Econometrics Notesheet

Michael B. Nattinger

May 5, 2021

1 GMM

- Given moment equations $E[g_i(\beta)] = 0$, where β is k dimensional and g is l dimensional, if $l = k$ then we have exact identification from the method of moments estimator: $\frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta}_{mm}) = 0$.
- If $l > k$ then we have overidentification. We need the sample moment equations g to be small in some sense. We define the criterion as follows:
- $J(\beta) = n\bar{g}_n(\beta)'W\bar{g}_n(\beta)$. $\hat{\beta}_{gmm} = \arg\min_{\beta} J(\beta)$.
- In general the efficient GMM uses the weight matrix Ω^{-1} where $\Omega = E[g_i(\beta)g_i(\beta)']$
- For IV go straight to the textbook starting at pdf page 435.
- For Wald testing see pdf page 441.
- For restricted GMM see pdf page 442-4.
- Best way to conduct inference of a restriction is via the distance test. See 445-447.
- A separate test is overidentification test. This tests if the assumptions of the model are valid. See page 447.
- To conduct inference via bootstrapping one needs to recenter s.t. the moment equations have value zero. See 451.

2 DiD

- This is just common sense so I'm not going to write notes.

3 Non-parametric regression

- Binned means estimator: Take mean of observations within each bin.
- Kernel regression: essentially just a generalized binned mean estimator, with observations near a point weighted more highly.
- Nadaraya-Watson estimator is the following: $\hat{m}_{nw}(X) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}$
- Local linear is a further generalization. Estimate WLS for each point.
- Smoothing bias is a concern. Boundary bias is a concern for NW, less so for LL.

- Asymptotic MSE and asymptotic integrated MSE are on pdf page 694. Can take FOC wrt h to solve for optimal bandwidth.
- the Epanechnikov kernel minimizes AIMSE. The efficiency loss of using other kernels are really small though. Gaussian in a sense is actually better owing to its smoothness.
- Generally although there is an 'optimal' bandwidth in theory, in practice it is a better idea to use CV to choose.

4 Series regression

- Another way to approximate a (possibly) nonlinear conditional expectation function is to make a series expansion. Estimate via OLS.
- Generally we use a quardratic expansion of order p, then this has number of parameters $K = p + 1$ due to the constant.
- Can also use splines. These have join points called knots.
- Linear: $m_k(x) = \beta_0 + \beta_1 x + \beta_2(x - \tau)1\{x \geq \tau\}$.
- Linear is continuous, quadratic has continuous first derivative, etc.
- Variance formulas are pdf pages 738-739.

5 Regression Discontinuity

- If a treatment is assigned as $D = 1\{X \geq c\}$ then $\bar{\theta} = m(c+) - m(c-)$.
- Can often calculate m as through local-linear technique.
- Equivalent form to LL using rectangular bandwidth: Estimate $Y = \beta_0 + \beta_1 X + \beta_3(X - c)D + \theta D + e$ on the subsample of observations such that $|X - c| \leq h$.
- Fuzzy RD: $\bar{\theta} = \frac{m(c+)-m(c-)}{p(c+)-p(c-)}$.

6 M-Estimators

- $\hat{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^n \rho(Y_i, X_i, \theta)$ for some objective (or criterion) function.
- Let $\theta_0 = \arg\min E[\rho(Y, X, \theta]$. If θ_0 is unique then θ is identified.
- Asymptotic details are given in pages 777-780.

7 NLLS

- If you want to estimate a particular functional form of a conditional expectation function then you estimate parameters of the functional form to be the argmin of the sum of squared errors.
- If part of the problem is linear than can use nested minimization with OLS on the inside.
- Asymptotics page 788-789.

8 QR

- Generalization of least absolute deviation. Just tilt the absolute value lines.
- let $\rho_\tau(x) = x(\tau - 1\{x < 0\})$ be the tilted absolute loss function.
- Then $\beta_\tau = \arg\min_b E[\rho_\tau(Y - X'b)]$ is the best linear quantile predictor.
- Requires $Q_\tau[e|X] = 0$.

9 Binary Choice

- $Y = P(X) + e, e = 1 - P(X) w.p. P(X), e = -P(X) w.p. 1 - P(X)$.
- Probit: $P(x) = \Phi(X'\beta)$
- Logit: $P(x) = \Lambda(x'\beta) = (1 + exp(-x'\beta))^{-1}$
- all of the other binary choice models are trash
- $Y^* = X'\beta + e, e \sim G(e), Y = 1\{Y^* > 0\}$
- $Y = 1 \iff Y^* > 0 \iff X'\beta + e > 0 \Rightarrow P(Y = 1|X) = P(e > -X'\beta) = 1 - G(-X'\beta) = G(X'\beta)$ where the last equality holds iff $G(\cdot)$ is symmetric around 0.
- note that scale of variance of e and β are not uniquely identified so standardize variance of e as a normalization to achieve identification.
- Estimate these models by maximum likelihood. Helpful math if this is relevant is on pdf page 826.
- Let $P(Y = 1|X = x) = G(x'\beta)$. $\frac{\partial}{\partial x} P(x) = \beta g(x'\beta)$.
- Average marginal effect $AME = \beta E[g(X'\beta)]$.

10 Multiple Choice

- Multinomial logit is $P_j(x) = \frac{exp(x'\beta_j)}{\sum_{l=1}^{L} exp(x'\beta_l)}$
- Again log likelihood is estimation method. See 842 if necessary.
- Conditional logit is very slightly different: $P_j(x) = \frac{exp(x_j'\gamma)}{\sum_{l=1}^{L} exp(x_l'\gamma)}$
- Conditional logit can be a combination of the two: $P_j(w, x) = \frac{exp(w'\beta_j + x_j'\gamma)}{\sum_{l=1}^{L} exp(x'\beta_l + x_l'\gamma)}$
- Again cond'l logit use maximum likelihood.
- Log likelihood of these models and average marginal effect 844.
- Problem: independence of irrelevant alternatives: $\frac{P_j(W,X|\theta)}{P_l(W,X|\theta)} = \frac{exp(W'\beta_j + X_j'\gamma)}{exp(W'\beta_l + X_l'\gamma)}$
- Nested logit fixes this. This basically is logit on categories, and then within the categories you have a nested logit for the individual items. 847-849.
- Mixed logit is conditional logit which allows the coefficients γ on the alternative-varying regressors to be random across individuals. This is estimated by MCMC. 850.

11 Censoring

- Model: $Y^* = X'\beta + e, E|X \sim N(0,\sigma^2), Y = \max(Y^*, 0). Y^\# = Y$ if $Y > 0$, or missing if $Y = 0$.
- Y is censored, $Y^\#$ is truncated. Truncated is worse in terms of bias than censored, generally.
- $P(Y^* < 0|X) = P(e < -X'\beta|X) = \Phi\left(-\frac{X'\beta}{\sigma}\right)$
- $m^*(X) = X'\beta$
- $m(X) = X'\beta\Phi\left(\frac{X'\beta}{\sigma}\right) + \sigma\phi\left(\frac{X'\beta}{\sigma}\right)$
- $m^\#(X) = X'\beta + \sigma\lambda\left(\frac{X'\beta}{\sigma}\right)$
- Under a lot of assumptions $\beta_{BLP} = \beta(1-\pi)$ where π is the censoring probability. See 865.
- Tobit estimator (note it is mixed continuous/discrete measure as opposed to continuous density):
- $F(y|x) = 0, y < 0; \Phi\left(\frac{y-x'\beta}{\sigma}\right), y \geq 0$. Conditional 'density' and other details are page 866.
- Estimate by maximum likelihood. 866-867.
- CLAD is basically LAD (QR for median quantile) with censoring added in:
- $\hat{\beta}_{CLAD} = \arg\min_\beta \frac{1}{n}\sum_{i=1}^n |Y_i - \max\{X_i'\beta, 0\}|$
- This works well because the conditional quantiles are unaffected by censoring, so long as the conditional quantiles are above 0.
- CQR is QR with censoring: $\hat{\beta}_{CQR} = \arg\min_\beta \frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_i - \max\{X_i'\beta, 0\})$ where ρ_τ is from the tilted absolute loss function from the QR section.
- Heckman's model: $Y^* = X'\beta + e, S^* = Z'\gamma + u, S = 1\{S^* > 0\}, Y = Y^*$ if $S = 1$ and missing otherwise. $(e,u)' \sim N(0,V)$ where the off-diagonal element of V are not necessarily 0. Normalize $u : \sigma_u^2 = 1$.
- Estimate via maximum likelihood. Details of this model are 872-874.
- Nonparametric selection is similar to Heckman but with non-specified functions in equations for Y^*, S^*.

12 Model selection

- AIC, BIC, and Cross-Validation are the criterion we discussed. Minimizing these criterion is good (small squared error, minimize negative log likelihood).
- CV is the sum of squared leave-one-out prediction errors.
- Linear: $BIC = n + nlog(2\pi\hat{\sigma}^2) + Klog(n), AIC = n + nlog(2\pi\hat{\sigma}^2) + 2K$
- ML estimation: $BIC = -2l_n(\hat{\theta}) + Klog(n), AIC = -2l_n(\hat{\theta}) + 2K$.

- BIC is appropriate for parametric models estimated by ML and is used to select the model with the highest approximate probability of being the true model.
- Under a diffuse prior and other standard regularity conditions then $-2log(p(Y)) = BIC + O(1)$. (e.g. approximately selects ML-estimated model which would be most likely under a Bayesian setting with a flat prior).
- AIC selects the model whose estimated density is closest to the true density. It also is designed for parametric models estimated by maximum likelihood.
- Details of AIC, BIC are pages 880-885.
- Mallows criterion was also mentioned in machine learning chapter, it is appropriate for linear estimators of homoskedastic regression models. 886
- K-fold cross validation : split your data up into 'folds' (subsamples) and treat each fold as a hold-out sample.
- BIC tends to select fewer variables than AIC, CV. As a result, asymptotically BIC will kick-out all variables with nonzero true parameters while AIC, CV do not.

13 Machine Learning

- Ridge regression has a dual representation:
- $\hat{\beta}_{ridge} = (X'X + \lambda I_p)^{-1}X'Y, \lambda > 0$
- $= \min_{\beta'\beta \leq \tau}(Y - X\beta)'(Y - X'\beta), \tau > 0$
- $\tau = Y'X(X'X + \lambda I_p)^{-1}(X'X + \lambda I_p)^{-1}X'Y$
- Typically choose λ via CV.
- This shrinks parameters but DOES NOT go to corner solutions in general.
- Statistical properties/asymptotics page 936-937.
- LASSO regression has a dual representation:
- $\hat{\beta}_{lasso} = \arg\min_\beta (Y - X\beta)'(Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|, \lambda > 0$
- $= \min_{|\beta| \leq \tau}(Y - X\beta)'(Y - X'\beta), \tau > 0$
- Typically choose λ via k-fold CV (bc computationally expensive in each iteration, so wouldn't want to do straight-up CV).
- This shrinks parameters and TYPICALLY DOES go to corner solutions in general.
- Elastic net is somewhere in between (literally, in terms of objective function):
- $\hat{\beta}_{EN} = \arg\min_\beta (Y - X\beta)'(Y - X\beta) + \lambda(\alpha \|\beta_j\|_2^2 + (1-\alpha)) \|\beta_j\|_1)$
- Can jointly select parameters λ, α via CV/K-fold CV
- Regression trees: split the sample into subsamples (branches) via splits (nodes). Increasing the number of branches is growing a tree, pruning is decreasing the number of branches.
- Go through all of your available variables and find the split in each one which yields the lowest squared error across groups. Choose the split which results in lowest squared error.

- Generally you keep growing a tree until you can split no more, then go through and prune. Prune a branch if pruning decreases (improves) Mallows criterion.
- Bagging is bootstrap aggregating. Bootstrap and take as expectation the mean of the conditional expectation across bootstrap samples.
- Problem: correlation of branches across bootstrap samples. Solution: Random forests.
- RF: Like bagging, but for each bootstrap sample you randomly choose a subset of your x variables to use in the splitting ($p/3$ is typical.)