Effect of Gun Ownership on Homicide Rates

In this segment, we consider inference for modern non-linear regression.

Recall the inference question: **How does the predicted value of Y change if we increase the regressor D by a unit, holding other regressors Z fixed?**

Here we answered this question within the context of the partially linear model, which reads:

$$Y = \beta D + g(Z) + \epsilon ,$$

- Where the conditional expectation of epsilon given Z and D equals zero
  i.e $E [\epsilon | Z, D] = 0$
- Y is the outcome variable, D is the regressor of interest, and Z is the high dimensional vector of other regressors or features called controls.

The coefficient β provides the answer to the inference question.

In this segment, we will discuss estimation and confidence intervals for β.

We have a case study in which we examine the effect of gun ownership on homicide rates.

In order to proceed, we can rewrite the partial linear model in the partialled-out form:

$$\tilde{Y} = \beta \tilde{D} + \epsilon, \quad \mathrm{E}(\epsilon \tilde{D}) = 0, \tag{1}$$

where $\tilde{Y}$ and $\tilde{D}$ are the residuals left after predicting Y and D using Z, namely,

$$\tilde{Y} := Y - \ell(Z), \quad \tilde{D} := D - m(Z),$$

where $\ell(Z)$ and $m(Z)$ are defined as conditional expectations of Y and D given Z:

Our decomposition can now be recognized as a normal equation for the population regression of $\tilde{Y}$ on $\tilde{D}$.

This implies the Frisch-Waugh-Lovell theorem for the partially linear model, which states that:

Theorem (Frisch-Waugh-Lovell for Partially Linear Model)

The population regression coefficient β can be recovered from the population linear regression of $\tilde{Y}$ on $\tilde{D}$:

$$\beta = \arg\min_b \mathrm{E}(\tilde{Y} - b\tilde{D})^2 = (\mathrm{E}\tilde{D}^2)^{-1}\mathrm{E}\tilde{D}\tilde{Y},$$

where β is uniquely defined if D can not be perfectly predicted by Z, i.e. $\mathrm{E}\tilde{D}^2 > 0$.

The coefficient is the solution to the best linear prediction problem where we predict $\tilde{Y}$ by a linear function of $\tilde{D}$.

We can give an explicit formula for β which is given by the ratio of the two averages that you see in the formula. We also see that β is uniquely defined if D is not perfectly predicted by Z. So $\tilde{D}$ has a non-zero variance. The theorem asserts that β can be interpreted as a regression coefficient or residualized Y on residualized D, where the residuals are defined by taking out the conditional expectation of Y and D given Z from Y and D.

Here we recall that the conditional expectations of Y and D given Z are the best predictors of Y and D using Z.

Now we proceed to set up an estimation procedure for β. Our estimation procedure in the sample will mimic the partialling-out procedure in the population.

We have n observations on Y_i , D_i , Z_i .

We randomly split the data into two halves: one half will serve as an auxiliary sample which will be used to estimate the best predictors of Y and D given Z. And then estimate the residualized Y and residualized D. Another half will serve as the main sample and will be used to estimate the regression coefficient β.

Let A denote the set of observation names in the auxiliary sample and M is the set of observation names in the main sample.

Our algorithm proceeds in three steps.

Step 1:
Using auxiliary samples, we employ modern nonlinear regression methods to build estimators $\widehat{l}(Z)$ and $\widehat{m}(Z)$ of the best predictors l(Z) and m(Z). Then using the main sample, we obtain the estimates or the residuals quantities:

$$\check{Y}_i = Y_i - \widehat{\ell}(Z_i), \quad \check{D}_i = D_i - \widehat{m}(Z_i), \quad \text{for each } i \in M,$$

and then using ordinary least squares of $\check{Y}_i$ on $\check{D}_i$ obtain the estimate of β, denoted by $\widehat{\beta}^1$ and defined by the formula:

$$\widehat{\beta}^1 = \arg\min_b \sum_{i \in M} (\check{Y}_i - b\check{D}_i)^2.$$

Step 2:
We reverse the roles of the auxiliary and main samples. Then repeat step 1 and obtain another estimate of β denoted by $\widehat{\beta}^2$.

Step 3:
We take the average of the two estimates from steps 1 and 2 obtaining the final estimate.

$$\hat{\beta} = \frac{1}{2}\hat{\beta}^1 + \frac{1}{2}\hat{\beta}^2$$

Using the formula the following result can be shown:

Theorem (Inference)

If estimators $\hat{\ell}(Z)$ and $\hat{m}(Z)$ provide approximation to the best predictors $\ell(Z)$ and $m(Z)$ that is of sufficiently high quality, then the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on $\hat{\beta}$, and

$$\hat{\beta} \overset{a}{\sim} N(\beta, V/n)$$

where $V = (\mathrm{E}\tilde{D}^2)^{-1}\mathrm{E}(\tilde{D}^2\epsilon^2)(\mathrm{E}\tilde{D}^2)^{-1}$.

In other words, we can say that the estimator β concentrates in a $\sqrt{V/n}$ neighborhood of β with deviations controlled by the normal law.

Now the standard error of $\hat{\beta}$ is $\sqrt{\hat{V}/n}$, where $\hat{V}$ is an estimator of V.

The result implies that the confidence interval [$\hat{\beta}$ -2$\sqrt{\hat{V}/n}$, $\hat{\beta}$ +2$\sqrt{\hat{V}/n}$] covers the true value of β for most realizations (approximately 95% of the realizations) of the data sample. In other words, if our data sample is not extremely unusual, the interval covers the truth.

Given that we can use a wide variety of methods for estimation of l(Z) and m(Z) , it is natural to try to choose the best one using the data splitting. In the construction we described, we use the auxiliary sample A to estimate predictive models using modern non-linear regression methods. We can, in principle, use the main sample M as the validation or test sample to choose the best model for predicting Y and the best model for predicting D, following the procedures we explained in the previous segment.

Or we can also use the main sample M to aggregate the predictive models for Y and aggregate the predictive models for D using least squares or Lasso following the procedures we explained in the previous segment.

Corollary:

The previous inferential results continue to hold if the best or aggregated prediction rules are used as estimators $\widehat{m}(Z)$ and $\widehat{l}(Z)$ of m(Z) and l(Z) in the algorithm we presented above. The required condition for this is that the number of rules we aggregate over or choose from is not "too large" relative to the overall sample size.
We provide a precise statement of the required condition in the supplementary course material.

We next consider the case study where we investigate the problem of estimating the effect of gun ownership on homicide rates in the United States.
For this purpose, we will estimate the partially linear model:

$$Y_{j,t} = \beta D_{j,(t-1)} + g(Z_{j,t}) + \epsilon_{j,t}$$

Here $Y_{j,t}$ **is the log homicide rate** in county j at time t.

$D_{j,t-1}$ **is the log fraction of suicides committed with a firearm** in county j at time t-1, which we use as a **proxy for gun ownership.**

And $Z_{j,t}$ **is a set of demographic and economic characteristics** of county j at time t.

The parameter β here is the effect of gun ownership on homicide rates, controlling for county level, demographic and economic characteristics.

To account for heterogeneity across counties and time trends in all variables, we have removed from them the county-specific and time-specific effects.

Let us now describe the data sources:

The sample covers 195 large United States counties between the years 1980 through 1999, giving us 3900 observations.

Control variables $Z_{j,\ t}$ are from the U.S Census Bureau and contain demographic and economic characteristics of the counties such as age distribution, income distribution, crime rates, federal spending, home ownership rates, house prices, educational attainment, voting patterns, employment statistics, and migration rates.

As a summary statistic, we first look at a simple regression of the outcome on the main regressor without controls (i.e $Y_{j,\ t}$ on $D_{j,\ t-1}$ without controls). The point estimate is 0.282 with the confidence interval ranging from 0.17 to 0.39. This suggests that gun ownership rates are related to gun homicide rates. If gun ownership increases by 1% relative to a trend, then the predicted gun homicide rate goes up by 0.28% without controlling for counties' characteristics.

Since our goal here is to estimate the effect of gun ownership after controlling for a rich set of county characteristics, we next include the controls and estimate the model by an array of the modern regression methods that we have learned.

We present the results in a table where the first column shows the method we use to estimate m(Z) and l(Z). The second column shows the estimated effect and the final column shows the 95% confidence interval for the effect. The table shows the estimated effects of the lagged gun ownership rate on the gun homicide rate as well as the 95% confidence bands for these effects.

	Estimate	% 95 Confidence Interval
Least Squares (no controls)	0.282	[0.170 0.394]
Least Squares	0.227	[0.115 0.339]
Lasso	0.242	[0.124 0.360]
Post-Lasso	0.249	[0.133 0.365]
CV Lasso	0.189	[0.073 0.305]
CV Ridge	0.211	[0.099 0.323]
CV Elnet	0.197	[0.081 0.313]
Random Forest	0.252	[0.078 0.426]
Boosted Trees	0.190	[0.057 0.323]
Pruned Tree	0.152	[-0.013 0.317]
Neural Network	0.291	[0.081 0.501]
Best	0.244	[0.130 0.358]

We first focus on the Lasso method. The estimated effect is about 0.25. This means a 1% increase in the gun ownership rate (as measured by the proxy), leads to a predicted quarter percent increase in gun homicide rates. The 95% confidence interval for the effect ranges from 0.12 to 0.36. These are similar point estimates and confidence intervals to those obtained by the least squares method.

The random forest also gives similar estimates. However, the confidence interval for this method is somewhat wider, covering the range from 0.7 to 0.42.

Next, in order to construct the best estimates for β, we evaluate the performance of predictors $\widehat{m}(Z)$ and $\widehat{l}(Z)$ estimated by different methods on the auxiliary samples using the main samples. Then we pick the methods giving the lowest average MSE.

In our case, ridge regression and the random forest give the best performances in predicting the outcome and the main regressor respectively.

We then use the best methods as predictors and estimation procedures as described above. The resulting estimate of the gun ownership effect is 0.24 and is similar to that

of Lasso. And the confidence interval is somewhat tighter, now ranging from point 0.13 to 0.35.

Let us summarize. In this segment, we have discussed the use of modern non-linear regression methods for inference. The procedure relies on sample splitting in order to avoid overfitting, which may be hard to control theoretically. We apply these inference methods to the case study where **we estimated the effect of gun ownership rates on the gun homicide rates in the United States.**