Theory of Linear Models

Steven Gilmour
King's College London

January – February 2020

Least Squares Estimation

Any solution $\hat{\boldsymbol{\beta}}$ of the normal equations

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$$

minimises $S = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$.

For any $p \times 1$ vector $\mathbf{c}$, $\mathbf{c}'\hat{\boldsymbol{\beta}}$ is defined to be a least squares estimator of $\mathbf{c}'\boldsymbol{\beta}$

Properties of Least Squares Estimators

Theorem

The following conditions are equivalent:

- $\mathbf{c}'\hat{\boldsymbol{\beta}}$ *is linear in* $\mathbf{Y}$ *and an unbiased estimator of* $\mathbf{c}'\boldsymbol{\beta}$;
- $\mathbf{c}'\hat{\boldsymbol{\beta}}$ *is unique for all solutions of the normal equations;*
- $\mathbf{c}$ *is in the vector space defined by the columns of* $\mathbf{X}'\mathbf{X}$;
- $\mathbf{c}$ *is in the vector space defined by the columns of* $\mathbf{X}'$;
- *There exists a linear function of* $\mathbf{Y}$ *with expectation* $\mathbf{c}'\boldsymbol{\beta}$.

If any of the conditions of Theorem 1 hold, we say that $\mathbf{c}'\boldsymbol{\beta}$ is estimable.

Example (simple linear regression): This theorem implies that if there are unique least squares estimators of the parameters β_0 and β_1, these estimators will be unbiased and linear in $\mathbf{Y}$.

In the case where $x_i = x \; \forall i = 1, \ldots, n$, $\beta_0 + \beta_1 x$ is estimable but, for example, β_1 is not estimable.

Theorem (Gauss-Markov Theorem)

If $\mathbf{c}'\boldsymbol{\beta}$ *is estimable, then* $\mathbf{c}'\hat{\boldsymbol{\beta}}$ *has minimum variance in the class of unbiased estimators which are linear in* $\mathbf{Y}$*.*

We call $\mathbf{c}'\hat{\boldsymbol{\beta}}$ the best linear unbiased estimator (BLUE) of $\mathbf{c}'\boldsymbol{\beta}$.

Example (simple linear regression): The estimator

$$\tilde{\beta}_1 = (Y_2 - Y_1)/(x_2 - x_1)$$

is an unbiased estimator of β_1 and is linear in $\mathbf{Y}$.

The Gauss-Markov Theorem shows that this estimator has higher variance than the least squares estimator (if $n > 2$).

Note that it is easy to find biased estimators with smaller variance than least squares estimators, e.g. $\tilde{\beta}_1 = 0$ has variance zero.

There are other (more or less) sensible methods of estimation, e.g. L_1-norm, ridge regression methods, maximum likelihood, empirical Bayes, subjective Bayes. They all produce biased estimators (except in a few special cases in which they are equivalent to least squares).

Best linear unbiased estimation is a very traditional interpretation of what it means for an estimator to be "good". Nevertheless, it is difficult to argue against it on general grounds.

The particular emphasis on unbiasedness suggests that it is particularly suited to situations in which multiple studies will be carried out.

Theorem

All linear functions of $\boldsymbol{\beta}$ are estimable if and only if rank$(\mathbf{X}) = p$.

If rank$(\mathbf{X}) = p$, i.e. $\mathbf{X}$ is of full rank, then

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

is the unique solution of the normal equations.

If rank$(\mathbf{X}) < p$, then unbiased estimators do not exist for some functions of the parameters. They are said to be inestimable or confounded.

Example (simple linear regression): $\text{rank}(\mathbf{X}) = 2$, unless $x_i = x;\ \forall i = 1, \ldots, n$, so this is the only case in which inestimable functions exist.

Otherwise

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

and

$$\hat{\boldsymbol{\beta}} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 \sum Y_i - \sum x_i \sum x_i Y_i \\ n\sum x_i Y_i - \sum x_i \sum Y_i \end{bmatrix}.$$

Example (one-way analysis of variance): Consider the model for t treatments,

$$\mu_{ri} = \beta_0 + \tau_r,$$

where observation i has treatment r.

If we write the model in full with $t+1$ parameters $\boldsymbol{\beta}' = [\beta_0 \ \tau_1 \ \cdots \ \tau_t]$, we have

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

and it is immediately obvious that $\text{rank}(\mathbf{X}) = t < t + 1$.

Hence there some functions of the parameters which are inestimable.

If we use only t parameters, $\boldsymbol{\beta}' = [\beta_0 \ \tau_2 \ \cdots \ \tau_t]$, then $\text{rank}(\mathbf{X}) = t$ and all linear functions of the parameters are estimable.

In general how can we determine which functions are estimable and which are not?

Generalized Inverse

Definition

A generalized inverse of an $m \times n$ matrix $\mathbf{A}$ is any matrix $\mathbf{A}^-$ such that $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$.

There is not, in general, a unique generalized inverse, but if $\mathbf{A}$ is a nonsingular square matrix, then $\mathbf{A}^{-1}$ is the unique generalized inverse.

If $\mathbf{X}'\mathbf{X}$ is singular, we can find a particular solution of the normal equations by finding a generalized inverse of $\mathbf{X}'\mathbf{X}$. Different generalized inverses can lead to different solutions.

One way to find a generalized inverse of $\mathbf{X}'\mathbf{X}$ is to append some rows to $\mathbf{X}$.

If $\text{rank}(\mathbf{X}) = q < p$ and $\mathbf{X}^*$ is a $(p-q) \times p$ matrix such that

$$\text{rank}\left(\begin{bmatrix} \mathbf{X} \\ \mathbf{X}^* \end{bmatrix}\right) = p,$$

then $(\mathbf{X}'\mathbf{X} + \mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$.

This can be thought of as using pseudo-data to allow estimation.

Example (simple linear regression): If $x_i = x;\ i = 1, \ldots, n$, then we can append the row

$$\mathbf{X}^* = [1\ x^*],$$

where $x^* \neq x$.

Then

$$\mathbf{X}'\mathbf{X} + \mathbf{X}^{*\prime}\mathbf{X}^* = \begin{bmatrix} n+1 & nx + x^* \\ nx + x^* & nx^2 + x^{*2} \end{bmatrix}$$

and

$$(\mathbf{X}'\mathbf{X})^- = \frac{1}{(n+1)(nx^2 + x^{*2}) - (nx + x^*)^2} \begin{bmatrix} nx^2 + x^{*2} & -(nx + x^*) \\ (-nx + x^*) & n+1 \end{bmatrix}.$$

Example (one-way analysis of variance): To fit the model

$$\mu_{ri} = \beta_0 + \tau_r,$$

we need one additional row and we can use $\mathbf{X}^* = [1\ 1\ \cdots\ 1]$.

Consider the case of two treatments, each replicated twice. Then

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix},$$

and

$$\mathbf{X}^{*\prime}\mathbf{X}^* = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

so that

$$\mathbf{X}'\mathbf{X} + \mathbf{X}^{*\prime}\mathbf{X}^* = \begin{bmatrix} 5 & 3 & 3 \\ 3 & 3 & 1 \\ 3 & 1 & 3 \end{bmatrix}.$$

Hence, a generalized inverse is

$$(\mathbf{X}'\mathbf{X})^{-} = \begin{bmatrix} 2 & -3/2 & -3/2 \\ -3/2 & 3/2 & 1 \\ -3/2 & 1 & 3/2 \end{bmatrix}.$$

Since

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} Y_1 + Y_2 + Y_3 + Y_4 \\ Y_1 + Y_2 \\ Y_3 + Y_4 \end{bmatrix},$$

a least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} (Y_1 + Y_2 + Y_3 + Y_4)/2 \\ -(Y_3 + Y_4)/2 \\ -(Y_1 + Y_2)/2 \end{bmatrix}.$$

Theorem

$\mathbf{c}'\boldsymbol{\beta}$ is estimable if and only if $\mathbf{c}'\{\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X})\} = \mathbf{0}$.

This allows us to find estimable functions of parameters in cases where the normal equations do not have a unique solution.

Example (one-way analysis of variance): Consider first estimating τ_1. i.e. $\mathbf{c}' = [0\ 1\ 0]$.

$$
\begin{aligned}
(\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X}) &= \begin{bmatrix} 2 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix} \\
\Rightarrow \mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X}) &= \begin{bmatrix} -1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\
\Rightarrow \mathbf{c}'\{\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X})\} &= [1\ 1\ 1]
\end{aligned}
$$

and so τ_1 is not estimable.

Consider instead estimating $\tau_2 - \tau_1$, i.e. $\mathbf{c}' = [0 \ -1 \ 1]$.

In this case

$$\mathbf{c}'\{\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X})\} = [0 \ 0 \ 0]$$

and so $\tau_2 - \tau_1$ is estimable.

The estimator is

$$\mathbf{c}'\hat{\boldsymbol{\beta}} = (Y_3 + Y_4)/2 - (Y_1 + Y_2)/2.$$

This is why overparameterised models, such as this one, can be used directly. It is possible, but not necessary, to impose some constraints on the parameters, e.g. $\tau_1 = 0$. This example shows that whatever constraints we impose, we will always get the same estimator for $\tau_2 - \tau_1$.

Variance and Covariance of Estimators

Theorem 1 told us that if $\mathbf{c}'\beta$ is estimable, then $\mathbf{c}$ is in the vector space defined by $\mathbf{X}'\mathbf{X} \Rightarrow \mathbf{c} = \mathbf{X}'\mathbf{X}\gamma$, for some vector γ.

Hence,

$$\begin{aligned} V(\mathbf{c}'\hat{\beta}) &= V\{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\} \\ &= \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'V(\mathbf{Y})\mathbf{X}\{(\mathbf{X}'\mathbf{X})^{-}\}'\mathbf{c} \\ &= \sigma^2\gamma'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{c} \\ &= \sigma^2\gamma'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{c} \\ &= \sigma^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}. \end{aligned}$$

Similarly for a $p \times 1$ vector $\mathbf{d}$,

$$Cov(\mathbf{c}'\hat{\beta}, \mathbf{d}'\hat{\beta}) = \sigma^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{d}.$$

Estimating σ^2

In order to do interval estimation or hypothesis testing, we need to be able to estimate σ^2.

We will use the following results.

Theorem

$$\mathbf{X}' = \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'.$$

Theorem

$$\mathbf{Y}'\{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\}\mathbf{Y} = (\mathbf{Y} - \mathbf{X}\beta)'\{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\}(\mathbf{Y} - \mathbf{X}\beta).$$

Theorem

Let $\mathbf{Z}$ *be an* $n \times 1$ *random vector and* $\mathbf{A}$ *an* $n \times n$ *constant matrix. Then*

$$E\left[\{\mathbf{Z} - E(\mathbf{Z})\}'\mathbf{A}\{\mathbf{Z} - E(\mathbf{Z})\}\right]$$

$$= \sum_{i=1}^{n} a_{ii} V(Z_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_{ij} Cov(Z_i, Z_j).$$

Using the results above, we find

Theorem

$$E\left\{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right\} = (n-r)\sigma^2,$$

where $r = rank(\mathbf{X})$.

Hence $S^2 = SS_R/(n-r)$ is an unbiased estimator of σ^2, where $SS_R = (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})$ is the residual sum of squares.

Note that this result applies whether or not the design matrix is of full rank.

It can also be shown that S^2 is a consistent estimator of σ^2.

Results on the optimality of S^2 as an estimator of σ^2 are not available in general, but are under certain assumptions, e.g. normality.

Minimum Variance Unbiased Estimation

If we adopt a fully parametric approach, then we can often find stronger results.

Theorem

If $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ *then, for any estimable function* $\mathbf{c}'\boldsymbol{\beta}$, $\mathbf{c}'\hat{\boldsymbol{\beta}}$ *is the unique minimum variance unbiased estimator of* $\mathbf{c}'\boldsymbol{\beta}$.

Note that we are no longer restricted to *linear* estimators, i.e. under normality, there can be no nonlinear unbiased estimator better than the least squares estimator.

Theorem

If each ϵ_i *has the same distribution, which has all moments finite, and* $\mathbf{c}'\hat{\boldsymbol{\beta}}$ *is the minimum variance unbiased estimator of* $\mathbf{c}'\boldsymbol{\beta}$, *then* ϵ_i *has a normal distribution.*

Notes:

- Theorem 11 says that the normal distribution is the *only* (nice) distribution for which the least squares estimators are minimum variance unbiased estimators.
- Theorems 10 and 11 together suggest that the additional assumption of normality is worth a great deal, in the sense that without this assumption, we know that we are using inferior estimators.
- However, for many other distributions, the minimum variance linear unbiased estimator will be almost as good as the minimum variance unbiased estimator.
- Overall, least squares estimation can be seen to be quite robust to distributional assumptions.

Theorem

If $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ *then* S^2 *is the minimum variance unbiased estimator of* σ^2.

Maximum Likelihood Estimation

If $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ then the likelihood is

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{Y}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right\}.$$

It is immediately obvious that the likelihood is maximised by $\hat{\boldsymbol{\beta}}$.

It is also easy to show that the MLE of σ^2 is

$$\hat{\sigma}^2 = SS_R/n.$$

Note that

$$\hat{\sigma}^2 = \frac{n-r}{n}S^2$$

and is a biased estimator of σ^2.

For other distributions, $\hat{\boldsymbol{\beta}}$ is not, in general, the MLE of $\boldsymbol{\beta}$.

Minimum Mean Square Error Estimation

It is easy to show that, among estimators of σ^2 of the form SS_R/m, that which minimises the mean square error is

$$\tilde{\sigma}^2 = \frac{SS_R}{n - r + 2}.$$

If rank$(\mathbf{X}) = p$ and $p \geq 3$, it can be shown that the James-Stein shrinkage estimator,

$$\tilde{\boldsymbol{\beta}} = \left\{ 1 - \frac{(p-2)\tilde{\sigma}^2}{\hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}}} \right\} \hat{\boldsymbol{\beta}},$$

of $\boldsymbol{\beta}$ has uniformly lower $tr\{\mathbf{X}'\mathbf{X}MSE(\tilde{\boldsymbol{\beta}})\}$ than $\hat{\boldsymbol{\beta}}$.

This might seem like a convincing argument against least squares estimators. However, note that $\hat{\beta}_r$ can have lower MSE than $\tilde{\beta}_r$

L_1-Norm Regression

If the ϵ_i are assumed to be independent and to have double exponential distributions, then the maximum likelihood estimator is equivalent to the L_1-norm or least absolute deviations (LAD) regression estimator, which minimises $\sum_{i=1}^{n} |\epsilon_i|$.

L_1-norm estimation is more robust to outliers than least squares (L_2-norm), in the sense that an extreme observation is much more likely from the double exponential than from the normal distribution.

We can also use L_d-norm regression for other values of d, e.g. $d = 1.5$, or we can find MLEs for other distributions for ϵ_i.

M-Estimators

An M-estimator of $\boldsymbol{\beta}$ is obtained by defining a function $\rho(u)$ and minimising

$$\sum_{i=1}^{n} \rho\left(\frac{\epsilon_i}{S}\right),$$

where S is some estimator of σ.

$\rho(u) = u^2$ is least squares and $\rho(u) = |u|$ is L_1-norm estimation.

Various other recommendations include Huber's function,

$$\rho(u) = \begin{cases} \frac{u^2}{2} & -a \leq u \leq a; \\ a|u| - \frac{a^2}{2} & \text{otherwise,} \end{cases}$$

and Tukey's biweight function,

$$\rho(u) = \begin{cases} \frac{u^2}{2} - \frac{u^4}{4a^2} & -a \leq u \leq a; \\ \frac{a^2}{4} & \text{otherwise.} \end{cases}$$

These lead to weighted least squares estimators, but with the weights depending on $\boldsymbol{\beta}$. Observations with large standardised residuals are downweighted.

It is also possible to choose a weight function directly, without defining $\rho(u)$.

Ranked Residuals Regression

One example of this, using the Wilcoxon score function, estimates $\boldsymbol{\beta}$ by minimising

$$\sum_{i=1}^{n} \left(\frac{i}{n+1} - \frac{1}{2} \right) \epsilon_{[i]},$$

where $\epsilon_{[i]}$ are the ranked deviations.

This method gives most weight to the large residuals.

Other score functions have been suggested.

Least Median Squares

The least median squares estimator of $\boldsymbol{\beta}$ minimises Median(ϵ_i^2).

This involves finding the narrowest strip which covers half of the observations, the estimate being in the middle of this strip.

PCR and PLS

Principal component regression (PCR) involves replacing the columns of $\mathbf{X}$ (usually centred and scaled) with their principal components and using the latter as the explanatory variables.

Usually, we can use a relatively small number of principal components to explain most of the variation in $\mathbf{y}$.

Partial least squares or projection to latent structures (PLS) regression represents a compromise between PCR and OLS.

In fact, it is possible to define continuum regression which represents a smooth transition between PCR and OLS.

Opinion: PCR is entirely meaningless except, perhaps, in calibration situations, i.e. when the "responses" were fixed and the "explanatory variables" were observed and we want to do the reverse regression.

Ridge Regression

If $\text{rank}(\mathbf{X}) < p$ or $\text{rank}(\mathbf{X}) = p$ but (numerically) there are near linear dependencies among the columns of $\mathbf{X}$, the method of ridge regression has been recommended.

The ridge regression estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_R = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}.$$

If the response data and columns of $\mathbf{X}$ are centred, then the ridge regression estimator can also be written as the least squares estimator with response vector $[\mathbf{Y}'\ \mathbf{0}']'$ and design matrix $[\mathbf{X}'\ \sqrt{\lambda}\mathbf{I}]'$.

Ridge regression appears to have some similarity to the use of a generalized inverse, but the philosophy is different:

- A generalized inverse allows us to estimate the estimable functions, when not all functions are estimable.
- Ridge regression allows us to estimate inestimable functions.

Ridge regression estimators have

$$Bias(\hat{\boldsymbol{\beta}}_R) = \lambda(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}$$

and

$$V(\hat{\boldsymbol{\beta}}_R) = \sigma^2(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}$$

and so we gain lower variances by increasing the bias.

Ridge regression is controversial, but is particularly useful when there is strong prior knowledge that the elements of $\boldsymbol{\beta}$ should not be "too large".

It can be shown that there exists a $\lambda > 0$ for which the total mean square error of $\hat{\beta}_R$ is smaller than that of $\hat{\beta}$, but this value depends on β and so cannot be found.

In practice λ is usually chosen (graphically) by comparing $\hat{\beta}_R$ for various λ and choosing the smallest value for which the estimates are stable.

Bayesian Estimation

The natural conjugate prior is inverse gamma for σ^2 and conditionally multivariate normal for $\boldsymbol{\beta}|\sigma^2$. Let the prior expectation and variance be $\boldsymbol{\beta}_P$ and $\sigma^2\mathbf{A}$ respectively.

This leads to a conditionally multivariate normal posterior for $\boldsymbol{\beta}|\sigma^2, \mathbf{Y}$ with variance matrix

$$\sigma^2(\mathbf{A}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$$

and expectation

$$\sigma^2(\mathbf{A}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\left(\frac{1}{\sigma^2}\mathbf{A}^{-1}\boldsymbol{\mu}_P + \mathbf{X}'\mathbf{Y}\right).$$

In the noninformative prior case, where $\mathbf{A} = c\mathbf{I}$ with $c \to \infty$, this reduces to least squares.

Of course, the Bayesian approach is very flexible and can meet all the objectives of the methods described above, for example:

- Appropriate distributional assumptions can be made to allow for possible outliers, e.g. Y_i can be assumed to have a t distribution, or a double exponential distribution.
- Estimation is always possible, no matter what the rank of $\mathbf{X}$ is.
- Informative priors centred at 0 can be used to express the prior belief that the elements of $\boldsymbol{\beta}$ will not be too large.
- If it is expected that many coefficients will be close to zero, priors can be used which are mixtures of normal distributions, or Bayesian variable selection methods can be used.

Some Personal Opinions of Estimation Methods

Most of the methods described here have their supporters and many of them are widely used in practice, so many statisticians will disagree with the following.

The only estimation methods I would *ever* recommend are least squares and Bayesian methods using informative subjective priors.

Least squares estimation is fully understood, is quick and easy to implement and has good frequentist properties. However, one must recognise the limitations of frequentist-like conclusions.

If one wants more detail than least squares can provide, e.g. for formal decision making, then I believe there is no sensible alternative to Bayesian methods. However, I believe these must be used properly, with great care taken in specifying the model, priors chosen very carefully to reflect the subjective views of each person involved in interpreting the conclusions and a careful investigation of the posteriors.

I believe that everything else, including "automatic" Bayesian analyses, are worthless. The numbers obtained from the analysis seem to be entirely meaningless. At best, they are attempts at approximating the fully Bayesian analysis. However, it must be better to describe the fully Bayesian analysis and then choose any necessary approximations on a rational basis.