

Problem Set 2

Due: February 28th, 2020

Problem sets are due at 5PM in the GSIs mailbox. You may work in groups, but each student should turn in their own write-up (including a narrated/commented and executed Jupyter Notebook). Please use markdown boxes within your Jupyter notebook for narrative answers to the questions below.

1 Frisch-Waugh Theorem and Residual Regression

Consider the long linear regression

$$\mathbb{E}^*[Y|X, W] = X'\beta_0 + W'\gamma_0 \quad (1)$$

with X a $K \times 1$ vector (which includes a constant) and W a $J \times 1$ vector (which *does not* include a constant). We also have the short linear regression

$$\mathbb{E}^*[Y|X] = X'b_0 \quad (2)$$

as well as the auxiliary (multivariate regression):

$$\mathbb{E}^*[W|X] = \Pi_0 X. \quad (3)$$

1. Construct the $J \times 1$ residual vector:

$$V = W - \Pi_0 X$$

and show that

$$\mathbb{E}^*[Y|X, V] = \mathbb{E}^*[Y|X] + \mathbb{E}^*[Y|V] - \mathbb{E}[Y].$$

Interpret your result [10 sentences].

2. Let $\mathbb{E}^*[Y|V] = V'\eta_0$ and show that

$$\mathbb{E}^*[Y|X, W] = X'b_0 + V'\gamma_0$$

and hence that $\gamma_0 = \eta_0$. Interpret your result [10 sentences].

2 Linear regression

This question uses the comma delimited dataset `nlsy79extract.csv`.

1. Load the dataset into a pandas dataframe called `nlsy79`. Use `HHID_79` and `PID_79` as the multi-indices for the dataframe.
2. Drop any cases where `core_sample` equals zero.
3. Drop any units where `male` equals zero.

4. Create a variable called `earnings_in_2000` which is the average of `real_earnings_1997`, `real_earnings_1999`, `real_earnings_2001`, `real_earnings_2003`. When computing this variable average over all non-missing values; for example if earnings is observed in just two of the four years listed above, average over the two years it is observed.
5. Drop all variables except `HGC_Age28`, `live_with_mom_at_14`, `live_with_dad_at_14`, `usborn`, `hispanic`, `black`, `AFQT`, `HGC_Fath79`, `HGC_Moth79`, and `earnings_in_2000`.
6. Finally retain only complete cases (you can use “`dropna()`” for this).
7. Those units that remain constitute your estimation sample. Use “`describe()`” to print out some basic summary statistics for your estimation sample. Write a short paragraph about your dataset.

Define `LogEarn` to be the natural logarithm of `earnings_in_2000`.

1. Compute the least squares fit of `LogEarn` onto a constant and `HGC_Age28`. You may use Python’s `StatsModels` OLS implementation for computation and standard error construction.
2. Estimate the parameters of the following linear regression model by the method of least squares

$$\mathbb{E}^*[\text{LogEarn} | X] = \alpha_0 + \beta_0 \text{HGC_Age28} + \gamma_0 \text{HGC_Age28} \times (\text{AFQT} - 50) + \delta_0 \text{AFQT}$$

where $X = (\text{HGC_Age28}, \text{HGC_Age28} \times (\text{AFQT} - 50), \text{AFQT})'$.

- (a) Plot your estimate of $\beta_0 + \gamma_0 (\text{AFQT} - 50)$ as a function of `AFQT` (for `AFQT` from zero to one hundred).
 - (b) Construct an asymptotic point-wise confidence band for $\beta_0 + \gamma_0 (\text{AFQT} - 50)$ and plot it on the same figure.
 - (c) Interpret β_0 and γ_0 and discuss your estimates of them.
3. Additionally condition on `live_with_mom_at_14`, `live_with_dad_at_14`, `usborn`, `hispanic`, `black`, `HGC_Fath79`, and `HGC_Moth79`. Do the estimated coefficients on `HGC_Age28` and its interaction with `AFQT` change? Explain.