Ec240a, Fall 2024

Professor Bryan Graham

Problem Set 2

Due: November 15th, 2024

Problem sets are due at 5PM. The GSI will provide instructions on how to turn in your problem set. You may work in groups, but each student should turn in their own write-up (including a "printout" of a narrated/commented and executed Jupyter Notebook if applicable). Please also e-mail a copy of any Jupyter Notebook to the GSI (if applicable).

1 Linear regression

[a] Consider the L^2 vector space $\mathcal{H}$ endowed with the covariance inner product $\langle h_1, h_2 \rangle = \mathbb{E}[h_1 h_2]$ and norm $\|h\| = \sqrt{\mathbb{E}[h^2]}$. Consider the subspace

$$\mathcal{L} = \{X'b \,:\, b \text{ is a } K \times 1 \text{ vector of real numbers}\}$$

and $\mathbb{E}^*[Y|X] = X'\beta_0$ denote the projection of Y onto $\mathcal{L}$. Assume that (i) $\mathbb{E}[Y^2] < \infty$, (ii) $\mathbb{E}\left[\|X\|^2\right] < \infty$ and (iii) $\mathbb{E}[(\alpha' X)^2] > 0$ for any non-zero $\alpha \in \mathbb{R}^K$. Let $X'b$ denote a generic element of $\mathcal{L}$ and $U = Y - X'\beta_0$ the projection error. Show that

$$\mathbb{E}\left[(Y - X'b)^2\right] = \mathbb{E}[U^2] + 2(\beta_0 - b)' \mathbb{E}[XU] + (\beta_0 - b)' \mathbb{E}[XX'](\beta_0 - b). \quad (1)$$

[b] Show that if $\mathbb{E}[XU] = 0$ (you may assume X includes a constant), then

$$\mathbb{E}\left[(Y - X'b)^2\right] \geq \mathbb{E}[U^2]$$

with strict inequality unless $b = \beta_0$. Why is $\mathbb{E}^*[Y|X = x]$ often called the mean squared error minimizing linear predictor of Y given $X = x$?

[c] (PYTHAGOREAN RULE) Show that

$$\mathbb{V}(Y) = \mathbb{V}(Y - \mathbb{E}^*[Y|X]) + \mathbb{V}(\mathbb{E}^*[Y|X]).$$

Interpret this expression?

[d] Let $X_1, ..., X_K$ be a set of regressors with the property that $\mathbb{C}(X_k, X_l) = 0$ for all $k \neq l$. Show that

$$\mathbb{E}^*[Y|X_1, \ldots, X_K] = \sum_{k=1}^{K} \mathbb{E}^*[Y|X_k] - (K-1)\mathbb{E}[Y].$$

HINT: First show that

$$\mathbb{E}^*[\mathbb{E}^*[Y|X_k]|X_l] = \mathbb{E}[Y]$$

for every $k \neq l$. Second verify the orthogonality conditions

$$\mathbb{E}[UX_l] = 0$$

for $U = \left(Y - \sum_{k=1}^{K} \mathbb{E}^* \left[Y | X_k \right] + (K-1) \mathbb{E}[Y]\right)$ and $l = 1, \ldots, K$.

[e] Under the same conditions as in part (c) above show that

$$\mathbb{E}^* \left[Y | X_1, \ldots, X_K \right] = \mathbb{E}[Y] + \sum_{k=1}^{K} \frac{\mathbb{C}(Y, X_k)}{\mathbb{V}(X_k)} (X_k - \mathbb{E}[X_k])$$

and hence that the proportion of variance 'explained' equals

$$1 - \frac{\mathbb{V}(U)}{\mathbb{V}(Y)} = \sum_{k=1}^{K} \rho_k^2$$

for $\rho_k = \frac{\mathbb{C}(Y, X_k)}{\mathbb{V}(X_k)^{1/2} \mathbb{V}(Y)^{1/2}}$.

2 Shrinkage

Let $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ be N independent measurements of the same outcome, each distributed

$$Y_i \sim \mathcal{N}\left(\mu, \sigma_i^2\right).$$

Let $\mathbf{c}$ be an $N \times 1$ vector of constants. Consider estimates of μ in the family

$$\hat{\mu} = \mathbf{c}'\mathbf{Y}. \tag{2}$$

[a] Show that the sample mean $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$ is a member of (2).

[b] Show that the mean squared error minimizing choice of $\mathbf{c}$ is

$$\mathbf{c} = \mu^2 \left(\operatorname{diag}\left\{\sigma_1^2, \ldots, \sigma_N^2\right\} + \mu^2 \iota_N \iota_N'\right)^{-1} \iota_N$$

with ι_N an $N \times 1$ vector of ones and $\operatorname{diag}\left\{\sigma_1^2, \ldots, \sigma_N^2\right\}$ denoting a diagonal matrix.

[c] Further show that the i^{th} element of $\mathbf{c}$ is

$$c_i = \frac{\mu^2}{\sigma_i^2} \frac{\left[\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right]^{-1}}{\left[\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right]^{-1} + \mu^2}.$$

HINT: For A an invertible matrix, u and v column vectors and b a scalar:

$$(A + buv')^{-1} = A^{-1} - \frac{b}{1 + bv'A^{-1}u} A^{-1}uv'A^{-1}.$$

[d] Assume that $\sigma_i^2 = \sigma^2$ for all $i = 1, \ldots, N$. Show that in this case the mean squared error minimizing estimate of μ is

$$\hat{\mu} = \frac{\mu^2}{\frac{\sigma^2}{N} + \mu^2} \bar{Y}$$

Prove that this estimate converges in mean square to μ (and hence also converges in probability).

[e] The estimate in part [d] is infeasible. Assume that σ^2 is known and consider the feasible estimator

$$\hat{\mu} = \left(1 - \frac{\frac{\sigma^2}{N}}{\bar{Y}^2}\right)\bar{Y}.$$

Provide a justification for this estimate. *Argue* that $\hat{\mu} \xrightarrow{p} \mu$. Do you think its mean squared error will be lower than that of the sample mean's in finite samples? Why?

[f] Rebut the assertion that "the sample mean's day has come and gone".

3 Calorie demand

Download the **RPS_calorie_data.out** dataset from the course webpage. For this problem set you will require only two columns of the dataset, those with the headings **Y0tc** and **X0te**. The first equals the log of total calories consumed in a household and the latter equals the log of total expenditure. To learn more about the dataset see Section 4 of Graham & Powell (2012). Subramanian & Deaton (1996) provide more background on calorie demand analysis.

1. Let Y denote log calories and X denote log expenditure. Assume that

$$m(x) = \mathbb{E}[Y|X = x] = \sum_{k=1}^{K} \alpha_k g_k(x)$$

   where $g_k(x) = x^{k-1}$.

2. Using the power series basis described above and the Gram-Schmidt algorithm construct a new basis that is orthogonal to the design points (set $K = 12$). Let W_i denote the $K \times 1$ vector of orthonormal basis functions for the i^{th} household. Compute the least squares fit

$$m(X_i) = W_i'\hat{\theta}$$

   with

$$\hat{\theta} = \left[\sum_{i=1}^{N} W_i W_i'\right]^{-1} \times \left[\sum_{i=1}^{N} W_i Y_i\right].$$

   Plot this function onto a scatter of the unsmoothed data.

3. Now use the shrinkage estimator of Efromovich (1999) as described in lecture to estimate $m(X_i)$. Plot this function onto a scatter of the unsmoothed data. Comment on your findings.

4. Now compute the soft threshold estimate of $m(X_i)$ (as defined above). Plot this function onto a scatter of the unsmoothed data.

References

Efromovich, S. (1999). *Nonparametric Curve Estimation.* New York: Springer.

Graham, B. S. & Powell, J. L. (2012). Identification and estimation of average partial effects in “irregular” correlated random coefficient panel data models. *Econometrica*, 80(5), 2105 – 2152.

Subramanian, S. & Deaton, A. (1996). The demand for food and calories. *Journal of Political Economy*, 104(1), 133 – 162.