Problem Set 2

Econ 4676: Big Data and Machine Learning for Applied Economics

Daniel Gómez Salazar[*], Lucas Gómez Tobón[†], José Daniel Palacio Murillo [‡]

1 Theory Exercises

1. Suppose you have the following spatial model $y = \rho Wy + X\beta + WX\theta + \epsilon$ with $|\rho| < 1$ this is sometimes known as the Spatial Durbin Model

   (a) First consider the following scenario $\beta = \theta = 0$.

      i. **Write the Likelihood function. Can you find a closed form for the parameter estimators? Don't forget to be specific on the assumptions you make.**

$$y = \rho Wy + \epsilon$$

We take into consideration the following asumptions:

- $|\rho| < 1$
- $\epsilon \sim N(0, \sigma^2)$
- w is exogenous

With this being said we can now define y:

$$\begin{aligned} y - \rho Wy &= \epsilon \\ (I_n - \rho W)y &= \epsilon \\ y &= (I_n - \rho W)^{-1}\epsilon \end{aligned}$$

Therefore, the likelihood function can be defined as:

$$L(y) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \cdot |det(v(y))|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(y - E(y))' v(y)^{-1}(y - E(y))\right)$$

[*]d.gomezs@uniandes.edu.co
[†]l.gomezt@uniandes.edu.co
[‡]jd.palacio@uniandes.edu.co

To fin the MLE we need $E[y]$ and $v(y)$:

$$E[y] = E\left[(I_n - \rho W)^{-1} E(\epsilon)\right]$$
$$\text{we assume that } \rho \text{ is given and } \epsilon \overset{i.i.d}{\sim} N(0, \sigma^2)$$
$$\Rightarrow E[y] = 0$$

Now we find $v(y)$

$$v(y) = E[yy'] - E[y]E[y]'$$
$$E[yy'] = E\left[\left(\left(I_n - \rho W\right)^{-1} \epsilon\right) \cdot \left(\left(I_n - \rho W\right)^{-1} \epsilon\right)'\right]$$
$$E[yy'] = E\left[\left(I_n - \rho W\right)^{-1} \epsilon\epsilon' \left(\left(I_n - \rho W\right)^{-1}\right)'\right]$$
$$E[yy'] = E\left[\left(I_n - \rho W\right)^{-1} \left(\left(I_n - \rho W\right)^{-1}\right)' \sigma^2\right]$$
$$E[yy'] = \underbrace{\left[(I_n - \rho W)'(I_n - \rho W)\right]^{-1}}_{\Omega} \sigma^2$$

$$v(y) = \sigma^2 \Omega$$

Now, it is possible to define the likelihood function:

$$L(\rho, \sigma^2, y) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \cdot \left|\sigma^2\Omega\right|^{-\frac{1}{2}} \cdot exp\left(-\frac{1}{2}(y-0)'\left(\sigma^2\Omega\right)^{-1}(y-0)\right)$$

$$L(\rho, \sigma^2, y) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \cdot \left|\sigma^2\Omega\right|^{-\frac{1}{2}} \cdot exp\left(-\frac{1}{2\sigma^2}\left((I_n - \rho W)^{-1}\epsilon\right)'\Omega^{-1}\left((I_n - \rho W)^{-1}\epsilon\right)\right)$$

$$L(\rho, \sigma^2, y) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \cdot \left|\sigma^2\Omega\right|^{-\frac{1}{2}} \cdot$$
$$exp\left(-\frac{1}{2\sigma^2}\left(\epsilon' \underbrace{\left(I_n - \rho W)'\right)^{-1} \left(I_n - \rho W\right)'}_{\text{I}} \underbrace{\left(I_n - \rho W\right)\left(I_n - \rho W\right)^{-1}}_{\text{I}} \epsilon\right)\right)$$

$$L(\rho, \sigma^2, y) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \cdot \left|\sigma^2\Omega\right|^{-\frac{1}{2}} \cdot exp\left(-\frac{1}{2\sigma^2}\epsilon'\epsilon\right)$$

We use the log function:

$$l(\sigma^2, \rho, y) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\left|\sigma^2\Omega\right|\right) - \frac{1}{2\sigma^2}\left((I_n - \rho W)y\right)'\left(I_n - \rho W\right)y$$

note that $|\sigma^2\Omega| = \sigma^{2n}|\Omega|$, also $|\Omega| = |(I_n - \rho W)|^{-2}$

$$l(\rho, \sigma^2, y) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}(n)\ln\left(\sigma^2\right) - \frac{n}{2}(-2)\ln(|I - \rho W|) - \frac{1}{2\sigma^2}\left((I_n - \rho W)y\right)'(I_n - \rho W)y$$

$$l(\rho, \sigma^2, y) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) + n\ln(|I - \rho W|) - \frac{1}{2\sigma^2}\left((I_n - \rho W)y\right)'(I_n - \rho W)y$$

Ord (1975) showed that

$$|I - \rho W| = \prod_{i=1}^{n}(1 - \rho W_i), \text{where } W_i \text{ is the eigenvalue of i.}$$

$$l(\rho, \sigma^2, y) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) + n\Sigma\ln(1 - \rho W_i) - \frac{1}{2\sigma^2}\left((I_n - \rho W)y\right)'(I_n - \rho W)y$$

$$l(\sigma^2, y|\rho) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) + n\Sigma\ln(1 - \rho W_i) - \frac{1}{2\sigma^2}\left((I_n - \rho W)y\right)'(I_n - \rho W)y$$

- $\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}\left((I_n - \rho W)y\right)'(I_n - \rho W)y = 0$

  $$\frac{1}{2}\frac{1}{\sigma^4}\left((I_n - \rho W)y\right)'(I_n - \rho W)y = \frac{n}{2}\frac{1}{\sigma^2}$$

  $$\boxed{\sigma^2(\rho) = \frac{1}{n}\left((I_n - \rho W)y\right)'(I_n - \rho W)y} = \frac{\epsilon'\epsilon}{n}$$

- $\frac{\partial l}{\partial \rho} = n\Sigma\frac{-W_i}{1 - \rho W_i} + \frac{1}{\sigma^2}\left(y'(I - \rho W)'Wy\right) = 0$

Since ρ cannot be derived analytically, ρ mmust be obtained from an explicit maximization of a concentrated log-likelihood function using numerical optimization:

$$l(\rho) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2(\rho)) + n\Sigma\ln(1 - \rho W_i) - \frac{1}{2\sigma^2(\rho)}\left((I_n - \rho W)y\right)'(I_n - \rho W)y$$

$$\sigma^2_{ML}(\hat{\rho}) = \frac{1}{n}\left((I_n - \hat{\rho} W)y\right)'(I_n - \hat{\rho} W)y$$

ii. **Suppose instead you use MCO, would you obtain the same estimates?**

$$y = (I_n - \rho W)^{-1}\epsilon$$

We now minimize the squared error:

$$\min_{\rho} \quad e'e = \left((I_n - \rho W)y\right)'(I_n - \rho W)y$$

Is not possible to fin a closed form for the estimates $\hat{\rho}_{OLS}$ and $\hat{\sigma}^2_{OLS}$. Therefore:

$$\hat{\sigma}^2_{OLS} \neq \hat{\sigma}^2_{ML} \quad ; \quad \hat{\rho}_{OLS} \neq \hat{\rho}_{ML}$$

$$[\rho]: \quad -2y'(I - \hat{\rho}W)'Wy = 0$$

$$y'(I - \hat{\rho}W')Wy = 0$$

$$(y' - \hat{\rho}y'W')Wy = 0$$

$$y'Wy - \hat{\rho}y'W'Wy = 0$$

$$\hat{\rho}y'W'Wy = y'Wy$$

$$\hat{\rho}_{OLS} = y'Wy(y'W'Wy)^{-1}$$

We also have:

$$\hat{\sigma}^2_{OLS} = \frac{e'e}{n-k} = \frac{((I_n - \rho W)y)'\,(I_n - \rho W)y}{n-k} \neq \hat{\sigma}^2_{ML}$$

(b) Now consider that $\rho = 0$, and let's proceed as before:

i. **Write the Likelihood function. Can you find a closed form for the parameter estimators? Don't forget to be specific on the assumptions you make.**

$$y = X\beta + WX\theta + \epsilon; \quad \epsilon \overset{i.i.d}{\sim} N(0, \sigma^2)$$

$$y = z\gamma + \epsilon; \quad z = [X, WX] \text{and} \gamma = [\beta, \theta]$$

Therefore:

$$L(y) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \cdot |det(v(y))|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(y - E(y))'v(y)^{-1}(y - E(y))\right)$$

$$L(y) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \cdot |det(v(y))|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(y - z\gamma)'v(y)^{-1}(y - z\gamma)\right)$$

$$\Rightarrow l(y) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right)) - \frac{1}{2\sigma^2}(y - z\gamma)'(y - z\gamma)$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}(y - z\gamma)'(y - z\gamma) = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}(y - z\gamma)'(y - z\gamma)$$

Concentrated likelihood

$$l^c(y|\gamma, z) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}ln\left(\frac{1}{n}(y - z\gamma)'(y - z\gamma)\right) - \frac{n}{2}$$

$$[\gamma]: \quad \frac{n}{(y-z\hat{\gamma})'(y-z\hat{\gamma})} \cdot z'(y-z\hat{\gamma}) = 0$$

$$z'y = z'z\hat{\gamma}$$

$$\boxed{\hat{\gamma_{ML}} = (z'z)^{-1}(z'y)}$$

$$\boxed{\hat{\sigma_{ML}}^2 = \frac{1}{n}\left(y - z(z'z)^{-1}(z'y)\right)'\left(y - z(z'z)^{-1}(z'y)\right)} = \frac{\epsilon'\epsilon}{n}$$

ii. **Suppose instead you use MCO, would you obtain the same estimates?**

In that case we have:

$$\min_{\gamma} \quad e'e = (y-z\gamma)'(y-z\gamma)$$

$$[\gamma]: \quad -2z'(y-z\gamma) = 0$$

$$\boxed{\hat{\gamma}_{OLS} = (z'z)^{-1}(z'y)} \Rightarrow \quad \hat{\gamma}_{ML} = \hat{\gamma}_{OLS}$$

$$\boxed{\hat{\sigma}^2_{OLS} = \frac{\epsilon'\epsilon}{n-k-1}} \Rightarrow \quad \hat{\sigma}^2_{ML} \neq \hat{\sigma}^2_{OLS}$$

And it can be proved as it follows:

$$y = z\gamma + \epsilon \Rightarrow \quad \epsilon = y - z\gamma$$

$$\hat{\epsilon} = y - z\hat{\gamma} \Rightarrow \quad \hat{\epsilon} = y - z(z'z)^{-1}z'y \Rightarrow \quad \hat{\epsilon} = (I - z(z'z^{-1}z)y)$$

$$\hat{\epsilon} = My, \text{where } M \text{ is an idempotent matrix}$$

$$var(\epsilon) = E\left[(\epsilon - E[\epsilon])'(\epsilon - E[\epsilon])\right] = E[\epsilon'\epsilon], \text{since } E(\epsilon) = 0$$

$$E[\epsilon'\epsilon|z] = E[y'M'My|z] = E[y'My|z]$$

The scalar $\epsilon'M\epsilon$ is a $1x1$ matrix, so its equal to its trace. By using the result on cyclic perutations

$$E[tr(\epsilon'M\epsilon)|z] = E[tr(M\epsilon\epsilon')|z]$$

Since M is function of z:

$$tr\left(E[\epsilon\epsilon'|z]\right) = tr(M\sigma^2 I) = \sigma^2 tr(M)$$

$$tr\left((I - z(z'z)^{-1}z)\right) = tr(I) - tr(z(z'z)^{-1}z) = n-k-1$$

$$E[\epsilon'\epsilon|z] = (n-k-1)\sigma^2$$

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{\epsilon'\epsilon}{n-k-1}$$

2. Consider the regression model $y = X\beta + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I)$ furthermore assume that β has a normal prior, i.e. $\beta \sim N(0, \tau^2 I)$.

   (a) **Find the posterior distribution.**

   The posterior distribution is built as it follows:

$$\pi(\beta|y, X) = \frac{f(y, X|\beta)p(\beta)}{m(y, X)}$$

$$\pi(\beta|y, X) = \frac{f(y|X, \beta)f(X|\beta)p(\beta)}{m(y, X)}$$

   With the asumption that $f(X|\beta) = f(x)$:

$$\pi(\beta|y, X) = f(y|X, \beta)p(\beta)\frac{f(X)}{m(y, X)}$$

$$\pi(\beta|y, X) \propto \underbrace{f(y|X, \beta)}_{\text{likelihood}}\underbrace{p(\beta)}_{\text{prior}}$$

   We have the following distribution for y:

$$L(y|\beta, \sigma^2, X) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(-\frac{1}{2\sigma^2}(y_i - x_i\beta)\right)\right)$$

$$L(y|\beta, \sigma^2, X) = (2\pi\sigma^2)^{\frac{n}{2}}exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right)$$

   We also have the prior:

$$p(\beta) = \left(2\pi\tau^2\right)^{-\frac{1}{2}}exp\left(-\frac{1}{2\tau^2}(\beta'\beta)\right)$$

   Which in turn gives the posterior distribution for β:

$$\pi(\beta|y, X) \propto (2\pi\sigma^2)^{\frac{n}{2}}exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right)\left(2\pi\tau^2\right)^{-\frac{1}{2}}exp\left(-\frac{1}{2\tau^2}(\beta'\beta)\right)$$

$$\pi(\beta|y, X) \propto exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}(y - X\beta)'(y - X\beta) + \frac{1}{\tau^2}\beta'\beta\right)\right)$$

   With this given, we can find the following distribution of the posterior:

$$\Sigma = \left(\frac{1}{\sigma^2}X'X + \frac{1}{T}\right), \text{where } T = \tau^2 I$$

$$\mu = \frac{1}{\sigma^2}\Sigma X'y \Rightarrow \mu = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X'X + \frac{1}{T}\right)X'y$$

Implying:

$$\beta|y, X \sim N(\mu, \Sigma)$$

(b) **Compare it with the ridge formula we saw in class.**

The following formula is the ridge formula for $\hat{\beta}$:

$$\hat{\beta}_R = (X'X + \lambda I)X'y$$

Which is similar to the estimation using a Bayesian estimation approach:

$$\hat{\beta}_{Bayesian} = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X'X + \frac{1}{T}\right)X'y$$

Their main differences lie in the inclusion of both variance parameters within the Bayesian estimation unlike the Ridge estimator which only includes the Lagrange multiplier.

(c) **What is the relationship between λ in the ridge model and σ^2 and τ^2?**

As we see in the previous point, the Bayesian estimation method allows us to see how the shrinkage of the estimation towards the prior or the data depends on the variance observed in the data σ^2 and the variance of the prior τ^2.

$$\lambda \approx \frac{\sigma^2}{\tau^2}$$

3. **Centered Ridge. Suppose that $\bar{x} = 0$, i.e. the data has been centered. Show that the parameters that minimize $R(\beta, \beta_0) = (y - X\beta - \beta_0\iota)'(y - X\beta - \beta_0\iota) + \lambda\beta'\beta$ are $\beta_0 = \bar{y}$ and $\beta = (X'X + \lambda I)^{-1}X'y$**

We define the model as it follows:

$$y = \beta_0\iota + X\beta + \varepsilon$$

where ι is a n-vector of 1s and X is an $n \times (k-1)$ matrix of observations. This specification would take the variables to the mean, but since the mean for each $x = 0$ it doesn't change the specification nor we need to use the M_ι matrix.

To find the estimator one should solve the least squares problem subject to the restriction in which $\beta'\beta = 0$, which can be expressed as a Lagrangian:

$$\left(\hat{\beta}_0, \hat{\beta}\right) = \underset{\beta_0,\beta}{argmin}\left[(y - X\beta - \beta_o\iota)'(y - X\beta - \beta_o\iota) + \lambda(\beta'\beta)\right]$$

We expand the equation:

$$\left(\hat{\beta}_0, \hat{\beta}\right) = \underset{\beta_0,\beta}{argmin}\left[y'y - y'X\beta - y'\beta_o\iota - \beta'X'y + \beta'X'X\beta - \iota'\beta_0'y + \iota'\beta_0'\beta_0\iota + \lambda\beta'\beta\right]$$

Now we find the first order conditions that allow us to find the estimators:

$$[\beta_0]: \quad -2\iota'y + 2\iota'\iota\beta_0 = 0$$

$$\iota'\iota\beta_0 = \iota'y$$

where $\iota'\iota' = 1$ and $\iota'y$ calculates the mean of the variable y:

$$\hat{\beta}_0 = \bar{y}$$

$$[\beta]: \quad -2X'y + 2(X'X + \lambda I)\beta = 0$$
$$(X'X + \lambda I)\beta = X'y$$
$$\hat{\beta} = (X'X + \lambda I)^{-1}X'y$$

4. **Suppose that we have the following regression model $y = X\beta + \epsilon$, and decide to do the following: Augment the centered matrix X with p additional rows with $\sqrt{\lambda}$, and augment y with zeros. Show that this procedures renders the ridge regression estimates, is there a link to the leverage statistic?**

We define the model as it follows in order to find the estimator β:

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} X \\ \sqrt{\lambda} \end{pmatrix}\beta + \epsilon \tag{1}$$

The following equation shows the related estimator for the model:

$$\hat{\beta} = \left(\begin{pmatrix} X & \sqrt{\lambda} \end{pmatrix}\begin{pmatrix} X \\ \sqrt{\lambda} \end{pmatrix}\right)^{-1}\left(\begin{pmatrix} X & \sqrt{\lambda} \end{pmatrix}\begin{pmatrix} y \\ 0 \end{pmatrix}\right) \tag{2}$$

$$\hat{\beta} = (X'X + \lambda I)^{-1}X'y$$

One can also take into consideration the H matrix also called the projection matrix which contains in the diagonal the λ used to construct the ridge estimator, showing the connection between the leverage statistic and the ridge regression.

$$H = X(X'X)^{-1}X' = \begin{pmatrix} X \\ \sqrt{\lambda} \end{pmatrix}(X'X + \lambda I)^{-1}\begin{pmatrix} X & \sqrt{\lambda} \end{pmatrix}$$

leading to the $\hat{y}$:

$$\hat{y} = \begin{pmatrix} X \\ \sqrt{\lambda} \end{pmatrix}(X'X + \lambda I)X'y$$

Therefore we can assume that the ridge estimator follows the same procedure as the leverage statistic in which we want to use the sensibility of the data to each observation in contrast to the ridge intuition that wants to reduce the overfitting through introducing new information to the model.

5. **Reducing elastic net to lasso. Suppose that you have the following functions $EL(\beta) = (y - X\beta)^2 + \lambda_2\beta^2 + \lambda_1|\beta|$ and $L(\beta) = (\tilde{y} - \tilde{X}\beta)^2 + c\lambda_1|\beta|$ where $c = (1+\lambda_2)^{\frac{-1}{2}}$ show that these two problems are equivalent when $\tilde{y}$ and $\tilde{X}$ are the augmented data versions of the previous exercise.**

The estimator for the reduced version of the problem is a lasso estimator which follows the form:

$$\hat{\beta} = \hat{\beta}_{Pseudo-OLS} - \frac{c\lambda_1}{2}$$

However the $\hat{\beta}_{Pseudo-OLS}$ is the same we had in the previous point, a ridge estimator thanks to the augmented form of both y and X.

$$\hat{\beta} = (X'X - \lambda_2 I)X'y - \frac{c\lambda_1}{2}$$

Or which can be rewritten as

$$\hat{\beta} = \frac{\hat{\beta}_{OLS}}{1+\lambda_2} - \frac{c\lambda_1}{2} = \frac{1}{1+\lambda_2}\left(\hat{\beta}_{OLS} - \frac{\lambda_1}{2}\right)$$

Now that we have the estimator we can compare it to the elastic net one:

$$\hat{\beta}_{EL} = \frac{\left(\hat{\beta}_{OLS} - \frac{\lambda_1}{2}\right)}{1+\lambda_2}$$

We can then conclude that both of this approaches yield the same result, therefore reducing the elastic net to lasso.