

# **ECON 8000/9000 Empirical Energy Econ**

## **Topic 07: Best Practice for Empirical Work**

**Christy Zhou**

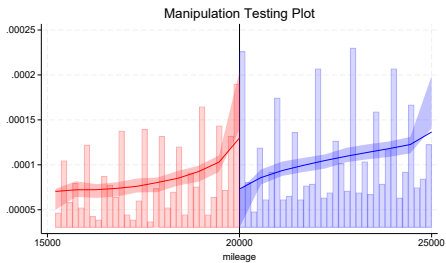
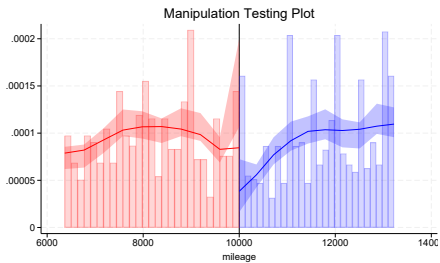
**April 23, 2026**

# Outline

- ▶ PS3
- ▶ Student Presentations
- ▶ Best Practice for Empirical Project

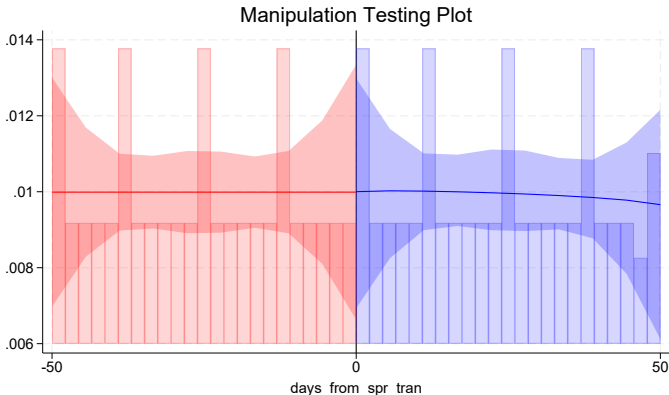
# PS3. Q1 Manipulation Test on Running Var Mileage

## Non-overlapping CIs



# PS3. Q2 Manipulation Test on Running Var Time $t$

Overlapping CIs doesn't imply passing the test for RDiT



- ▶ Notice the uniform distribution in histogram (except weekly spike)



# PS3. Q2 Part 10

## Student Ideas for RD or RDiT

- ▶ "A possible RDiT would be to look at electric vehicles purchased following the Biden-Harris administration plan to build a national network of electric vehicle charging stations by 2030. There might be a jump in purchases as people react to the easing of the concern that electric vehicles may not be able to travel long distances. It would be interesting to see if this announcement lead to an increase in the transition of combustion engines to electric vehicles."
- ▶ "I think that if you could find high-frequency data regarding wages and employment over a long period of time, you could study a minimum wage reform using an RDiT approach. The implementation date would serve as the cutoff and you could look at how employment varies around the policy implementation. This would be especially helpful in many international settings where the countries lack a pure control group and cannot use a DiD approach."
- ▶ "Does the switch to/from Daylight Saving Time affect residential electricity consumption patterns?"
- ▶ "What is the impact of renewable energy subsidy expiration dates on solar panel installation rates?"
- ▶ "An example could be the implementation of a renewable energy subsidy and seeing if it causes a jump in renewable energy building permits when the policy goes into effect"

# PS3. Q2 Part 10

## Student Ideas for RD or RDiT

- ▶ "For example, one could observe the effect that narcotic stimulants can have on local energy consumption. An RDIT could be used to analyze the change in energy demand in an area immediately following a known supply shock of methamphetamines (or other stimulants such as cocaine or crack) to see if residential utility consumption patterns change."

# Outline

- ▶ PS3 ✓
- ▶ Student Presentations
- ▶ Best Practice for Empirical Project

# Outline

- ▶ PS3 ✓
- ▶ Student Presentations ✓
- ▶ Best Practice for Empirical Project

# Best Practices

## Resource

### Best Practices

- ▶ The GOAT of Best Practice: Gentzkow and Shapiro (2014) “Practitioner’s Guide”
- ▶ Recent Update: Ristosvka (2019) “Coding for Economists”

### Some Good Websites for Best Practices

- ▶ Julian Reif ([link](#))
- ▶ Daniel Sullian ([link](#))
- ▶ Grant McDermott ([link](#))

# My Recommendation

## Top Principals

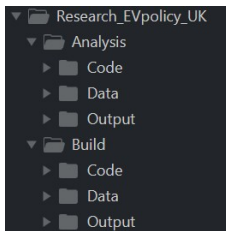
### Priority

- ▶ A good folder structure
- ▶ Automate your code to avoid error-prone repetition
- ▶ Keep a good README for data cleaning

### Additional comments on readability

- ▶ Try to write your code in a way that it is readable even if you write no comments
- ▶ Try to name your variable in a way that you can understand (meaning & unit) even if you do not create a variable label (in Stata) or create a codebook (for other settings w/o the variable label feature). You should still maintain a codebook, though.

# Folder Structure



- ▶ You don't have to follow my preferred folder structure, but key principles:
- ▶ Write your data cleaning code and estimation code in different folders. Designate a path for your final dataset(s) that you will directly use for estimation.
- ▶ Designate a path to store all your outputs: visualization, estimation tables, etc.
- ▶ README is most important for data cleaning (Build folder). Some people also prefer to maintain one for analysis.

# Maintain a README-file for Data Cleaning

```

FOLDERS
├── Research_EVpolicy_UK
│   ├── Analysis
│   │   ├── Code
│   │   ├── Data
│   │   └── Output
│   └── Build
│       └── Code
│           ├── 00_01a_ihs_sales_cleanwide.do
│           ├── 00_01b_ihs_sales_price_modelvariant_panelDF.do
│           ├── 00_01c_ihs_sales_price_extract_identifiers.do
│           ├── 00_01d_ihs_premerge_p_to_q_1_to_1_checkmatch.do
│           ├── 00_01e_ihs_premerge_p_to_q_1_to_1_genccrosswalks.do
│           ├── 00_01f_import_imputedprice_merge_allbuyer.do
│           ├── 00_01g_mergeprice_household_fleet.do
│           ├── 00_02a_aggrg_ihs_model_to_add_subsidyeligibility.do
│           ├── 00_02b_import_ihs_eligibility.do
│           ├── 00_02c_ihsdata_merge_eligibility.do
│           ├── 00_02d_mergesubsidy_household_fleet.do
│           ├── 00_03a_comcar_price_autodownload_daily.py
│           ├── 00_03b_comcar_price_appenddaily_allbuyers.py
│           ├── 00_03c_comcar_price_csvappend.do
│           ├── 00_03h_comcar_aggrg_model_to_recordgotvmining.do
│           ├── 00_03i_comcar_prep_eligibletiming_formerga.do
│           ├── 00_04a_comcar_mergesubsidyinfo.do
│           ├── 00_05a_comcar_n_ihs_premerge_prep.do
│           ├── 00_05b_import_crosswalk_comcar_n_ihs.do
│           ├── 00_05c_Aggregate_Comcar_n_ihs_pre_m_to_1_merge.do
│           ├── 00_06a_ihsdata_add_comcar_attr.do
│           ├── 00_06b_RegionalData_Filterzero_AcrossLocations.do
│           ├── 00_07a_Comcardata_add_ihsdata.do
│           ├── 01_01a_Extract_Arealist_ihsdata.do
│           ├── 01_01b_Extract_ONS_annual_demographics_by_Areado
│           └── README.md
│               ├── subsidy_trend.pdf
│               ├── subsidy_trend_onlycar.pdf
│               └── subsidy_trend_onlyvan.pdf
│   ├── Data
│   ├── Output
│   ├── Literature
│   └── MeetingNote
└── README.md
1 ## Final Datasets + Steps and documentation
2
3
4
5 #1. Where to find final dataset
6
7
8
9
10
11
12
13 #2. Final Set #1: IHS national-level product X month data with Comcar attributes merged in:
14
15 1. Where is it produced: 00_06a.do-file.
16 2. How many versions: Three: (i) allbuyers, (ii) household, (iii) fleet
17 3. Future update possibility: Unclear any important vars to be merged in
18
19
20
21 #3. Final Set #2 : IHS regional-level product X month data with Comcar attributes merged in:
22
23 1. Where is it produced: 00_06a.do-file.
24 2. How many versions: Three: (i) allbuyers, (ii) household, (iii) fleet
25 3. Future update possibility: UK postal-area demographics to be merged in
26
27
28
29 #3. Final Set #3: Comcar national-level data with IHS sales and sales-per-obs merged in:
30
31 1. Where is it produced: 00_07a.do-file.
32 2. How many versions: Three: (i) All 10+ years, (ii) a 3.5 years sample matched with IHS panel, (iii) 3.5 years
33 3. Future update possibility: Unclear.
34
35
36
37 # Step 00: Prep files (download + initial cleanup)
38
39 Note for Download:
40
41 0. Link: https://download.comcar.co.uk/; ID: tkiso@abdn.ac.uk/; PW: ColesLaw-Raider6-Activity
42
43 1. Step 00-01: Christy cleans IHS sales data + IHS price data
44 Note: So far we will only use the national + regional spreadsheet. Won't use the price spreadsheet
45
46 -- 00-01a: [Stata] Clean IHS sales spreadsheet (national/regional sales, national price) in wideformat
47 Don't drop any obs yet to examine side-by-side in *00-01h.do* forward
48 Output: "Delivery_UK_201801_202206...wideformat.dta" files by national/regional/price
49
50 -- 00-01b: [Stata] 1> Sales data reshaped into long format:
51 Output: national sales at modelvariant X year X month: N = 28196, modelvariant = 374
52 Output: regional sales at modelvariant X year X month: N = 28196, modelvariant = 374
53 >> Price data long format tsfill. To clean then into then aggregate to modelvariant X year X month
54 Output: national price at modelvariant_sub X year X month: N = 118196, modelvariant_sub = 5588
55 Output: national price at modelvariant X year X month: N = 7974, modelvariant = 351
56
57 -- 00-01c: [Stata] Examine 3 datasets from IHS in terms of (i) identifiers and (ii) variation
58 This file doesn't produce data cleaning
59
60 -- 00-01d: [Stata] Examine merge-ability for 3 IHS datasets @ crosswalk level
61 -- 00-01e: [Stata] Merge national all buyer with national price and impute using slightly boarder definition
62 Add additional aggregated identifier and price for imputation
63 -- 00-01f: [Stata] Impute manual imputation and correction. Generate matched P=Q for IHS national all buyers, calla
64 -- 00-01g: [Stata] Redo 01e-11f for (1) national data in (1b) household and (1c) fleet, with (1a) produced in 01f,
65 (2) regional data in (2a) all buyers, (2b) household, and (2c) fleet
66
67 Summary of output:
68 - Produced in (01f) for national all buyers, and (01g) for other national and regional level:
69 - The paths for all buyers and households are are:
70
71 --> "$mybuild_output/MatchIHS_IHSSales_IHSPrice_national_allbuyer.dta" // from 01-1f
72 --> "$mybuild_output/MatchIHS_IHSSales_IHSPrice_national_household.dta" // from 01-1g
73 --> "$mybuild_output/MatchIHS_IHSSales_IHSPrice_regional_allbuyer.dta" // from 01-1g
74 --> "$mybuild_output/MatchIHS_IHSSales_IHSPrice_regional_household.dta" // from 01-1d

```

# Maintain a README-file for Data Cleaning

```

1 # Steps and documentation
2
3 # Step 00: Prep files (download + initial cleanup)
4
5 1. Step 00-1: Download ACS files and unzip
6
7   - 00-1a: [Python] Download ACS 1-year survey files in zip/rar format from ACS website
8   - 00-1b: [Python] Unzip those ACS 1-year files and generate new directory/path
9   - 00-1c: [Python] Download ACS 3-yr and 5-yr files (in case 1yr not enough)
10  - 00-1d: [Python] Unzip those ACS 3-yr and 5-year files (in case 1yr not enough)
11
12
13 2. Step 00-2: Prep Czone files from Dorn's files
14
15   - 00-2a: [R] Clean up Dorn's crosswalk to a mergeable format for later use
16     - Czone to Puma (Link ACS to Puma)
17     - Czone to State (Czone goes across state border. This is most likely state by cz)
18
19   - 00-2b: [R] Merge CZ's Longitude and Latitude
20
21
22 3. Step 00-3: Clean up David Papp's SOC and NAICS files + BEA files
23
24   (1) 3a-3h: SOC and NAICS
25   (2) 3a-3p: BEA files for energy intensity
26
27   - 00-3a: [R] Clean up NAICS level brownness index to a mergeable file for future use
28   - 00-3b: [R] Clean up SOC and greenness files (likely abandon this file to "archive"
29     since Ron will work on SOC from his end using DMNET files)
30   - 00-3c: [R] Clean up SOC and brownness files
31     (Dorn) Clean up SOC and greenness files from DMNET directly
32   - 00-3d: [R] Aggregate from [01_1e]'s output, personal data, into SOCP and year
33   - 00-3f: [R] Develop cross-walk to 6, 4, 3 digit SOC
34   - 00-3g: [R] Aggregate 'SOC_Greenness_States' cleaned by Ron in 00-3d.R to 'SOC_2010_6digit' for merge to other
35   - 00-3h: [R] Output: Develop SOCP-by-year Greenness, ready to merge later in '01_11.R'
36
37   - 00-3i: [R] Output: Develop SOCP-by-year Greenness, ready to merge later in '01_11.R'
38
39   - 00-3a: [R] (Ron) Clean the upstream/downstream industries based on energy use
40   - 00-3p: [R] Prepare energy intensity, ready to merge later in '01_11.R'
41
42   - 00-3r: [R] Ron: Prepare an (manufacturing) industry definition of tradability
43   - 00-3s: [R] Prep Ron's file to mergeable format
44
45
46 4. Step 00-4:
47
48   (1) 4a-4c: Clean up EIA data to Czone level
49   (2) 4a-4j: Additional variables to capture clean energy demand
50   (3) 4l-4m: BEA transfer payment data
51   (4) 4u-4z: CBP data
52   (5) 4s-4t: Migration flows data
53
54   - 00-4a: [Stata] Generate capacity and nat generation at Czone level (by year) + add number plants (for R1)
55   - 00-4b: [Stata] Add buffer to count renewable generation
56   - 00-4c: [R] (Ron) Calculate the neighbour's solar + wind capacity for each CZ.
57   - 00-4d: [R] Output: Combine capacity and generation files to prep to merge later (including neighbour cap)
58
59   - 00-4h: [R] (Ron) Calculate the average electricity prices at Czone level (by year)
60   - 00-4i: [R] Prep energy price at CZ level to mergeable data
61
62   - 00-4l: [R] (Ron) Clean BEA Benefit data
63   - 00-4e: [R] Prep BEA data to mergeable level
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
  
```

# Use Automation to Reduce Error

```

1  cap log close
2  cap eststo clear
3  pad
4  version
5  timer on 100
6  clear all
7  mata: mata clear
8  set more off
9  set trace off
10 set linesize 100
11 //set memory allow
12 set matsize 10000
13 set seed 321456
14 timer clear
15
16
17 * My Build Directory
18 global mydirectory
19 //global mydirectory
20 global mybuild
21 global mybuild_code
22 global mybuild_output
23 global mybuild_data
24
25
26 global myanalysis
27 global myanalysis_code
28 global myanalysis_output
29 global myanalysis_data
30
31
32 * Pre-Run Set Parameters
33 global folder
34 capture cd
35
36
37
38 *****
39 * Test if Treatment is different by timing
40 * Part 1. Prep
41 *****
42 * 1. Load data
43 // use "$mybuild_output/On_Time_Performance_addEstimate_final.dta", clear
44 use "$mybuild_output/On_Time_Performance_merged_withweather_final_b3match.dta", clear
45
46
47 * 2. Data Cleaning
48 do "$myanalysis_code/automate_datacleaning.do"
49
50
51 * 3. Pre-Run Program: Automate My Two-way FE Regression
52 do "$myanalysis_code/automate_systfe.do"
53
54
55 * 4. Pre-Run Program: Automate Lates Table
56 do "$myanalysis_code/automate_sylatextable.do"
57
58
59 *****
60 * Part 2. Estimation
61 *****
62 global x_treat "Upgrade_o_ever"
63 global x_ctrl ""
64

```

# Other things

Length of scripts... this would be up to your preference

- ▶ I would avoid writing lengthy code whenever I can
- ▶ So that each script/file (data cleaning or estimation) usually serves one and only one purpose

The "product" of the build folder includes all final datasets

- ▶ If there are more than 1 final dataset, then I would make it clear in README (Some ppl also maintain a Google Doc for collaborative work)
- ▶ If there is exactly one final dataset, then naturally the last step creates the final dataset