Quant II

Lab 11: Distributional analysis

Giacomo Lemoli

April 20, 2023

Housekeeping

- Homework 4 tips

Today's plan

- Distributional effects (material from last year's lectures)
- Quantile regression
- Distribution regression
- Application

Motivation

- So far: focused on average causal effects

Motivation

- So far: focused on average causal effects
- Other objects of interest: effects for other features of the potential outcome distributions

Motivation

- So far: focused on average causal effects
- Other objects of interest: effects for other features of the potential outcome distributions
- Examples: inequality in the wage distribution, changes in electoral outcomes

Concepts review

- CDF: $F_Y(y) = Pr(Y \leq y) = \mathbb{E}[\mathbb{I}(Y \leq y)]$

Concepts review

- CDF: $F_Y(y) = Pr(Y \leq y) = \mathbb{E}[\mathbb{I}(Y \leq y)]$
- Quantile function: $Q_\tau(Y) = \mathbb{F}_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$

Concepts review

- CDF: $F_Y(y) = Pr(Y \leq y) = \mathbb{E}[\mathbb{I}(Y \leq y)]$
- Quantile function: $Q_\tau(Y) = \mathbb{F}_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$
- Quantile at $\tau =$ value of Y such that τ of observations are less than Y

Concepts review

- CDF: $F_Y(y) = Pr(Y \leq y) = \mathbb{E}[\mathbb{I}(Y \leq y)]$
- Quantile function: $Q_\tau(Y) = \mathbb{F}_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$
- Quantile at τ = value of Y such that τ of observations are less than Y
- If τ=0.9: 9^{th} decile or 90% percentile

Quantile regression

- Our focus so far: CEF

$$E[Y_i \mid X_i] = \int_y y dF_{Y|X}(y \mid X_i)$$

Quantile regression

- Our focus so far: CEF

$$E[Y_i \mid X_i] = \int_y y dF_{Y|X}(y \mid X_i)$$

- Minimizes MSE of the prediction:

$$E[Y_i \mid X_i] = \arg\min_{m(X_i)} E[(Y_i - m(X_i))^2] =$$

Quantile regression

- Our focus so far: CEF

$$E[Y_i \mid X_i] = \int_y y dF_{Y|X}(y \mid X_i)$$

- Minimizes MSE of the prediction:

$$E[Y_i \mid X_i] = \arg\min_{m(X_i)} E[(Y_i - m(X_i))^2] =$$

- OLS estimator is the solution in the sample analogue when we use a linear specification $m(X_i) = X_i'\beta$

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} (Y_i - X_i'\beta)^2$$

Quantile regression

- Replace CEF with CQF (conditional quantile function) for τ:
  $Q_\tau(Y_i \mid X_i) = \mathbb{F}^{-1}_{Y|X}(\tau \mid X_i)$

Quantile regression

- Replace CEF with CQF (conditional quantile function) for τ: $Q_\tau(Y_i \mid X_i) = \mathbb{F}^{-1}_{Y|X}(\tau \mid X_i)$
- Solves another minimization problem

$$Q_\tau(Y_i \mid X_i) = \arg\min_{q(X_i)} \mathbb{E}[\rho_\tau(Y_i - q(X_i))]$$
$$\rho_\tau(u) = u(\tau - \mathbb{I}(u \leq 0))$$

Quantile regression

- Sample analogue with linear specification for $q(X_i)$:

$$\hat{\beta}_\tau = \arg\min_{\beta_\tau} \frac{1}{N} \sum_{i=1}^{N} \rho_\tau(Y_i - X_i'\beta_\tau)$$

Quantile regression

- Sample analogue with linear specification for $q(X_i)$:

$$\hat{\beta}_\tau = \arg\min_{\beta_\tau} \frac{1}{N} \sum_{i=1}^{N} \rho_\tau(Y_i - X_i'\beta_\tau)$$

- $\hat{\beta}_\tau$ generally asymptotically normal
- Bootstrap to compute SE and confidence intervals

Quantile regression

- Interpretation: QR coefficient estimates a change in the quantile of the outcome distribution

Quantile regression

- Interpretation: QR coefficient estimates a change in the quantile of the outcome distribution
- Not for units whose outcome lies at the quantile

Quantile regression

- Interpretation: QR coefficient estimates a change in the quantile of the outcome distribution
- Not for units whose outcome lies at the quantile
- Angrist and Pischke (p. 281): positive effect on lowest decile of wage distribution means that "the poors" are less poor than in the counterfactual; not that someone poor is now less poor

Quantile regression

- Interpretation: QR coefficient estimates a change in the quantile of the outcome distribution
- Not for units whose outcome lies at the quantile
- Angrist and Pischke (p. 281): positive effect on lowest decile of wage distribution means that "the poors" are less poor than in the counterfactual; not that someone poor is now less poor
- Another useful application: censored data

Quantile regression

- Interpretation: QR coefficient estimates a change in the quantile of the outcome distribution
- Not for units whose outcome lies at the quantile
- Angrist and Pischke (p. 281): positive effect on lowest decile of wage distribution means that "the poors" are less poor than in the counterfactual; not that someone poor is now less poor
- Another useful application: censored data
  - E.g. effects at 90^{th} percentile identifiable even if top 5% of data are censored

Quantile regression: implementation

- Stata: `qreg` and sister commands

Quantile regression: implementation

- Stata: `qreg` and sister commands
  - `sqreg` and `bsqreg` for bootstrapped standard errors

Quantile regression: implementation

- Stata: `qreg` and sister commands
  - `sqreg` and `bsqreg` for bootstrapped standard errors
- R: `quantreg`

Quantile regression: implementation

- Stata: `qreg` and sister commands
  - `sqreg` and `bsqreg` for bootstrapped standard errors
- R: `quantreg`
  - Several methods available, different SEs through `summary`

Working example

The effect of military repression on support for democracy: the Chilean 1988 plebiscite

AMERICAN JOURNAL of POLITICAL SCIENCE

The Geography of Repression and Opposition to Autocracy

Maria Angélica Bautista University of Chicago
Felipe González Pontificia Universidad Católica de Chile
Luis R. Martínez University of Chicago
Pablo Muñoz FGV EPGE Brazilian School of Economics and Finance
Mounu Prem Universidad del Rosario

Abstract: *State repression is a prominent feature of nondemocracies, but its effectiveness in quieting dissent and fostering regime survival remains unclear. We exploit the location of military bases before the coup that brought Augusto Pinochet to power in Chile in 1973, which is uncorrelated to precoup electoral outcomes, and show that counties near these bases experienced more killings and forced disappearances at the hands of the government during the dictatorship. Our main result is that residents of counties close to military bases both registered to vote and voted "No" to Pinochet's continuation in power at higher rates in the crucial 1988 plebiscite that bolstered the democratic transition. Potential mechanisms include informational frictions on the intensity of repression in counties far from bases and shifts in preferences caused by increased proximity to the events. Election outcomes after democratization show no lasting change in political preferences.*

Quantile regression in Stata

Estimation with qreg

```
. esttab est_25 est_50 est_75, keep(DMilitaryPresence) se
```

	(1) VoteShareNo	(2) VoteShareNo	(3) VoteShareNo
DMilitaryP~e	2.744	2.097	1.707
	(1.773)	(1.699)	(1.738)
N	276	276	276

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Quantile regression in Stata

Bootstrapping option

```
      (1)   (2)   (3)
VARIABLES   q25   q50   q75

DMilitaryPresence 2.74378*    2.09668     1.70705
      (1.589)     (1.519)     (2.040)

Observations      276   276   276
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
```

Quantile regression in R

```r
library(quantreg); library(haven); library(dplyr); library(ggplot2)

# Import data
d <- read_dta("FinalDatasetForReplication.dta")

# Prepare
d <- d %>% filter(MainSample == 1)

# Quantile regression
set.seed(10)
mod <- rq(VoteShareNo ~ DMilitaryPresence + share_allende70 + share_alessandri70 +
            lnDistStgo + lnDistRegCapital + Pop70_pthousands + sh_rural_70 + factor(IDProv),
          tau = c(0.25, 0.5, 0.75), data = d)

# Results (bootstrap SE)
est <- summary(mod, se = "boot")
out <- do.call("rbind", lapply(est, function(x) c(x$tau,
                                 x$coefficients["DMilitaryPresence",])))
out
```

```
##              Value Std. Error   t value  Pr(>|t|)
## [1,] 0.25 2.743785   1.664754 1.6481617 0.1006065
## [2,] 0.50 2.096675   1.517925 1.3812773 0.1684573
## [3,] 0.75 1.707053   1.979950 0.8621697 0.3894407
```

Distribution regression

- Alternative approach to distributional effects

Distribution regression

- Alternative approach to distributional effects
- Instead of estimating changes in quantiles ("x-axis"), we estimate changes in density ("y-axis")

Distribution regression

- Alternative approach to distributional effects
- Instead of estimating changes in quantiles ("x-axis"), we estimate changes in density ("y-axis")
- CDF: $F(y \mid x) = Pr(Y \leq y) \mid X = x) = \mathbb{E}[\mathbb{I}(Y \leq y) \mid X = x]$

Distribution regression

- Alternative approach to distributional effects
- Instead of estimating changes in quantiles ("x-axis"), we estimate changes in density ("y-axis")
- CDF: $F(y \mid x) = Pr(Y \leq y) \mid X = x) = \mathbb{E}[\mathbb{I}(Y \leq y) \mid X = x]$
- Model $\mathbb{E}[\mathbb{I}(Y \leq y) \mid X = x]$ at selected values of y

Implementation

- Create dummies for being below/above some value in the outcome distribution

Implementation

- Create dummies for being below/above some value in the outcome distribution
- For interpretability reasons use the inverse of the CDF $1 - F(y) = \mathbb{I}[y_i \geq y]$

Implementation

- Create dummies for being below/above some value in the outcome distribution
- For interpretability reasons use the inverse of the CDF $1 - F(y) = \mathbb{I}[y_i \geq y]$
- Use dummies as outcomes in separate regressions

Implementation

- Create dummies for being below/above some value in the outcome distribution
- For interpretability reasons use the inverse of the CDF $1 - F(y) = \mathbb{I}[y_i \geq y]$
- Use dummies as outcomes in separate regressions
- OLS coefficient: treatment effect on the share of units in that support region

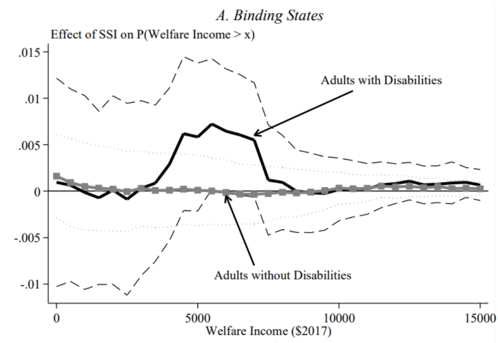
Implementation

- Create dummies for being below/above some value in the outcome distribution
- For interpretability reasons use the inverse of the CDF $1 - F(y) = \mathbb{I}[y_i \geq y]$
- Use dummies as outcomes in separate regressions
- OLS coefficient: treatment effect on the share of units in that support region
- With continuous treatment one can also use logit or probit

Distribution regression

Characterize the changes in the distribution. From Goodman-Bacon and Schmidt (2020):

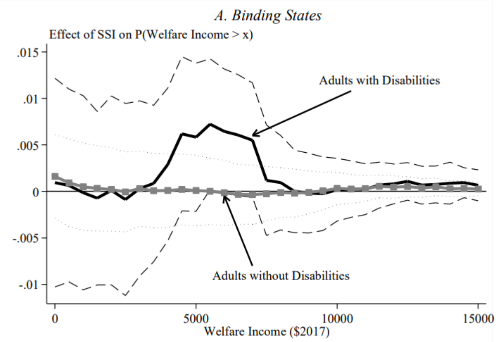
Figure 6. The Effect of SSI on the Distribution of Welfare Income for Adults With and Without Disabilities, 1970 and 1980 Censuses



Distribution regression

Characterize the changes in the distribution. From Goodman-Bacon and Schmidt (2020):

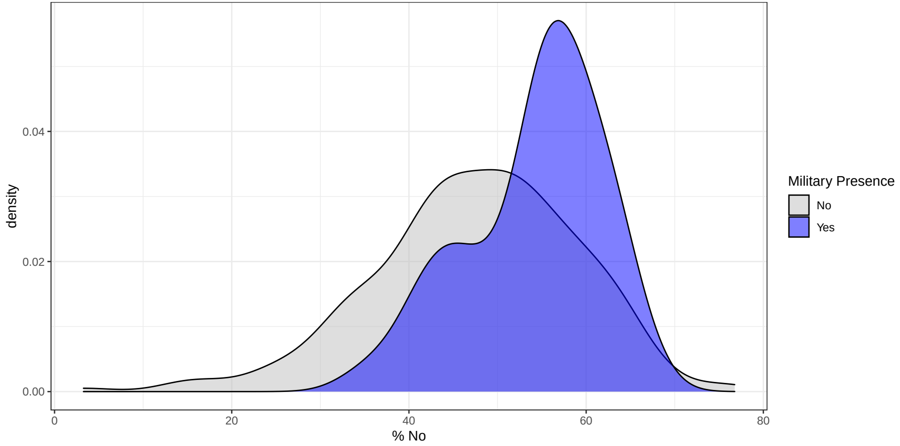
Figure 6. The Effect of SSI on the Distribution of Welfare Income for Adults With and Without Disabilities, 1970 and 1980 Censuses



Application: Chilean referendum

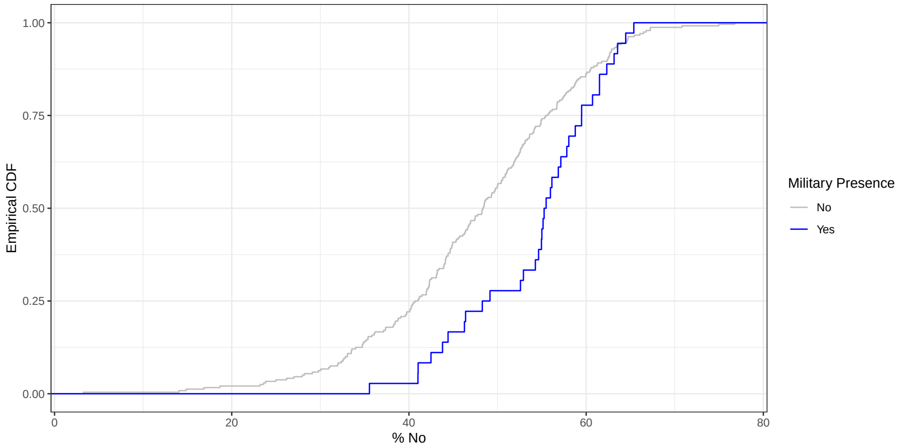
First visualization: PDF

```
ggplot(d, aes(x=VoteShareNo,  group=DMilitaryPresence, fill=factor(DMilitaryPresence))) +
  geom_density(alpha=0.5) +
  scale_fill_manual(values = c("grey", "blue"), name = "Military Presence", labels=c("No", "Yes")) +
  labs(x="% No") + theme_bw()
```



Another visualization: CDF

```
ggplot(d, aes(x=VoteShareNo, group=DMilitaryPresence, color=factor(DMilitaryPresence))) +
  stat_ecdf() + scale_color_manual(values = c("grey", "blue"), name = "Military Presence", labels=c("No", "Yes"
  labs(x="% No", y="Empirical CDF") + theme_bw()
```



Distribution regression

```
distr_reg <- function(i) {
  library(fixest); library(dplyr)
  d <- d %>% mutate(dum = case_when(VoteShareNo >= i ~ 1,
                                    VoteShareNo < i  ~ 0))
  fit <- feols(dum ~ DMilitaryPresence + share_allende70 + share_alessandri70 +
                   lnDistStgo + lnDistRegCapital + Pop70_pthousands +
                   sh_rural_70 | IDProv, data = d, weights = ~Pop70, vcov = "hetero")
  coef <- coefficients(fit)["DMilitaryPresence"]
  ll <- confint(fit)["DMilitaryPresence", "2.5 %"]
  ul <- confint(fit)["DMilitaryPresence", "97.5 %"]

  cbind(i, coef, ll, ul)
}

out <- as.data.frame(do.call("rbind", lapply(as.list(seq(10, 70, 10)),
                                             function(x) distr_reg(x))))
```

Distribution regression

```r
ggplot(out, aes(x=i, y = coef)) + geom_point() + geom_line() +
  geom_line(aes(x=i, y=ll), colour = "blue", linetype="dashed") +
  geom_line(aes(x=i, y=ul), colour = "blue", linetype="dashed") +
  geom_hline(yintercept=0, col="red", linetype = "dotted") +
  scale_x_continuous(breaks = seq(10,70,10)) + labs(x="% No", y="Effect") + theme_bw()
```