Effective Programming Practices for Economists
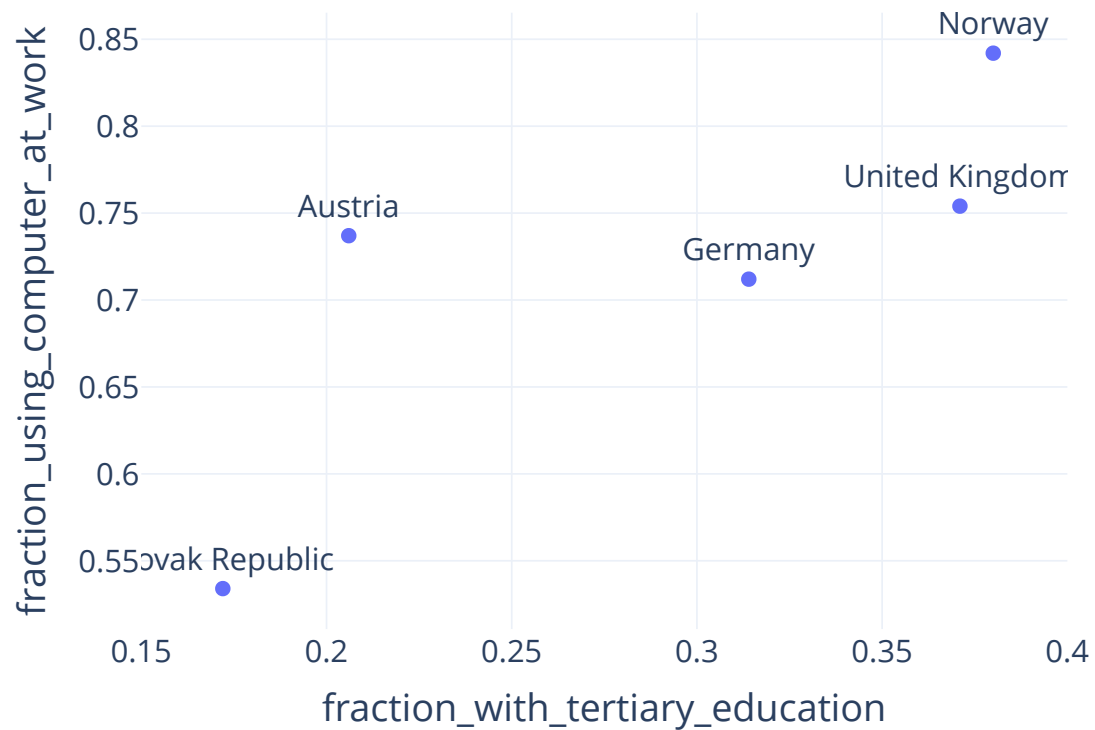
Data Analysis in Python

Running regressions using statsmodels

Janoś Gabler, Hans-Martin von Gaudecker, and Tim Mensinger

Example

country	fraction_with_tertiary_education	fraction_using_computer_at_work
Slovak Republic	0.172	0.534
Austria	0.206	0.737
Germany	0.314	0.712
United Kingdom	0.371	0.754
Norway	0.38	0.842

Example

Importing conventions

- Plain statsmodels

```
import statsmodels.api as sm
```

- Formula interface

```
import statsmodels.formula.api as smf
```

The formula interface

```
>>> model = smf.ols(
...     data=df,
...     formula="fraction_using_computer_at_work ~ fraction_with_tertiary_education",
... )
```

- Use a regression model implemented in `statsmodels.formula.api`

- `data` is a dataframe, `formula` is a string

- Separate left-hand side and right-hand by `~`

The formula interface

```
>>> model = smf.ols(
...     data=df,
...     formula="fraction_using_computer_at_work ~ fraction_with_tertiary_education",
... )
```

- Intercept is implicit for OLS
- Right hand-side can contain lots of mathematical expressions
    - `+`, `**`, `*`, `:` for sums, powers, interactions
    - `C()` for categorical variables
    - `np.log()` for logarithms (and any similar functions)

Model objects

```
>>> model = smf.ols(
...     data=df,
...     formula="fraction_using_computer_at_work ~ fraction_with_tertiary_education",
... )
>>> model
<statsmodels.regression.linear_model.OLS at 0x7fb56c905250>
```

- Almost always, the next step is to call the `.fit()` method on the model object.