Details on Model Diagnostics

This document discusses some of the mathematical details of the model diagnostics - leverage, standardized residuals, and Cook's distance. We assume the reader knowledge of the matrix form for multiple linear regression.Please see Matrix Form of Linear Regression for a review.

Introduction

Suppose we have n observations. Let the i^{th} be $(x_{i1}, \ldots, x_{ip}, y_i)$, such that $x_{i1}, \ldots, x_{ip}$ are the explanatory variables (predictors) and y_i is the response variable. We assume the data can be modeled using the least-squares regression model, such that the mean response for a given combination of explanatory variables follows the form in (1).

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1}$$

We can write the response for the i^{th} observation as shown in (2)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{2}$$

such that ϵ_i is the amount y_i deviates from $\mu\{y|x_{i1}, \ldots, x_{ip}\}$, the mean response for a given combination of explanatory variables. We assume each $\epsilon_i \sim N(0, \sigma^2)$, where σ^2 is a constant variance for the distribution of the response y for any combination of explanatory variables $x_1, \ldots, x_p$.

Matrix Form for the Regression Model

We can represent the (1) and (2) using matrix notation. Let

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{3}$$

Thus,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

Therefore the estimated response for a given combination of explanatory variables and the associated residuals can be written as

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{4}$$

Hat Matrix & Leverage

Recall from the notes **Matrix Form of Linear Regression** that $\hat{\boldsymbol{\beta}}$ can be written as the following:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{5}$$

Combining (4) and (5), we can write $\hat{\mathbf{Y}}$ as the following:

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \end{aligned} \tag{6}$$

We define the **hat matrix** as an $n \times n$ matrix of the form $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Thus (6) becomes

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \tag{7}$$

The diagonal elements of the hat matrix are a measure of how far the predictor variables of each observation are from the means of the predictor variables. For example, h_{ii} is a measure of how far the values of the predictor variables for the i^{th} observation, $x_{i1}, x_{i2}, \ldots, x_{ip}$, are from the mean values of the predictor variables, $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$. In the case of simple linear regression, the i^{th} diagonal, h_{ii}, can be written as

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

We call these diagonal elements, the **leverage** of each observation.

The diagonal elements of the hat matrix have the following properties:

- $0 \leq h_i i \leq 1$
- $\sum_{i=1}^{n} h_{ii} = p + 1$, where p is the number of predictor variables in the model.
- The mean hat value is $\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p+1}{n}$.

Using these properties, we consider a point to have **high leverage** if it has a leverage value that is more than 2 times the average. In other words, observations with leverage greater than $\frac{2(p+1)}{n}$ are considered to be **high leverage** points, i.e. outliers in the predictor variables. We are interested in flagging high leverage points, because they may have an influence on the regression coefficients.

When there are high leverage points in the data, the regression line will tend towards those points; therefore, one property of high leverage points is that they tend to have small residuals. We will show this by rewriting the residuals from (4) using (7).

$$\begin{aligned} \mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{H}\mathbf{Y} \\ &= (1 - \mathbf{H})\mathbf{Y} \end{aligned} \tag{8}$$

Note that the identity matrix and hat matrix are **idempotent**, i.e. $\mathbf{II} = \mathbf{I}$, $\mathbf{HH} = \mathbf{H}$. Thus, ($\mathbf{I} - \mathbf{H}$ is also idempotent. These matrices are also symmetric. Using these properties and (8), we have that the variance-covariance matrix of the residuals $\boldsymbol{e}$, is

$$
\begin{aligned}
Var(\mathbf{e}) &= \mathbf{e}\mathbf{e}^T \\
&= (1-\mathbf{H})Var(\mathbf{Y})^T(1-\mathbf{H})^T \\
&= (1-\mathbf{H})\hat{\sigma}^2(1-\mathbf{H})^T \\
&= \hat{\sigma}^2(1-\mathbf{H})(1-\mathbf{H}) \\
&= \hat{\sigma}^2(1-\mathbf{H})
\end{aligned} \tag{9}
$$

where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1}$ is the estimated regression variance. Thus, the variance of the i^{th} residual is $Var(e_i) = \hat{\sigma}^2(1-h_{ii})$. Therefore, the higher the leverage, the smaller the variance of the residual. Because the expected value of the residuals is 0, we conclude that points with high leverage tend to have smaller residuals than points with lower leverage.

Standardized Residuals

In general, we standardize a value by shifting by the expected value and rescaling by the standard deviation (or standard error). Thus, the i^{th} standardized residual takes the form

$$
std.res_i = \frac{e_i - E(e_i)}{SE(e_i)}
$$

The expected value of the residuals is 0, i.e. $E(e_i) = 0$. From (9), the standard error of the residual is $SE(e_i) = \hat{\sigma}\sqrt{1-h_{ii}}$. Therefore,

$$
std.res_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \tag{10}
$$

Cook's Distance

Cook's distance is a measure of how much each observation influences the model coefficients, and thus the predicted values. The Cook's distance for the i^{th} observation can be written as

$$
D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})^T(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{(p+1)\hat{\sigma}} \tag{11}
$$

where $\hat{\mathbf{Y}}_{(i)}$ is the vector of predicted values from the model fitted when the i^{th} observation is deleted. Cook's Distance can be calculated without deleting observations one at a time, since (12) below is mathematically equivalent to (11).

$$
D_i = \frac{1}{p+1} std.res_i^2 \left[\frac{h_{ii}}{(1-h_{ii})}\right] = \frac{e_i^2}{(p+1)\hat{\sigma}^2(1-h_{ii})}\left[\frac{h_{ii}}{(1-h_{ii})}\right] \tag{12}
$$