In-class project

EC 421

Edward Rubin
Winter 2021

Prologue

Schedule

Last Time

Instrumental variables

Today

Extra-credit prediction competition
Submissions due by midnight (Pacific) tonight.

Upcoming

- **Last problem** set due by midnight (PST) on Friday
- **Review** in lecture on Thursday (review materials are on Canvas)
- **Final** next week: Canvas **8AM Pacific, Tuesday, 16 March 2021**

Prediction

Prediction

Intro

Most tasks in econometrics boil down to one of two goals:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

1. **Prediction:** Accurately and dependably predict/forecast y using on some set of explanatory variables—doesn't need to be x_1 through x_k. Focuses on $\hat{y}$. β_j doesn't really matter.

2. **Causal estimation:**[†] Estimate the actual data-generating process—learning about the true, population model that explains how y changes when we change x_j—focuses on β_j. Accuracy of $\hat{y}$ is not important.

† Often called *causal identification*.

Prediction

Competition

Today your job is to figure out the model that best **predicts** the outcome `y`.

Specifics

- **Train** (build) your model using the `train.csv` dataset.
  - 15 predictors (regressors): `x1`, `x2`, ... `x15`
  - outcome: `y`
- **Predict** the outcome for the `test.csv` data set.
- **Submit** a CSV of your predictions.
  The CSV should only have a column of your preditions.

Reward: Better preditions = more extra-credit points.

Prediction

In R

Fit a model and then use `predict()` to predict onto `newdata`.

```
# Load packages
library(pacman)
p_load(tidyverse)
# Load datasets
train_df = read_csv("train.csv")
test_df = read_csv("test.csv")
# Fit a model
my_model = lm(y ~ x1 + x15 + x1:x15, data = train_df)
# Predict onto new dataset
my_predictions = predict(my_model, newdata = test_df)
# Save as CSV
write_csv(
  as.data.frame(my_predictions),
  "my-predictions.csv"
)
```