Problem Set 2

Alexei Onatski

Problem 1. This problem is based on Box (1953) "Non-normality and tests on variances," *Biometrika* 40, 318-335. In that paper, Box coined the term "robustness". Let $Y_{11}, ..., Y_{1n_1}$ and $Y_{21}, ..., Y_{2n_2}$ be two independent samples, each sample being i.i.d. with cumulative distribution function $G_j(y)$, mean μ_j and variance σ_j^2, $j = 1, 2$. The sample means and variances are $\bar{Y}_j = n_j^{-1} \sum_{i=1}^{n_j} Y_{ji}$ and $s_i^2 = (n_j - 1)^{-1} \sum_{i=1}^{n_j} \left(Y_{ji} - \bar{Y}_j\right)^2$. Suppose that you would like to test for $H_0 : \sigma_1^2 = \sigma_2^2$. The usual test (based on the assumption that the data are normally distributed) is to compare s_1^2/s_2^2 to an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

(a) Consider the logarithm of the normal theory test statistic s_1^2/s_2^2, standardized by sample sizes:

$$T = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \left[\log s_1^2 - \log s_2^2\right].$$

Prove that asymptotically as n_1 and n_2 go to infinity, the usual test is equivalent to comparing T to a $N(0, 2)$ distribution. (Hint: as $n_1, n_2 \to \infty$, an F distribution with $n_1 - 1$ and $n_2 - 1$ would put all mass at 1 because both its 'numerator' and 'denominator' converge in probability to one. But what about small deviations of the 'numerator' and 'denominator' from unity? Can you use a CLT to figure out how these deviations behave, and therefore derive an approximate distribution of the lograithm of F, multiplied by $\left(\frac{n_1 n_2}{n_1+n_2}\right)^{1/2}$?)

(b) Now, suppose that

$$G_1(y) = F_0\left(\left(y - \mu_1\right)/\sigma_1\right) \text{ and } G_2(y) = F_0\left(\left(y - \mu_2\right)/\sigma_2\right),$$

where F_0 is a non-Gaussian cdf with $\int y \mathrm{d}F_0(y) = 0$ and $\int y^2 \mathrm{d}F_0(y) = 1$. Show that as n_1 and n_2 go to infinity, under the null hypothesis, T converges in distribution to $N(0, \kappa - 1)$, where κ is the kurtosis of F_0, that is

$$\kappa = \int y^4 \mathrm{d}F_0(y)$$

is the i-th central moment of F_0. (Hint: you might want to use the fact that $Var\left\{\left(\frac{Y_{ji}-\mu_j}{\sigma_j}\right)^2\right\} = \kappa - 1$)

(c) Using the result from (b), demonstrate that, if the populations have kurtosis greater than 3, comparison of s_1^2/s_2^2 to an F distribution is asymptotically equivalent to comparing an $N(0, \kappa - 1)$ random variable to an $N(0, 2)$ distribution. What would the true asymptotic level of a nominal $\alpha = 0.05$ one-sided test would be if $\kappa = 5$?

(d) The file wage.xlsx contains data on hourly wages for 3296 working individuals. Variable "male" equals 1 for males and 0 for females. Suppose that we would like to test a hypothesis that the population variance of the logarithm of wage for males equals that for females against the alternative that the variance for females is larger than the variance for males.The above discussion suggests that a test robust to non-normality of the population would compare $T/\sqrt{\hat{\kappa} - 1}$ to $N(0, 1)$, where

$$\hat{\kappa} = \frac{(n_1 + n_2) \sum_{j=1}^{2} \sum_{i=1}^{n_j} \left(Y_{ji} - \bar{Y}_j\right)^4}{\left[\sum_{j=1}^{2} \sum_{i=1}^{n_j} \left(Y_{ji} - \bar{Y}_j\right)^2\right]^2}$$

is an estimate of κ. Conduct such a test, then perform the standard (normal theory) test based on s_1^2/s_2^2 and compare the results.

Problem 2. Consider a linear regression model

$$y_i = x_i'\beta + \epsilon_i, \quad i = 1, ..., n.$$

Suppose that the large sample OLS assumptions hold. That is

- (y_i, x_i) are i.i.d. across $i = 1, \ldots, n$
- $E(x_i x_i')$ has full rank
- $E(\epsilon_i | x_i) = 0$
- $Var(\epsilon_i | x_i) = \sigma^2$
- the fourth moments of ϵ_i and of the components of x_i are finite

Consider the ridge regression estimator of β, $\hat{\beta}_r = (X'X + \lambda I_k)^{-1} X'Y$ with fixed $\lambda > 0$.

(a) Is $\hat{\beta}_r$ a conditionally unbiased estimator of β? Is $\hat{\beta}_r$ consistent for β?

(b) Find the asymptotic distribution of $\sqrt{n}(\hat{\beta}_r - \beta)$ as $n \to \infty$.

Problem 3. This problem is based on Hastie et al (2022) "Surprises in High-dimensional Ridgeless Least Squares Interpolation", Annals of Statistics 50, pp.949-986. It is related to a fascinating "double descent" phenomenon in machine learning, recently pointed out by Belkin et al (2019) "Reconciling modern machine-learning practice and the classical bias-variance trade-off" Proc. Matl. Acad. Sci. USA 116.

Let X and ϵ be, respectively, an $n \times p$ matrix and an $n \times 1$ vector with i.i.d. $N(0,1)$ elements. Consider a linear regression model

$$Y = X\beta + \epsilon.$$

Let $\hat{\beta}$ be standard OLS if $n \geq p$. If $n < p$, let us define it as $\hat{\beta} = (X'X)^{+} X'Y$, where $(X'X)^{+}$ is the so called Moore-Penrose pseudo-inverse of $X'X$. The Moore-Penrose pseudo-inverse is defined in terms of the eigenvalue-eigenvector pairs $(\lambda_i, v_i), i = 1, ..., p$ of $X'X$. Note that when $n < p$, $\lambda_{n+1} = ... = \lambda_p = 0$, so that

$$X'X = \lambda_1 v_1 v_1' + ... + \lambda_n v_n v_n'.$$

The Moore-Penrose pseudo-inverse is simply

$$(X'X)^{+} = \frac{1}{\lambda_1} v_1 v_1' + ... + \frac{1}{\lambda_n} v_n v_n'.$$

It can be shown that, if $n < p$, $(X'X)^{+} X' = X'(XX')^{-1}$ so that $\hat{\beta}$ is the minimum ℓ_2 norm least squares estimator derived in class. In particular, $X\hat{\beta}$ exactly equals Y, so that the regression "interpolates" (exactly fits) the data.

We would like to explore the risk $\hat{\beta}$

$$R_X(\hat{\beta}, \beta) = E(\|\hat{\beta} - \beta\|^2 | X) = E((\hat{\beta} - \beta)'(\hat{\beta} - \beta) | X),$$

in the limit as both n and p go to infinity (big data).

(a) Establish the risk decomposition into bias and variance part:

$$R_X(\hat{\beta}, \beta) = \underbrace{\|E(\hat{\beta}|X) - \beta\|^2}_{B_X(\hat{\beta},\beta)} + \underbrace{\text{trace}\left[Var(\hat{\beta}|X)\right]}_{V_X(\hat{\beta},\beta)}.$$

(b) Let $\Pi = v_{n+1} v_{n+1}' + ... + v_p v_p'$ if $p > n$ and $\Pi = 0$ if $p \leq n$. Show that

$$B_X(\hat{\beta}, \beta) = \beta' \Pi \beta = \|\Pi \beta\|^2, \quad \text{and} \quad V_X(\hat{\beta}, \beta) = \text{trace}\left[(X'X)^{+}\right].$$

Note that for $n \geq p$, these are the usual formulas for the squared Euclidean norm of the OLS bias (zero for $n \geq p$) and the trace of the variance of the OLS estimator (when $\sigma^2 = 1$). For $n < p$, vector $\Pi\beta$ is sometimes called the non-identifiable part of β. Why do you think this name is used?

(c) Using large Random Matrix Theory results, it is possible to show that as $n, p \to \infty$ so that $p/n \to \gamma \neq 1$,

$$R_X(\hat{\beta}, \beta) \xrightarrow{p} \begin{cases} \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1 \\ \|\beta\|^2 \frac{\gamma - 1}{\gamma} + \frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases}$$

Verify this result using simulations for $n = 300$ and $p = 100 + 30j$ with $j = 0, 1, 2, ..., 20$. According to your simulations, which part of the above formula for the case $\gamma > 1$ corresponds to bias and which part corresponds to variance?

(d) Suppose that $n = 300$, $p = 200$ and you know Y and X. By the Gauss-Markov theorem, OLS is the best unbiased estimator, so you compute $\hat{\beta}$ to estimate β. Your friend, who is a machine learning geek, suggests that you should, instead, run minimum ℓ_2 norm least squares regression of Y on X and W, where W is an 300×800 matrix of additional regressors with all entries of W being i.i.d. $N(0, 1)$, independent from X and ϵ (so, you suspect that your friend is a lunatic because these additional regressors are clearly rubbish). Using the theoretical formula for $R_X(\hat{\beta}, \beta)$ from (e), compare the risk of your OLS estimator and the estimator proposed by your friend, assuming that $\|\beta\| = 1$. What do you conclude? [Hint: your friend's model is accommodated by the above framework with all the coefficients on W equal to zero, so you can use the formula from (e) for the comparison of the two estimators.] If you do not believe the theoretical formula, do simulations for the comparison.

Problem 4 In the discussion of OLS under serial correlation, we assumed that (y_t, x_t) is strictly stationary. In particular, the variance-covariance matrix of (y_t, x_t) stays constant for $t = 1, 2, ..., T$. Does this mean that we are considering serial correlation without heteroskedasticity? Discuss briefly.

Problem 5 Consider a regression model with only constant as the explanatory variable, that is,

$$y_t = \beta + \varepsilon_t.$$

Suppose that ε_t is serially correlated. Precisely, let it satisfy the autoregression of order 1, AR(1),

$$\varepsilon_t = \rho\varepsilon_{t-1} + \zeta_t, \qquad |\rho| < 1, \qquad \zeta_t \overset{\text{i.i.d.}}{\sim} N(0, 1).$$

(a) Represent ε_t is the infinite moving average, MA(∞) form:

$$\varepsilon = c_0\zeta_t + c_1\zeta_{t-1} + c_2\zeta_{t-2} + ...$$

What is the value of the long-run variance of ε_t?

(b) Let $\hat{\beta}$ be the OLS estimator of β from the regression of y_t, $t = 1, ..., T$ on constant only. What is the asymptotic distribution of $\sqrt{T}(\hat{\beta} - \beta)$ as $T \to \infty$?

(c) Using your favourite computer language/package, simulate $y_1, ..., y_{100}$ for $\beta = 0$ and three choices of ρ: $\rho = 0, 0.5, 0.95$. For each of the obtained three datasets, report the OLS estimate of β and the default values of t-statistics for testing the hypothesis that $\beta = 0$. Briefly discuss.

(d) For each of the three simulated datasets, compute Newey-West standard errors (with $G = 4$) and the corresponding t-statistics. Compare with the t-statistics from (c).

(e) For each of the three simulated datasets, compute the Kiefer-Vogelsang-Bunzel t-statistics based on the fixed-b approach (with b=1).

(f) Simulate Brownian motion (for example, by simulating 1000 observations of random walk) many times (say 2000), so that you have 2000 Brownian motions. Using these simulations, approximate the p-values corresponding to the Kiefer-Vogelsang-Bunzel t-statistic reported in (e). Compare with the default and the Newey-West results.