Chapter 4 Lab

Tianwei Liu

Oct 7

Preparation

```
require(knitr)
require(haven)
require(car)
library(readxl)

opts_chunk$set(echo = TRUE)
options(digits = 3)
```

Load the data in Chapter4_Lab_ChildPoverty.xlsx (or Chapter4_Lab_ChildPoverty.RData). The codebook for the data is in the "codebook" tab of Chapter4_Lab_ChildPoverty.xlsx.

```
load("Chapter4_Lab_ChildPoverty.RData")
```

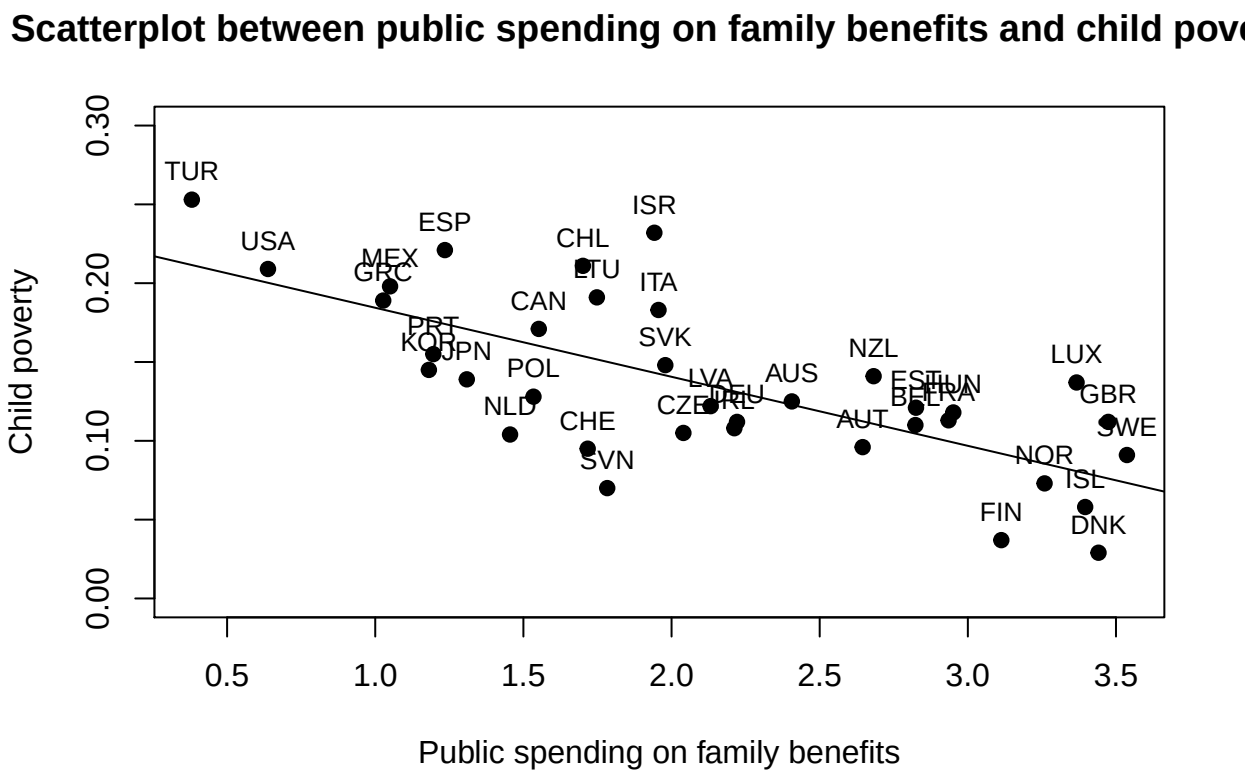
(a) Estimate a model in which public spending on family benefits explains child poverty. What is the coefficient family spending? Is it statistically significant?

```
my_reg1 <- lm(child_poverty ~ family_spending, data = dta)
summary(my_reg1)
```

```
##
## Call:
## lm(formula = child_poverty ~ family_spending, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.0801 -0.0254  0.0056  0.0218  0.0888
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.22820    0.01724   13.24  5.7e-15 ***
## family_spending -0.04379    0.00749   -5.85  1.4e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0386 on 34 degrees of freedom
## Multiple R-squared:  0.501,  Adjusted R-squared:  0.487
## F-statistic: 34.2 on 1 and 34 DF,  p-value: 1.37e-06
```

The coefficient of family spending on child poverty is -0.044, which is negative. It means that public spending on family benefits is negatively correlated with child poverty; greater public spending on families decreases child poverty. An one-unit increase in spending on family benefits will decrease child poverty by 0.044 units, hodling all other conditions equal. It's p-value is very small so it is statistically significant (pr<0.001).

(b) Create a scatterplot that corresponds to the above model. Include the OLS line, label the axes and label each point with the country name.



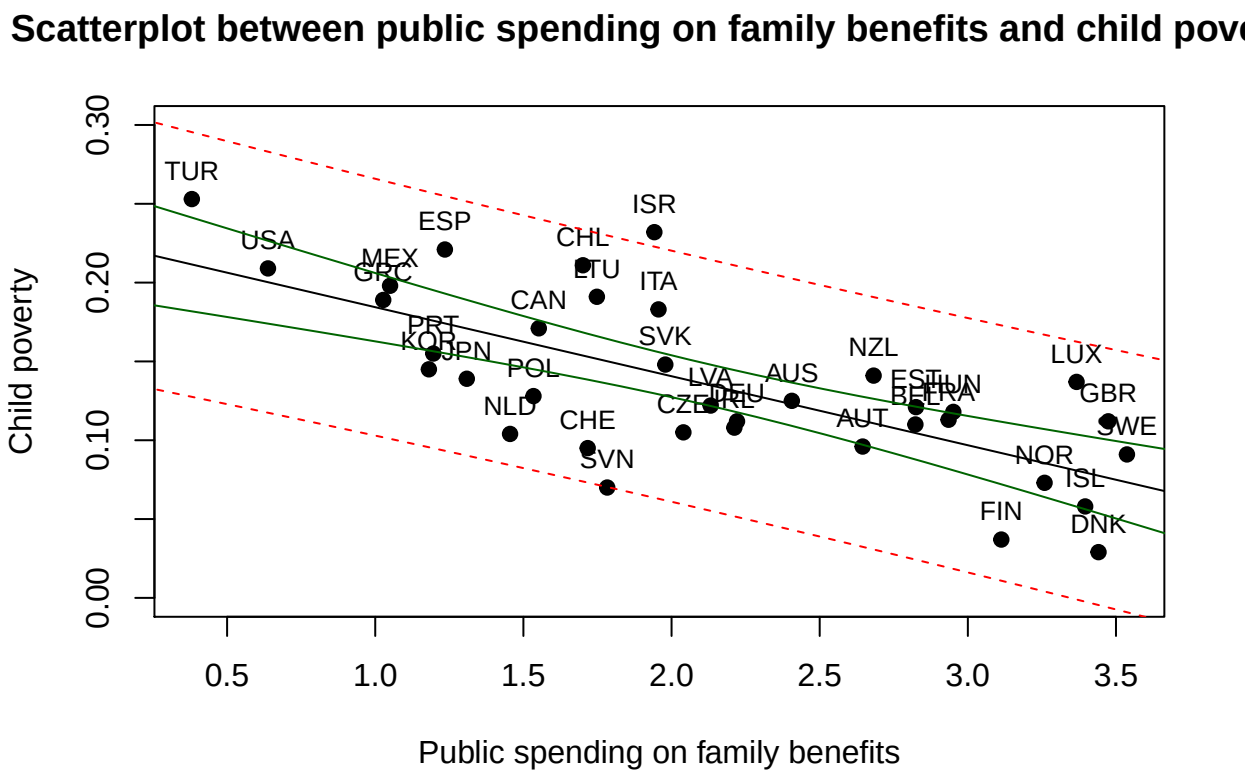
(c) Add the prediction and confidence intervals (as defined in class) to the above figure.

```
new.dat <- data.frame(family_spending=seq(0, 4, length.out=dim(dta)[1]))
#calculate prediction interval
Prediction.interval = predict(my_reg1, newdata = new.dat, interval = 'prediction', level = 0.95)
#calculate confidence interval
Confidence.interval = predict(my_reg1, newdata = new.dat, interval = 'confidence', level = 0.95)

## Plot the prediction and confidence intervals
## The original plot
plot (dta$family_spending, dta$child_poverty, main = "Scatterplot between public spending on family ben
abline (my_reg1)
text(dta$family_spending, dta$child_poverty, labels = dta$country, cex = 0.8, pos = 3)

## And we need to add lines for prediction and confidence intervals
lines(x = new.dat$family_spending,
      y = Confidence.interval[, "lwr"],
      col= "darkgreen")
lines(x = new.dat$family_spending,
      y = Confidence.interval[, "upr"],
      col= "darkgreen")
lines(x = new.dat$family_spending,
```

```
      y = Prediction.interval[, "lwr"],
      col= "red", lty = 2)
lines(x = new.dat$family_spending,
      y = Prediction.interval[, "upr"],
      col= "red", lty = 2)
```



(d) Estimate a model in which child poverty is a function of unemployment. Very briefly note the sign and significance of the coefficient on unemployment.

```
my_reg2 <- lm(child_poverty ~ unemploy, data = dta)
summary(my_reg2)
```

```
##
## Call:
## lm(formula = child_poverty ~ unemploy, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1027 -0.0286 -0.0089  0.0280  0.1101
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.10555    0.01767    5.97  9.3e-07 ***
## unemploy     0.00365    0.00193    1.89    0.067 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.052 on 34 degrees of freedom
## Multiple R-squared:  0.0955, Adjusted R-squared:  0.0689
## F-statistic: 3.59 on 1 and 34 DF,  p-value: 0.0667
```

The coefficient of unemployment on child poverty is 0.004, which indicates that unemployment is positively correlated with child poverty. However, we must note that this coefficient is not statistically significnat.

(e) What other factors should be included in multivariate model in order to help amerliorate endogeneity? Run a model with all other available and relevant independent variables. Very briefly comment on the coefficients on family_spending and unemployment.

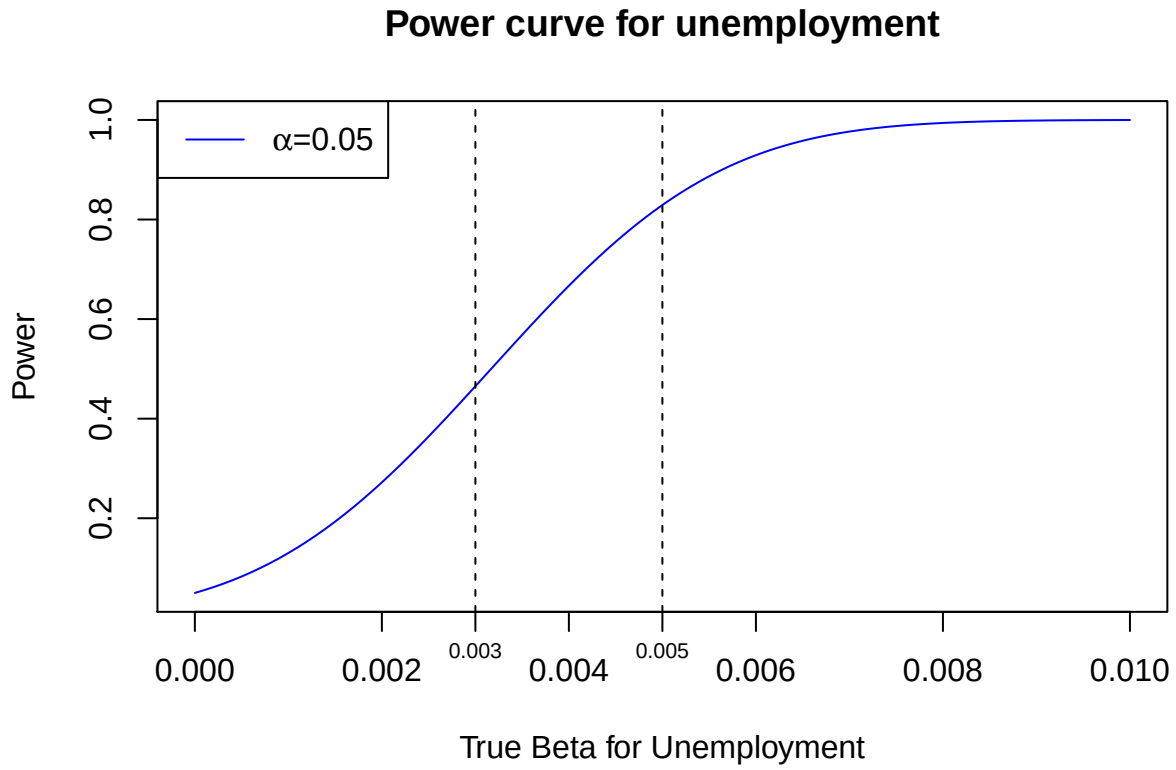
```
my_reg3 <- lm (child_poverty ~ family_spending + child_vac + unemploy + gdppercap, data = dta)
summary (my_reg3)
```

```
##
## Call:
## lm(formula = child_poverty ~ family_spending + child_vac + unemploy +
##     gdppercap, data = dta)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08387 -0.02282  0.00113  0.02478  0.08988
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.11e-01   2.96e-01    1.39     0.17
## family_spending -4.09e-02   8.53e-03   -4.80  3.8e-05 ***
## child_vac       -2.04e-03   3.11e-03   -0.65     0.52
## unemploy         1.55e-03   1.61e-03    0.97     0.34
## gdppercap       -1.41e-07   4.60e-07   -0.31     0.76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0396 on 31 degrees of freedom
## Multiple R-squared:  0.523,  Adjusted R-squared:  0.461
## F-statistic: 8.48 on 4 and 31 DF,  p-value: 9.59e-05
```

To amerliorate endogeneity, we can include child vaccination and GDP per capita in addition to spending on family benefits and unemployment. When we run this multivariate model, we can notice that coefficient on family_spending is -0.04 with p-value at 3.8*10^-5, which is statistically significant, meaning that a unit increase in public spending on family benefits can improve child poverty by 0.04. The coefficient on unemployment is 0.00155 and it is not statistically significant.

(f) Create a power curve for the *unemploy* variable for values of beta from 0 to 0.1. Use the actual standard error of the unemploy coefficient from the model explaining child poverty in the model you just ran. Sketch the power curve for one-sided $\alpha = 0.05$. Approximately what is the probability of rejecting the null is if the true value of β on unemploy is 0.003? What if the true beta on unemploy is 0.005?

```r
# Define the sequence for values of TrueBeta
TrueBeta <- seq(0, 0.01, by = 0.0001)
se <- sqrt(vcov(my_reg2)[2, 2])
# Draw the power curve for one-sided alpha = 0.05
power_0.05 <- 1 - pnorm(qnorm(0.95, mean = 0, sd = 1), mean = TrueBeta/se, sd = 1)
plot(TrueBeta, power_0.05, type = "l", main = "Power curve for unemployment", xlab = "True Beta for Une
## add legend
legend ("topleft", c(expression(paste(alpha, "=0.05"))), lty = 1, col = "blue")
# add vertical lines for x = 0.003 and x= 0.005
abline (v = 0.003, lty = 2)
abline (v = 0.005, lty = 2)
# label 0.003 and 0.005 on the x-axis
axis(1, at=0.003, labels = 0.003, cex.axis=0.7, mgp = c(2,.4,0))
axis(1, at=0.005, labels = 0.005, cex.axis=0.7, mgp = c(2,.4,0))
```



Approximately, the probability of rejecting the null hypothesis if the true bata is 0.003 is 0.45. The probability of rejecting the null hypothesis if the true beta is 0.005 is roughly 0.8.

(g) Do the above exercise, but suppose that the standard error on the unemployment variable is half what it was above. See if you can create a single plot with two power curves, one for each standard error.

```r
TrueBeta <- seq(0, 0.01, by = 0.0001)
se <- sqrt(vcov(my_reg2)[2, 2])
# Draw the power curve for one-sided alpha = 0.05
```

```r
power_0.05 <- 1 - pnorm(qnorm(0.95, mean = 0, sd = 1), mean = TrueBeta/se, sd = 1)
# Create a new power curve for half se
power_0.05_halfse <- 1 - pnorm(qnorm(0.95, mean = 0, sd = 1), mean = TrueBeta/(se/2), sd = 1)
# First plot the graph with the line for SE
plot(TrueBeta, power_0.05, type = "l", main = "Power curve for unemployment", xlab = "True Beta for Une
## Plot the line for half SE
lines (TrueBeta, power_0.05_halfse, lty = 2, col="red" )
## add legend for SE and Half SE
legend ("topleft", c(expression(paste("SE")), expression(paste("Half SE"))), lty = c(1,2), col = c("blu
# add vertical lines for x = 0.003 and x= 0.005
abline (v = 0.003, lty = 2)
abline (v = 0.005, lty = 2)
# label 0.003 and 0.005 on the x-axis
axis(1, at=0.003, labels = 0.003, cex.axis=0.7, mgp = c(2,.4,0))
axis(1, at=0.005, labels = 0.005, cex.axis=0.7, mgp = c(2,.4,0))
```