Chapter 2

Classical Regression Analysis Using OLS

2.A Formulation of the Model

Consider the desire to investigate the relationship between some response variable y, and k regressor variables x_1, x_2, ..., x_k. In general, the function we wish to estimate is the mean function m such that

$$y_i = m(x_{1i}, x_{2i}, ..., x_{ki}) + \varepsilon_i ,$$

where x_{ji} is the i^{th} realization of the regressor variable x_j. It is assumed that the values of the explanatory variables are measured without error. The linear assumption referred to in the previous section implies that the form for the underlying model is

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \ldots + \beta_k \cdot x_{ki} + \varepsilon_i \quad i = 1,...,n.$$

This can also be written as $y_i = x_i' \boldsymbol{\beta} + \varepsilon_i$, where $x_i' = (1\ x_{1i}\ x_{2i}\ \ldots\ x_{ki})$ is a (1xR) vector of observed regressor information for the i^{th} subject, $\boldsymbol{\beta}$ is a (Rx1) vector of unknown regression parameters, and R = k+1 is the number of parameters in the model.

Traditionally, an assumption that accompanies the linear model is that the random errors are independent and follow a normal distribution with mean zero and common variance σ^2. We also make these assumptions in this section, but they will be much different in subsequent sections. In matrix notation, the linear model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} , \qquad (2.A.1)$$

where $\mathbf{y}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ are all (R×1) vectors and $\mathbf{X}$ is an (n×R) model matrix. The i^{th} row of $\mathbf{X}$ represents the values of all the regressor variables for observation i.

Since the model in (2.A.1) involves unknown parameters, one goal in regression analysis is to obtain the coefficient estimates $\hat{\boldsymbol{\beta}}$. The method of least squares selects the set of coefficients that produce the fitted model that has the smallest sum of squared residuals. In matrix notation, it is the vector $\hat{\boldsymbol{\beta}}$ that minimizes the quantity $(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$, which leads to $\hat{\boldsymbol{\beta}}^{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. These coefficient estimates are optimal (Uniformly Minimum Variance Unbiased Estimators (UMVUE), Best Linear Unbiased Estimators (BLUE)) under the assumptions that the model is correct and that the error terms are independent and identically distributed normal random variables.

The inherently non-robust properties of ordinary least squares (OLS) are consequences of the use of the quadratic loss function. Relatively large residuals (which may result from either outliers or a misspecified model leading to lack-of-fit) can have a large impact on the objective function and thus exert a large amount of influence on the fitted values.

2.B The HAT Matrix

The estimated values of the dependent variable Y may be written as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{ols} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}^{ols}\mathbf{y} = \begin{bmatrix} \mathbf{h}_1' \\ \mathbf{h}_2' \\ \vdots \\ \mathbf{h}_n' \end{bmatrix}\mathbf{y},$$

where $\mathbf{h}_i' = (h_{i1}\ h_{i2}\ \ldots\ h_{in})$ is a (1xn) vector and $\mathbf{H}^{OLS}$ is the "Hat" matrix for OLS. The Hat matrix can be thought of as a transformation matrix that transforms the y values into $\hat{y}$ values. Note that $\hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j$, which indicates that $\hat{y}_i$ is just a weighted sum (average) of the observations. Note also that h_{ii} is the weight that y_i receives when calculating the predicted

value $\hat{y}_i$ and h_{ij} is the weight that y_j receives when calculating $\hat{y}_i$. If k = 1 (the simple linear regression case) then

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

This indicates that the magnitude of h_{ij} (that is, $|h_{ij}|$) is increasing in $|x_j - \bar{x}|$ when finding the predicted value $\hat{y}_i$. That is, observations farthest from x_i receive the most weight. This reflects the assumption that the simple linear regression model is correct and that responses at the most remote locations be given the largest weight in the calculation of the estimated responses. It also indicates that h_{ii} is a (standardized) measure of the distance of x_i from the rest of the data. In the multivariate case, the h_{ii} value is analogous to Mahalanobis distance.

Note the following properties of the OLS Hat matrix, which shed some light on the usefulness of it and some of its components.

1. It is idempotent ($(\mathbf{H}^{OLS})^2 = \mathbf{H}^{OLS}$) and symmetric (n×n)
2. Trace($\mathbf{H}^{OLS}$)= $\sum_{i=1}^{n} h_{ii}^{(OLS)}$ = , the number of parameters in the assumed underlying model.
3. $-1 \le h_{ij}^{(OLS)} \le 1$
4. $\frac{1}{n} \le h_{ii}^{(OLS)} \le 1$
5. $\sum_{j=1}^{n} h_{ij}^{(OLS)} = 1$, for all i = 1, 2, …, n (the sum of the weights used to compute the i^{th} predicted value $\hat{y}_i$ is 1).
6. $\mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H}^{OLS})\mathbf{y} = \mathbf{e}$, the vector of residuals.
7. Var($\mathbf{e}$)= $\sigma^2 (\mathbf{I} - \mathbf{H}^{OLS})$ under the assumption of constant variance.

8. $s^2 = \hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - R} = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - \text{trace}(\mathbf{H}^{OLS})}$, an unbiased estimate of σ^2.

2.C Outliers

An observation, say (x_i, y_i), with high leverage, or with a large h_{ii} value, has the potential for a large influence on the parametric fit. It is referred to as an outlier in the x-direction. This means that y_i will have large weight for all those observations for which x_i is far from the x-values that correspond to those observations, giving it a large amount of influence on several of the fitted values. Note that the average h_{ii} value is R/n, and as a rule of thumb an observation is considered to be a high leverage point if $h_{ii} > 2R/n$ (if its leverage value is more than twice the average leverage value).

Leverage, in and of itself, is not inherently an undesirable characteristic. The classification of an observation as good or bad depends on the value of the *dependent* variable at that location. If the value of y follows the general trend of the data, then that observation would be considered a good point because it reinforces the trend via its influence on the individual fits. If the value of y does not follow the general trend of the data, then that observation, called a high influence point (hip), would be considered a "bad" observation, since its large influence is detrimental to the overall fit.

For clarity, then, an outlier can be defined as any observation whose true error value is large relative to the set of errors for the entire sample. If the leverage at that point is large, then it will be referred to as a high influence point.

2.D Outlier Diagnostics in the Linear Regression Model

Residuals are the key to identifying outliers in the regression situation. For example, it is common for a high influence point to have a small residual because of the large influence it exerts

on the overall fit. Thus one might look for high leverage and a small residual in order to identify a hip. A low leverage outlier will exert a certain amount of influence over the fit, but not as much as a hip. Therefore, one would expect a low leverage outlier to have a small to moderate h_{ii} value, while having a relatively large residual. The conclusion is that one must consider both leverage values and residual values when utilizing a diagnostic tool for identification of extreme observations.

In the list of properties of $\mathbf{H}^{OLS}$, it follows from point 7 that $Var(e_i) = \sigma^2 \cdot (1-h_{ii})$, and $Cov(e_i,e_j) = -h_{ij} \cdot \sigma^2$. Note that the farther x_i is from the center of the data, the larger the value of h_{ii} becomes, which indicates a smaller variance for e_i. Thus, since e_i could be used as an estimate for ε_i, and $E(\varepsilon_i) = 0$, the e_i's can be thought of as better estimates of zero if the corresponding h_{ii} is "large". Note that also, under the assumption of normally distributed errors, it follows that $e_i \sim N(0, \sigma^2 \cdot (1-h_{ii}))$.

Studentized Residuals

A common goal when analyzing residuals is to distinguish between observations that are high leverage, outliers, or hips. The studentized residuals can provide this information, and are merely the residuals standardized by their estimated standard deviation, which are given by

$$r_i = \frac{y_i - \hat{y}_i}{s\sqrt{1-h_{ii}}} = \frac{e_i}{se(e_i)},$$

where $se(e_i)$ is the standard error of the i^{th} residual. This standardization removes the effect of the "x" position of the data point. Large values of r_i indicate that the response y_i is an outlier or perhaps the model is misspecified.

R-Student Residuals and DFFITS

The R-student residuals (externally studentized residuals) were formed after the realization that an outlier will inflate the value of s^2, causing an artificial deflation in its corresponding

studentized residual. Thus a "leave-one-out" method of calculating the variance estimate was used to form the R-student residuals given by

$$t_i = \frac{y_i - \hat{y}_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}},$$

where s_{-i} denotes the sample standard deviation that is calculated in the absence of the i^{th} observation, $(y_i, \mathbf{x}_i')$.

It is also common to determine the amount of influence of an observation by way of cross-validation, that is observing the difference in estimated coefficient values or fitted values of the response variable with and without a particular observation being used in the calculation of those fits. This is the single point deletion scheme as introduced above in calculating s_{-i}. Notation for a fitted value for y_i that does not use y_i in calculating the fit is given by $\hat{y}_{i,-i}$, and the corresponding single point deletion residual, called a PRESS residual, is given by $e_{i,-i} = y_i - \hat{y}_{i,-i}$. By some algebraic calculations, it can be shown that the PRESS residual $e_{i,-i} = \frac{e_i}{1-h_{ii}}$.

The influence of an observation on the fit at that location can be measured by the DFFITS statistic, which is similar to the r-student residual and is defined as

$$\begin{aligned} \text{DFFITS}_i &= \frac{\hat{y}_i - \hat{y}_{i,-i}}{se(\hat{y}_i)} = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}} \\ &= t_i * \left[\frac{h_{ii}}{1-h_{ii}}\right]^{\frac{1}{2}}. \end{aligned}$$

It is evident from the relationship of DFFITS to t_i that an observation's t_i value is either exaggerated or deflated according to the size of its leverage value. Since either a small residual or a small leverage value can cause an observation's DFFITS value to be small, its primary use is in detecting high influence points.

2.E Discussion

The purpose of this chapter is to give a brief review of classical methods (for a more detailed discussion on any of the topics mentioned above, see Myers (1990)). The non-robustness of least squares regression to either model misspecification or outlyingness is quite obvious from the definition of the estimators. However, the notation and the procedures outlined in the development of the least squares estimators prove useful in the subsequent sections.

The chapter is also meant to give short discussion on developments such as outlier diagnostics. A diagnostic based on the method developed in the current research will have a similar theoretical basis as these and is described briefly in Chapter 7. Many of the areas that have been given extensive attention in the classical regression setting remain unexplored in the setting with which we are concerned. We will address this issue in subsequent chapters.