Notes on Statistical Decision Theory with Applications to Regression for Econometrics II

Tim Armstrong

last updated: January 22, 2020

- Consider a statistical model in which we observe Y with distribution $f(\cdot;\theta)$, where θ is an unknown parameter taking values in a parameter space Θ.
- We use the notation E_θ and P_θ to denote expectations and probabilities when Y is distributed according to the parameter θ.
- We will discuss some decision theoretic concepts for formalizing the notion that a particular estimator or test is "optimal" or "preferred" to another estimator. We will then apply this to answer the question of when or in what sense ordinary least squares (OLS) is optimal in the regression model.

  Formally, we will consider the *linear regression model with fixed design*:

  $$\underset{n\times 1}{Y} = \underset{n\times k}{X}\underset{k\times 1}{\theta} + \underset{n\times 1}{\varepsilon} \quad \varepsilon \sim N(0, \sigma^2 \underset{n\times n}{I})$$

  where we treat $X = (x_1, \ldots, x_n)'$ as fixed (nonrandom) and σ^2 as known. This can be written as

  $$Y \sim N(X\theta, \sigma^2 I),$$

  so we obtain our setup with $f(y;\theta)$ given by the multivariate normal density with mean $X\theta$ and variance $\sigma^2 I$. We can take Θ to be $\mathbb{R}^k$, or we can consider restricted parameter spaces, such as restricting the magnitude of $\|\theta\|$. The OLS estimator is given by $\hat{\theta}_{OLS} = (X'X)^{-1}X'Y$.

  - Since we treat $X = (x_1, \ldots, x_n)'$ as nonrandom, our treatment can be related to

the treatment in Hansen's text (in which X is random) by conditioning on X in his setup. Note also that we use Y in place of Hansen's y and ε in place of Hansen's e.

- We are faced with a set $\mathcal{A}$ of possible actions. If we choose an action $a \in \mathcal{A}$ and the parameter is given by θ, we incur a loss

$$L(\theta, a),$$

where $L(\theta, a)$ is called the *loss function*. A *decision rule* is a mapping $\delta(Y)$ that takes the data Y to an action in $\mathcal{A}$. The *risk function* of the decision rule δ is

$$R(\theta, \delta) = E_\theta L(\theta, \delta(Y)) = \int L(\theta, \delta(y)) f(y; \theta)\, dy.$$

- Examples of decisions and loss functions:
  - In the problem of *estimation*, the action space $\mathcal{A}$ is identical to the parameter space Θ, and the decision function $\delta(Y)$ is called an *estimator*, often written $\hat{\theta} = \hat{\theta}(Y)$. A common choice is *squared error* loss, given by (in the case where θ is scalar)

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

which leads to *mean squared error (MSE)* as the risk function:

$$R(\theta, \hat{\theta}) = E_\theta \left[(\hat{\theta} - \theta)^2 \right]$$

Of course, other loss functions are possible, for example, *absolute error loss* $L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$.

    - Often, θ is multivariate, and we are interested in a scalar function $T(\theta)$ (e.g. in the linear regression model, we are often interested in a particular coefficient, say θ_1, viewing the rest of the variables as controls). Then, we can phrase the estimation problem as coming up with an estimator $\hat{T}$ for $T(\theta)$, and we can define squared error loss as $L(\theta, \hat{T}) = (\hat{T} - T(\theta))^2$ (and similarly for absolute error, etc.).

  - In the problem of *hypothesis testing*, we are interested in determining whether

$\theta \in H_0$ or $\theta \in H_1$, where H_0 and H_1 are called the *null hypothesis* and *alternative hypothesis*. The action space is $\{0, 1\}$, with 1 denoting rejection of the null and 0 denoting failure to reject. The decision function $\delta(Y)$ is called a test, and is often denoted $\phi(Y)$. We can define the loss function as

	$\phi(Y) = 0$	$\phi(Y) = 1$
$\theta \in H_0$	0	L_I
$\theta \in H_1$	L_{II}	0

where L_I and L_{II} are relative weights given to type I and II errors respectively.

- The risk function $R(\theta, \delta)$ defines a partial ordering on decision functions δ. If $R(\theta, \delta) \leq R(\theta, \tilde{\delta})$ for all θ, then δ is preferred to $\tilde{\delta}$. If δ is preferred to $\tilde{\delta}$ and the inequality is strict for some θ, then δ is strictly preferred to δ. We say that δ is *admissible* if no other decision function $\tilde{\delta}$ is strictly preferred to δ.

- Since this is only a partial ordering, we will typically not be able to come up with a "uniformly optimal" decision. In other words, we will typically have decisions δ and $\tilde{\delta}$ with $R(\theta, \delta) > R(\theta, \tilde{\delta})$ for some values of θ and $R(\theta, \delta) < R(\theta, \tilde{\delta})$ for other values of θ. Two ways of addressing this...

- The *minimax risk* of a decision δ over the parameter space Θ is given by

$$R_{\text{minimax}}(\Theta, \delta) = \sup_{\theta \in \Theta} R(\theta, \delta).$$

  A decision that optimizes minimax risk is said to be *minimax*.

- The *Bayes risk* of a decision δ for a *prior* $\pi(\theta)$ on θ is given by

$$R_{\text{Bayes}}(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)\, d\theta.$$

  A decision that optimizes Bayes risk for a prior $\pi(\theta)$ is said to be a *Bayes decision* for prior $\pi(\theta)$.

  – We can interpret Bayes risk as coming from a setup where θ is drawn randomly from $\pi(\theta)$, and Y is then drawn from $f(y; \theta)$. The prior can be interpreted as representing the researcher's beliefs.

- We can view statistical decision theory as modeling the empirical researcher as a rational individual making a choice under uncertainty, using the framework typically taught in a first year graduate microeconomic theory course (see, e.g., Ch. 6 of Mas-Colell et al., 1995).
  - The loss $L(\theta, \delta)$ plays the role of the (negative of the) Bernoulli utility function.
  - The Bayes risk $R_{\text{Bayes}}(\pi, \delta)$ plays the role of the (negative of the) expected utility from the decision δ when uncertainty in θ is characterized by $\pi(\theta)$.
  - In contrast to the direct interpretation of Bayes risk as (the negative of) expected utility, the minimax criterion is perhaps more influential in statistics than in microeconomic theory. However, it has interpretations that are closely related to concepts encountered in microeconomic theory:
    * The minimax decision (choosing δ to minimize $R_{\text{minimax}}(\Theta, \delta)$) arises from a zero-sum game against nature in which the empirical researcher chooses δ and nature chooses θ. The researcher's payoff is $-R(\theta, \delta(Y))$, and nature's payoff is $R(\theta, \delta(Y))$.
    * Minimax can be used to formalize the idea of *ambiguity aversion* (see Gilboa and Schmeidler, 1989).
    * The minimax decision can be motivated as a way for a group of individuals with different priors to agree on a single decision. A decision with low minimax risk will have low Bayes risk regardless of the prior.
  - Much of statistical decision theory was developed in parallel with related topics in microeconomic theory, with some of the same people working on both topics. See early books on these topics such as Blackwell and Girshick (1954) and Savage (1972).

- Consider the hypothesis testing setup. The *size* of the test ϕ is given by the worst-case risk over the null when L_I (the weight on type I error) is given by 1:

$$\text{size}(\phi) = \sup_{\theta \in H_0} R(\theta, \phi) = \sup_{\theta \in H_0} E_\theta \phi(Y).$$

  The classical approach is to take a given α (often $\alpha = .05$) and to restrict attention to tests with size bounded by α. This can be considered a partial minimax approach, since type I error is handled with a minimax (worst-case) criterion.

With this approach, tests satisfying this criterion are compared according to type II error, with type I error not playing any further role so long as the size is controlled. For $\theta \in H_1$, we refer to the rejection probability $E_\theta \phi(Y)$ (one minus the risk with $L_{II} = 1$) as the *power* of the test at θ. As with the general problem of optimizing risk, the problem of "maximizing power subject to size control" does not, in general, have a unique solution, and we can resolve this choice using Bayes or minimax. In certain cases, the same test ϕ maximizes power subject to the size constraint *simultaneously* at each $\theta \in H_1$. Such a test is called *uniformly most powerful (UMP)*. When a UMP test exists, we don't have to worry about how to trade off power over different parts of H_1: the UMP test is optimal regardless.

- Often, additional considerations such as unbiasedness are imposed when comparing estimators. Unbiasedness is difficult to motivate from a decision theoretic perspective as being of interest per se, although it becomes useful if estimates are averaged over multiple studies as a form of meta-analysis.

- The concepts of admissibility, Bayes risk and minimax risk are related by two important theorems (we summarize them here without giving formal conditions):

  - Under regularity conditions, a decision is admissable if and only if there exists a prior π such that it is Bayes.
  - Under regularity conditions, there exists a prior π such that the minimax decision is the Bayes decision for π. This prior is called *least favorable*.

- Suppose that a decision δ has constant risk function: $R(\theta, \delta)$ does not depend on θ. Then, if δ is admissible, it is also minimax (admissability means that any other decision δ' has $R(\theta, \delta') > R(\theta, \delta)$ for some θ; since the right hand side is constant over θ, the sup of the left hand side over θ must be greater than the sup of the right hand side over θ). However, the converse is not true (a minimax decision with constant risk function need not be admissible).

Optimality of OLS for Linear Regression

- The Gauss-Markov Theorem states that $\hat{\theta}_{OLS} = (X'X)^{-1}X'Y$ minimizes variance among unbiased estimators that are linear in Y (i.e. estimators of the form $\underset{k \times n}{A(X)} \underset{n \times 1}{Y}$ such that $E_\theta A(X)Y = \theta$ for all θ). A common (and very reasonable) critique of this

theorem is that it is difficult to motivate restricting attention to linear unbiased estimators: if there is a biased and/or nonlinear estimator that performs better for all θ, we would probably want to use it! This section discusses some optimality results for OLS that are not subject to this critique, as well as criteria under which OLS is suboptimal.

- Let $T\theta$ be a linear functional of θ, where $\underset{k\times 1}{T}$ is a row vector (e.g. if $T = (1, 0, \ldots, 0)$, then $T\theta = \theta_1$, the first coefficient). In the fixed design homoskedastic normal model with known error variance, the OLS estimator $\hat{\theta}_{OLS} = (X'X)^{-1}X'Y$ has two important optimality properties for estimation and inference on $T\theta$:
  - $T\hat{\theta}_{OLS}$ is an admissable minimax estimator for $T\theta$ (under general conditions on the loss function).
  - Consider the null hypothesis $H_0 : T\theta \leq T_0$, where $T_0 \in \mathbb{R}$ is given. The uniformly most powerful (UMP) test of H_0 is a z-test based on $T\hat{\theta}_{OLS}$: it rejects when $T\hat{\theta}_{OLS} > T_0 + z_{1-\alpha} \cdot \text{se}(T\hat{\theta}_{OLS})$ where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the $N(0,1)$ distribution and $\text{se}(T\hat{\theta}_{OLS}) = \sigma\sqrt{T(X'X)^{-1}T'}$.
- On the other hand, OLS is suboptimal according to other criteria:
  - OLS is suboptimal for the Bayes criterion $R_{\text{Bayes}}(\pi, \hat{\theta})$ unless we take π to be an improper prior (an improper prior refers to the case where π integrates to ∞; to accomodate this case, we have to define Bayes risk as a limit of Bayes risk for proper priors).
  - If we consider a restricted parameter space $\widetilde{\Theta} \subsetneq \mathbb{R}^k$, then OLS is suboptimal for $R_{\text{minimax}}(\widetilde{\Theta}, \hat{\theta})$ (minimax risk over $\widetilde{\Theta}$) for typical choices of $\widetilde{\Theta}$.
  - Suppose that, rather than looking at risk of a single functional $T\theta$, we look at overall risk for θ with loss function $L(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|^2 = \sum_{j=1}^{k}(\hat{\theta}_j - \theta_j)^2$. Then $\hat{\theta}_{OLS}$ is inadmissable when $k \geq 3$.

0.1 Optimality for One-sided Testing

- To prove the testing optimality result, let ϕ be a test of $H_0 : T\theta \leq T_0$ with size α and let $K_\phi(\theta) = E_\theta\phi(Y)$ denote its power function.

- Step 1: let $\theta_0 \in H_0$ and $\theta_1 \in H_1$. Let $\phi_{NP,\theta_0,\theta_1}$ denote the test that maximizes power at θ_1 subject to the constraint that the rejection probability at θ_0 is bounded by α (but without requiring size control over all of H_0). Argue that $K_\phi(\theta_1) \leq K_{\phi_{NP,\theta_0,\theta_1}}(\theta_1)$.

- Step 2: By the Neyman-Pearson lemma (covered in Econometrics I), $\phi_{NP,\theta_0,\theta_1}$ is the likelihood ratio test for θ_0 vs θ_1. Derive $\phi_{NP,\theta_0,\theta_1}$ and compute its power $K_{\phi_{NP,\theta_0,\theta_1}}(\theta_1)$ at θ_1.

- Step 3: Minimize $K_{\phi_{NP,\theta_0,\theta_1}}(\theta_1)$ over $\theta_0 \in H_0$. By Step 1, this gives an upper bound on $K_\phi(\theta_1)$.

- Step 4: Let ϕ_{OLS} be the test described above based on $T\hat{\theta}_{OLS}$. Derive $K_{\phi_{OLS}}(\theta_1)$ and show that it is identical to the bound derived in Step 3.

- Step 5: Since θ_1 was arbitrary, the result follows: for all $\theta_1 \in H_1$ and any level α test ϕ, $K_{\phi_{OLS}}(\theta_1) \geq K_\phi(\theta_1)$.

- Suppose that we are interested in the first component θ_1 (here, we change notation by using θ_1 to refer to the first component, rather than the whole parameter vector of an alternative parameter value). We wish to perform inference on θ_1 while controlling size regardless of $\theta_2, \ldots, \theta_k$. However, we are only allowing for arbitrary values of $\theta_2, \ldots, \theta_k$ as a precaution: we only care about having good power when $\theta_2, \ldots, \theta_k$ are zero.

  This might lead us to attempt to form a test where we only use the first regressor and throw away the rest of the regressors (i.e. forming a test based on the short regression), perhaps with some additional pre-testing procedure in which we add other regressors if their parameters are statistically significant in the long regression. The above result tells us that such a procedure cannot improve on a procedure where we always use the long regression while controlling size, even if it "turns out" that all of the other coefficients are indeed equal to zero.

0.2 Admissibility and Minimaxity of OLS for Linear Functionals

- In this section, we sketch a proof of admissibility and minimaxity of $T\hat{\theta}_{OLS}$ as an estimator of $T\theta$ for squared error loss $L(\theta, \hat{T}) = (\hat{T} - T\theta)^2$ (the result also holds for loss functions of the form $L(\theta, \hat{T}) = \ell(\hat{T} - T\theta)$ where ℓ satisfies certain conditions). First, note that $R(\theta, T\hat{\theta}_{OLS})$ is constant in θ (since the distribution of $T\hat{\theta}_{OLS} - T\theta$

does not depend on θ). Thus, it suffices to show admissibility (as noted in a comment above, any admissible estimator with constant risk is also minimax).

- By unbiasedness of OLS, $R(\theta, T\hat{\theta}_{OLS}) = var_\theta(T\hat{\theta}_{OLS}) = \sigma^2 T(X'X)^{-1}T'$. Suppose that, for some alternative estimator $\hat{T}$ and some θ^*, we have $R(\theta^*, \hat{T}) < \sigma^2 T(X'X)^{-1}T'$. We will show that there exists θ such that $R(\theta, \hat{T}) > \sigma^2 T(X'X)^{-1}T'$, thereby showing admissibility of $T\hat{\theta}_{OLS}$.

- Step 1: Consider the *one-dimensional submodel* $\widetilde{\Theta}_a = \{\theta^* + ta | t \in \mathbb{R}\}$, where a is a vector in $\mathbb{R}^k$ with $Ta \neq 0$. For $\theta = \theta^* + ta$ in this submodel, we have

$$Y - X\theta^* \sim N(tXa, \sigma^2 I)$$

  so that the problem of estimating θ can be reduced to estimating t in this model. Let B be a $(n-1)\times n$ matrix such that (Xa, B') is invertible and $BXa = 0$ (i.e. all of the rows of B are orthogonal to Xa). Let $\tilde{Y} = (Xa)'(Y - X\theta^*)/(aX'Xa)$ and $Z = B(Y - X\theta^*)$. Then $(\tilde{Y}, Z)$ is a one-to-one transformation of Y with $\tilde{Y} \sim N(t, \sigma^2/a'X'Xa)$, $Z \sim N(tBXa, \sigma^2 BB') = N(0, \sigma^2 BB')$ and Z independent of $\tilde{Y}$.

- Step 2: For the estimator $\hat{T} = \hat{T}(Y)$ with lower risk than $T\hat{\theta}_{OLS}$ at θ^*, we have, for $\theta^* + ta$ in the submodel,

$$\begin{aligned} &R(\theta, \hat{T}) = E_\theta(\hat{T} - T\theta)^2 = E_{\theta^*+ta}(\hat{T} - T(\theta^* + ta))^2 = E_{\theta^*+ta}(\hat{T} - T\theta^* - tTa))^2 \\ &= (Ta)^2 E_{\theta^*+ta}((\hat{T} - T\theta^*)/Ta - t)^2 = (Ta)^2 \tilde{R}(t, \hat{t}_a) \end{aligned} \tag{1}$$

  where $\tilde{R}(t, \tilde{t}_a)$ is the risk of the estimator $\hat{t} = (\hat{T} - T\theta^*)/Ta$ for t. Since $\hat{t}$ can be written as a function of $\tilde{Y}, Z$ where Z is independent noise that does not depend on t, $\tilde{R}(t, \hat{t})$ corresponds to the risk function for a (possibly randomized) estimator of t based on $\tilde{Y} \sim N(t, \sigma^2/(a'X'Xa))$. By Example 2.8 (starting on p. 324) in Lehmann and Casella (1998), $\tilde{Y}$ is admissable for t in this setting, so $\tilde{R}(t, \hat{t})$ cannot be less than or equal to $\sigma^2/(a'X'Xa)$ for all t with strict inequality for some t. In particular, if $\tilde{R}(0, \hat{t}) < \sigma^2/(a'X'Xa)$, then there must be some t such that $\tilde{R}(t, \hat{t}) > \sigma^2/(a'X'Xa)$. Combining this with (1), it follows that,

$$\text{if } R(\theta^*, \hat{T}) < \sigma^2(Ta)^2/(a'X'Xa), \text{ then } \sup_\theta R(\theta, \hat{T}) > \sigma^2(Ta)^2/(a'X'Xa). \tag{2}$$

- Step 3: To obtain the sharpest bound in (2), we maximize $(Ta)^2/(a'X'Xa)$ over a. The solution sets a proportional to $(X'X)^{-1}T'$ (this can be seen by noting that the maximization problem is equivalent to maximizing Ta subject to a bound on $a'X'Xa$ and taking first order conditions for the Lagrangian for this problem), which gives $(Ta)^2/(a'X'Xa) = (T(X'X)^{-1}T')^2/(T(X'X)^{-1}X'X(X'X)^{-1}T') = T(X'X)^{-1}T'$. Plugging this into (2), it follows that, if $R(\theta^*, \hat{T}) < \sigma^2 T(X'X)^{-1}T'$, then there exists θ with $R(\theta, \hat{T}) > \sigma^2 T(X'X)^{-1}T'$. This is exactly what we needed to show.

0.3 Suboptimality for Bayes with Proper Prior and Minimax with Restricted Parameter Space

- Let $\widetilde{\Theta}$ be a convex subset of $\mathbb{R}^n$. A theory of minimax affine estimators (estimators of the form $c(X) + a(X)Y$) has been developed for this case which derives the minimax affine estimator as the solution to a convex optimization problem and shows that the minimax affine estimator is near minimax among all estimators (see Donoho, 1994). As an example, the minimax affine estimator when $\widetilde{\Theta} = \{\theta | \|\theta\| \leq C\}$ (we restrict the magnitude of θ by bounding its Euclidean norm) turns out to be a *ridge regression* estimator, which takes the form $(X'X + \lambda I)^{-1}X'Y$ where λ is a constant that depends on C (see Sections 4.1.2 and D.2 of Armstrong and Kolesár, 2016).

- Under a Bayes criterion, the optimal estimator under squared error loss is the posterior mean (the mean of the conditional distribution of θ given the observed value of Y, calculated with $\theta \sim \pi(\theta)$ and $Y|\theta \sim f(Y;\theta)$). As an example, if $\pi(\theta)$ is the $N(0, \sigma A^{-1})$ distribution, then the optimal estimate under squared error loss is $(X'X + A)^{-1}X'Y$. When A is a multiple of the identity matrix, this again gives the ridge regression estimator. See Rossi et al. (2012, Section 2.8.1).

- The ridge regression estimator is an example of a *shrinkage estimator*. For other priors π or parameter spaces $\widetilde{\Theta}$, other optimal estimators can be derived, and these can often be interpreted as using different forms of "shrinkage" or "regularization." These ideas come up often in the fields of *high dimensional* and *nonparametric* statistics.

0.4 Inadmissability for Estimation of the Entire Vector θ when $k \geq 3$

- Consider the case where $X = I$, so that we are simply interested in estimating a vector of normal means based on a single observation of each variable (equivalently, we can view these observations as sample means of each variable): $Y_j \sim N(\beta_j, \sigma^2)$ for $j = 1, \ldots, k$. The inadmissability of the estimator $(Y_1, \ldots, Y_k)$ for $(\beta_1, \ldots, \beta_k)$ when $k \geq 3$ is a classic result due to Stein (1956). See the first chapter of Efron (2012) for an intuitive discussion of this phenomenon and proof.

- This phenomenon is related to the ideas of *adaptive estimation* (in the example above of minimax with constrainted parameter space $\widetilde{\Theta} = \{\theta | \|\theta\| \leq C\}$, this corresponds, roughly speaking, to estimating C) and *empirical Bayes* (an example of this is letting $\Sigma = \tau I$ in the example above and estimating τ).

References

Armstrong, T. B. and M. Kolesár (2016): "Optimal inference in a class of regression models," *arXiv:1511.06028 [math, stat]*.

Blackwell, D. and M. A. Girshick (1954): *Theory of Games and Statistical Decisions*, John Wiley & Sons, Incorporated.

Donoho, D. L. (1994): "Statistical Estimation and Optimal Recovery," *The Annals of Statistics*, 22, 238–270.

Efron, B. (2012): *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Cambridge: Cambridge University Press.

Gilboa, I. and D. Schmeidler (1989): "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, 18, 141–153.

Lehmann, E. L. and G. Casella (1998): *Theory of Point Estimation*, New York: Springer, 2nd edition ed.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995): *Microeconomic Theory*, New York: Oxford University Press, 1 edition ed.

Rossi, P. E., G. M. Allenby, and R. McCulloch (2012): *Bayesian Statistics and Marketing*, John Wiley & Sons.

Savage, L. J. (1972): *The Foundations of Statistics*, New York, NY: Dover, 2 ed.

Stein, C. (1956): "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, The Regents of the University of California.