17 Shrinkage

17.1 Mallows Averaging and Shrinkage

Suppose there are two models or estimators of $g = E(y \mid X)$

(1) $g_0 = 0$

(2) $\hat{g}_1 = X\hat{\beta}$

Given weights $(1 - w)$ and w an averaging estimator is $\hat{g} = wX\hat{\beta}$.

The Mallows criterion is

$$\begin{aligned} C(w) &= \mathbf{w}'\hat{\mathbf{e}}'\hat{\mathbf{e}}\mathbf{w} + 2\hat{\sigma}^2\mathbf{w}'\mathbf{K} \\ &= \begin{pmatrix} 1-w & w \end{pmatrix} \begin{pmatrix} y'y & y'\hat{e} \\ \hat{e}'y & \hat{e}'\hat{e} \end{pmatrix} \begin{pmatrix} 1-w \\ w \end{pmatrix} + 2\hat{\sigma}^2 wk \\ &= (1-w)^2 y'y + \left(w^2 + 2w(1-w)\right)\hat{e}'\hat{e} + 2\hat{\sigma}^2 wk \\ &= (1-w)^2\left(y'y - \hat{e}'\hat{e}\right) + \hat{e}'\hat{e} + 2\hat{\sigma}^2 wk \end{aligned}$$

The FOC for minimization is

$$\frac{d}{dw}C(w) = -2(1-w)\left(y'y - \hat{e}'\hat{e}\right) + 2\hat{\sigma}^2 k = 0$$

with solution

$$\hat{w} = 1 - \frac{k}{F}$$

where

$$F = \frac{y'y - \hat{e}'\hat{e}}{\hat{\sigma}^2}$$

is the Wald statistic for $\beta = 0$. Imposing the constraint $\hat{w} \in [0, 1]$ we obtain

$$\hat{w} = \begin{cases} 1 - \dfrac{k}{F} & F \geq k \\ \\ 0 & F < k \end{cases}$$

The Mallow averaging estimator thus equals

$$\hat{\beta}^* = \hat{\beta}\left(1 - \frac{k}{F}\right)_+$$

where $(a)_+ = a$ if $a \geq 0$, 0 else.

This is a Stein-type shrinkage estimator.

17.2 Loss and Risk

A great reference is *Theory of Point Estimation*, 2nd Edition, by Lehmann and Casella.

Let $\hat{\theta}$ be an estimator for θ, $k \times 1$. Suppose $\hat{\theta}$ is an (asymptotic) sufficient statistic for θ so that any other estimator can be written as a function of $\hat{\theta}$. We call $\hat{\theta}$ the "usual"estimator.

Suppose that $\sqrt{n}\left(\hat{\theta} - \theta\right) \to_d N(0, \mathbf{V})$. Thus, approximately,

$$\hat{\theta} \sim_a N(\theta, \mathbf{V}_n)$$

where $\mathbf{V}_n = n^{-1}\mathbf{V}$. Most of Stein-type theory is developed for the exact distribution case. It carries over to the asymptotic setting as approximations. For now on we will assume that $\hat{\theta}$ has an exact normal distribution, and that $\mathbf{V}_n = \mathbf{V}$ is known. (Equivalently, we can rewrite the statistical problem as local to θ using the "Limits of Experiments" theory.

Is $\hat{\theta}$ the best estimator for θ, in the sense of minimizing the risk (expected loss)?

The risk of $\tilde{\theta}$ under weighted squared error loss is

$$\begin{aligned} R(\theta, \tilde{\theta}, \mathbf{W}) &= E\left(\left(\tilde{\theta} - \theta\right)' \mathbf{W} \left(\tilde{\theta} - \theta\right)\right) \\ &= \operatorname{tr}\left(\mathbf{W} E\left(\left(\tilde{\theta} - \theta\right)\left(\tilde{\theta} - \theta\right)\right)'\right) \end{aligned}$$

A convenient choice for the weight matrix is $\mathbf{W} = \mathbf{V}^{-1}$. Then

$$\begin{aligned} R(\theta, \hat{\theta}, \mathbf{V}^{-1}) &= \operatorname{tr}\left(\mathbf{V}^{-1} E\left(\left(\tilde{\theta} - \theta\right)\left(\tilde{\theta} - \theta\right)\right)'\right) \\ &= \operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{V}\right) \\ &= k. \end{aligned}$$

If $\mathbf{W} \neq \mathbf{V}^{-1}$ then

$$\begin{aligned} R(\theta, \hat{\theta}, \mathbf{W}) &= \operatorname{tr}\left(\mathbf{W} E\left(\left(\tilde{\theta} - \theta\right)\left(\tilde{\theta} - \theta\right)\right)'\right) \\ &= \operatorname{tr}\left(\mathbf{W}\mathbf{V}\right) \end{aligned}$$

which depends on $\mathbf{WV}$.

Again, we want to know if the risk of another feasible estimator is smaller than $\operatorname{tr}(\mathbf{WV})$.

Take the simple (or silly) estimator $\tilde{\theta} = 0$. This has risk

$$R(\theta, 0, \mathbf{W}) = \theta'\mathbf{W}\theta.$$

Thus $\tilde{\theta} = 0$ has smaller risk than $\hat{\theta}$ when $\theta'\mathbf{W}\theta < \operatorname{tr}(\mathbf{WV})$, and larger risk when $\theta'\mathbf{W}\theta > \operatorname{tr}(\mathbf{WV})$. Neither $\hat{\theta}$ nor $\tilde{\theta} = 0$ is "better" in the sense of having (uniformly) smaller risk! It is not enough to ask that one estimator has smaller risk than another, as in general the risk is a function depending

on unknowns.

As another example, take the simple averaging (or shrinkage) estimator

$$\tilde{\theta} = w\hat{\theta}$$

where w is a fixed constant. Since

$$\tilde{\theta} - \theta = w\left(\hat{\theta} - \theta\right) - (1-w)\theta$$

we can calculate that

$$\begin{aligned} R(\theta, \tilde{\theta}, \mathbf{W}) &= w^2 R(\theta, \tilde{\theta}, \mathbf{W}) + (1-w)^2 \theta' \mathbf{W} \theta \\ &= w^2 \operatorname{tr}(\mathbf{W}\mathbf{V}) + (1-w)^2 \theta' \mathbf{W} \theta \end{aligned}$$

This is minimized by setting

$$w = \frac{\theta' \mathbf{W} \theta}{\operatorname{tr}(\mathbf{W}\mathbf{V}) + \theta' \mathbf{W} \theta}$$

which is strictly in (0,1). [This is illustrative, and does not suggest an empirical rule for selecting w.]

17.3 Admissibile and Minimax Estimators

For reference.

To compare the risk functions of two estimators, we have the following concepts.

Definition 1 *$\hat{\theta}$ weakly dominates $\tilde{\theta}$ if $R(\theta, \hat{\theta}) \leq R(\theta, \tilde{\theta})$ for all θ*

Definition 2 *$\hat{\theta}$ dominates $\tilde{\theta}$ if $R(\theta, \hat{\theta}) \leq R(\theta, \tilde{\theta})$ for all θ, and $R(\theta, \hat{\theta}) < R(\theta, \tilde{\theta})$ for at least one θ.*

Clearly, we should prefer an estimator if it dominates the other.

Definition 3 *An estimator is admissible if it is not dominated by another estimator. An estimator is inadmissible if it is dominated by another estimator.*

Admissibility is a desirable property for an estimator.

If the risk functions of two estimators cross, then neither dominates the other. How do we compare these two estimators?

One approach is to calculate the worst-case scenerio. Specifically, we define the maximum risk of an estimator $\tilde{\theta}$ as

$$\bar{R}(\tilde{\theta}) = \sup_{\theta} R\left(\theta, \tilde{\theta}\right)$$

We can think: Suppose we use $\tilde{\theta}$ to estimate θ. Then what is the worst case, how bad can than this estimator do?

For example, for the usual estimator, $R\left(\theta, \hat{\theta}, \mathbf{W}\right) = \mathrm{tr}\left(\mathbf{WV}\right)$ for all θ, so

$$\bar{R}(\hat{\theta}) = \mathrm{tr}\left(\mathbf{WV}\right)$$

while for the silly estimator $\tilde{\theta} = 0$

$$\bar{R}(0) = \infty$$

The latter is an example of an estimator with unbounded risk. To guard against extreme worst cases, it seems sensible to avoid estimators with unbounded risk.

The minimium value of the maximum risk $\bar{R}(\tilde{\theta})$ across all estimators $\delta = \delta(\hat{\theta})$ is

$$\inf_{\delta} \bar{R}(\delta) = \inf_{\delta} \sup_{\theta} R(\theta, \delta)$$

where

Definition 4 *An estimator $\tilde{\theta}$ of θ which minimizes the maximum risk*

$$\inf_{\delta} \sup_{\theta} R(\theta, \delta) = \sup_{\theta} R(\theta, \tilde{\theta})$$

is called a minimax estimator.

It is desireable for an estimator to be minimax, again as a protection against the worst-case scenerio.

There is no general rule for determining the minimax bound. However, in the case $\hat{\theta} \sim N(\theta, I_k)$, it is known that $\hat{\theta}$ is mimimax for θ.

17.4 Shrinkage Estimators

Suppose $\hat{\theta} \sim N(\theta, \mathbf{V})$

A general form for a shrinkage estimator for θ is

$$\hat{\theta}^* = \left(1 - h(\hat{\theta}'\mathbf{W}\hat{\theta})\right)\hat{\theta}$$

where $h : [0, \infty) \to [0, \infty)$. Sometimes this is written as

$$\hat{\theta}^* = \left(1 - \frac{c(\hat{\theta}'\mathbf{W}\hat{\theta})}{\hat{\theta}'\mathbf{W}\hat{\theta}}\right)\hat{\theta}$$

where $c(q) = qh(q)$.

This notation includes the James-Stein estimator, pretest estimators, selection estimators, and the Model averaging estimator of section 17.1. Pretest and selection estimators take the form

$$h(q) = 1(q < a)$$

where $a = 2k$ for Mallows selection, and a is the critical value from a chi-square distribution for a pretest estimator.

We now calculate the risk of $\hat{\theta}^*$. Note

$$\hat{\theta}^* - \theta = \left(\hat{\theta} - \theta\right) - h(\hat{\theta}'\mathbf{W}\hat{\theta})\hat{\theta}$$

Thus

$$\left(\hat{\theta}^* - \theta\right)' \mathbf{W} \left(\hat{\theta}^* - \theta\right) = \left(\hat{\theta} - \theta\right)' \mathbf{W} \left(\hat{\theta} - \theta\right) + h(\hat{\theta}'\mathbf{W}\hat{\theta})^2\hat{\theta}'\mathbf{W}\hat{\theta} - 2h(\hat{\theta}'\mathbf{W}\hat{\theta})\hat{\theta}'\mathbf{W}\left(\hat{\theta} - \theta\right)$$

Taking expectations:

$$R(\theta, \hat{\theta}^*, \mathbf{W}) = \operatorname{tr}(\mathbf{WV}) + E\left[h(\hat{\theta}'\mathbf{W}\hat{\theta})^2\hat{\theta}'\mathbf{W}\hat{\theta}\right] - 2E\left[h(\hat{\theta}'\mathbf{W}\hat{\theta})\hat{\theta}'\mathbf{W}\left(\hat{\theta} - \theta\right)\right]$$

To simplify the second expectation when h is continuous we use:

Lemma 1 *(Stein's Lemma) If $\eta(\theta) : \mathbb{R}^k \to \mathbb{R}^k$ is absolutely continuous and $\hat{\theta} \sim N(\theta, \mathbf{V})$ then*

$$\mathrm{E}\left(\eta(\hat{\theta})'\left(\hat{\theta} - \theta\right)\right) = \mathrm{E}\operatorname{tr}\left(\frac{\partial}{\partial\theta}\eta(\hat{\theta})'\mathbf{V}\right).$$

Proof: Let

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{V}^{-1}\mathbf{x}\right)$$

denote the $\mathrm{N}(\mathbf{0}, \mathbf{V})$ density. Then

$$\frac{\partial}{\partial\mathbf{x}}\phi(\mathbf{x}) = -\mathbf{V}^{-1}\mathbf{x}\phi(\mathbf{x})$$

and

$$\frac{\partial}{\partial\mathbf{x}}\phi(\mathbf{x} - \theta) = -\mathbf{V}^{-1}(\mathbf{x} - \theta)\phi(\mathbf{x} - \theta).$$

By multivariate integration by parts

$$\begin{aligned}
\mathrm{E}\left(\eta(\hat{\theta})'\left(\hat{\theta} - \theta\right)\right) &= \int \eta(\mathbf{x})'\mathbf{V}\mathbf{V}^{-1}(\mathbf{x} - \theta)\ \phi(\mathbf{x} - \theta)\,(\mathbf{dx}) \\
&= \int \operatorname{tr}\left(\frac{\partial}{\partial\theta}\eta(\mathbf{x})'\mathbf{V}\phi(\mathbf{x} - \theta)\right)(\mathbf{dx}) \\
&= \mathrm{E}\operatorname{tr}\left(\frac{\partial}{\partial\theta}\eta(\hat{\theta})'\mathbf{V}\right)
\end{aligned}$$

as stated. ■

Let $\eta(\theta)' = h\left(\theta'\mathbf{W}\theta\right)\theta'\mathbf{W}$, for which

$$\frac{\partial}{\partial\theta}\eta(\theta)' = h(\theta'\mathbf{W}\theta)\mathbf{W} + 2\mathbf{W}\theta\theta'\mathbf{W}h'(\theta'\mathbf{W}\theta)$$

and

$$\operatorname{tr} \frac{\partial}{\partial \theta} \eta(\theta)' \mathbf{V} = \operatorname{tr}(\mathbf{WV})\, h(\theta' \mathbf{W} \theta) + 2\theta' \mathbf{WVW} \theta h'(\theta' \mathbf{W} \theta)$$

Then by Stein's Lemma

$$E\left[h(\hat{\theta}' \mathbf{W} \hat{\theta}) \hat{\theta}' \mathbf{W}\left(\hat{\theta} - \theta\right)\right] = \operatorname{tr}(\mathbf{WV})\, E h(\hat{\theta}' \mathbf{W} \hat{\theta}) + 2E\left[(\hat{\theta}' \mathbf{WVW} \hat{\theta}) h'(\hat{\theta}' \mathbf{W} \hat{\theta})\right]$$

Applying this to the risk calculation, we obtain

Theorem.

$$\begin{aligned} R(\theta, \hat{\theta}^*, \mathbf{W}) &= \operatorname{tr}(\mathbf{WV}) + E\left[h(\hat{\theta}' \mathbf{W} \hat{\theta})^2 \hat{\theta}' \mathbf{W} \hat{\theta}\right] - 2\operatorname{tr}(\mathbf{WV})\, E h(\hat{\theta}' \mathbf{W} \hat{\theta}) - 4E\left[(\hat{\theta}' \mathbf{WVW} \theta) h'(\hat{\theta}' \mathbf{W} \hat{\theta})\right] \\ &= \operatorname{tr}(\mathbf{WV}) + E\left[c(\hat{\theta}' \mathbf{W} \hat{\theta}) \frac{\left(c(\hat{\theta}' \mathbf{W} \hat{\theta}) - 2\operatorname{tr}(\mathbf{WV}) + 4\dfrac{\hat{\theta}' \mathbf{WVW} \hat{\theta}}{\hat{\theta}' \mathbf{W} \hat{\theta}}\right)}{\hat{\theta}' \mathbf{W} \hat{\theta}} - 4\frac{\hat{\theta}' \mathbf{WVW} \hat{\theta}}{\hat{\theta}' \mathbf{W} \hat{\theta}} c'(\hat{\theta}' \mathbf{W} \hat{\theta})\right] \end{aligned}$$

where the final equality uses the alternative expression $h(q) = c(q)/q$.

We are trying to find cases where $R(\theta, \hat{\theta}^*, \mathbf{W}) < R(\theta, \hat{\theta}, \mathbf{W})$. This requires the term in the expectation to be negative.

We now explore some special cases.

17.5 Default Weight Matrix

Set

$$\mathbf{W} = \mathbf{V}^{-1}$$

and write

$$R(\theta, \hat{\theta}^*) = R(\theta, \hat{\theta}^*, \mathbf{V}^{-1})$$

Then

$$R(\theta, \hat{\theta}^*) = k + E\left[c(\hat{\theta}' \mathbf{V}^{-1} \hat{\theta}) \frac{\left(c(\hat{\theta}' \mathbf{V}^{-1} \hat{\theta}) - 2k + 4\right)}{\hat{\theta}' \mathbf{V}^{-1} \hat{\theta}} - 4c'(\hat{\theta}' \mathbf{V}^{-1} \hat{\theta})\right].$$

Theorem 1 *For any absolutely continuous and non-decreasing function $c(q)$ such that*

$$0 < c(q) < 2(k-2) \tag{1}$$

then

$$R(\theta, \hat{\theta}^*) < R(\theta, \hat{\theta}),$$

the risk of $\hat{\theta}^*$ is strictly less than the risk of $\hat{\theta}$. This inequality holds for all values of the parameter θ.

Note: Condition (1) can only hold if $k > 2$. (Since k , the dimension of θ, is an integer, this means $k \geq 3$.)

Proof. Let

$$g(q) = \frac{c(q)\left(c(q) - 2\left(k-2\right)\right)}{q} - 4c'(q)$$

For all $q \geq 0$, $g(q) < 0$ by the assumptions. Thus $Eg\left(q\right)$ for any non-negative random variable q. Setting $q_k = \hat{\theta}'\mathbf{V}^{-1}\hat{\theta}$,

$$R(\theta, \hat{\theta}^*) = k + Eg\left(q_k\right) < k = R(\theta, \hat{\theta})$$

which proves the result.

It also useful to note that

$$q_k = \hat{\theta}'\mathbf{V}^{-1}\hat{\theta} \sim \chi^2_k\left(\psi\right)$$

a non-central chi-square random variable with k degrees of freedom and non-centrality parameter

$$\psi = \theta'\mathbf{V}^{-1}\theta$$

17.6 James-Stein Estimator

Set $c(q) = c$, a constant. This is the James-Stein estimator

$$\hat{\theta}^* = \left(1 - \frac{c}{\hat{\theta}'\mathbf{V}^{-1}\hat{\theta}}\right)\hat{\theta} \tag{2}$$

Theorem 2 *If* $\hat{\theta} \sim N(\theta, \mathbf{V})$, $k > 2$, *and* $0 < c < 2(k-2)$, *then for (2),*

$$R(\theta, \hat{\theta}^*) < R(\theta, \hat{\theta})$$

the risk of the James-Stein estimator is strictly less than the usual estimator. This inequality holds for all values of the parameter θ.

Since the risk is quadratic in c, we can also see that the risk is minimized by setting $c = k - 2$. This yields the classic form of the James-Stein estimator

$$\hat{\theta}^* = \left(1 - \frac{k-2}{\hat{\theta}'\mathbf{V}^{-1}\hat{\theta}}\right)\hat{\theta}$$

17.7 Positive-Part James-Stein

If $\hat{\theta}'\mathbf{V}^{-1}\hat{\theta} < c$ then

$$1 - \frac{c}{\hat{\theta}'\mathbf{V}^{-1}\hat{\theta}} < 0$$

and the James-Stein estimator over-shrinks, and flips the sign of $\hat{\theta}^*$ relative to $\hat{\theta}$. This is corrected by using the positive-part version

$$\begin{aligned}\hat{\theta}^+ &= \left(1 - \frac{c}{\hat{\theta}'\mathbf{V}^{-1}\hat{\theta}}\right)_+ \hat{\theta} \\ &= \begin{cases} \left(1 - \frac{c}{\hat{\theta}'\mathbf{V}^{-1}\hat{\theta}}\right)\hat{\theta} & \hat{\theta}'\mathbf{V}^{-1}\hat{\theta} \geq c \\ 0 & \text{else} \end{cases}\end{aligned}$$

This bears some resemblance to selection estimators.

The positive-part estimator takes the shrinkage form with

$$c(q) = \begin{cases} c & q \geq c \\ q & q < c \end{cases}$$

or

$$h(q) = \begin{cases} \frac{c}{q} & q \geq c \\ 1 & q < c \end{cases}$$

In general the positive-part version of

$$\hat{\theta}^* = \left(1 - h(\hat{\theta}'\mathbf{W}\hat{\theta})\right)\hat{\theta}$$

is

$$\hat{\theta}^+ = \left(1 - h(\hat{\theta}'\mathbf{W}\hat{\theta})\right)_+ \hat{\theta}$$

Theorem. For any shrinakge estimator, $R(\theta, \hat{\theta}^+) < R(\theta, \hat{\theta})$

The proof is a bit technical, so we will skip it.

17.8 General Weight Matrix

Recall that for general $c(q)$ and weight $\mathbf{W}$ we had

$$R(\theta, \hat{\theta}^*, \mathbf{W}) = \operatorname{tr}(\mathbf{WV}) + E\left[c(\hat{\theta}'\mathbf{W}\hat{\theta})\frac{\left(c(\hat{\theta}'\mathbf{W}\hat{\theta}) - 2\operatorname{tr}(\mathbf{WV}) + 4\frac{\hat{\theta}'\mathbf{WVW}\hat{\theta}}{\hat{\theta}'\mathbf{W}\hat{\theta}}\right)}{\hat{\theta}'\mathbf{W}\hat{\theta}} - 4\frac{\hat{\theta}'\mathbf{WVW}\hat{\theta}}{\hat{\theta}'\mathbf{W}\hat{\theta}}c'(\hat{\theta}'\mathbf{W}\hat{\theta})\right]$$

Using a result about eigenvalues and setting $h = \mathbf{W}^{-1/2}\theta$

$$\begin{aligned}
\frac{\hat{\theta}'\mathbf{W}\mathbf{V}\mathbf{W}\hat{\theta}}{\hat{\theta}'\mathbf{W}\hat{\theta}} &\leq \max_{\theta} \frac{\theta'\mathbf{W}\mathbf{V}\mathbf{W}\theta}{\theta'\mathbf{W}\theta} \\
&= \max_{h} \frac{h'\mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}h}{h'h} \\
&= \lambda_{\max}(\mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}) \\
&= \lambda_{\max}(\mathbf{W}\mathbf{V})
\end{aligned}$$

Thus if $c'(q) \geq 0$,

$$\begin{aligned}
R(\theta, \hat{\theta}^*, \mathbf{W}) &\leq \operatorname{tr}(\mathbf{W}\mathbf{V}) + E\left[c(\hat{\theta}'\mathbf{W}\hat{\theta})\frac{\left(c(\hat{\theta}'\mathbf{W}\hat{\theta}) - 2\operatorname{tr}(\mathbf{W}\mathbf{V}) + 4\lambda_{\max}(\mathbf{W}\mathbf{V})\right)}{\hat{\theta}'\mathbf{W}\hat{\theta}}\right] \\
&< \operatorname{tr}(\mathbf{W}\mathbf{V})
\end{aligned}$$

the final inequality if

$$0 < c(q) < 2\left(\operatorname{tr}(\mathbf{W}\mathbf{V}) - 2\lambda_{\max}(\mathbf{W}\mathbf{V})\right) \tag{3}$$

When $W = V$, the upper bound is $2(k-2)$ so this is the same as for the default weight matrix.

Theorem 3 *For any absolutely continuous and non-decreasing function $c(q)$ such that (3) holds, then*

$$R(\theta, \hat{\theta}^*, \mathbf{W}) < R(\theta, \hat{\theta}, \mathbf{W}),$$

the risk of $\hat{\theta}^*$ is strictly less than the risk of $\hat{\theta}$.

17.9 Shrinkage Towards Restrictions

The classic James-Stein estimator shrinks towards the zero vector. More generally, shrinkage can be towards restricted estimators, or towards linear or non-linear subspaces.

These estimators take the form

$$\hat{\theta}^* = \hat{\theta} - h(\left(\hat{\theta} - \tilde{\theta}\right)' \mathbf{W} \left(\hat{\theta} - \tilde{\theta}\right)) \left(\hat{\theta} - \tilde{\theta}\right)$$

where $\hat{\theta}$ is the unrestricted estimator (e.g. the long regression) and $\tilde{\theta}$ is the restricted estimator (e.g. the short regression).

The classic form is

$$\hat{\theta}^* = \hat{\theta} - \left(\frac{r-2}{\left(\hat{\theta} - \tilde{\theta}\right)' \hat{\mathbf{V}}^{-1} \left(\hat{\theta} - \tilde{\theta}\right)}\right)_1 \left(\hat{\theta} - \tilde{\theta}\right)$$

where $(a)_1 = \max(a, 1)$, $\hat{\mathbf{V}}$ is the covariance matrix for $\hat{\theta}$, and r is the number of restrictions (by the restriction from $\hat{\theta}$ to $\tilde{\theta}$).

This estimator shrinks $\hat{\theta}$ towards $\tilde{\theta}$, with the degree of shrinkage depending on the magnitude of $\left(\hat{\theta}-\tilde{\theta}\right)$.

This approach works for nested models, so that $\left(\hat{\theta}-\tilde{\theta}\right)' \hat{\mathbf{V}}^{-1}\left(\hat{\theta}-\tilde{\theta}\right)$ is approximately (non-central) chi-square.

It is unclear how to extend the idea to non-nested models, where $\left(\hat{\theta}-\tilde{\theta}\right)' \hat{\mathbf{V}}^{-1}\left(\hat{\theta}-\tilde{\theta}\right)$ is not chi-square.

17.10 Inference

We discussed shrinkage estimation.

Model averaging, Selection, and Shrinkage estimators have non-standard non-normal distributions.

Standard errors, testing, and confidence intervals need development.