Appendix

Two Alternative Difference-in-Difference Estimators

# The Least Squares Estimator as a solution to a Maximum Entropy Problem

Consider the following linear model:

Subject to the following assumptions:

1. (errors have zero mean)
2. (errors have constant variance)
3. (errors are orthogonal to the covariates)

Note that assumption 1 is not restrictive for a linear model. Any linear model where can be recast to an alternative model with by adding an intercept.

Define the entropy of the probability distribution as:

where the integration is over the support of (usually the entire ).

Let us find a function that maximizes subject to the assumptions (constraints) described above. Using the method of Lagrange multipliers:

(where we have added the constraint that the probability distribution integrate to 1)

Taking derivatives w.r.t. and setting them equal to zero:

Which implies p is of the form:

Using the constraints to solve for the we obtain:

Which is the standard normal distribution

We can rewrite and use Maximum Likelihood Estimation to obtain the standard Least Squares estimators:

Similarly, if we replace assumption 2 with:

Then we obtain the Weighted Least Squares (WLS estimator):

where is a vector of weights.

(Using matrix notation, the above can be extended to more than one covariate).

Difference-in-difference can be expressed as:

where is a dummy variable for the treatment group, is a dummy variable for post-treatment period and is the error term. Since this model is of the same form as described earlier, it can be estimated using the method described above.

# Maximum Entropy Linear Regression with Noisy Constraints

Let us modify the above constraints, to reflect that they might be measured with error (that is the constraints may be noisy). We now have:

1. (where is a small noise term)
2. (where is a small noise term around )

I modify the Lagrangian to add penalty terms for squared deviations from the ideal constraints.

Using the method of Lagrange Multipliers, we obtain:

Which has a solution of the form:

Where is a normalization constant (the partition function) so that the probability distribution is a proper probability distribution.

We can use the above expression for as the likelihood function:

and obtain parameter estimates using Maximum Likelihood Estimation:

Where is the variance of , is the covariance of and and is a penalty term. Note that this is simply the ridge regression coefficient with well-studied properties.

# An Info-Metrics Difference-in-Differences estimator

Consider two groups (treatment and control group ). At time both groups are untreated. At time there is an intervention where the treatment group receives the treatment while the control group remains untreated.

Traditionally, the treatment effect is calculated by regressing

where: is the response of interest, is the average treatment estimator (parameter of interest), is a parameter capturing differences in the mean response between the treatment and control group and is a parameter capturing the time trend, and are indicator variables for treatment group and post-treatment time period.

Instead of using Least Squares, we now wish to estimate these parameters using the info-metrics framework.

Let be the parameters to be estimated. We now consider these parameters to be discrete random variables with support space for the parameter. Let be the number of points in the support space. Then

Require to be proper probabilities (that is and ).

Similarly, the errors can be represented as:

Where are the error support points, is the total number of support points, and is a proper probability (in the same sense as above).

We wish to maximize

Subject to

We obtain the following solutions

Where and are the partition functions for the parameters and the error respectively.

We can further let the error where is the known variance-covariance matrix. For example, if allow heteroskedasticity then we obtain: