ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

Part 5.4 | Model Selection

Model Selection

How do we know if adding a variable improves our model?

- *Adding variables reduces SSE (better fit to the data)*
- *But is the improvement real or just fitting noise?*
- *We need a way to test whether the improvement is statistically significant*

R^2: Proportion of Variation Explained

How much of the variation does our model capture?

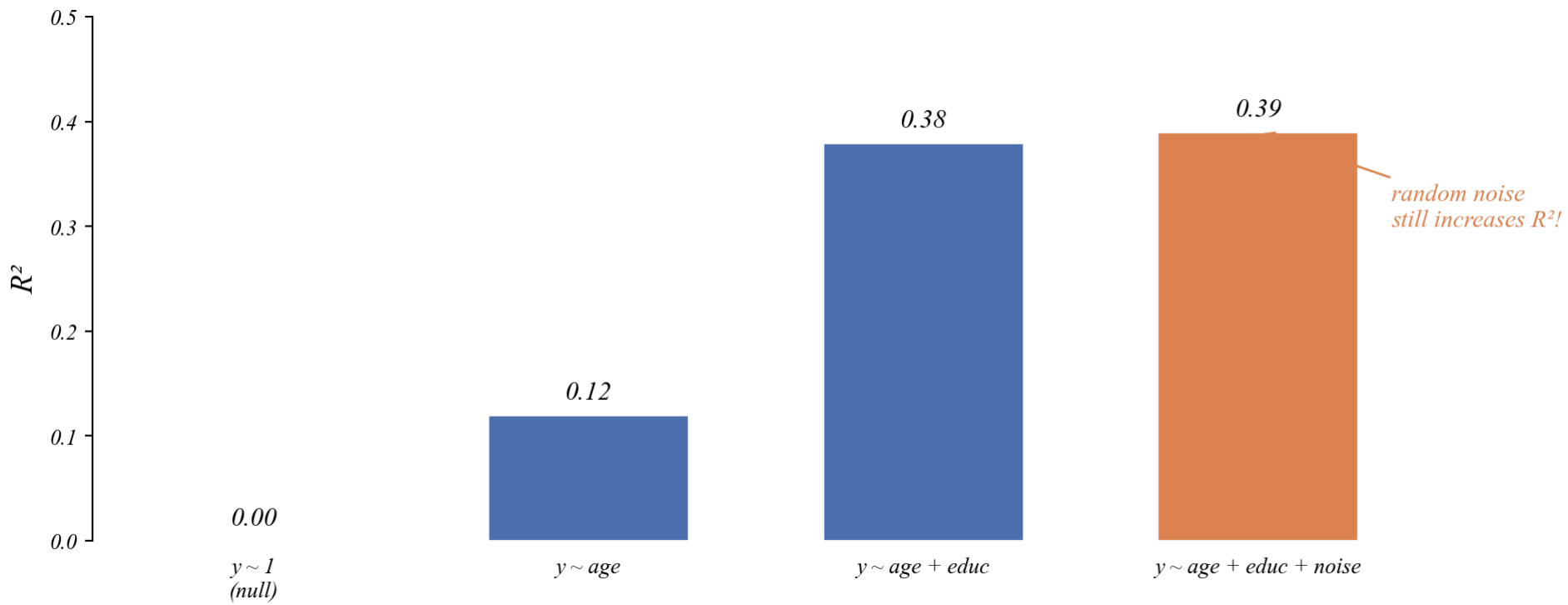
$$R^2 = 1 - \frac{SSE}{SST}$$

- *SST: total variation (SSE of the null model—just the mean)*
- *SSE: leftover variation after fitting the model*

- $R^2 = 0$*: model does no better than the mean*
- $R^2 = 1$*: model predicts perfectly*

The Problem with R^2

R^2 always goes up when you add variables.



> *even adding random noise will reduce SSE a little*

> *so how do we know if the improvement is real?*

The F-Test

Is the reduction in SSE larger than we'd expect by chance?

Compare two models:

- ***Restricted***: *fewer variables* → *higher* SSE_R
- ***Full***: *more variables* → *lower* SSE_F

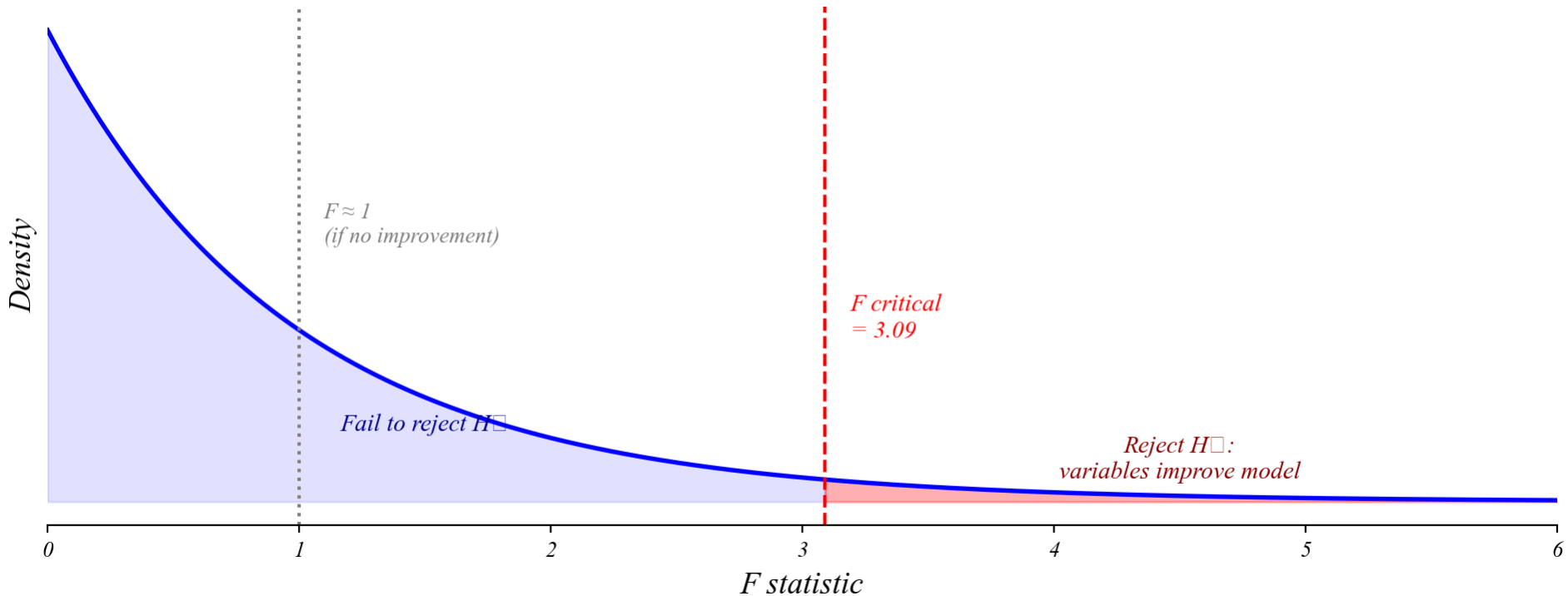
$$F = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F}$$

> *numerator: average SSE reduction per variable added*

> *denominator: average remaining error*

The F-Test

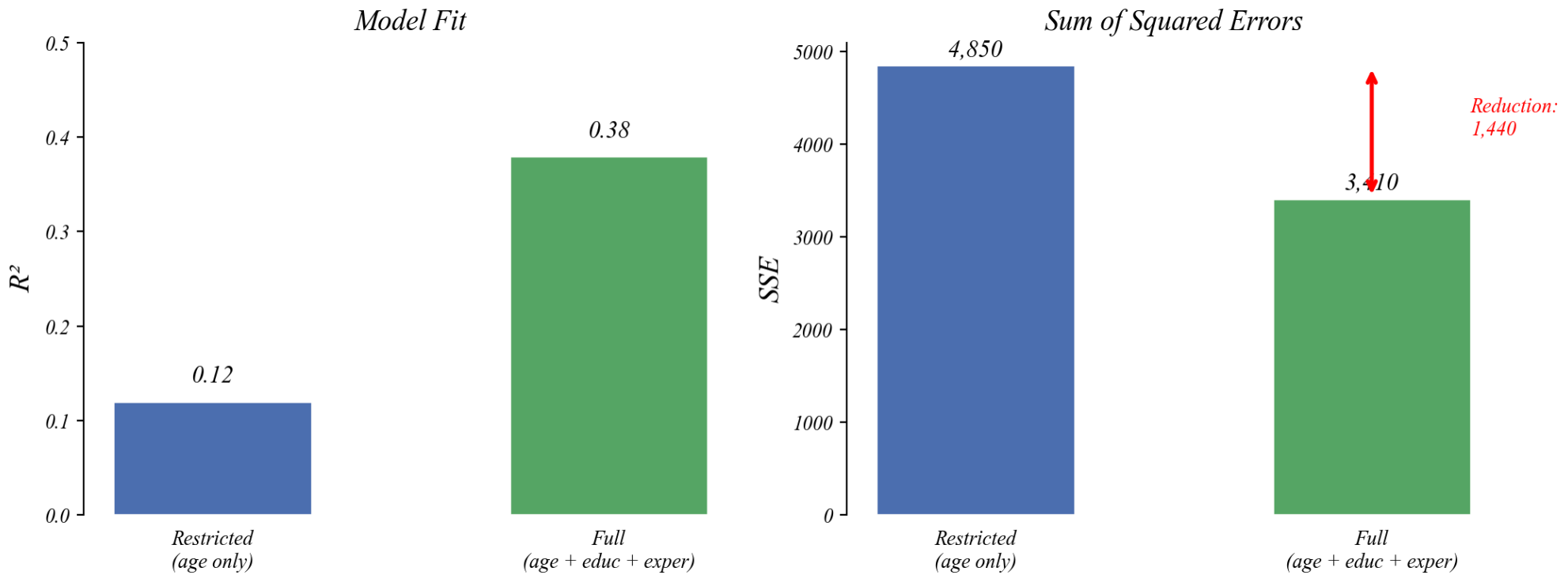
If the new variables are just noise, F ≈ 1. If meaningful, F is large.



> *large F means the SSE reduction is unlikely due to chance*

Example: Wage Model

Does adding education and experience improve predictions?



> R^2 increased by 0.26—but is that significant?

Example: Wage Model

The F-test tells us if the improvement is statistically significant.

```python
from scipy import stats

# Model comparison
sse_r = 4850      # restricted model (age only)
sse_f = 3410      # full model (age + educ + exper)
k = 2             # variables added
n = 100           # sample size

# F-statistic
f_stat = ((sse_r - sse_f) / k) / (sse_f / (n - 4))
p_value = 1 - stats.f.cdf(f_stat, k, n - 4)
```

```
F-statistic: 20.27
p-value: 0.0000
```

> $p < 0.05$: education and experience significantly improve the model

Summary

Model selection uses the same logic as hypothesis testing.

- ***R^2** tells us how much variation the model explains*
- ***R^2 always increases** when adding variables—even useless ones*
- ***F-test** asks: is the SSE reduction statistically significant?*
- *Same idea as t-tests, but for groups of variables*