Introduction to Econometrics: Recitation 11

Seung-hun Lee

1 Big Data

Big data can mean so many things: it can simply mean data with large observations or large number of control variables. In general instances, atypical data such as text data, and satellite imagery are referred to as big data. In this class, we focus on the instance where there are many control variables, specifically on what to do if there are many control variables relative to the number of observations. Additionally, we focus on trying to get the best way to predict a result that is currently not in the dataset.

1.1 Prediction with many X's: Why we don't want too many X's

In this section, we will focus on the issue of **shrinkage**. I will introduce Ridge estimation, LASSO estimation, and principal component method. Before introducing these models, I will talk about the regression framework that will be used throughout the discussion

$$Y_i^* = \beta_1 X_{1i}^* + ... + \beta_k X_{ki}^* + u_i^*$$

where X_{1i}^* is the "standardized" version of X_{1i}'s, defined as

$$X_{1i}^* = \frac{X_{1i} - \bar{X}_1}{\sigma_{X_1}}$$

and we define Y_i^* as $Y_i - \bar{Y}$ - the "demeaned" version. Note that we CANNOT have a constant term β_0 here because if we demean β_0, which is the same for all i's, they vanish.

We then define the mean squared prediction error, or MSPE as the expected value of the squared error made by predicting Y for an observation not in the dataset. In maths,

$$MSPE = E[Y^{OS} - \hat{Y}^{OS}]^2$$

where the $\hat{Y}^{OS}$ is obtained from the coefficients of β's made from the in-sample prediction

$$\hat{Y}_i^{OS} = \hat{\beta}_1 X_{1i}^{OS} + ... + \hat{\beta}_k X_{ki}^{OS}$$

and the Y^{OS} is the realized value of Y outside of the sample

$$Y_i^{OS} = \beta_1 X_{1i}^{OS} + ... + \beta_k X_{ki}^{OS} + u_i^{OS}$$

With these notations, we can define the prediction error as

$$Y_i^{OS} - \hat{Y}_i^{OS} = (\beta_1 - \hat{\beta}_1) X_{1i}^{OS} + ... + (\beta_k - \hat{\beta}_k) X_{ki}^{OS} + u_i^{OS}$$

Define $\sigma_u^2 = E[u_i^{OS}]^2$, then we can write MSPE as

$$MSPE = \sigma_u^2 + E[(\beta_1 - \hat{\beta}_1) X_{1i}^{OS} + ... + (\beta_k - \hat{\beta}_k) X_{ki}^{OS}]$$

Ideally, the smallest possible MSPE would be σ_u^2, referred to as the **oracle prediction**. Which happens if your predicted $\hat{\beta}$'s match the actual β. Unfortunately, we cannot predict β's perfectly. and the more predictors we have, we generally end up having larger MSPE. As such having a large set of X's will not be a good idea (Thus, the **problem of many predictors**). This is where Ridge, LASSO, and Principal Component method comes in.

1.2 Big data methods

The goal here is to find other estimators that does not increase the MSPE compared to the rate in which the same rises in OLS estimators. The idea here is to reduce the σ_u^2, the variance from the residual sums of squares, at the expense of introducing a small bit of bias. We do this by providing a penalty for having a model with large number of regressors (what we formally call 'shrinkage'). All of the methods I introduced above involve these tricks.

1.3 Ridge regression

The Ridge regression minimizes a 'penalized' sum squared of residuals in the following sense

$$\hat{\beta}_{Ridge} = \arg\min_{\beta_1,..,\beta_k} \left[\sum_{i=1}^{n} (Y_i - \beta_1 X_{1i} - .... - \beta_k X_{ki})^2 + \lambda_{Ridge} \sum_{j=1}^{k} \beta_j^2 \right]$$

Here, the last term on the right hand side is the penalty we are introducing. Note that for any term in which β_j is not zero, the penalty term is also nonzero. Also note that we are introducing some degree of bias (compared to OLS). However, the gain is that the variance term σ_u^2 will be reduced. Lastly, you need to capture that the variance and the bias that determines the MSPE moves in a trade-off relation - If you want to decrease one component, you increase the other (but not at a 1-for-1 rate). The Ridge estimator that comes from here minimizes MSPE by working through reducing the variance term to the extent that the bias term does not rise too drastically.

1.4 LASSO regression

LASSO regression also minimizes a penalized sum squared residuals. The difference is that the penalty term takes a different form:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta_1,...,\beta_k} \left[\sum_{i=1}^{n} (Y_i - \beta_1 X_{1i} - .... - \beta_k X_{ki})^2 + \lambda_{LASSO} \sum_{j=1}^{k} |\beta_j| \right]$$

In essence, the idea is similar with the Ridge regression - you find an estimator that minimizes MSPE by reducing σ_u^2 at the expense of some amount of increase in the bias.

The difference between the two lies in the degree of shrinkage. The difference is stark when the OLS estimates are virtually close to zero. When the OLS estimates are small, the LASSO shrinks those estimates all the way to 0. While Ridge also shrinks those coefficients close to 0, they do not exactly set them to 0.

1.5 Principal component

Say that you have a giant k number of regressors. In the principal component method, you are using a linear combination of some subset of k variables. At the end, you end up using $p < k$ number of regressors and effectively 'collapsing' the model to a smaller dimension.

So how do you determine the right linear combinations? For a two-variable case, the linear combination looks something like

$$aX_1 + bX_2$$

where a, b are some real numbers. The optimal value of a and b is determined by

$$\max var(aX_1 + bX_2) \text{ s.t. } a^2 + b^2 = 1$$

Note that we want to solve the *maximization* problem in this case as we want the X's to explain more of the variation that we observe in the data. You can understand $a^2 + b^2 = 1$ as a regularization method.

So for a general case, you would be solving the following problem to get the j'th principal component PC_j

$$\max var\left(\sum_{i=1}^{K} a_{ji}X_i\right) \text{ s.t. } \sum_{i=1}^{k} a_{ji}^2 = 1$$

with another condition being that $corr(PC_j, PC_{j-1}) = 0$. This additional condition is here because we want to minimize the overlapping amount of information across different principal components.

So we still need to determine how many principal component to use. Scree plot gives you an idea about how much each new principal component adds to the variation that is explained by additional principal component. However, this does not give rigorous cutoff. For formal approach, m-fold cross validation approach can be used. You split the sample into m subsets of equal size. Then, one of them becomes your 'test' sample and the rest becomes an out-sample. You will derive a first estimate of MSPE. Repeat this for each of the m subsets until you get m estimates of MSPE. The right number of principal components minimizes the averages of these MSPEs. Note that determining the λ's in LASSO and Ridge can also be done by using the m-fold cross validation.

2 Time series estimation

Now we shift our attention to the **time series** data, where we collect data on same observational unit i for multiple time periods. Primary uses for time series include forecasting, modeling risks, and analyzing dynamic causal effects. Time series differs from previous models in that errors are likely to be autocorrelated and thus require different ways to calculate the standard error. Moreover, we introduce time lagged variables into the independent variables.

Some terms need to be cleared out. Let Y_t be the time series data captured at certain period t - GDP, for instance. **Lags** are characterized as Y_{t-1} and **leads** are defined as Y_{t+1}. You will also be seeing a lot of $\Delta Y_t \equiv Y_t - Y_{t-1}$, which is the **first difference** at time t.

2.1 AR and ADL Models

We first consider the pth order **autoregressive model**, or AR(p) for short. This class of models refers to any model in which Y_t is regressed against its own lagged values. They take the following form

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} + u_t \tag{2}$$

each coefficients in front of $Y_{t-j}, j \in \{1, .., p\}$ indicate whether they are useful for forecasting future values of Y_t. To see if they are individually useful for forecasting purposes, we test whether $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$.

To simply the discussion of the properties of AR(p) models, we restrict to the case where $p = 1$. I first focus on the difference between 'predicted' values and the 'forecasts'. Predicted values refer to in-sample fitting. Forecasts focus on future values which are out of the sample. Suppose we have data of up to T periods. Suppose we want to predict the value of Y at period $T + 1$, which is out of sample. Keeping the following notation in mind would be useful

- Y_{T+1} : True value of Y at $T + 1$
- $\hat{Y}_{T+1|T}$: Forecast of Y_{T+1} based on the estimated coefficients obtained from regressing data that we have up to period T. That is, $\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T$
- $Y_{T+1|T}$: Forecast of Y_{T+1} based on the population coefficients up to period T. That is, $Y_{T+1|T} = \beta_0 + \beta_1 Y_T$

Then we are ready to define the **forecast error**. Forecast error refers to how much $\hat{Y}_{T+1|T}$ deviates from the actual Y_{T+1} value, or $Y_{T+1} - \hat{Y}_{T+1|T}$. The best forecast is the one that minimizes the forecast error.

We can conduct a time series regression where Y_{t-j} are not the only independent variable. For a better forecast, it makes sense to control for other variables, denoted as X, that captures the movement of Y. Such models are called **autoregressive distributed lag model**, or ADL(p, q) for short. They have p lags of dependent variable and q lags for X variable. It takes the following form:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} + \delta_1 X_{t-1} + ... + \delta_q X_{t-q} + u_t \tag{3}$$

ADL models are similar in form to AR models, except that we include lagged variables for X.

So what is the right amount of lags for Y and X variables? One way to answer this is to check for the various values of the **information criteria**. There are two types

$$AIC : \ln\left(\frac{SSR(p,q)}{T}\right) + (K)\frac{2}{T}$$
$$BIC : \ln\left(\frac{SSR(p,q)}{T}\right) + (K)\frac{\ln T}{T}$$

where $K = 1 + p + q$. We want to find the right combination of (p, q) that minimizes the two criteria above. Increasing p and q decreases $SSR(p, q)$ as more of the variation is explained by the right hand side variables. How much it decreases can differ depending on the combination of p and q (Trial and error!). However, any increase in p and q raises the rightmost terms in the above two information criteria. Lastly, BIC usually penalizes complexity higher than AIC since $2 < \ln T$ as $T \to \infty$.

To see whether X's help us predict Y, we use the **Granger causality test**. It is a joint hypothesis test that states, in the null, that none of the X's are a useful predictor above and beyond what is predicted by lagged Y's. In equation (3), the hypotheses can be written as

$$H_0 : \delta_1 = ... = \delta_q = 0, \;\; H_1 : \neg H_0$$

If the null hypothesis is rejected, we say that X *Granger-causes* Y. Note that this refers to the idea that X contains information that help us (marginally) predict Y

(or the idea of correlation). It does not imply that X causes Y, as regression usually captures the idea of causality[1].

2.2 Stationarity

Ideally, we want the distribution of of time series to be **stationary**. We say $(Y_{t+1}, .., Y_{t+s})$ is stationary if the distribution of $(Y_{t+1}, .., Y_{t+s})$ does not depend on t. In other words, the distribution of Y does not change over time. This is when we can use an in-sample data for an accurate analysis of prediction and dynamic causal relations. When there is a trend or a break in the movement of the data (or any change in underlying parameters), the data is nonstationary. If the data is nonstationary, we need to detrend the data.

A trend is defined as a persistent, long-term movement in the data. There are two types of trends. **Deterministic trend** is a nonrandom function of time, ($Y_t = \alpha t^2$). **Stochastic trend** is random, and time-variant. A random walk defined as $Y_t = Y_{t-1} + u_t$ is a type of stochastic trend. In this particular setting, the best predictor of Y_{T+1} given data at T is Y_T itself. Moreover, we can check that $var(u_t)$ is

$$Y_t = Y_{t-1} + u_t = Y_{t-2} + u_{t-1} + u_t = ... = Y_0 + u_1 + ... + u_t$$

If $Y_0 = 0$, then assuming u_t has no serial correlation, we get that $var(Y_t) = t\sigma_u^2$. Therefore, Y_t is not stationary. However, if you difference Y_t you will be able to see that the above process is stationary. This is because

$$\Delta Y_t = Y_{t-1} - Y_{t-1} + u_t = u_t$$

and $E[\Delta Y_t] = 0$.

One of the most well-known nonstationary time series is a random walk process. A random walk is equal to AR(1) of $Y_t = \beta_1 Y_{t-1} + u_t$ with $\beta_1 = 1$. Also, for $\beta_1 > 1$ we have a case where Y_t values blow up and not become stationary.

So how do we test for stationarity? Informally, we can always do an 'eyeball' test: In a graph, do you see some sort of trend - say upward or downward - across

[1]Clive Granger (Nobel Prize in Economics winner at the year 2003), who came up with Granger causality test, argued that causality in economics could be tested for by measuring the ability to predict the future values of a time series using prior values of another time series. However, idea of "true causality" is still being contested and the above statement could fall into *post hoc ergo propter hoc* fallacy of assuming that one thing preceding another can be used as a proof of causation.

long span of time? For a formal test of stationarity, we need to use a Dickey-Fuller test. We go back to our random walk to find out how. Suppose we have $Y_t = \beta_1 Y_{t-1} + u_t$. After multiple iterations, we can re-write Y_t as

$$Y_t = \beta_1^T Y_{t-T} + \beta_1^{T-1} u_{t-T+1} + ... + \beta_1 u_{t-1} + u_t$$

If it is the case that $|\beta_1| < 1$, then Y_{t-T} has no long term effect. If otherwise, they do. So to check whether the data is stationary, we see if there exists a **unit root** or not. In a generalized $AR(p)$ setup, we have

$$\begin{aligned}
Y_t &= \beta_0 + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} + u_t \\
&= \beta_0 + \beta_1 Y_{t-1} - \beta_2 Y_{t-1} + \beta_2 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + u_t \\
&= \beta_0 + (\beta_1 + \beta_2) Y_{t-1} - \beta_2 \Delta Y_{t-1} - \beta_3 Y_{t-2} - \beta_3 Y_{t-1} + \beta_3 Y_{t-1} + \beta_3 Y_{t-2} + \beta_3 Y_{t-3} + ... \\
&= \beta_0 + (\beta_1 + \beta_2 + \beta_3) Y_{t-1} - (\beta_2 + \beta_3) \Delta Y_{t-1} - \beta_3 \Delta Y_{t-2} + .. + \beta_p Y_{t-p} + u_t \\
&= \beta_0 + (\beta_1 + ... + \beta_p) Y_{t-1} - (\beta_2 + ... + \beta_p) \Delta Y_{t-1} - ... - \beta_p \Delta Y_{t-p+1} + u_t
\end{aligned}$$

By further subtracting Y_{t-1} from both sides, we get

$$\Delta Y_t = \beta_0 + (\beta_1 +, .. + \beta_p - 1) Y_{t-1} + ... + u_t = \beta_0 + \delta Y_{t-1} + ... + u_t$$

where $\delta = \beta_1 +, .. + \beta_p - 1$. To check for stationarity, we set the hypothesis as

$$H_0 : \delta \geq 0 \ (\beta_1 + ... + \beta_p \geq 1), \ H_1 : \delta < 0$$

If there is a unit root in the model, then $\delta = 0$ and the data is nonstationary. Dickey Fuller test uses this idea to check for the existence of stationarity in the data. In a simple $AR(1)$, this reduces down to

$$H_0 : \beta_1 \geq 1, \ H_1 : \beta_1 < 1$$

Note that the critical value of this test differs depending on whether we are assuming a drift (intercept only) or drift & time trend (intercept and a term that is a function of t).

Another pattern in which the stationarity can be broken is through a structural break. Assume a ADL(1,1) structure, but that we know when the structural break occurs (e.g. Oil shock and the oil price). Suppose that the shock happens at year

τ Let $D_t(\tau) = 1$ if year $t \geq \tau$ and 0 otherwise. Then we write the equation as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \gamma_1 D_t(\tau) + \gamma_2 D_t(\tau) Y_{t-1} + \gamma_3 D_t(\tau) X_{t-1} + u_t$$

To check for the existence of a structural break, all we need is a joint hypothesis of the following form:

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0,\ H_1 : \neg H_0$$

This is the idea behind the **Chow test**. If the data of the structural break is unknown, we can do a **Quandt Likelihood Ratio test**, which effectively carries out multiple Chow tests and finds the time point where structural break most likely happened, if it occurred.

2.3 Dynamic causal analysis

We are now interested in whether our independent variable X has cause changes in Y over time. **Dynamic causal effect** is intended to capture the effect of X on Y over time. To analyze this, **Distributed lag model** comes in handy. They can be written as

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + ... + \beta_p X_{t-p} + u_t$$

The interpretation of the coefficients are as follows: β_0 captures the contemporaneous impact of X on Y, holding past values of X constant. $\beta_j, j \in [1, p]$ captures the impact of X from j period(s) ago on Y, holding X from other periods constant. We can analyze the cumulative effect by summing over multiple β's. If we are interested in the effect of X 1 period ago, we can check $\beta_0 + \beta_1$ and so on. Specifically, we can write

$$\begin{aligned} Y_t &= \alpha + \beta_0 X_t + \beta_1 X_{t-1} + u_t \\ &= \alpha + \beta_0 X_t - \beta_0 X_{t-1} + \beta_0 X_{t-1} + \beta_1 X_{t-1} + u_t \\ &= \alpha + \beta_0 \Delta X_t + (\beta_0 + \beta_1) X_{t-1} + u_t \end{aligned}$$

However, for this to be accurately measured, we need following assumptions.

- (Sequential) Exogeneity: $E[u_t | X_t, X_{t-1}, ..., X_1] = 0$. Or that error terms should not be correlated with current and past values of X
- Stationarity: Y and X should have stationary distributions and (Y_t, X_t) and (Y_{t-j}, X_{t-j}) becomes independent as j gets large.

- Y and X has nonzero finite moments
- There is no perfect multicollinearity

Key focus is on the first assumption. Sequential exogeneity implies that the error term should not be correlated with independent variables - past and present. If otherwise, the estimates of β's will be biased. There is a more stronger assumption that this, called **strict exogeneity**, that states that the error term should not be correlated with X for not only past and present, but also the future. In the above, setup, sequential exogeneity is sufficient.

In addition, we need to consider a different type of standard errors. Given that there is a possibility that autocorrelation can exist, we need a standard error that takes into account autocorrelation and heteroskedasticity. This is known as **heteroskedasticity and autocorrelation consistent** errors (HAC errors). The take-away is that standard errors in the typical STATA output can be wrong and we need to take a slightly different approach.