Recitation 11

Seung-hun Lee

Columbia University

Big data

What is Big data?

- 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.

Characterizing MSPE

Mean squared prediction error

- Regression format is

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

- X_{1i}^* is the "standardized" version of X_{1i}'s, Y_i^* 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.
- MSPE: the expected value of the squared error made by predicting Y for an observation not in the dataset.

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

  - $\hat{Y}^{OS}$: obtained from the coefficients of β's made from the in-sample
  - Y^{OS} : realized value of Y outside of the sample

Characterizing MSPE

Mean squared prediction error

- 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

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

- Oracle prediction: the smallest possible MSPE, σ_u^2
- However, we cannot predict β's perfectly.
- The more predictors we have, we generally end up having larger MSPE $\rightarrow$ need to reduce X's.
- Need Ridge, LASSO, and Principal Component method comes in

Methods

In principle

- Goal: Find other estimators that does not increase the MSPE compared to the rate in which the same rises in OLS estimators.
- Idea: Reduce the σ_u^2, the variance from the residual sums of squares, at the expense of introducing a small bit of bias.
- How: Providing a penalty for having a model with large number of regressors (what we formally call 'shrinkage').

Ridge

Estimation

- Minimize a 'penalized' sum squared of residuals

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

- $\lambda_{Ridge} \sum_{j=1}^{k} \beta_j^2$: penalty for complexity.
- Introduce bias so that the variance term σ_u^2 will be reduced.
- Variance and the bias in MSPE moves in a trade-off relation
- Ridge estimator minimizes MSPE by reducing the variance term to the extent that the bias term does not rise too drastically.

LASSO

Estimation

- Penalty term takes a different form:

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

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

Principal component

Estimation

- You are using a linear combination of some subset of k variables so that you end up with $p < k$ number of regressors ('collapsing' the model)
- Solve 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 $corr(PC_j, PC_{j-1}) = 0$
- Solve the *maximization*: We want the X's to explain more of the variation
- $\sum_{i=1}^{k} a_{ji}^2 = 1$: Regularization method
- $corr(PC_j, PC_{j-1}) = 0$: We want to minimize the overlapping amount of information across different principal components.

m-fold cross validation

Idea

- Goal: Select the optimal λ penalty parameters for Ridge/LASSO and find right amount of principal components
- 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 until you get m estimates of MSPE.
- The right parameter values minimizes the averages of these MSPEs.

Time series

Setup

- Collect data on same observational unit i for multiple time periods.
- Primary uses: Forecasting, modeling risks, and analyzing dynamic causal effects
- Time series differs in that errors are likely to be autocorrelated and thus require different ways to calculate the standard error.
- Let Y_t be the time series data captured at certain period t - GDP
- **Lags** are characterized as Y_{t-1} and **leads** are defined as Y_{t+1}.
- $\Delta Y_t \equiv Y_t - Y_{t-1}$: The **first difference** at time t.

AR and ADL

Model

- $AR(p)$: Y_t is regressed against its own lagged values by p times:

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

  - Each coefficient β_k indicates how past values are useful in forecasting
- $ADL(p, q)$: p lags of dependent variable and q lags for X variable

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

AR and ADL

Model

- Right amount of p and q minimizes the following **information criteria**

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

  where $K = 1 + p + q$

- **Granger causality**: Test that helps us see whether X is useful in predicting Y

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

  If the null hypothesis is rejected, we say that X *Granger-causes* Y

Stationarity

Idea

- Stationary: 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
- Nonstationary: When there is a trend or a break in the movement of the data (or any change in underlying parameters),
- Trends
  - **Deterministic trend** is a nonrandom function of time, ($Y_t = \alpha t^2$)
  - **Stochastic trend** is random, and time-variant distribution, such as the random walk $Y_t = Y_{t-1} + u_t$ (You can check that $var(Y_t) = t\sigma_u^2$)
  - Any other case where $\beta_1 > 1$ is also nonstationary

Stationarity

Testing for this in AR(1)

- Eyeball test: Check graphically
- Dickey-Fuller test: Check for the existence of a ‘unit root’ by testing

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

- See notes to have an idea of what to do in an $AR(p)$ case

Stationarity

Testing structural breaks

- Assume a ADL(1,1) structure, but that we know when the structural break occurs 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 structural break, test 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 structural break is unknown, we can do a **Quandt Likelihood Ratio test** that implements multiple Chow tests and finds the point where structural break most likely happened, if it occurred.

Dynamic Causal analysis

Causal analysis

- **Dynamic causal effect** captures the effect of X on Y over time.
- Write the distributed lag model as

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

- 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

Dynamic Causal analysis

Causal analysis

- Cumulative effect: Cumulative effect can be captured by summing over multiple β's
- Specifically, we can write

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

- 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

Dynamic Causal analysis

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 takeaway is that standard errors in the typical STATA output can be wrong and we need to take a slightly different approach.
- Use `newey` in STATA