3.3 LSE's as Minimum Deviance Type Estimators

The criterion of least squares leads to best estimators among the class of linear unbiased estimators which is the Gauss-Markov Theorem. However, there are questions about whether the resulting estimators remain optimal, or even have any good properties when the assumptions of the general linear model are not true. This leads to a discussion of robustness and resistance.

Definition: Robustness is loosely defined as the property of statistical procedures to be relatively insensitive to distributional assumptions.

Definition: Resistance is loosely defined as the property of statistical procedures to be relatively insensitive to changes in a small fraction of the data.

example: The simplest linear model is

$$E(Y_i) = \mu_i \quad i = 1, 2, \ldots, n$$

where the Y_i are uncorrelated with common variance σ^2. Under this model the best linear unbiased estimator for μ is $\hat{\mu} = \bar{Y}$. If the underlying distribution is normal then the estimator is also minimum variance unbiased. However, if the distribution is of the form

$$f_Y(y) = \left(1 - \frac{1}{k}\right) \phi(y - \mu) + \frac{1}{k} c(y - \mu)$$

where c denotes the standard Cauchy density and ϕ denotes the standard normal density then the expected value of $\bar{Y}$ is not defined for any value of k. Thus the sample mean is not robust.

Let $\bar{Y}_{(i)}$ denote the mean of $Y_1, Y_2, \ldots, Y_n$ with the ith observation removed. Then since

$$\bar{Y} = \left(1 - \frac{1}{n}\right) \bar{Y}_{(i)} + \frac{Y_i}{n}$$

we see that $\bar{Y}$ is not resistant.

- It is useful to consider LSE's as a member of another larger class of estimators within which LSE's are no longer optimal.
- This serves to put the Gauss-Markov result in perspective and illustrates how information about the response distribution beyond the first two moments may be used to construct estimators that perform better than LSE's. These are so-called robust estimators.
- Assume a general form of dependence of the expected value of $\mathbf{Y}$ on $\mathbf{X}$ and $\boldsymbol{\beta}$.
    - We will denote this expected value in short-hand notation as $\widehat{\mathbf{Y}}(\boldsymbol{\beta})$.
    - Thus, for any $p \times 1$ vector $\mathbf{b}$, $\widehat{\mathbf{Y}}(\mathbf{b})$ will denote the expected value of $\mathbf{Y}$ under the assumption that $\mathbf{b}$ is the true parameter value.

To define an estimator, we now choose a function that represents a measure of discrepancy or **deviance** between two vectors in $\mathbf{R}^n$.

- This function will be used to determine the deviance of $\widehat{\mathbf{Y}}(\mathbf{b})$ from $\mathbf{Y}$ for given $\mathbf{b}$, denoted dev $(\mathbf{Y}; \widehat{\mathbf{Y}}(\mathbf{b}))$
- An estimator of $\boldsymbol{\beta}$ will be defined as the value (or set of values) of $\mathbf{b}$ at which dev $(\mathbf{Y}; \widehat{\mathbf{Y}}(\mathbf{b}))$ attains a minimum.
- This class of estimators, indexed by possible choices of deviance functions, can be called minimum deviance estimators.

Clearly, reasonable deviance functions should have properties similar to distance functions defined in $\mathbf{R}^n$ corresponding to choices of inner products.

- In fact, for any choice of inner product in $\mathbf{R}^n$, the resulting distance function would be a perfectly good candidate for a deviance function.
- However, deviance functions need not satisfy all the properties of distance functions. For instance, there is no reason why deviance functions should be symmetric in their arguments.

- By making some general assumptions about the distribution of $\mathbf{Y}$, we can make reasonable restrictions on the form of the deviance function.
  - If we consider the components of $\mathbf{Y}$ to be independent, then we might consider deviance functions that are represented as the sum of deviances for each component of Y, i.e.

$$\text{dev}\left(\mathbf{Y}; \widehat{\mathbf{Y}}(\mathbf{b})\right) = \sum_{i=1}^{n} \text{dev}_i\left(Y_i; \widehat{Y}_i(\mathbf{b})\right)$$

  - If the distribution of Y is a location-scale family i.e.

$$f_{Y_i}(y_i) = f\left(\frac{y_i - \widehat{Y}_i(\mathbf{b})}{\sigma}\right)$$

    then $\text{dev}_i(Y_i; \widehat{Y}_i(\mathbf{b}))$ may be taken to be the scaled residual.

- Clearly, LSE's are a member of the class of minimum deviance estimators with deviance function

$$\text{dev}_{LSE}\left(\mathbf{Y}, \widehat{\mathbf{Y}}(\mathbf{b})\right) = \|\mathbf{Y} - \widehat{\mathbf{Y}}(\mathbf{b}))\|^2 = \sum_{i=1}^{n}[y_i - \hat{y}_i(\mathbf{b})]^2$$

- Maximum likelihood estimators are also a member of this class with deviance function defined as minus one times the log-likelihood function.

- As a special case, suppose

$$\mathbf{Y} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

  Then the deviance function corresponding to the MLE is (up to an additive constant)

$$\text{dev}_{MLE}\left(\mathbf{Y}, \widehat{\mathbf{Y}}(\mathbf{b})\right) = \frac{1}{2\sigma^2}\|\mathbf{Y} - \widehat{\mathbf{Y}}(\mathbf{b})\|^2$$

  - Therefore, with an assumption of normality, the MLE and LSE of $\boldsymbol{\beta}$ are identical.
  - Thus, at normality, the LSE is not only BLUE but also has the optimality properties associated with the MLE for normal data (i.e. it has minimum variance among all unbiased estimators whether they are linear or not).

As we let the underlying distribution of $\mathbf{Y}$ move away from normality, the MLE retains some optimality properties.

- The MLE has minimum asymptotic variance among the class of regular, $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$.
    - A regular $\sqrt{n}$ consistent estimator is one which converges to the true value at rate $\sqrt{n}$ and, when properly normalized, converges in law to a Gaussian random variable.
- Therefore, since away from normality, LSE's and MLE's no longer coincide, the LSE is dominated (in large samples) by the MLE.
- This dominance may be explained by the MLE's use of the shape of the response distribution to define the measure of deviance.
    - For example, if the response distribution has a long right tail, then an observation in a fixed distance above its expected value should not be counted as discrepant as an observation the same distance below its expected value.
    - Or, if the response distribution had heavy tails (i.e. prone to "outliers"), a few observations far away from their expected values might not necessarily be counted as overly discrepant.
- With these ideas in mind, we note that the LSE's measure of deviance is well suited for symmetric distributions with relatively "light" tails.

Because of the sensitivity of LSE's to heavy tailed response distributions, other minimum deviance type estimators have been proposed whose deviance functions are designed to retain some of the optimality properties of LSE's at normality but which are less sensitive to heavy tails.

- The deviance function for these so called robust regression estimators are defined by

$$\text{dev}\left(\mathbf{Y};\widehat{\mathbf{Y}}(\mathbf{b})\right) = \sum_{i=1}^{n} \rho\left(\frac{Y_i - \widehat{Y}_i(\mathbf{b})}{\sigma}\right)$$

  where the function $\rho(x)$ has a form similar to x^2 (i.e. the deviance for LSE) for x near zero but for x farther away from zero, $\rho(x) < x^2$.

- Typically a symmetric interval about zero is designated as $[-c, c]$ and $\rho(x)$ is defined separately for $x \in [-c, c]$ and for $x \notin [-c, c]$.

  - One example is Huber's deviance function with trimming constant H, $\rho_H(\cdot)$;

$$\rho_H(x) = \begin{cases} \frac{x^2}{2} & |x| \leq H \\ H|x| - \frac{H^2}{2} & |x| > H \end{cases}$$

  - Another example is Tukey's biweight deviance function with trimming constant B, $\rho_B(\cdot)$

$$\rho_H(x) = \begin{cases} \frac{B^2}{2}\left\{1 - \left[1 - \left(\frac{x}{B}\right)^2\right]^2\right\} & |x| \leq B \\ \frac{B^2}{2} & |x| > B \end{cases}$$

- Note that as the trimming constants become large, then $\rho_H(\cdot)$ and $\rho_B(\cdot)$ become quadratic functions over a wide interval around zero.
- Thus, for large H and or B, these estimators correspond to LSE's, and so, at normality, these estimators should be nearly fully efficient.
- If H and B are chosen to be small, then a greater proportion of the data have a non quadratic deviance function applied to it and so the efficiency at normality will be low.
- The trimming constants are typically chosen so that they are far enough away from zero so that the asymptotic relative efficiency (ARE) of the resulting estimator at normality is high (95% say).
- Thus, the price for using a robust estimator as expressed in terms of loss of efficiency at normality is set at a predetermined, small amount.

- These estimators are called M-estimators after maximum likelihood estimators.
- Also note that unlike LSE's, the deviance functions for these M-estimators require knowledge of σ^2. In practice a robust estimator of scale (such as the median absolute deviation (MAD)) is used in place of σ in the deviance function.

3.4 Weighted Least Squares and Maximum Likelihood

Let $\mathbf{Y}$ be WS$(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$ where

$$\mathbf{V} = \text{diag}\,(V_1, \ldots, V_N) \text{ where } V_i > 0; \;\; 1 \leq i \leq n$$

- Then we know that the LSE of $\mathbf{X}\boldsymbol{\beta}$, $\mathbf{Xb}$, is unique and is given by the formula
$$\mathbf{Xb} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{Y}$$
- It is easy to verify that $\mathbf{Xb}$ is still unbiased under this model and that
$$\text{var}\,(\mathbf{Xb}) = \sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$$
- Therefore $\mathbf{Xb}$ is an estimator of $\mathbf{X}\boldsymbol{\beta}$ with some reasonable properties.
- However, the deviance function corresponding to least squares does not use all of the information available about the distribution of $\mathbf{Y}$ and hence might be improved upon.

- Specifically, the V_i's are not used in the LSE.
  - Thus if $V_1 >> V_2$ and

$$Y_1 - \widehat{\mathbf{Y}}_1(\mathbf{b}) = Y_2 - \widehat{\mathbf{Y}}_2(\mathbf{b})$$

    we would not want to consider $\widehat{\mathbf{Y}}_1(\mathbf{b})$ as equally discrepant as $\widehat{\mathbf{Y}}_2(\mathbf{b})$.

- One reasonable approach for incorporating the information in the V_i's is to define the deviance function as

$$\text{dev}\,(\mathbf{Y}; \widehat{\mathbf{Y}}(\mathbf{b})) = \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{Y}_i(\mathbf{b})}{\sqrt{V_i}} \right)^2$$

  - That is we apply the least squares deviance function to standardized residuals.
  - One might also view this deviance function as the least squares deviance applied to the transformed model:

$$\widetilde{\mathbf{Y}} = \mathbf{V}^{-\frac{1}{2}}\mathbf{Y} \;\; ; \;\; \widetilde{\mathbf{X}} = \mathbf{V}^{-\frac{1}{2}}\mathbf{X}$$

    where

$$\mathbf{V}^{-\frac{1}{2}} = \text{diag}\,\left(\frac{1}{\sqrt{V_1}}, \ldots, \frac{1}{\sqrt{V_n}} \right)$$

    so that

$$\widetilde{\mathbf{Y}} \sim \text{WS}(\widetilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

- Since after the transformation, the model has the uncorrelated, homoscedastic form for var $(\mathbf{Y})$, we may apply results from previous sections to determine the BLUE (or Gauss-Markov estimator) of $\mathbf{X}\boldsymbol{\beta}$ by minimizing the deviance function:

$$\begin{aligned} \|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\mathbf{b}\|^2 &= \|\mathbf{V}^{-\frac{1}{2}}(\mathbf{Y} - \mathbf{X}\mathbf{b})\|^2 \\ &= (\mathbf{Y} - \mathbf{X}\mathbf{b})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{b}) \end{aligned}$$

- This last form for the deviance function function, provides the motivation for defining the Weighted Least Squares Estimator (WLSE) of $\boldsymbol{\beta}$ for general, known positive definite $\mathbf{V}$ as the value or set of values, $\mathbf{b}$, at which the function

$$(\mathbf{Y} - \mathbf{X}\mathbf{b})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{b})$$

  is minimized.

- The weighted least squares estimate is easily seen to be

$$\begin{aligned} \mathbf{b} &= (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\tilde{\mathbf{y}} \\ &= [(\mathbf{V}^{-1/2}\mathbf{X})^T(\mathbf{V}^{-1/2}\mathbf{X})]^{-1}(\mathbf{V}^{-1/2}\mathbf{X})^T(\mathbf{V}^{-1/2}\mathbf{y} \\ &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \end{aligned}$$

- There is another interpretation of the WLSE that is of interest.
  - Recall that one way to define an inner product in $\mathbf{R}^n$ other than the standard inner product is to define

    $$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y}$$

    for any fixed, positive definite matrix $\mathbf{A}$ where $(\mathbf{x}, \mathbf{y})$ denote vectors in $\mathbf{R}^n$.
  - Since $\mathbf{V}^{-1}$ is positive definite, define the inner product

    $$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{V}^{-1}} = \mathbf{x}^T \mathbf{V}^{-1} \mathbf{y}$$

  - Note that the WLSE has a deviance function that may be interpreted as the squared length of the vector $\mathbf{Y} - \mathbf{X}\mathbf{b}$ except with the length function now defined in terms of this different inner product:

    $$\text{dev}\,(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2_{\mathbf{V}^{-1}} = \langle \mathbf{Y} - \mathbf{X}\mathbf{b}, \mathbf{Y} - \mathbf{X}\mathbf{b} \rangle_{\mathbf{V}^{-1}}$$

  - Thus, the WLSE of $\mathbf{X}\mathbf{b}$ may be considered as the projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$ except now the projection operator used is that corresponding to to a different inner product.

- To gain insight into this non-standard type of projection, recall that the equation

$$f(\boldsymbol{\eta}) = (\mathbf{y} - \boldsymbol{\eta})^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\eta}) = c^2$$

  defines an ellipsoid in $\mathbf{R}^n$ with center $\mathbf{Y}$, principal axes corresponding to the eigenvectors of $\mathbf{V}$ and lengths of semi principal axes $c\sqrt{\lambda_1}, \ldots, c\sqrt{\lambda_n}$ where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{V}$.

- To find the projection of $\mathbf{Y}$ onto a subspace, $\mathbf{Sp}(\mathbf{X})$, we can consider the set of points

$$\{\boldsymbol{\eta} \in \mathbf{R}^n \ : \ f(\boldsymbol{\eta}) = c^2\}$$

  starting at small values of c and then letting c increase until a point on the ellipsoid just touches the subspace.

- The projection with respect to $\langle \cdot\,, \cdot \rangle_{\mathbf{V}^{-1}}$ of $\mathbf{Y}$ onto $S(\mathbf{X})$ is the vector whose tip corresponds to this first point of intersection of the ellipsoids with $\mathbf{Sp}(\mathbf{X})$.

- Finally, we note that under the assumption that $\mathbf{Y}$ is MVN$(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$ we have the deviance function corresponding to the MLE for $\boldsymbol{\beta}$ given by:

$$\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2_{\mathbf{V}^{-1}} + \text{constants}$$

  - Thus, just as the LSE and the MLE are identical under the assumption that $\mathbf{Y}$ is MVN$(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ so are the WLSE and the MLE under the assumption that $\mathbf{Y}$ is MVN$(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$

3.4.1 Non-Robustness of the Weighted Least Squares Estimate

Let Y_1 and Y_2 be independent with Y_1 having pdf given by

$$f_{Y_1}(x) = \begin{cases} \frac{1}{2} & \theta - 1 \leq x \leq \theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

and Y_2 having pdf given by

$$f_{Y_2}(x) = \begin{cases} \frac{1}{4} & \theta - 2 \leq x \leq \theta + 2 \\ 0 & \text{otherwise} \end{cases}$$

Given observations y_1 and y_2 the likelihood for for θ is given by

$$\text{lik}\,(\theta; y_1, y_2) = \frac{1}{8}\, I_{\theta-1,\theta+1}(y_1) I_{\theta-2,\theta+2}(y_2)$$

where

$$I_{a,b}(y) = \begin{cases} 1 & a \leq y \leq b \\ 0 & \text{otherwise} \end{cases}$$

Thus if $y_1 = 1$ we have that

$$\theta - 1 \leq 1 \leq \theta + 1 \implies \theta \leq 2 \text{ and } \theta \geq 0$$

while if $y_2 = 3.8$ we have that

$$\theta - 2 \leq 3.8 \leq \theta + 2 \implies \theta \leq 5.8 \text{ and } \theta \geq 1.8$$

and it follows that

$$\text{lik}\,(\theta; y_1 = 1, y_2 = 3.8) = \frac{1}{8}\, I_{1.8,2}(\theta)$$

The assumptions on Y_1 and Y_2 satisfy the general linear model so that we have

$$E\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \theta$$

Since

$$\text{var}\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{4}{3} \end{bmatrix}$$

the Gauss Markov Theorem gives the BLUE as

$$\hat{\theta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{4}{3} \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 3.8 \end{bmatrix}$$

It is easy to check that

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} = \frac{15}{4}$$

and that

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \frac{23.4}{4}$$

so that

$$\hat{\theta} = \frac{23.4}{15} = 1.56$$

which is not a value of θ consistent with the likelihood function.

Therefore the Gauss-Markov Theorem can produce estimates which are inconsistent with the likelihood function.

3.5 Estimating Functions

Given n independent observations $y_1, y_2, \ldots, y_N$ each having the same pdf $f(y; \boldsymbol{\theta})$ the estimation problem of statistics is to use $\mathbf{y}$ to estimate $\boldsymbol{\theta}$. Three popular methods of estimation are:

(1) Method of Moments. Here we calculate

$$\mu_r^{'}(\boldsymbol{\theta}) = E(Y^r) \quad r = 1, 2, \ldots$$

and

$$\widehat{\mu}_r^{'} = \frac{1}{N} \sum_{i=1}^{N} y_i^r$$

We then solve the equations

$$\mu_r^{'}(\boldsymbol{\theta}) = \widehat{\mu}_r^{'}$$

to obtain the estimate of $\boldsymbol{\theta}$.

(2) Least Squares. Here we choose $\boldsymbol{\theta}$ to minimize

$$\sum_{i=1}^{N}[y_i - \mu_i(\boldsymbol{\theta})]^2$$

which leads to solving equations of the form

$$\sum_{i=1}^{N}[y_i - \mu_i(\boldsymbol{\theta})]\frac{\partial \mu_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

for $\boldsymbol{\theta}$. We know that if $\mu_i(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\mathbf{x}_i$ i.e. $\mu_i(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$ then least squares produces minimum variance estimators among the class of all linear unbiased estimators.

(3) Maximum Likelihood. Here we choose $\boldsymbol{\theta}$ to maximize

$$\sum_{i=1}^{N}\ln[f(y_i;\boldsymbol{\theta})]$$

which leads to equations of the form

$$\sum_{i=1}^{N}\frac{\partial \ln[f(y_i;\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} = 0$$

to be solved for $\boldsymbol{\theta}$. These equations are called the **score equations**.

All three methods lead to solving equations of the form

$$\sum_{i=1}^{N} g(y_i; \boldsymbol{\theta}) = 0$$

which are called estimating equations and the function

$$g(\mathbf{y}; \boldsymbol{\theta}) = \sum_{i=1}^{N} g(y_i; \boldsymbol{\theta})$$

is called an estimating function. The simplest estimating function is of the form

$$\mathbf{t}(\mathbf{y}) - \boldsymbol{\theta}$$

which if $\mathbf{t}(\mathbf{Y})$ is unbiased for $\boldsymbol{\theta}$ leads to an unbiased estimator for $\boldsymbol{\theta}$.

We define an estimating function $g(\mathbf{y};\boldsymbol{\theta})$ to be unbiased if

$$E[g(\mathbf{y};\boldsymbol{\theta})] = 0$$

Note that, under relatively simple regularity conditions, each of the three methods of estimation described above are unbiased estimating functions. Note also that there is no guarantee that an unbiased estimating equation will yield an unbiased estimator, only that it will lead to an estimator. Properties of the estimator obtained from solving an estimating equation must be investigated.

Among the class of unbiased estimating functions Godambe defined an optimal unbiased estimating function to be an unbiased estimating function which minimizes

$$S(\boldsymbol{\theta}) = E\left[\left(\frac{g(y;\boldsymbol{\theta})}{E\left[\frac{\partial\, g(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]}\right)^2\right]$$

Godambe proved that for all unbiased estimating functions.

$$S(\boldsymbol{\theta})^{-1} \leq I(\boldsymbol{\theta})$$

where $I(\boldsymbol{\theta})$ is Fisher's Information.

If we use the estimating function

$$g(y;\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\partial \ln[f(y_i;\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}}$$

i.e. the score fuction for maximum likelihood, then we know that this fuction is an unbiased estimating function and that

$$S(\boldsymbol{\theta}) = -\left\{ E\left[\sum_{i=1}^{N} \frac{\partial^2 \ln[f(y_i;\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^2} \right] \right\}^{-1}$$

which is the inverse of Fisher's Information. It follows that the score function is an optimal unbiased estimating function. In the presence of nuisance parameters Godambe also established that the conditional score function is the optimal estimating function. Note that, as defined, optimal unbiased estimating functions require knowledge of the distribution of $\mathbf{y}$.

Wedderburn in 1974 proposed estimation of $\boldsymbol{\theta}$ by solving

$$\sum_{i=1}^{N} \frac{\partial \mu_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \, v_i^{-1}[y_i - \mu_i(\boldsymbol{\theta})] = \mathbf{0}$$

for the estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. He called his equations the quasi-score equations because of their similarity to the score equations of maximum likelihood.The y_i are assumed to be independent with

$$E(Y_i) = \mu_i(\boldsymbol{\theta}) \;\text{ and }\; \text{var}\,(Y_i) = v_i = \frac{w(\mu_i)}{\phi}$$

where ϕ is a nuisance scale parameter. If the distribution of the Y_i is one of the family of linear exponential distributions then the above eqautions are, in fact, the score equations for $\boldsymbol{\theta}$. They are thus referred to as linear estimating equations and the function as a linear estimating function. Godambe established that among all unbiased estimating functions which are linear in the responses i.e. of the form

$$\sum_{i=1}^{N} a_i(\boldsymbol{\theta})[y_i - \mu_i(\boldsymbol{\theta})]$$

Wedderburn's function was optimal. McCullagh showed that Wedderburn's estimator was asymptotically unbiased and asymptotically normal with minimum variance among all estimators obtained from solving estimating equations linear in the data.

In a fundamental paper Liang and Zeger extended the quasi-score function to deal with longitudinal data. They assumed N vectors of responses $\mathbf{z}_i$ where

$$\mathbf{z}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix} \quad ; \quad E(\mathbf{z}_i) = \boldsymbol{\mu}_i(\boldsymbol{\theta}) = \begin{bmatrix} \mu_{i1}(\boldsymbol{\theta}) \\ \mu_{i2}(\boldsymbol{\theta}) \\ \vdots \\ \mu_{in_i}(\boldsymbol{\theta}) \end{bmatrix} \quad ; \quad \text{var}\,(\mathbf{z}_i) = \mathbf{V}_i[\boldsymbol{\theta}, \boldsymbol{\alpha}, \phi]$$

The matrix $\mathbf{V}_i[\boldsymbol{\theta}, \boldsymbol{\alpha}, \phi]$ can be different for each observation but the parameter $\boldsymbol{\alpha}$ is assumed to be the same for each vector of observations. ϕ is a nuisance scale parameter. Their equations, called generalized estimating equations (GEE) are given by

$$\sum_{i=1}^{N} \left[\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]^T \{\mathbf{V}_i[\boldsymbol{\theta}, \boldsymbol{\alpha}), \phi]\}^{-1} [\mathbf{z}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})] = \mathbf{0}$$

Under weak regularity conditions they established that the estimator for $\boldsymbol{\theta}$ obtained by solving the GEE equations were asymptotically normal regardless of the form of the covariance matrix $\mathbf{V}_i$ used in solving the equations. Much work followed to find GEE which utilized information from the variances. This lead to quadratic unbiased estimating functions, etc. Work still proceeds.

The important fact is that all of these methods were developed to extend linear models to distributions which are not normal, where the variance is not constant and the observations are not independent.

3.6 Orthogonality in the Design Matrix

The matrix $\mathbf{X}$ in the general linear model is often called the **design matrix**.

- Suppose that the design matrix $\mathbf{X}$ can be partitioned into components

$$\mathbf{X} = [\mathbf{1} \;\; \mathbf{X}_1 \;\; \mathbf{X}_2]$$

where $\mathbf{1}$ is an $n \times 1$ vector each element equal to 1, $\mathbf{X}_1$ and $\mathbf{X}_2$ are $n \times q$ and $n \times (p-q)$ matrices respectively such that

$$[\mathbf{Sp}(\mathbf{D}_1\mathbf{X}_1)] \perp \mathbf{Sp}(\mathbf{D}_1\mathbf{X}_2) \text{ where } \mathbf{D}_1 = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$$

(i.e. $\mathbf{X}_1^T\mathbf{D}_1\mathbf{D}_1\mathbf{X}_2 = \mathbf{0}$) In this case some special results about least squares estimates of $\mathbf{X}\boldsymbol{\beta}$ are true.

- Let $\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ be $1 \times 1, p \times 1$ and $(p-q) \times 1$ vectors corresponding to a partition of the parameter $\boldsymbol{\beta}$. Then we may write the linear model as

$$\mathbf{y} \sim \mathrm{WS}(\beta_0\mathbf{1} + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \sigma^2\mathbf{I})$$

- To obtain the LSE's of $\boldsymbol{\beta}$, we require the projection matrix

$$\mathbf{P_X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$$

  - Writing $\mathbf{X} = [\mathbf{1}\ \mathbf{Z}]$ where $\mathbf{Z} = [\mathbf{X}_1\ \mathbf{X}_2]$ we see that a generalized inverse of $\mathbf{X}^T\mathbf{X}$ is

$$(\mathbf{X}^T\mathbf{X})^- = \begin{bmatrix} \frac{1}{n} + \frac{1}{n^2}\mathbf{1}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{D}_1\mathbf{Z})^-\mathbf{Z}^T\mathbf{1} & -\frac{1}{n}\mathbf{1}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{D}_1\mathbf{Z})^- \\ -\frac{1}{n}(\mathbf{Z}^T\mathbf{D}_1\mathbf{Z})^-\mathbf{Z}^T\mathbf{1} & (\mathbf{Z}^T\mathbf{D}_1\mathbf{Z})^- \end{bmatrix}$$

  - It follows that

$$\mathbf{P_X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T = \frac{1}{n}\mathbf{1}\mathbf{1}^T + \mathbf{D}_1\mathbf{Z}(\mathbf{Z}^T\mathbf{D}_1\mathbf{Z})^-\mathbf{Z}^T\mathbf{D}_1$$

  - Since

$$\mathbf{Z}^T\mathbf{D}_1\mathbf{Z} = \begin{bmatrix} \mathbf{X}_1^T \\ \mathbf{X}_2^T \end{bmatrix} \mathbf{D}_1 [\mathbf{X}_1\ \mathbf{X}_2] = \begin{bmatrix} \mathbf{X}_1^T\mathbf{D}_1\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2^T\mathbf{D}_1\mathbf{X}_2 \end{bmatrix}$$

  - $\mathbf{P_X}$ may be written as

$$\mathbf{P_X} = \mathbf{P_1} + \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1} + \mathbf{P}_{\mathbf{D}_1\mathbf{X}_2}$$

– This simple form is useful in several ways.

Suppose we are interested in whether $\boldsymbol{\beta}_2$ equals zero or not.

- Evidence that $\boldsymbol{\beta}_2 = \mathbf{0}$ might be whether the model $E[\mathbf{Y}] = \beta_0\mathbf{1} + \mathbf{X}_1\boldsymbol{\beta}_1$ provided about as good a fit to the data as the model

$$E[\mathbf{Y}] = \beta_0\mathbf{1} + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$$

- Since the length of the residual vector from any model fit is literally how close the fitted values from the model approximate the observed data, a comparison of the length (squared) of the residual vectors from the two models would yield evidence of whether $\boldsymbol{\beta}_2 = \mathbf{0}$ or not.

- Now

$$\begin{aligned} \|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2 &= \mathbf{Y}^T(\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1} - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_2})\mathbf{Y} \\ &= \mathbf{Y}^T\mathbf{D}_1\mathbf{Y} - \mathbf{Y}^T\mathbf{P}_{\mathbf{D}_1\mathbf{X}_1}\mathbf{Y} - \mathbf{Y}^T\mathbf{P}_{\mathbf{D}_1\mathbf{X}_2}\mathbf{Y} \\ &= \|(\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1})\mathbf{Y}\|^2 - \|\mathbf{P}_{\mathbf{D}_1\mathbf{X}_2}\mathbf{Y}\|^2 \end{aligned}$$

- Thus the difference in squared lengths of the residual vectors from the "full" model and the reduced model is simply

$$\|\mathbf{P}_{\mathbf{D}_1\mathbf{X}_2}\mathbf{Y}\|^2 = \|\mathbf{D}_1\mathbf{X}_2\mathbf{b}_2\|^2$$

- This quantity may be interpreted as the squared distance of the least squares estimate of $E(\mathbf{Y})$ (under the full model) from the subspace $\mathbf{Sp}(\mathbf{D}_1\mathbf{X}_1)$.

  - Note that the subspace $\mathbf{Sp}(\mathbf{D}_1\mathbf{X}_1)$ is simply the estimation space under the reduced model.

A very useful extension of this idea is when the design matrix can be partitioned into K components $\mathbf{X}_1, \ldots, \mathbf{X}_K$, all pairwise orthogonal.

- Then as above, we have the representation of the projection matrix

$$\mathbf{P}_{\mathbf{D}_1\mathbf{X}} = \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1} + \cdots + \mathbf{P}_{\mathbf{D}_1\mathbf{X}_K}$$

  with a corresponding partition of the squared length of $\mathbf{P}_{\mathbf{X}}\mathbf{Y} = \mathbf{X}\mathbf{b}$ as

$$\|\mathbf{X}\mathbf{b}\|^2 = \|\mathbf{1}b_0\|^2 + \|\mathbf{D}_1\mathbf{X}_1\mathbf{b}_1\|^2 + \|\mathbf{D}_1\mathbf{X}_2\mathbf{b}_2\|^2 + + \cdots + \|\mathbf{D}_1\mathbf{X}_K\mathbf{b}_K\|^2$$

- As we shall see later, tests of hypotheses about the corresponding $\boldsymbol{\beta}_i$'s may be performed independently as a consequence of this orthogonality.

- A result of orthogonality that is complementary to the above partitioning of $\|\mathbf{X}\mathbf{b}\|^2$, pertains to LSE's of estimable functions $\mathbf{L}^T\boldsymbol{\beta}$ which depend on $\boldsymbol{\beta}$ only through $\boldsymbol{\beta}_1$.

- Suppose $\boldsymbol{\beta}_1$ is the parameter of primary interest. Then noting that estimable functions of $\mathbf{L}^T\boldsymbol{\beta}$ have the representation $\mathbf{C}^T\mathbf{X}_1\boldsymbol{\beta}$, we see that the LSE for $\mathbf{L}^T\boldsymbol{\beta}_1$ from the model $E[\mathbf{Y}] = \beta_0\mathbf{1} + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$ is the same as the LSE for $\mathbf{L}^T\boldsymbol{\beta}_1$ assuming the (wrong) model

$$E[\mathbf{Y}] = \beta_0\mathbf{1} + \mathbf{X}_1\boldsymbol{\beta}_1$$

  (i.e., they are both equal to $\mathbf{P}_{\mathbf{D}_1\mathbf{X}_1}\mathbf{Y}$).

- Therefore, because of the orthogonality, we might be able to ignore the effects of $\mathbf{X}_2$ and still make the same inferences about $\boldsymbol{\beta}_1$ that we would if we carried $\mathbf{X}_2$ along is the data analysis.

Unfortunately, although the component of interest in $\mathbf{P}_\mathbf{X}\mathbf{Y}$ is the same as $\mathbf{P}_{\mathbf{D}_1\mathbf{X}_1}\mathbf{Y}$, the residual vectors $(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}$ and $(\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1})\mathbf{Y}$ are not.

- Thus, if estimation of σ^2 were based on $(\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1})\mathbf{Y}$, we would not obtain an unbiased estimate of σ^2.

- In fact

$$E(\mathbf{Y}^T(\mathbf{D} - \mathbf{P}_{\mathbf{D}\mathbf{X}_1})\mathbf{Y}) = (n - \text{rank}\,(\mathbf{X}_1))\sigma^2 + \|\mathbf{D}\mathbf{X}_2\boldsymbol{\beta}_2\|^2$$

- Therefore, simply ignoring $\mathbf{X}_2$ and behaving as if the true model were

$$\mathbf{y} \sim \text{WS}(\beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta}_1, \sigma^2\mathbf{I})$$

  will not result in the proper inference about $\boldsymbol{\beta}_1$

- Write the residual vector as

$$\begin{aligned}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y} &= (\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1} - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_2})\mathbf{Y} \\ &= (\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1})(\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_2})\mathbf{Y} \\ &= (\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1})\widetilde{\mathbf{Y}}\end{aligned}$$

  where $\widetilde{\mathbf{Y}} = (\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_2})\mathbf{Y}$

- Then we note that after the transformation to $\widetilde{\mathbf{Y}}$, the projection of $\widetilde{\mathbf{Y}}$ onto the orthogonal complement of $\mathbf{Sp}(\mathbf{D}_1\mathbf{X}_1)$ is the same as the residual vector from the least squares fit of the full model.
    - Also, the projection of $\widetilde{\mathbf{Y}}$ onto $\mathbf{Sp}(\mathbf{D}_1\mathbf{X}_1)$ corresponds to the LSE of $\mathbf{X}_1\mathbf{b}_1$ from the full model since

$$\begin{aligned}\mathbf{P}_{\mathbf{D}_1\mathbf{X}_1}\widetilde{\mathbf{Y}} &= \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1}[\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_2}]\mathbf{Y} \\ &= \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1}\mathbf{Y} - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1}\mathbf{P}_{\mathbf{D}_1\mathbf{X}_2}\mathbf{Y} \\ &= \mathbf{P}_{\mathbf{D}_1\mathbf{X}_1}\mathbf{Y}\end{aligned}$$

- Therefore, if we first project $\mathbf{Y}$ onto $\mathbf{Sp}^{\perp}(\mathbf{X}_2)$ and then, taking this vector as our response vector, behave as if

$$\tilde{\mathbf{y}} \sim \mathrm{WS}(\beta_0\mathbf{1} + \mathbf{X}_1\boldsymbol{\beta}_1, \sigma^2\mathbf{I})$$

  we will be making exactly the same inferences about $\boldsymbol{\beta}_1$ as if we performed the analysis on the more complicated model:

$$\tilde{\mathbf{y}} \sim \mathrm{WS}(\beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta}_1 + \mathbf{X}\boldsymbol{\beta}_2, \sigma^2\mathbf{I})$$

- The only adjustment required in the analysis using the reduced model is in total degress of freedom.
    - By making $\widetilde{\mathbf{Y}}$ orthogonal to $\mathbf{1}$ $\mathbf{X}_1$, we have used $\mathrm{rank}(\mathbf{X}_2)$ degrees of freedom.
    - Therefore the total number of degrees of freedom in the reduced analysis is $n - 1 - \mathrm{rank}\,(\mathbf{X}_2)$ instead of n.