204B Problem Set 5

Clinton Tepper

02/14/2017

914-589-5370

clinton.tepper@gmail.com

Problem 1

- For a linear regression, the metric becomes:

$$
\begin{aligned}
AIC =& -2logL\left(\beta_{MLE}\right)+2k+2 \\
AIC =& -2log\left[\frac{1}{\sqrt{2\sigma_{\varepsilon}^{2}\pi}}exp\left(-\frac{\left(Y-X\beta_{MLE}\right)'\left(Y-X\beta_{MLE}\right)}{2\sigma_{\varepsilon}^{2}}\right)\right]+2k+2 \\
AIC =& nlog\left(2\sigma_{\varepsilon}^{2}\pi\right)+\frac{\left(Y-X\beta_{MLE}\right)'\left(Y-X\beta_{MLE}\right)}{\sigma_{\varepsilon}^{2}}+2k+2
\end{aligned}
$$

- But $\frac{\left(Y-X\beta_{MLE}\right)'\left(Y-X\beta_{MLE}\right)}{\sigma_{\varepsilon}^{2}}=\frac{n\sigma_{\varepsilon}^{2}}{\sigma_{\varepsilon}^{2}}=n$, therefore $AIC=n\left(log\left(2\sigma_{\varepsilon}^{2}\pi\right)+1\right)+2k+2$.

Problem 2

- I start from $Cov\left(y,\ z\right)$, and work through the algebra:

$$
\begin{aligned}
Cov\left(y,\ z\right) =& E\left[\left(Z-E\left(Z\right)\right)\left(Y-E\left(Y\right)\right)\right] \\
Cov\left(y,\ z\right) =& E\left[ZY-YE\left(Z\right)-E\left(Y\right)Z+E\left(Y\right)E\left(Z\right)\right] \\
Cov\left(y,\ z\right) =& E\left[ZY\right]-E\left(Y\right)E\left(Z\right) \\
Cov\left(y,\ z\right) =& E\left[E\left[YZ|Z\right]\right]-pE\left(Y\right) \\
Cov\left(y,\ z\right) =& E\left[E\left[YZ|Z\right]\right]-pE\left[E\left(Y|Z\right)\right] \\
Cov\left(y,\ z\right) =& E\left[Y|Z=1\right]p-p^{2}E\left(Y|Z=1\right)-p\left(1-p\right)E\left(Y|Z=0\right) \\
Cov\left(y,\ z\right) =& E\left[Y|Z=1\right]p\left(1-p\right)-p\left(1-p\right)E\left(Y|Z=0\right) \\
Cov\left(y,\ z\right) =& \left(E\left[Y|Z=1\right]-E\left(Y|Z=0\right)\right)p\left(1-p\right)\checkmark
\end{aligned}
$$

Problem 3

- I adapted the code from the TSLS estimator, although I use the "sandwich" package to check the results

```
require(devtools);
require(DataAnalytics);
require(plyr);
require(ggplot2);
require(Matrix);
require(AER);
require(sandwich);

set.seed(11);
#allows for reproducability

#adapted from Rossi tsls funciton
#In: focal variables mX, exogeneous covariates mW, intstrumental variables mZ, y variables
#out: coeficients of the regression and a matrix of modified white standard errors
twoSLS = function(mX, mW, mZ, vY) {
    #assumes both Z and W are exogeneous
    mZa = cbind(mZ, mW);
    mRInv = backsolve(chol(crossprod(mZa)), diag(ncol(mZa)));
    mPZa = tcrossprod(mZa %*% mRInv);
    mXTild = cbind(mPZa %*% mX, mW);
    mXspXsinv = chol2inv(chol(crossprod(mXTild)));

    mXW = cbind(mX, mW);
    vB = mXspXsinv %*% crossprod(mXTild, vY);

    rownames(vB) = c(colnames(mX), colnames(mW));
    resid = vY - mXW %*% vB;
    mWhiteErrors = GetMWhiteErrors(mXTild, resid);
    colnames(mWhiteErrors) = rownames(vB)

    return(list(
    vCoef = vB,
    mMWSE = mWhiteErrors,
    mHSSE = rep(var(resid), ncol(mXW)) * mXspXsinv));
}

#in: X variables and residuals
#out: Modified White stnadard errors
GetMWhiteErrors = function(X, vResid) {
    #extract the temporary variables
    qrX = qr(X);
    mX = qr.X(qrX);
    mR = qr.R(qrX);
    mQ = qr.Q(qrX);

    mQRTInv = mQ %*% t(backsolve(mR, diag(ncol(mR))));
    return(t(mQRTInv * rep((vResid / (1 - rowSums(mQ * mQ))) ^ 2, ncol(mX))) %*% mQRTInv);
}
```

```r
#in: a regression object
#out: a matrix object for the coefficients
GetCoefAsMatrix = function(reg) matrix(coef(reg),
    nrow = length(coef(reg)),
    dimnames = list(names(coef(reg))));


tbAcemoglu = read.table("acemoglu.dat", header = TRUE);
iNumPts = nrow(tbAcemoglu);

mX = matrix(tbAcemoglu[["Exprop"]], nrow = iNumPts, dimnames = dimnames(tbAcemoglu["Exprop"]));
mW = cbind(matrix(tbAcemoglu[["Latitude"]], nrow = iNumPts, dimnames = dimnames(tbAcemoglu["Latitude"])),
    matrix(1, nrow = iNumPts));
mZ = matrix(log(tbAcemoglu[["Mort"]]), nrow = iNumPts, dimnames = dimnames(tbAcemoglu["Mort"]));
vY = tbAcemoglu$GDP;
colnames(mW)[ncol(mW)] = "Intercept";

spec = as.formula("GDP ~ Exprop + Latitude");
vOLS = lm(spec, data = tbAcemoglu);
vOLSCoef = GetCoefAsMatrix(vOLS);
vOLSResid = vY - cbind(mX, mW) %*% rbind(vOLSCoef[2], vOLSCoef[3], vOLSCoef[1]);
print("OLS Coeficients");
print(t(vOLSCoef));

print("OLS Homoskedastic SE");
print(sqrt(diag(vcov(vOLS))));

print("OLS Modified White Standard Errors")
mOLSMWSE = sqrt(diag(GetMWhiteErrors(cbind(mX, mW), vOLSResid)));
print(cbind(mOLSMWSE[3], mOLSMWSE[1], mOLSMWSE[2]));

#Prints in order of X, W, intercept
lSLSOut = twoSLS(mX = mX, mW = mW, mZ = mZ, vY = vY);

##Coeficients:
print("2SLS Coeficients")
print(t(lSLSOut$vCoef));

aerReg = ivreg(formula = GDP ~ Exprop + Latitude | Latitude + log(Mort), data = tbAcemoglu);

print("2SLS Modified White Standard Error");
print(sqrt(diag(vcovHC(aerReg, type = "HC3"))));
```

```
[1] "OLS Coeficients"
     (Intercept)    Exprop Latitude
[1,]    4.692963 0.4874712 1.013893
[1] "OLS Homoskedastic SE"
(Intercept)      Exprop    Latitude
 0.40473342  0.06449979  0.65304087
[1] "OLS Modified White Standard Errors"
          [,1]       [,2]      [,3]
[1,] 0.3438337 0.06388568 0.7570712
[1] "2SLS Coeficients"
        Exprop   Latitude Intercept
[1,] 0.9692382 -0.6696108  1.874402
[1] "2SLS Modified White Standard Error"
(Intercept)      Exprop    Latitude
  1.3249510   0.2154887   0.9541935
```

- As expected, the expropriation risk standard error increases in the instrumental regression, although the results are still likely to be significant.
- The OLS coeficient and standard error for exporpriation is reasonably close to what is shown in the paper, while the coefficient and standard error for lattitude differs.
- The coeficients in the instrumental case are reasonably close to the results presented in the paper, with moderate deviation in the standard error for lattitude.

Problem 4

A

- Angrist 1990 uses draft lottery eligibility an instrumental variable. Specifically, he uses 73 groups of five draft numbers (which can be viewed as a function of the eligibility instrument, and therefore is an instrument itself) for each race, cohort, and year. Lower draft numbers, typically < 95 (group 19) would typically have a higher likelihood of conscription. Thus the draft number is likely to be highly correlated with the likelihood of military service. Angrist argues that there is no ex-ante reason to expect that draft numbers would correlate in any way with earnings potential beyond the likelihood for a given draft number cohort to serve.
- Draft avoidance and evasion could lead to an impact on earnings other than through military service. Angrist convincingly responds to the issue of draft avoidance by pointing out that, given the popularity of the educational deferment, if anything avoidance would lead to a negative bias in calculating the effect of draft status on earnings.
- However, Angrist avoids a discussion on draft evasion. According to Baskir Strauss 1987, about 210,000 individuals were accused of draft evasion. Regardless of Presidents Ford and Carter's amnesty and pardons respectively, draft evasion remains a significant offense, and likely impaired the earnings of these individuals prior to any potential exoneration.[1] This would have likely inflated the impact of draft status on earnings.

[1] Baskir, Lawrence M., and William Strauss. 1978. Chance and circumstance: the draft, the war, and the Vietnam generation. New York: Knopf.

B

- In Table 2, Angrist defines p^e as the probability a cohort member holding veteran status conditional on draft eligibility, and defines p^n as the probability of a cohort member holding veteran status conditional on not being draft eligible.[2] . For whites, the difference between the two probabilities based on the DMDC/CWHS metric were approximately 14% and 16% for 1951 and 1952 respectively. With the SIPP data, the probability differences range from 16% to 10% from 1950 to 1952. These values suggest that draft eligibility significantly affected veteran status for whites over these years. For non-whites, the differences are smaller (about 6% for both the DMDC/CWHS metric and SIPP data). I therefore conclude that the draft eligibility instrument is strong in its ability to affect veteran status, particularly for whites.

C

- In Section 3, Angrist states that the focal variable may be correlated with the earnings equation due to unobserved components. He provides two examples: selection bias related to veteran self-selection and hidden variable bias related to the armed forces selection criteria. Angrist finds no evidence of a relationship between military service and earnings within the OLS results (p. 331).
  - Suppose self-selection bias distorted the results. A simplified decomposition of the selection bias term can be denoted as

    $$E\left(Y_{1i}|V_i=1\right)-E\left(Y_{0i}|V_i=0\right)=E\left(Y_{1i}|V_i=1\right)-E\left(Y_{0i}|V_i=1\right)+\left(E\left(Y_{0i}|V_i=1\right)-E\left(Y_{0i}|V_i=0\right)\right)$$
    $$E\left(Y_{1i}|V_i=1\right)-E\left(Y_{0i}|V_i=0\right)=\text{Effect of Veteran Status}+\text{Selection Bias}$$

    where V_i represents veteran status and Y_i earnings. Thus the veracity of the self-selection bias explanation depends on a tendency for self-selectees to earn more money even had they not enlisted. I do not find this explanation plausible, since enlistees (at least in modern times) tend to be from the lower end of the socioeconomic spectrum.
  - The hidden variable explanation based on military eligibility criteria seems more plausible. A hidden (to the econometrician) criteria in the selection process for service directly related to earnings ability could explain for an underestimate in the impairment of earnings in the OLS regression. That is, individuals with higher earnings potential might pass this criteria, resulting in a negative bias to the focal variable OLS coefficient.

[2] Angrist smooths the estimates for the SIPP data over three cohorts due to the small sample sizes