Project 2: Using Machine Learning and Neural Networks to Determine Highest Impact Factors for Household Income and Wealth

Gaurang Mohan, Roshan Ramakrishnan
Data: Panel Study of Income Dynamics (PSID), family heads, 1968–2021

Research Question

Which household and head characteristics are most strongly associated with household income?

How this extends Project 1

- Re-use cleaned PSID long file from Project 1
- Add **three new models**: **LASSO**, **Random Forest, and Neural Network** (vs. OLS baseline)
- Add one **new technique**: **permutation test** for income differences by sex of head and neural network feature importance

Data & Outcome

Unit of analysis

- One row = **household–year** (family head)

Sample

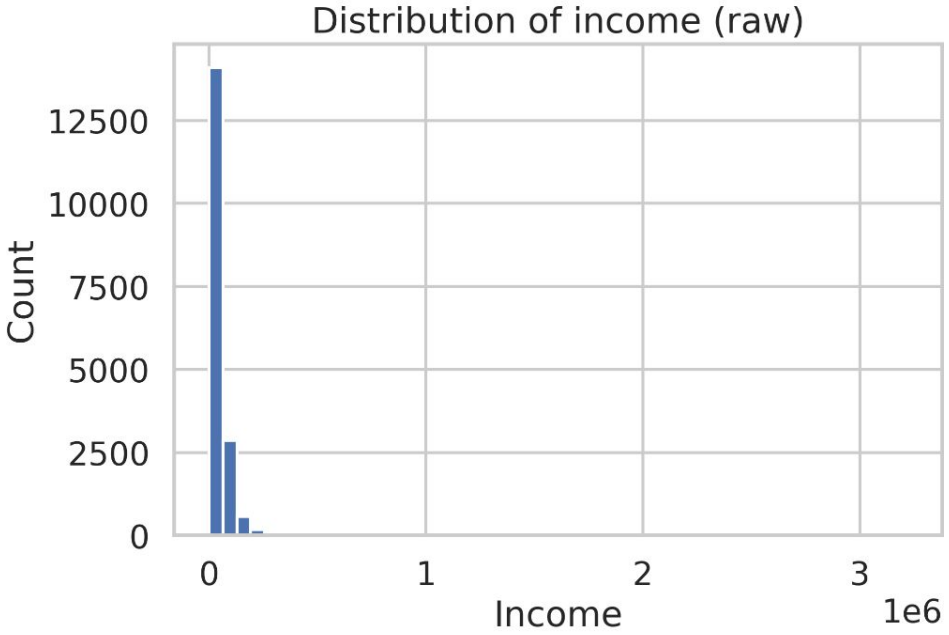
- After cleaning: ~**17,900 rows**, **9 predictors**

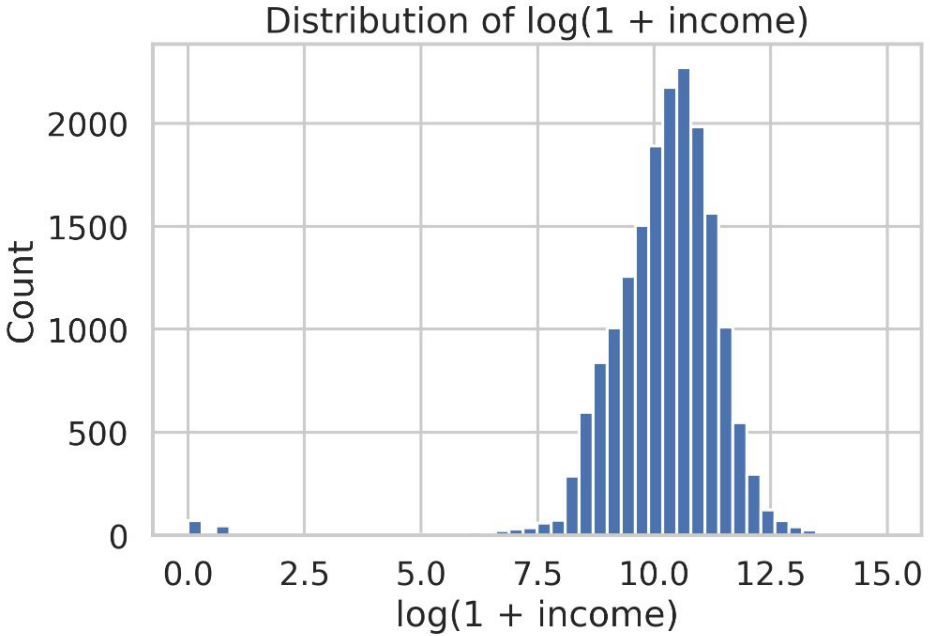
Outcome variable

- y=log(1+total family income)

- Log transform to reduce right skew and stabilize variance

Data & Outcome





Predictors & How Time is Encoded

Time & household structure

- `year` – **numeric** calendar year (linear time trend)
- `decade` – categorical (1960s, 1970s, …) ➜ **dummy variables** with 1960s as baseline
- `family_size` – number of people in family
- `own_or_rent` – tenure: own vs rent vs other (dummy-coded)

Head characteristics

- `head_age` – age of head
- `head_sex` – sex category of head (baseline = category 1)
- `head_race` – race category of head (baseline = category 1)
- `employment_now` – current employment status (several dummy categories)

Pre-Processing & Modeling Setup

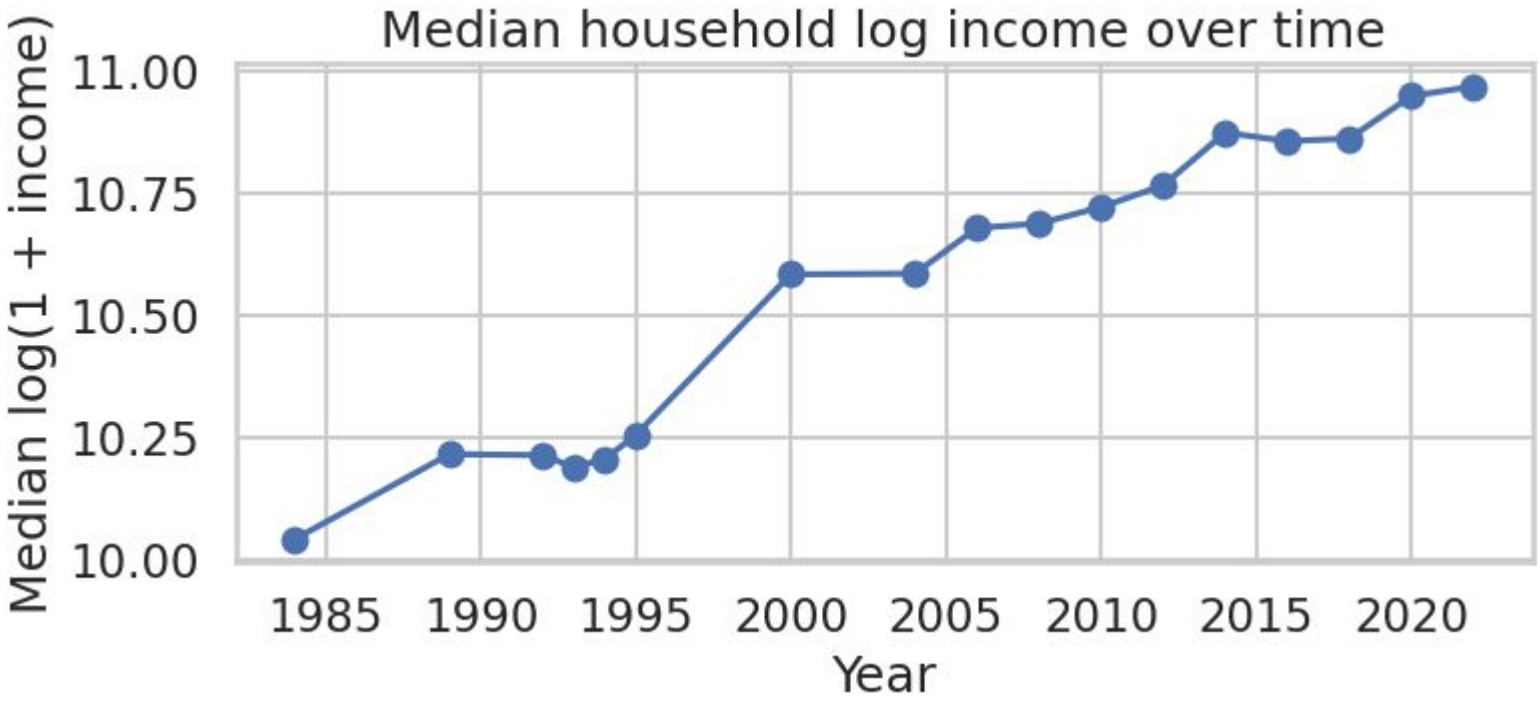
Pre-processing

- Drop rows with missing values in key fields, Create `log_income = log(1 + income)`
- One-hot encode: `decade`, `own_or_rent`, `head_sex`, `head_race`, `employment_now`
- Final feature matrix: **19 columns** (numeric + dummies)

Modeling

- Train / test split: **80% train, 20% test** (`random_state=42`)
- Models:
  - **OLS** linear regression (baseline)
  - **LASSO** regression with cross-validated α
  - **Random Forest** regressor (tree-based, non-linear)

Income Over Time (Exploratory)



Income Differences by Sex of Head (Permutation Test)

Question

- Is mean log income different between **sex category 1 vs 2** for the head?

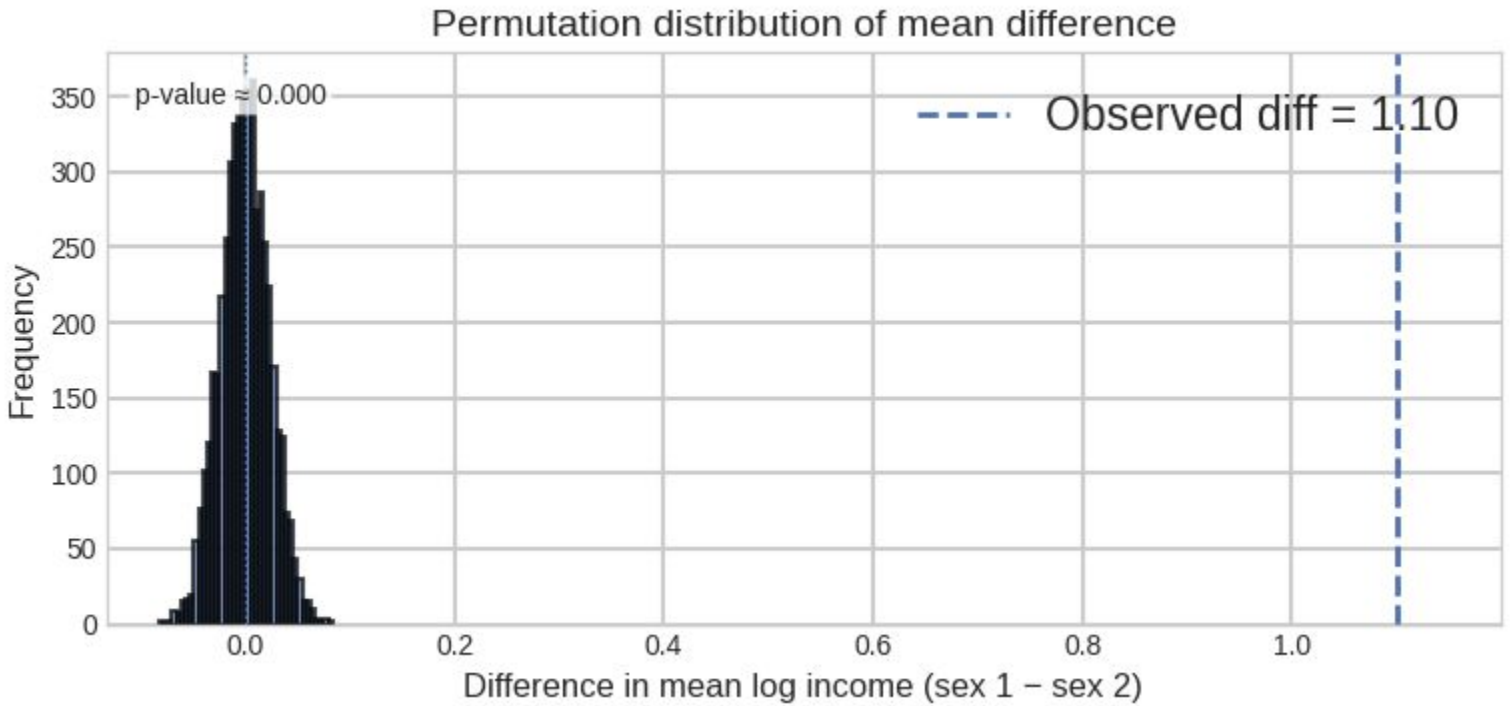
Approach

- Compute observed difference in mean log income (sex 1 – sex 2)
- Randomly shuffle sex labels many times; recompute difference each time

Results

- Observed difference ≃ **1.10 log points**, far to the right of permutation distribution
- Approximate permutation p-value **~ 0.000 →** strong evidence of a real difference

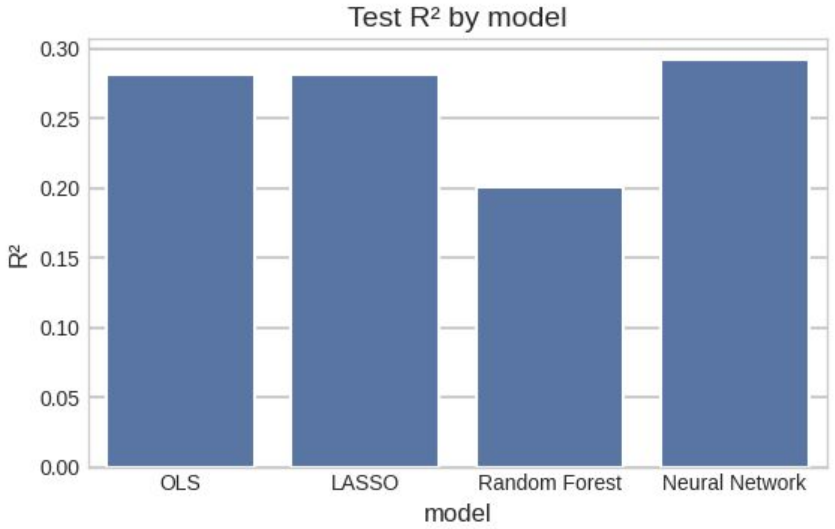
Income Differences by Sex of Head (Permutation Test)



Model Performance on Test Set

Metrics (log-income scale)

- **OLS**: RMSE ≃ **1.05**, R^2 ≃ **0.28**
- **LASSO**: RMSE ≃ **1.05**, R^2 ≃ **0.28**
- **Random Forest**: RMSE ≃ **1.10**, R^2 ≃ **0.20**
- **Neural Network**: RMSE ≃ **1.04**, R^2 ≃ **0.29**



Takeaways

- OLS and LASSO perform **very similarly** and **better than Random Forest**
- The neural network model performs **better than all the models** (albeit not by a substantial amount compared to the OLS and Lasso models)
- All models explain only about **28% of variation** → many unobserved factors remain

What do the linear models say?

Time effects

- year coefficient ≃ **0.012** ➜ ≃ **1.2% higher expected income per additional year**, holding other variables fixed

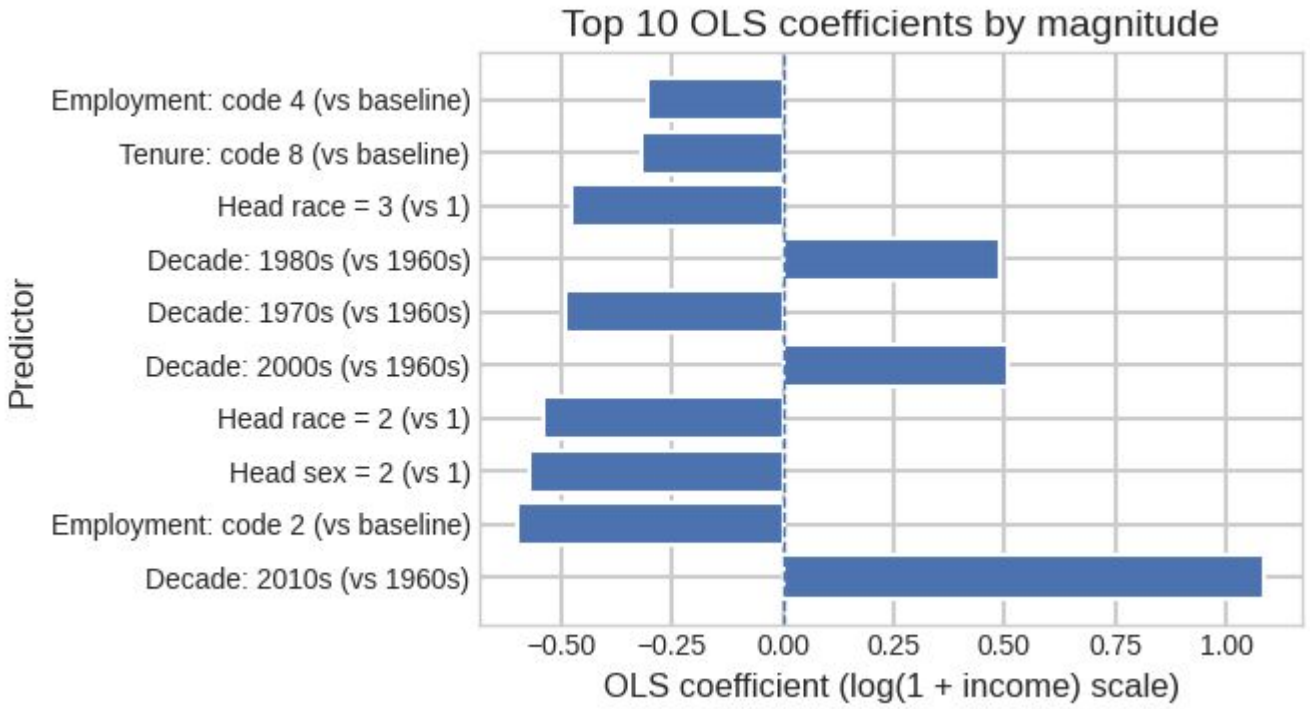
- Positive decade dummies (e.g., decade_2010s) ➜ higher incomes vs **1960s baseline**

Demographic & household effects

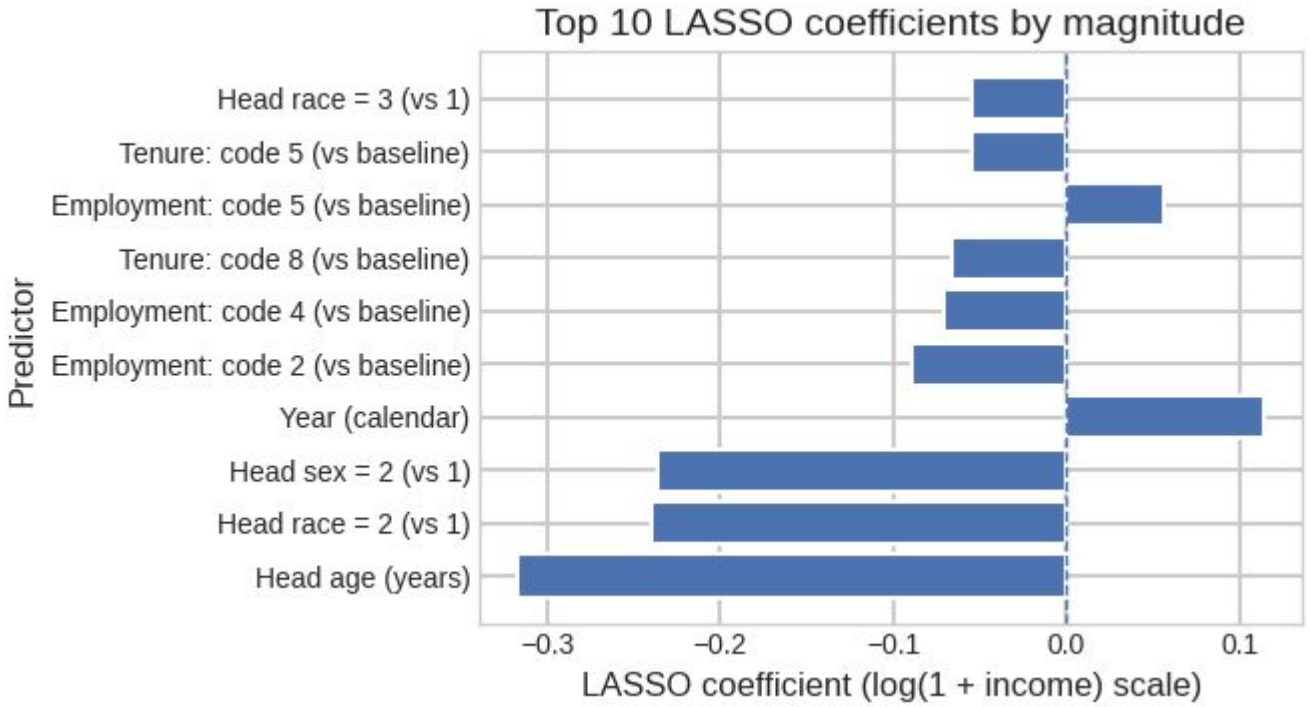
- Heads in **sex category 2** and some **race categories** have **lower** predicted log income than baseline groups

- Certain employment statuses and tenure categories (e.g., renting, non-employed) also associated with lower income

OLS


Top 10 OLS coefficients by magnitude
Predictor
Employment: code 4 (vs baseline)
Tenure: code 8 (vs baseline)
Head race = 3 (vs 1)
Decade: 1980s (vs 1960s)
Decade: 1970s (vs 1960s)
Decade: 2000s (vs 1960s)
Head race = 2 (vs 1)
Head sex = 2 (vs 1)
Employment: code 2 (vs baseline)
Decade: 2010s (vs 1960s)
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
OLS coefficient (log(1 + income) scale)

LASSO


Top 10 LASSO coefficients by magnitude
Predictor
Head race = 3 (vs 1)
Tenure: code 5 (vs baseline)
Employment: code 5 (vs baseline)
Tenure: code 8 (vs baseline)
Employment: code 4 (vs baseline)
Employment: code 2 (vs baseline)
Year (calendar)
Head sex = 2 (vs 1)
Head race = 2 (vs 1)
Head age (years)
−0.3
−0.2
−0.1
0.0
0.1
LASSO coefficient (log(1 + income) scale)

Random Forest: which predictors matter most?

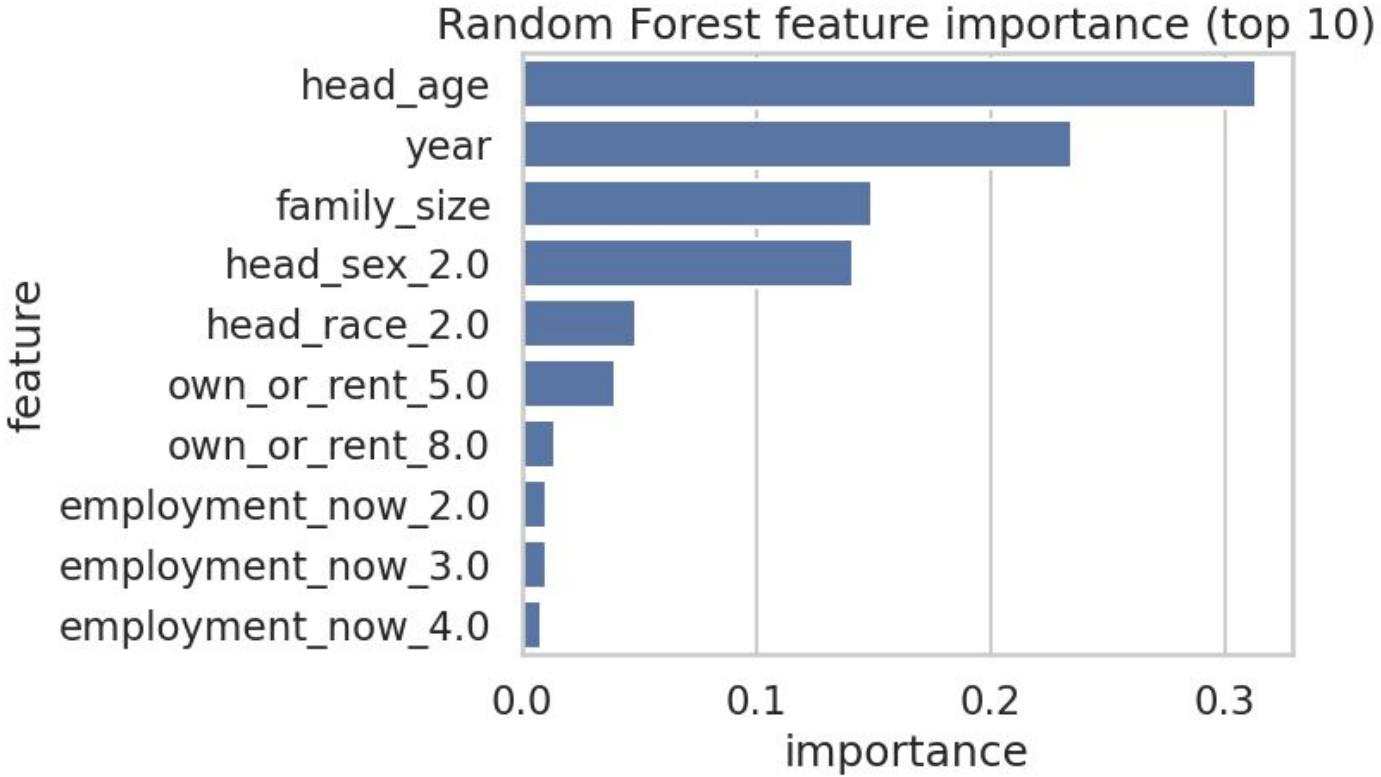
Top features

- **Head age**, **year**, and **family_size** have highest importance
- Sex, race, tenure, and employment dummies also contribute but less strongly

Interpretation

- Tree-based model confirms that **age + time + household structure** are key drivers
- Broad agreement with linear models on which variables carry the most signal

Random Forest: which predictors matter most?



Neural Network Predictor Impact (Permutation Importance)

Top features

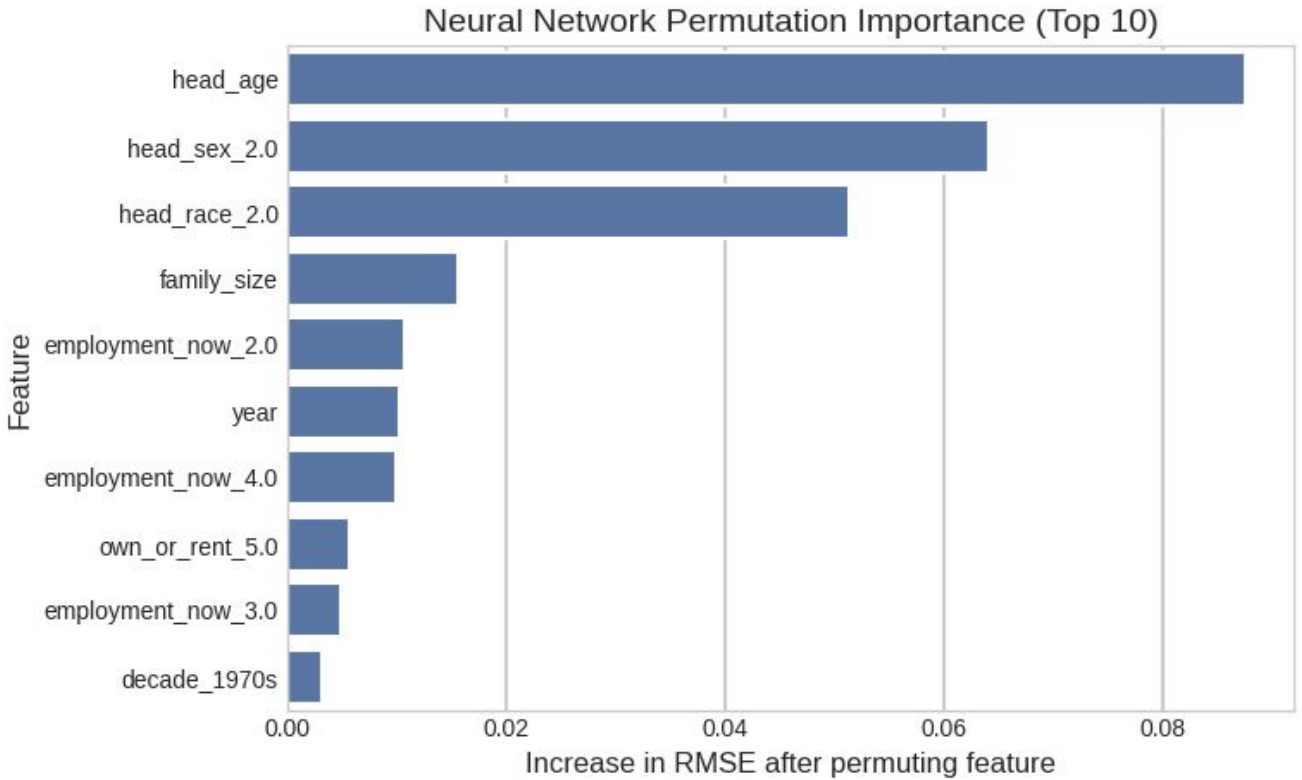
- **Head age**, **head sex 2.0**, and **head race 2.0** have highest importance
  Family size, employment dummies, tenure dummies, and year also contribute

Interpretation

- Neural network model confirms that **age + sex + race** are key drivers

- Broad agreement with linear and nonlinear (random forest) models on which features are most impactful

Neural Network Predictor Impact (Permutation Importance)

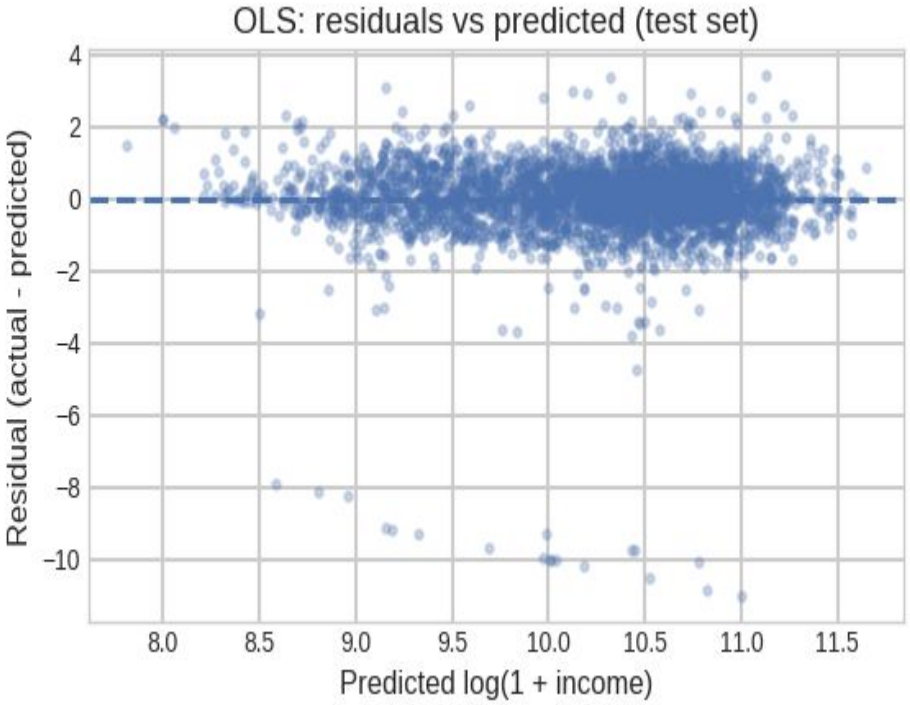


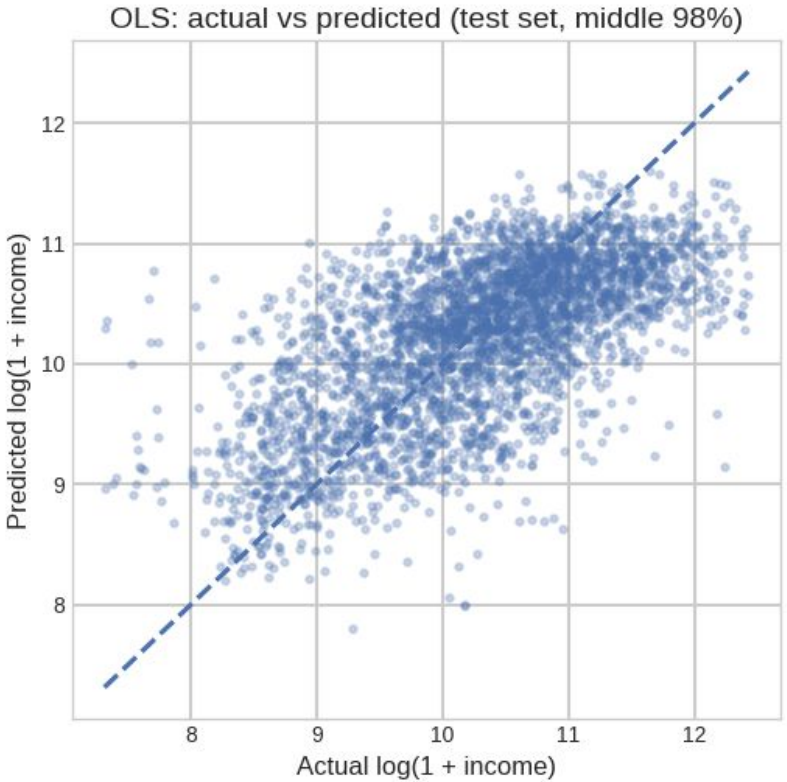
OLS diagnostics: how well do we fit?

Observations

- Points roughly follow 45° line ➔ model captures **overall trend**
- Under-prediction for highest incomes and over-prediction for lowest incomes
- Residuals mostly centered near zero but show **heteroskedasticity / nonlinear patterns** at extremes

OLS diagnostics: how well do we fit?





Summary of findings

Time trend: household log income rises over calendar years and newer decades
Head characteristics:

- Older heads tend to have **higher** incomes
- Sex and race categories show **systematic differences** in expected income

Household structure:

- Tenure and employment status are important predictors in all models

Model comparison:

- Neural network provides **best performance**
- OLS and LASSO provide **similar performance** to neural network model
- Random Forest is less accurate but agrees on which predictors matter

Limitations

Limitations

- $R^2 \approx 0.28 \rightarrow$ large role for unobserved factors (education, occupation, region, etc.)

- Linear model assumes **additive effects**; RF still relatively simple tuning

- Coding of sex and race is coarse; not a causal fairness analysis
- Model weights and interactions are difficult to interpret, even with permutation importance
- Neural network might memorize noise due to feature space only consisting of 9 predictors and small sample size of 17,900