401k_fractional coding challenge

Intro

Replicate the Quasi-Maximum Likelihood Estimation (QMLE) models from *Papke and Wooldridge (1996)*, Section 4, equations (22) and (23), using `k401.dta`. This involves reproducing the results in Table III and analyzing the relationship between the participation rate and other 401k variables.

Note: I interpret Table III columns 3, 4 as not tied to equations 22, 23 since they contain MRATE^2.

User Guide

git clone

```
git clone git@github.com:garthmortensen/401k_fractional.git
cd 401k_fractional
```

setup venv

```
python -m venv venv
source venv/bin/activate
```

install dependencies

```
pip install -r ./requirements.txt
```

run code

```
python 401k.py
```

test code

```
pytest .
```

run docker

```
podman build -t Dockerfile .
```

Overview

Dataset: `k401.dta`

prate	mrate	totemp	age	sole
0.658784	0.580822	353.0	7	0
0.843350	0.218458	4130.0	22	0
1.000000	0.767652	177.0	21	1

prate	mrate	totemp	age	sole
0.941003	0.365554	2309.0	11	0
0.830149	0.407965	2309.0	7	0
1.000000	1.174729	452.0	34	1

Variables Explained

- **`participation_rate (prate)`**: % of employees eligible for a 401(k) plan who have an active account, whether or not they contributed to it in the current year.
- **`match_rate (mrate)`**: Estimate of how much the employer contributes to the employee's 401k, vs. what the employee contributes.
- **`totemp`**: Total firm employees.
- **`age`**: The age of 401k plan.
- **`sole_plan`**: Binary where 1 means this is the firm's only pension fund.

Equations

22

```
E(part_rate x) = 1 + 2match_rate + 3log(emp) + 4log(EMP)^2 +
 5age + 6age^2 + 7sole_plan
```

23

```
E(part_rate x) = G( 1 + 2match_rate + 3log(emp) + 4log(emp)^2 +
 5age + 6age^2 + 7sole_plan)
```

Analysis

Assumptions: - total employees = 20,000 - On average, employees contribute 21% of salary to 401k - Employer contributes 7% of salary. - The 401k is not the only pension plan. - Age of account = 12 years. - Ignore other factors.

Conclusion

1. How do you think predictions from (22) and (23) will match with the employer's participation rate?

I expect match rate will play a key role in predicting participation, and per the paper, the most important role. Given we have a fair amount of observations and few variables, I expect a reasonable OLS fit.

2. Which model seems more reasonable and why?

Logit, since it captures values ranging [0, 1], which is appropriate for participation rate (a percentage).

3. 2-3 paragraphs summarizing your analysis aimed at a non-technical policy maker audience.

This repo seeks to reproduce part of Papke and Wooldridge's research paper *Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates*, using an iterative approach. In this initial iteration, the basic building blocks to enable reproduction are established:

```
1. version control (git)
1. language and library version control (virtual environment, requirements)
1. logging
1. unit testing
1. CI/CD (github actions)
1. containerization
1. econometric analysis
1. TODO comments to suggest improvements for subsequent iterations
```

The econometric analysis models participation rates of 401k plans, given `participation_rate`, `match_rate`, `totemp`, `age`, and `sole_plan` as well as transformations to capture non-linear behavior. Two regression models are fitted to the author's Stata dataset (`./inputs/k401k.dta`), one OLS (equation 22) and one Logit (equation 23). Once models are fitted, specific values are imported from a configuration file (`./inputs/assumed_df.yaml`), and predictions are made.

Summary statistics (`./output_tables/*.html`) and logs (`./logs/YYYYMMDD_HHMMSS_401k.log`) indicate `const` (=2.53), `mrate` (=0.55), `log_emp2` (=-3.18) and `age` (=0.04) are all statistically significant.

Links

Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates.

Data description.