# Stata Workshop

Dr. Vinci Chow
vincichow@cuhk.edu.hk

# What is This Workshop About?

- A hands-on tutorial on using Stata for **estimation** tasks

- Emphasize on practical usage instead of theory

- Notes and assignment:

  - Blackboard (ECON4901-WORKSHOP)

  - https://www.ticoneva.com/econ/stata-workshop/

- What is this *not* about?

  - Forecasting/making accuracy predictions—find a statistical learning/machine learning course for that

# Estimation vs Prediction

Estimation: want $\beta$ to be accurate

$$y = \alpha + \beta x$$

Prediction: want $y$ to be accurate

$$y = F(x)$$

Prediction: might not care what $F$ looks like

# Stata: Pros and Cons

✓ Widely use in economics

✓ Rich assortment of estimation procedures

✓ Consistent and easy-to-use syntax

✓ Excellent documentation

✓ Extendable. Many user-written packages online.

✕ Not free

✕ Before version 16, assumes user works with one dataset at a time

✕ Lack modern statistical learning procedures

# SCRP

High performance computing mini-cluster managed by the Department of Economics.

Features:

- Remote access to R, Stata, Python and MATLAB.

- Browser-based access to R, Python and Stata.

- Daily onsite and offsite user file backup.

- Loads of computing power.

Website: https://scrp.econ.cuhk.edu.hk/

# Account and Access

**Step 1: Account Creation**

Accounts should have been already been created for you.

**Step 2: Create Password**

Follow instructions in account creation email.

# Account and Access

**Step 3: Connect to SCRP**

**Option 1:** Web access for Python, R and Stata:

- https://scrp-login.econ.cuhk.edu.hk

- https://scrp-login-2.econ.cuhk.edu.hk

**Option 2:** SSH access:

- Windows – install MobaXterm.

**Option 3:** Remote Desktop

# Access Method Comparison

| Method | Browser | SSH | Remote Desktop |
|---|---|---|---|
| Pros | • Works on all devices<br>• No software installation required<br>• Fast | • GUI<br>• Fast | • GUI<br>• No software installation required on Windows |
| Cons | • No graphical user interface (GUI) | • Requires VPN<br>• Requires SSH software | • Requires VPN<br>• Uses a lot of memory |

# Demo

# Odds Ratio

$$odds\ ratio = \frac{\Pr(y = 1|\Delta x = 1)}{\Pr(y = 1|\Delta x = 0)}$$

Logit assumes this **<u>ratio</u>** is constant for all values of $x$.

Suppose the estimated ratio is $1.25$. Also assume that when $x = 5$, $\Pr(y) = 0.4$ and when $x = 9$, $\Pr(y) = 0.8$.

- In the first case, if $x$ increases by $1$ to $6$, an odds ratio of $1.25$ means $\frac{\Pr(y|x=6)}{\Pr(y|x=5)} = 1.25$, so $\Pr(y|x = 6) = 1.25 \times 0.4 = 0.5$. The increase is 10 percentage points.

- In the second case, if $x$ increases by $1$ to $9$, an odds ratio of $1.25$ means $\frac{\Pr(y|x=10)}{\Pr(y|x=9)} = 1.25$, so $\Pr(y|x = 10) = 1.25 \times 0.8 = 1$. The increase is 20 percentage points.

The **<u>percentage change</u>** is *not* constant!

A common practice is the report the percentage change at specific intervals. e.g. at the average $x$ value.

# Storage

SCRP has different types of storage for different purposes.

Key properties:

Home directory ~/ is reasonably fast and backed up daily.

/data is very fast but has no backup and no redundancy.

| File Type | Need | Recommended Location |
|---|---|---|
| Scripts and results | Reliable I/O Backup | ~/ |
| Large datasets | Fast I/O on multiple nodes | /data |
| Temporary files | Fast I/O on a single node | /tmp |

| Location | Quota | Speed (R/W) | Backup |
|---|---|---|---|
| ~/ | 2/10/20GB* | 350MB/s 150MB/s | Daily |
| /data | 5.8TB (shared) | 1500MB/s 550MB/s | No |
| /tmp | 200GB (shared) | 500MB/s 200MB/s | No |

*Quotas for undergraduate/postgraduate/faculty respectively.

# Using Compute Nodes