LECTURE 10: ESTIMATION, TESTING, CONFIDENCE INTERVALS

1. ESTIMATION: AS A BAYESIAN DECISION PROBLEM

- The "decision" is choice of a vector $\hat{\beta}$ that should be close to the unknown vector β.
- For example, if our loss function is $\mathcal{L}(\beta,\hat{\beta}) = (\beta-\hat{\beta})'A(\beta-\hat{\beta})$ for some positive definite A, then the optimal decision, as a function of observed data X, is $\hat{\beta} = \bar{\beta} \stackrel{\text{def}}{=} E[\beta \mid X]$.

2. PROOF:

$$
\begin{aligned}
&E[(\beta-\hat{\beta})'A(\beta-\hat{\beta})] \\
&\quad = E[(\beta-\bar{\beta})'A(\beta-\bar{\beta}) + 2(\beta-\bar{\beta})'A(\bar{\beta}-\hat{\beta}) + (\bar{\beta}-\hat{\beta})'A(\bar{\beta}-\hat{\beta}) \mid X] \\
&\quad = \operatorname{tr}\big(A\operatorname{Var}(\beta \mid X)\big) + (\bar{\beta}-\hat{\beta})'A(\bar{\beta}-\hat{\beta})\,.
\end{aligned}
$$

This latter expression consists of a first piece that does not depend on $\hat{\beta}$, and a second expression that, because A is p.d., is minimized at $\bar{\beta} = \hat{\beta}$, where it is zero. The middle term in the expansion drops out because $E[\beta-\bar{\beta} \mid X] = 0$ by construction.

There is a corresponding result for $\mathcal{L}(\beta,\hat{\beta}) = \sum \left|\hat{\beta}_i - \beta_i\right|$. With this loss function, it is optimal to set $\hat{\beta}$ so that $(\forall i)P[\beta_i > \hat{\beta}_i \mid X] = .5$, i.e. so that $\hat{\beta}$ is the vector of medians of the $\beta \mid X$ distribution. (This assumes the $\beta \mid X$ distribution has a pdf.)

3. UNBIASEDNESS

- You might think that this means $E[\beta \mid X] = \hat{\beta}$, but that would be a Bayesian analogue of unbiasedness. Unbiasedness is a non-Bayesian attribute for an estimator, meaning that it is a property of the distribution of the estimator as the observed data X varies randomly. $\hat{\beta}(X)$ is an **unbiased estimator for** β iff $(\forall\beta)E[\hat{\beta} \mid \beta] = \beta$.
- While this property is intuitively appealing, it is hard to give any formal argument that unbiasedness is a good property. In fact, we have the following result:
- *No finite-variance Bayesian posterior mean is unbiased, unless it is completely error free.*
- Proof, for the 1-dimensional case: Suppose instead $E[\hat{\beta} \mid \beta] = \beta$ and $E[\beta \mid X] = \hat{\beta}$. Then

$$
\begin{aligned}
E[\beta^2] &= E\big[\operatorname{Var}(\beta \mid X)\big] + E[\hat{\beta}^2] \Rightarrow E[\beta^2] > E[\hat{\beta}^2] \\
E[\hat{\beta}^2] &= E\big[\operatorname{Var}(\hat{\beta} \mid \beta)\big] + E[\beta^2] \Rightarrow E[\hat{\beta}^2] > E[\beta^2] \rightarrow\leftarrow
\end{aligned}
$$

Since we know that under some regularity conditions every admissible estimator will be Bayesian for some prior, this suggests that *usually* unbiased estimators are inadmissible under a quadratic loss function.

4. THE STEIN RESULT

- Having seen the result that unbiased estimates can't be admissible under mean-squared-error loss, it is perhaps surprising that the OLS estimator of β in the SNLM $Y = X\beta + \varepsilon$ is admissible — if X has no more than two columns. This reminds us that the regularity conditions that make Bayesian and admissible estimators equivalent classes are indeed restrictive. The OLS estimator is not a Bayes estimator under any proper prior, but it is admissible. (It is the limit of a sequence of Bayesian estimators.)
- When β is of dimension higher than two, OLS is not admissible. This was shown by Stein, who constructed an estimator (itself inadmissible) that dominates OLS under a mean squared error loss function.

5. CONSISTENCY

- $\hat{\beta}_T \xrightarrow[t\to\infty]{P} \beta \overset{\text{def}}{\Leftrightarrow} (\forall \varepsilon > 0) P\left[\left\|\hat{\beta}_T - \beta\right\| > \varepsilon\right] \xrightarrow[T\to\infty]{} 0$.
- This is **convergence in probability**.
- Note that β, like the $\hat{\beta}_T$'s, in this definition can be a random variable.
- Convergence in probability makes a statement about the sequence of joint distributions of $(\hat{\beta}_T, \beta)$.
- This contrasts with $\hat{\beta}_T \xrightarrow[T\to\infty]{\mathcal{D}} \beta$, which makes a statement about the sequence of marginal distributions of the $\hat{\beta}_T$'s.
- $\hat{\beta}_T \xrightarrow{P} \beta \Rightarrow \hat{\beta}_T \xrightarrow{\mathcal{D}} \beta$.
- A commonly occurring special case is that where β has a degenerate distribution making it a constant.
- $\hat{\beta}_T$ is a **consistent** sequence of estimators for β if $\hat{\beta}_T \xrightarrow{P} \beta$.

6. TESTING

Θ :	parameter space
$H_0 \subset \Theta$:	null hypothesis
$H_A \subset \Theta$:	alternative hypothesis
X :	observed data, taking values in Γ
$S : \Gamma \to \mathbb{R}^n$:	test statistic
$B \subset \mathbb{R}^n$:	rejection region
$\alpha = P[S \in B \mid H_0]$:	significance level, or size
$\gamma = P[S \in B \mid H_A]$:	power

7. CONNECTION TO A DECISION PROBLEM

- The definitions displayed here are only appropriate, strictly speaking, when H_A and H_0 are each a single point.
- If in addition $\Theta = \{H_0, H_A\}$ and the only data we observe is $Z = 1_{S(X)\in B}$, then knowledge of α and γ allows us to fully specify the likelihood. For $Z = 1$ the likelihood is just α at H_0 and γ at H_A, so with, e.g., equal prior probabilities on H_0 and H_A, the probabilities of H_0 and H_A when $Z = 1$ are $\alpha/(\alpha+\gamma)$ and $\gamma/(\alpha+\gamma)$. When $Z = 0$ (the null is not rejected) the probabilities are instead $(1-\alpha)/(2-\alpha-\gamma)$ and $(1-\gamma)/(2-\alpha-\gamma)$.

8. COMPOUND HYPOTHESES

- When an hypothesis is more than a single point in the parameter space, it is called a **compound** hypothesis. If H_0 is compound, the probability of rejection may vary over the parameter space. It is standard terminology in this case to say that the size or significance level of the test is

$$\max_{\theta \in H_0} P[S \in B \mid \theta] .$$

- It is *not* standard to make the corresponding change modification of the definition of power, which would make it

$$\min_{\theta \in H_A} P[S \in B \mid \theta] .$$

- If we made this modification to the notion of power, most standard cases would show power equal to significance level. Since in the point null, point alternative case a test with power equal to significance level is useless, it is clear that we don't want to define power as this minimum, since it would make useful tests look useless.

- However, it is not so clear why size should be maximized over H_0, while power is treated as a function of $\theta \in H_A$, instead of defining power as the minimum over H_A of the rejection probability, while significance level were treated as function of $\theta \in H_0$.
- A test is **unbiased** iff

$$\theta_0 \in H_0\,,\ \theta_A \in H_A \ \Rightarrow\ P[S \in B \mid \theta_0] \leq P[S \in B \mid \theta_1]\,.$$

- Unbiasedness of tests has nothing to do with unbiasedness of estimators. For point null, point alternative cases a biased test is ridiculous — it would be better to replace the rejection region with its complement, which would produce an unbiased test. In more complicated situations it is sometimes reasonable to use an easily computed or interpreted test that is biased. In such cases the portion of the parameter space over which the rejection probabilities have the wrong relative magnitudes is judged to be small.
- Example of a biased test:

$$\begin{aligned} H_0: &\qquad S \sim N(1,1) \\ H_A: &\qquad S \sim N(\gamma,\gamma^2)\,,\gamma \in (0,\infty) \\ B: &\qquad |S-1| > 1.96\,. \end{aligned}$$

This is a test with size .05 under the point null. However, only for $\gamma > 1$ is the probability of rejection greater under H_A. The probability of rejection, conditional on γ, is

$$P[S-1>1.96 \mid \gamma] + P[S-1<-1.96 \mid \gamma]$$
$$= 1-\Phi\left(\frac{2.96}{\gamma}+1\right)+\Phi\left(\frac{-.96}{\gamma}-1\right),$$

which is everywhere increasing in γ, and thus *smaller* for $\gamma < 1$ than for $\gamma = 1$.