$$IF_I \equiv \hat{\beta}_{(-I)} - \hat{\beta}$$

[1] Figure from Jacoby, William G. "Lecture 11: Outliers and Influential Data" of *Regression III: Advanced Methods* (http://polisci.msu.edu/jacoby/icpsr/regress3/).

and the distance function:

$$D_I \equiv D_I(\mathbf{X}^T\mathbf{X}, ps^2) = \frac{(IF_I)^T(\mathbf{X}^T\mathbf{X})(IF_I)}{ps^2}$$

In order to derive simpler and more intuitive versions of the formula, we utilize a result by Bingham (1977), who showed that:

$$IF_I = -(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_I^T(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{e}_I$$

where each of the $\mathbf{X}_I, \mathbf{H}_I$ are matrices consisting only of the rows/columns corresponding to the indexed observations I to be deleted. Using this formula, we have:

$$\begin{aligned} D_I &= \frac{[\mathbf{e}_I^T(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{X}_I(\mathbf{X}^T\mathbf{X})^{-1}](\mathbf{X}^T\mathbf{X})[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_I^T(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{e}_I]}{ps^2} \\ &= \frac{\mathbf{e}_I^T(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{H}_I(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{e}_I}{ps^2} \end{aligned}$$

Since $\mathbf{H}_I$ is a symmetric $m \times m$ matrix, we can diagonalize the matrix with the spectral decomposition: that is, there exists an $m \times m$ diagonal matrix $\mathbf{\Lambda}_I = \text{diag}(\lambda_i)$ and orthogonal matrix $\mathbf{\Gamma}_I$ such that $\mathbf{H}_I = \mathbf{\Gamma}_I^T\mathbf{\Lambda}_I\mathbf{\Gamma}_I$, where λ_i are the eigenvalues of the matrix $\mathbf{H}_I$. If all eigenvalues $\lambda_i < 1$, then we have:

$$\begin{aligned} D_I &= \frac{\mathbf{e}_I^T(\mathbf{\Gamma}_I^T\mathbf{\Gamma}_I - \mathbf{\Gamma}_I^T\mathbf{\Lambda}_I\mathbf{\Gamma}_I)^{-1}(\mathbf{\Gamma}_I^T\mathbf{\Lambda}_I\mathbf{\Gamma}_I)(\mathbf{\Gamma}_I^T\mathbf{\Gamma}_I - \mathbf{\Gamma}_I^T\mathbf{\Lambda}_I\mathbf{\Gamma}_I)^{-1}\mathbf{e}_I}{ps^2} \\ &= \frac{(\mathbf{\Gamma}_I\mathbf{e}_I)^T(\mathbf{I} - \mathbf{\Lambda}_I)^{-1}\mathbf{\Lambda}_I(\mathbf{I} - \mathbf{\Lambda}_I)^{-1}(\mathbf{\Gamma}_I\mathbf{e}_I)}{ps^2} \\ &= \frac{\mathbf{g}^T(\mathbf{I} - \mathbf{\Lambda}_I)^{-1}\mathbf{\Lambda}_I(\mathbf{I} - \mathbf{\Lambda}_I)^{-1}\mathbf{g}}{ps^2} \\ &= \frac{\sum_{i=1}^m g_i^2 \frac{\lambda_i}{(1-\lambda_i)^2}}{ps^2} \end{aligned}$$

where $\mathbf{g} = \mathbf{\Gamma}_I\mathbf{e}_I = (g_1, \ldots, g_m)$. Then, each g_i is a linear combination of the elements of $\mathbf{e}_I$, or the residuals, with:

$$\begin{aligned} \text{var}(\mathbf{g}) &= \text{var}(\mathbf{\Gamma}_I\mathbf{e}_I) \\ &= \mathbf{\Gamma}_I^T[\sigma^2(\mathbf{I} - \mathbf{H})I)]\mathbf{\Gamma}_I \\ &= \sigma^2\mathbf{\Gamma}\mathbf{\Gamma}^T(\mathbf{I} - \mathbf{\Lambda}_I)\mathbf{\Gamma}\mathbf{\Gamma}^T \\ &= \sigma^2(\mathbf{I} - \mathbf{\Lambda}_I) \end{aligned}$$

Thus, the g_i are uncorrelated with $\text{var}(g_i) = \sigma^2(1 - \lambda_{ii})$. Moreover, we can now standardize the g_i and define:

$$\tilde{r}_i = \frac{g_i}{s\sqrt{1 - \lambda_i}}$$

which leads us to the simplified form of the distance metric for the deletion of a subset as:

$$\boxed{D_I = \sum_{i=1}^{m} \frac{\tilde{r}_i^2}{p} \frac{\lambda_i}{1-\lambda_i}}$$

In this form, the resemblance of this distance metric to the one derived for the single-deletion case is clear: we place r_i by $\tilde{r}_i$, and the leverage values h_{ii} by the eigenvalues of the hat matrix computed using the subset of the data, λ_i. Moreover, we must sum over the m orthogonal directions in this metric since we are considering a deletion of m observations, whereas in the single-deletion case we simply had $m = 1$.

Finally, the empirical influence function metric can be extended to analyze changes in q linearly independent combinations of parameter elements β_i due to the removal of a subset of data. A quintessential example of such a case is where we are interested in a subset of the parameters rather than the entire parameter vector. Thus, consider $\hat{\psi} = \mathbf{L}\hat{\beta}$, where $\mathbf{L}$ is a $q \times p$ matrix of rank q. Then, the distance $D_I(\psi)$ between the full-data estimate $\hat{\psi}$ and the estimate with subset I removed $\hat{\psi}_{(-I)}$ is:

$$D_I(\psi) = \frac{(\hat{\psi} - \hat{\psi}_{(-I)})^T [\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1} (\hat{\psi} - \hat{\psi}_{(-I)})}{qs^2}$$

which is equivalent to the general form $D_I(\mathbf{M}, c)$ with the matrix and scale factor chosen such that $c = qs^2$ and $\mathbf{M} = \mathbf{L}^T[\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1}\mathbf{L}$:

$$\begin{aligned} D_I(\psi) &= \frac{(IF_I)^T \mathbf{L}^T [\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1} \mathbf{L}(IF_I)}{qs^2} \\ &= \frac{\mathbf{e}_I^T(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{X}_I(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{M}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_I^T(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{e}_I}{qs^2} \end{aligned}$$

again employing the formula by Bingham (1977). While this expression is not ripe for simplification in the general case, a reduction is possible in the special case in which we are interested in a subset of β. Suppose we are interested in the last q components of β; consequently, partition $\mathbf{X} = (\mathbf{X}_1 : \mathbf{X}_2)$ such that $\mathbf{X}_1$ consist of the first $p - q$ columns and $\mathbf{X}_2$ consists of the last q columns. Then, we have $\mathbf{L} = (\mathbf{0}_{q\times(p-q)} : \mathbf{I}_{q\times q})$. Thus, we have:

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{M}(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1} - \begin{pmatrix} (\mathbf{X}_1^T\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

which can be used to simplify the above expression for $D_I(\psi)$ as:

$$\begin{aligned} D_I(\psi) &= \frac{\mathbf{e}_I^T(\mathbf{I} - \mathbf{H}_I)^{-1}(\mathbf{H}_I - \mathbf{U}_I)(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{e}_I}{qs^2} \\ &= \frac{ps^2 D_I - \mathbf{e}_I^T(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{U}_I(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{e}_I}{qs^2} \end{aligned}$$

where $\mathbf{U} = \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T$. Thus, in the single-deletion case in which we are interested in a subset of the parameters β, we have $\mathbf{e}_I = e_i$ and:

$$\boxed{D_i(\psi) = \frac{r_i^2}{q}\frac{h_{ii} - u_{ii}}{1 - h_{ii}}}$$

This expression also resembles the formula for the simplified Cook's distance for the entire parameter β in the single-deletion case, except we are now scaling by the number of parameter elements of interest, q, and correcting the leverage term h_{ii} by the submatrix of the projection corresponding to the parameters not under investigation, u_{ii}.

4 Extending to Generalized Linear Models

While the particular methods developed and reviewed in this paper are applicable only to linear models, the conceptual framework underlying the diagnostics can be extended to generalized linear models. The analysis of influential observations in diagnosing model fit and issues in the data is again based on a combination of residuals and single case deletions. One point of difference in generalized linear models, as noted by Pregibon (1981), is that there do not exist natural, uniquely defined residuals; they can be defined on several scales and in varying manners. The two most often-utilized residuals are the components of the Pearson chi-squared statistic and the deviance residuals ($D(y, \hat{\mu}) = \sum_i d_i^2$).

Just as Cook's distance and empirical influence function were metrics based on the "distance" traveled by the estimated effect parameters $\hat{\beta}$ when a parameter was deleted $\hat{\beta}_{(-i)}$, generalized linear model diagnostics seek to quantify this distance for the fitted parameters. Pregibon (1981) exploits the weight matrix $\mathbb{W}$, which for canonical link functions is equivalent to a diagonal matrix consisting of the dispersion parameter functions $a(\phi)$. In turn, these $a(\phi) = w_i$ generally, where w_i is the weight given to a particular observation and is often related to the number of counts for the observation in grouped cases, yielding the equations:

$$\sum_{i=1}^{n} w_i(y_i - \mu_i)x_{ij} = 0$$

Pregibon's "one-step approximation" utilizes the fact that $w_i = 1$ for ungrouped data, which is often the case with quantitative or many-category predictors. Iteratively solving for the $\hat{\beta}_{(i)}$ with $w_i \in [0, 1]$ for the i^{th} observation and $w_j = 1$ for $j \neq i$, we can explore how this estimate varies with w; that is, analyze $\frac{\partial}{\partial w}\hat{\beta}_{(i)}(w)$. This provides a measure of how the estimate changes when we place less weight on the i^{th} observation.

In conclusion, such measures - which sometimes technically or computationally more involved due to the greater number of moving parts of generalized linear models - are based on the same fundamental idea as Cook's distance and empirical influence functions for linear models: diagnosing potentially problematic points in the data, or influential observations, by comparison of fitted estimates when less (or no) weight is placed on a particular observation. In this sense, the results and methods of this paper are conceptual forebears to diagnostics for generalized linear models, and can be extended accordingly.

References

Andrews, D. F., and D. Pregibon (1978). “Finding the Outliers that Matter.” *J. R. Statist. Soc. B* **40**: 85-93.

Beckman, R. J., and H. J. Trussell (1974). “The Distribution of an Aribtrary Studentized Residual and the Effects of Updating in Multiple Regression.” *J. Amer. Statist. Assoc.* **69**: 199-201.

Bingham, C. (1977). “Some Identities Useful in the Analysis of Residuals from Linear Regression.” *Technical Report No. 300*, School of Statistics, University of Minnesota.

Cook, R. D. (1977) “Detection of Influential Observation in Linear Regression.” *Technometrics* **19**: 15-18.

Cook, R. D., and S. Weisberg (1979). “Finding Influential Cases in Linear Regression – A Review.”

Cook, R. D., and S. Weisberg (1980). “Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression.” *Technometrics* **22**: 495-508.

Cook, R. D., and S. Weisberg (1982). *Residuals and Influence in Regression.* London & New York: Chapman & Hall.

Hager, W. W. (1989). “Updating the Inverse of a Matrix.” *SIAM Review* **31**: 221-239.

Hampel, F. R. (1974). “The Influence Curve and Its Role in Robust Estimation.” *J. Amer. Statist. Assoc.* **69**: 383-393.
Pregibon, D. (1981) “Logistic Regression Diagnostics.” *Ann. Stat.* **9**: 705-724.

Welsch, R. E., and E. Kuh (1977). “Linear Regression Diagnostics.” *NBER Working Paper* **173**.

Williams, D. A. (1987). “Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions.” *Appl. Statist.* **36**: 181-191.