R TUTORIAL 8

Siria Angino
Federica Romei

1. *Guns.dta* is a balanced panel of data on 50 US states, plus the District of Columbia (for a total of 51 states), by year for 1977-1999. It is a data frame containing 1,173 observations on 13 variables:

   - *state*: factor indicating state;
   - *year*: factor indicating year:
   - *violent*: violent crime rate (incidents per 100,000 members of the population);
   - *murder*: murder rate (incidents per 100,000);
   - *robbery*: robbery rate (incidents per 100,000);
   - *prisoners*: incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents; value for the previous year);
   - *afam*: percent of state population that is African-American, ages 10 to 64;
   - *cauc*: percent of state population that is Caucasian, ages 10 to 64;
   - *male*: percent of state population that is male, ages 10 to 29;
   - *population*: state population, in millions of people;
   - *income*: real per capita personal income in the state (US dollars);
   - *density*: population per square mile of land area, divided by 1,000;
   - *shall*: 1 if the state has a shall carry law in effect in that year, 0 otherwise.

   Some U.S. states have enacted laws that allow citizens to carry concealed weapons. These laws are known as "shall-issue" laws because they instruct local authorities to issue a concealed weapons permit to all applicants who are citizens, mentally competent and have not been convicted of a felony (some state have additional restrictions).

   Proponents argue that, if more people carry concealed weapons, crime will decline because criminals are deterred from attacking other people. Opponents argue that crime will increase because of accidental or spontaneous use of the weapon. In this exercise, you will analyze the effect of concealed weapons laws on violent crimes.

Estimate a regression of *ln(violent)* against *shall* (1) and one of *ln(violent)* against *shall, prisoners, density, income, population, afam, cauc* and *male* (2).

(a) Interpret the coefficient on *shall* in regression (1) and (2). Is this estimate large or small in a "real-world" sense? Does adding the control variables in regression (2) change the estimated effect of a shall-issue law in regression (1), as measured by statistical significance? As measured by "real- world" significance of the estimated coefficient?

Solution:

Create the variable *ln(violent*, the run the two regressions:

```
summary_rob(lm(lviolent~shall))

               Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.13492    0.01931 317.720   <2e-16
shall          -0.44296    0.04764  -9.299   <2e-16

--- Heteroskadasticity robust standard errors used
Residual standard error: 0.6174 on 1171 degrees of freedom
Multiple R-squared:  0.08664,Adjusted R-squared:  0.08586
F-statistic: 86.47 on 1 and Inf DF,  p-value: < 2.2e-16
```

From regression (1) it seems that states that enforced the shall law have 44

```
summary_rob(lm(lviolent~shall+prisoners+density+income+population+afam+cauc+male))

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.981738   0.613642   4.859 1.34e-06
shall       -0.368387   0.034985 -10.530  < 2e-16
prisoners    0.001613   0.000192   8.397  < 2e-16
density      0.026689   0.015011   1.778  0.07568
income       0.001205   0.007381   0.163  0.87032
population   0.042710   0.003212  13.297  < 2e-16
afam         0.080853   0.020274   3.988 7.08e-05
cauc         0.031200   0.009825   3.176  0.00153
male         0.008871   0.012317   0.720  0.47155

--- Heteroskadasticity robust standard errors used
Residual standard error: 0.4277 on 1164 degrees of freedom
```

```
Multiple R-squared:  0.5643,Adjusted R-squared:  0.5613
F-statistic: 729.9 on 8 and Inf DF,  p-value: < 2.2e-16
```

Again, the enforcement of the law seems to decrease a lot the percentage of violent crimes. Indeed states that enforced the shall law seem to have 36% less violent crimes than those that did not. The effect is huge from a real world perspective and encourage the enforcement of shall law.

(b) Suggest a variable that varies across states but plausibility varies little or not at all over time, and that could cause omitted variable bias in regression (2).

Solution:

One possible problem is the presence of state specific characteristics that influence both the adoption of the shall law and the rate of violent crimes. Some states governed by Conservative Party could be very concerned about security issues. This could increase the probability to adopt the shall law and to fight violent crimes, but this does not mean that the shall law is good at decreasing violent crimes.

(c) Do the results change when you add a fixed state effect? If so, which set of regression results is more credible, and why?

Solution:

Add the fixed state effect using the option "factor(state)" in the regression (most fixed state effect's coefficients are omitted):

```
summary_rob(lm(lviolent~shall+prisoners+density+income+population+
afam+cauc+male+factor(state)))

                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.037e+00  3.975e-01  10.156  < 2e-16
shall           -4.614e-02  2.052e-02  -2.249 0.024706
prisoners       -7.101e-05  1.021e-04  -0.695 0.486922
density         -1.723e-01  1.207e-01  -1.427 0.153878
income          -9.204e-03  6.942e-03  -1.326 0.185158
population       1.152e-02  1.027e-02   1.122 0.262030
afam             1.043e-01  1.737e-02   6.003 2.61e-09
```

```
cauc              4.086e-02  5.542e-03   7.373 3.24e-13
male             -5.027e-02  8.048e-03  -6.247 5.96e-10
factor(state)2    5.596e-02  8.148e-02   0.687 0.492318
factor(state)4    2.404e-01  9.061e-02   2.653 0.008084
...

Heteroskadasticity robust standard errors used
Residual standard error: 0.1607 on 1114 degrees of freedom
Multiple R-squared:  0.9411,Adjusted R-squared:  0.938
F-statistic: 1.99e+04 on 58 and Inf DF,  p-value: < 2.2e-16
```

We are running a regression with $n-1$ dummy variables, where n is the number of states. In doing this, we are trying to clean our regression from the immutable characteristic of the states. The coefficient of *shall* drops dramatically. Indeed states that enforce the shall law have 4% violent crimes less than states that did not. Moreover, the coefficient is no more significant at 1% level. This is the most significant regression among the others since we are accounting for (at least partially) the omitted variable bias. As you can see, the adjusted R^2 and the R^2 are incredibly high.

(d) Do the results change when you add fixed time effects? If so, which set of regression results is more credible, and why?

Solution:

Another threat to our results can arise from specific time features. Imagine that in a specific year crime drops for some reason and many states adopt the shall law. It can happen that in that specific year more citizens are concerned with security and decide to vote at the national level for tje Conservative Party, that pushes for more severe law against crime. To take into consideration this effect, we add fixed time effects, $T-1$ dummy variables where T is the number of years in the dataset.

```
summary_rob(lm(lviolent~shall+prisoners+density+income+population+
afam+cauc+male+factor(state)+factor(year)))

                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.972e+00  4.690e-01   8.469  < 2e-16
shall            -2.799e-02  2.010e-02  -1.393 0.163976
prisoners         7.599e-05  8.735e-05   0.870 0.384472
density          -9.155e-02  7.213e-02  -1.269 0.204609
```

```
income           9.587e-04  7.527e-03   0.127 0.898675
population      -4.754e-03  7.104e-03  -0.669 0.503466
afam             2.919e-02  2.219e-02   1.315 0.188662
cauc             9.250e-03  8.910e-03   1.038 0.299433
male             7.333e-02  1.969e-02   3.723 0.000207
factor(state)2  -1.474e-01  7.271e-02  -2.027 0.042874
factor(state)4   1.394e-01  7.523e-02   1.852 0.064233
...
factor(year)78   5.853e-02  3.581e-02   1.634 0.102487
factor(year)79   1.639e-01  3.389e-02   4.838 1.50e-06
...

Heteroskadasticity robust standard errors used
Residual standard error: 0.14 on 1092 degrees of freedom
Multiple R-squared:  0.9562,Adjusted R-squared:  0.953
F-statistic: 3.554e+04 on 80 and Inf DF,  p-value: < 2.2e-16
```

The coefficient of *shall* is not significantly different from zero at 10% significance level. It seems that there is no relation between the shall law and the rate of violent crimes. Moreover, crime seems to be explained by state and time specific factors, as all other variables are not statistically significant.

(e) Repeat the analysis using *ln(robbery)* and *ln(murder)* in place of ln(violent).
Solution:

Everything is similar to previous points:

```
lrobbery<-log(robbery)
lmurder<-log(murder)
summary_rob(lm(robbery~shall+prisoners+density+
income+population+afam+cauc+ male+factor(state)+factor(year)))

summary_rob(lm(lmurder~shall+prisoners+density+income+population+afam+
cauc+male+factor(state)+factor(year)))
```

(f) In your view, what are the most important remaining threats to the internal validity of this analysis?

Solution:

We cannot take into account state-time specific features that affect our dependent variable. Second, it could be that some states adopt the shall law to fight the increase in violent crimes; in this case, simultaneous causality arises, and this makes our results invalid.

(g) Based on your analysis, what conclusions would you draw about the effect of the shall law on criminal rates?

Solution:

This is up to you!

2. It is extremely important to know what are the determinants of wage. Assume that your boss give to you dataset *wage2.dta*. It contains 935 observations on the following variables:

- *wage*: monthly earnings;
- *hours*: average weekly hours;
- *IQ*: IQ score;
- *KWW*: knowledge of world work score;
- *educ*: years of education;
- *exper*: years of work experience;
- *tenure*: years with current employer;
- *age*: age in years;
- *married*: =1 if married;
- *black*: =1 if black;
- *south*: =1 if live in south;
- *urban*: =1 if live in SMSA;
- *sibs*: number of siblings;
- *brthord*: birth order;
- *meduc*: mother's education;
- *feduc*: father's education;
- *lwage*: natural log of wage.

(a) Your boss is very lazy and does not want a regression with too many regressors. He asks to you to regress *log(wage)* on 6 variables maximum, but there should be an interaction term, one variable squared and a dummy. You should be able to comment properly your output.

Solution:

Again, this is up to you!