Predicting Years of Education and Human Development Index Values Using 1994 Census Data Grouped by Controllable and Uncontrollable Attributes

Presented by Benton Stacy and Katarya Johnson-Williams

An inspirational quote…

“Who cares?”

\- Dr. Condori

Spring 2023

Who cares?

- United States Census Bureau
- What is census data used for?
    - Determine electoral bodies
    - Allocate funding for important initiatives
    - Gauge changes over time
        - Demographics, economics, etc.

Why did we do this?

- Learning!
  - About census data
    - Why does it matter?
  - Application of class concepts
    - Julia documentation 💀
  - Discovery
    - What new opportunities await?

Methodology

Our Data

- 1994 Census Data
    - From Irvine Machine Learning Repository
- 48,842 instances
- 15 attributes

1. Age
2. Work class
3. Education
4. Education number
5. Marital status
6. Occupation
7. Relationship
8. Race
9. Sex
10. Capital gain
11. Capital loss
12. Hours worked per week
13. Native country
14. Final weight
15. Income level (above or below $50,000)

What did we do?

- Created 10 Models
  - Initial Least Squares Solution
  - Used Cross Validation (5 Folds)
- Models 1-6 use arbitrary pairs of attributes
  - Predict years of education
- Models 7-10 use controllable/uncontrollable categories
  - Categories will be explained later
  - 7-8 predict years of education
  - 9-10 predict Human Development Index (HDI)

How did we do it?

- Threw out several attributes
  - Final weight, capital gain/loss, education number, and relationship
- Categorized attributes
  - Controllable/uncontrollable, etc.
- Removed rows with missing values
  - To save time!
- Added two attributes
  - Gross Domestic Product (GDP) for native countries
    - Secondary economic indicator
  - Human Development Index (HDI) for native countries
    - Secondary education indicator

Data Columns

1. Age – no change
2. Work Class – 3 columns (private, self-employed, government, *not working**)
3. Education – translated into years (1 year to 24 years)
4. Marital Status – 2 columns (single, married, *previously married**)
5. Occupation – 4 columns (engineering, business, technical, non-degree, *government**)
6. Race – 4 columns (White, Asian Pacific-Islander, Native-American, Black, *Other**)
7. Sex – no change (Boolean)
8. Hours Worked Per Week – no change
9. Native Country – 3 columns (North America, South America, Asia, *Europe**)
10. Income – no change (Boolean)
11. GDP – numerical
12. HDI – numerical

**Implicit Columns*

Discovery Models

Categories

Model 1: Age and Hours Worked Per Week

Model 2: Work Class and Occupation

Model 3: Marital Status and Native Country

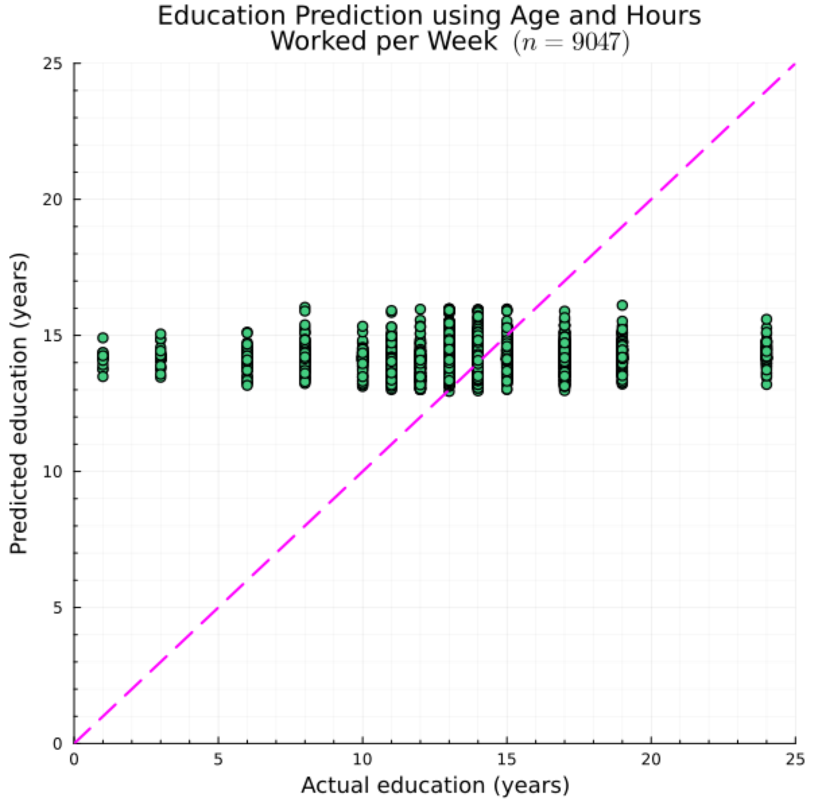
Model 4: Race and Sex

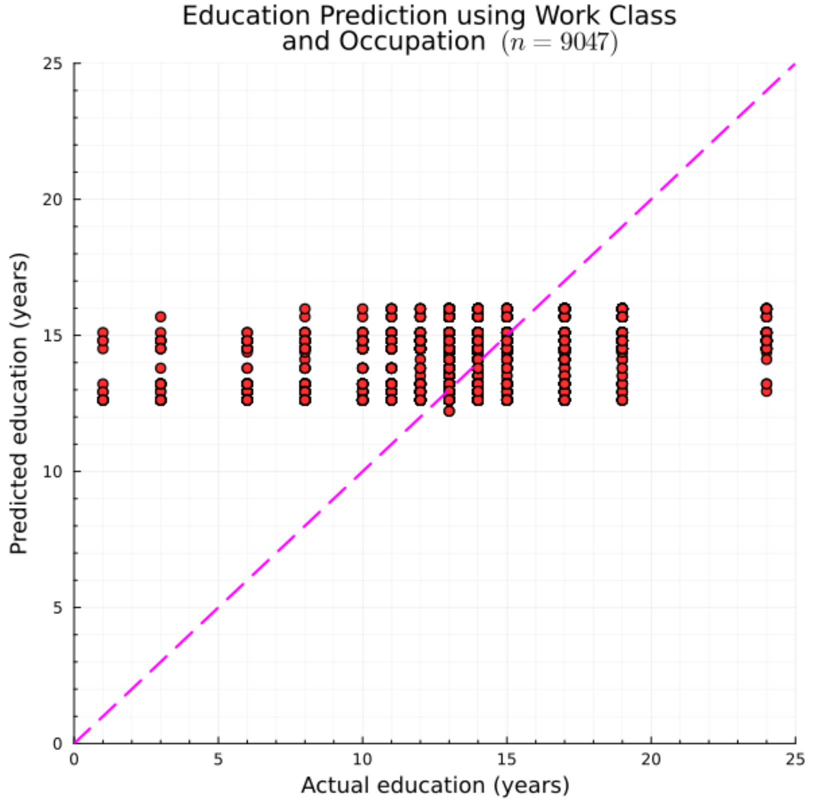
Model 5: Gross Domestic Product and Income

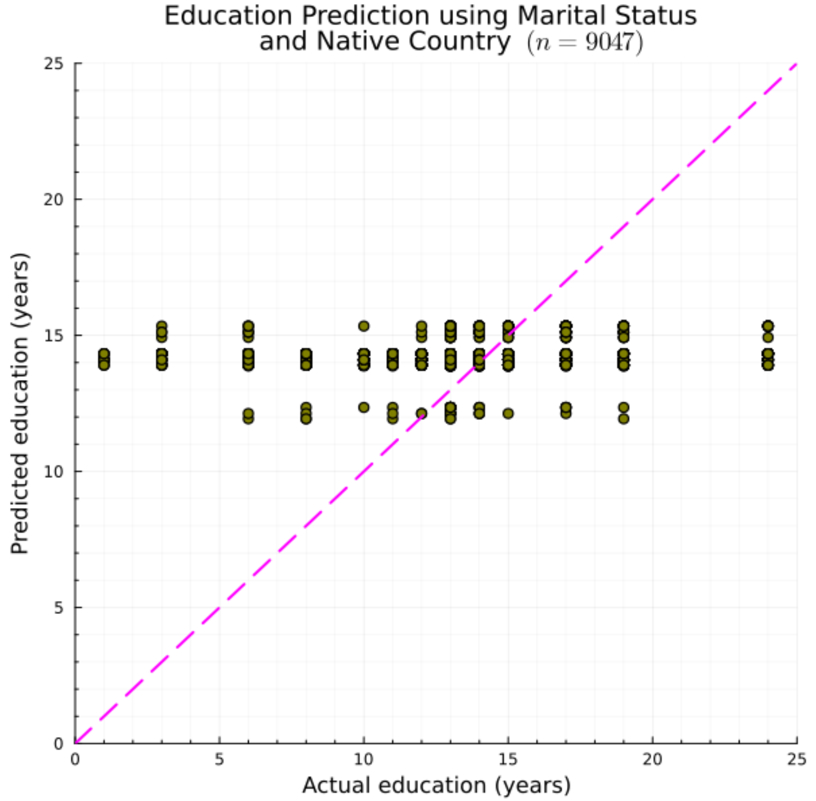
Model 6: Human Development Index and Income

Initial Least Squares

Discovery Models (Least Squares)

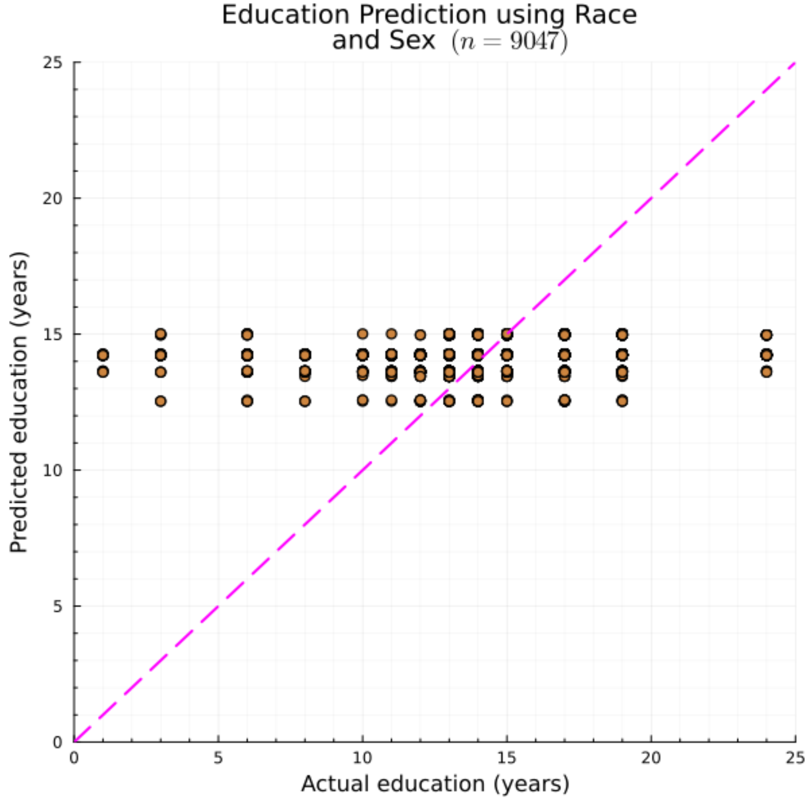


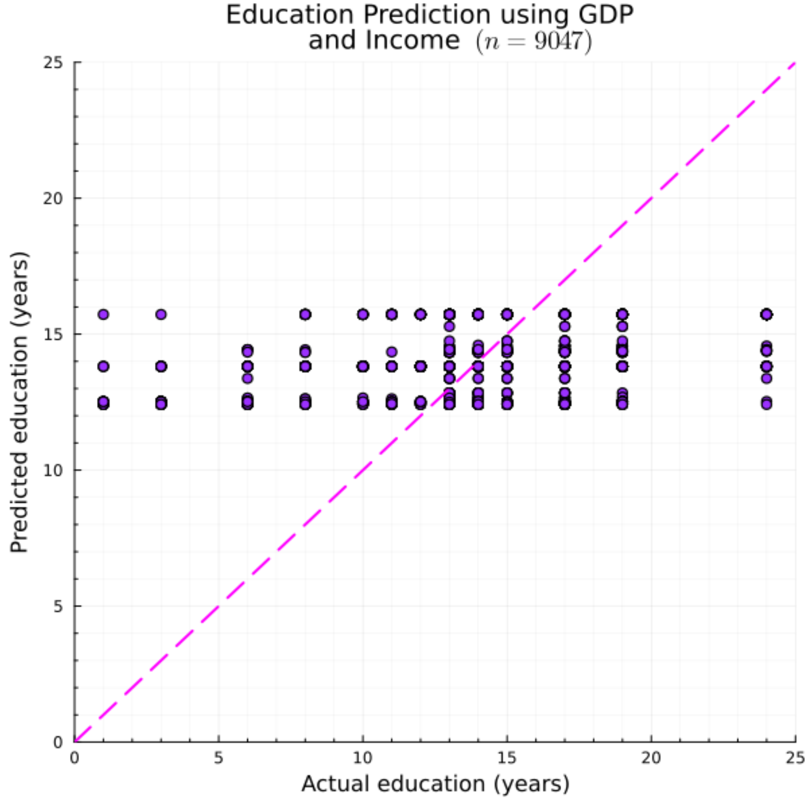


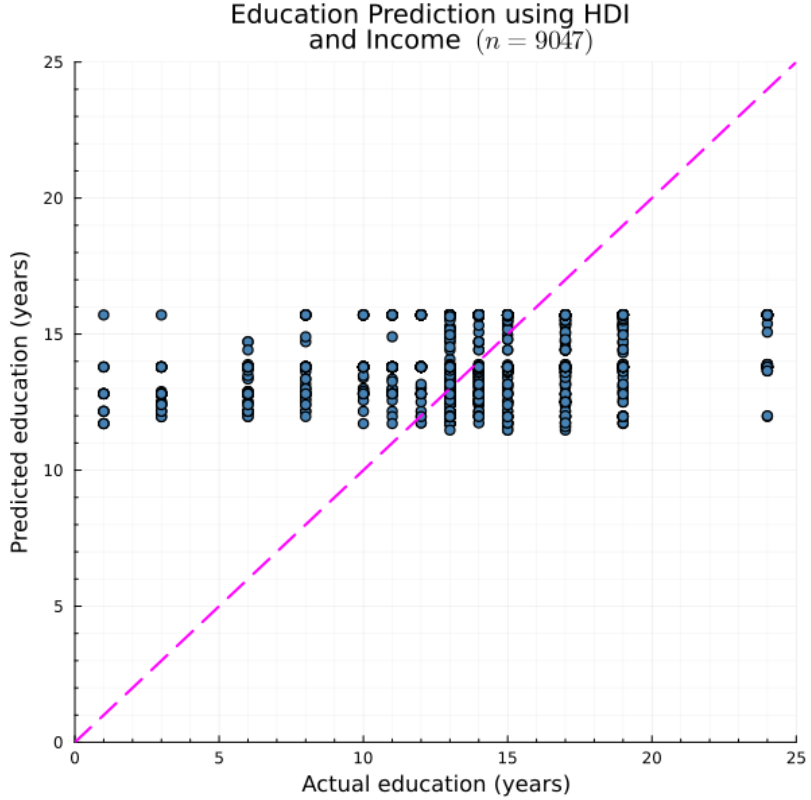


	Training RMS	Testing RMS
Model 1 (Green)	2.6751650113399617	2.679998454219119
Model 2 (Red)	2.502494041374889	2.512325126447655
Model 3 (Olive)	2.685490258184635	2.683340402831584

Discovery Models (Least Squares)



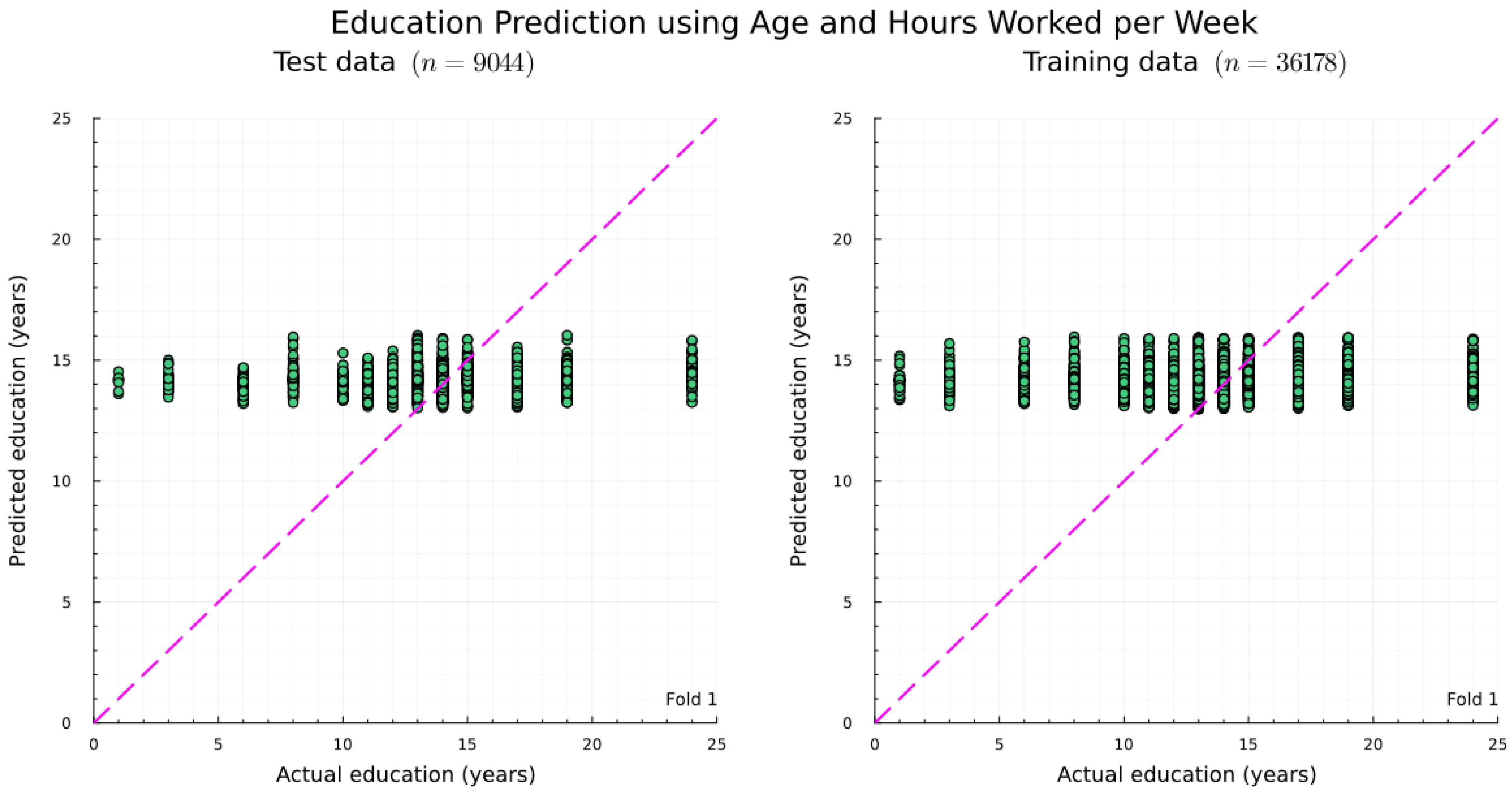




	Training RMS	Testing RMS
Model 4 (Orange)	2.68532773666033	2.6798834191143026
Model 5 (Purple)	2.53986984246196	2.5391713577388058
Model 6 (Cyan)	2.5438250170376357	2.5457038692366583

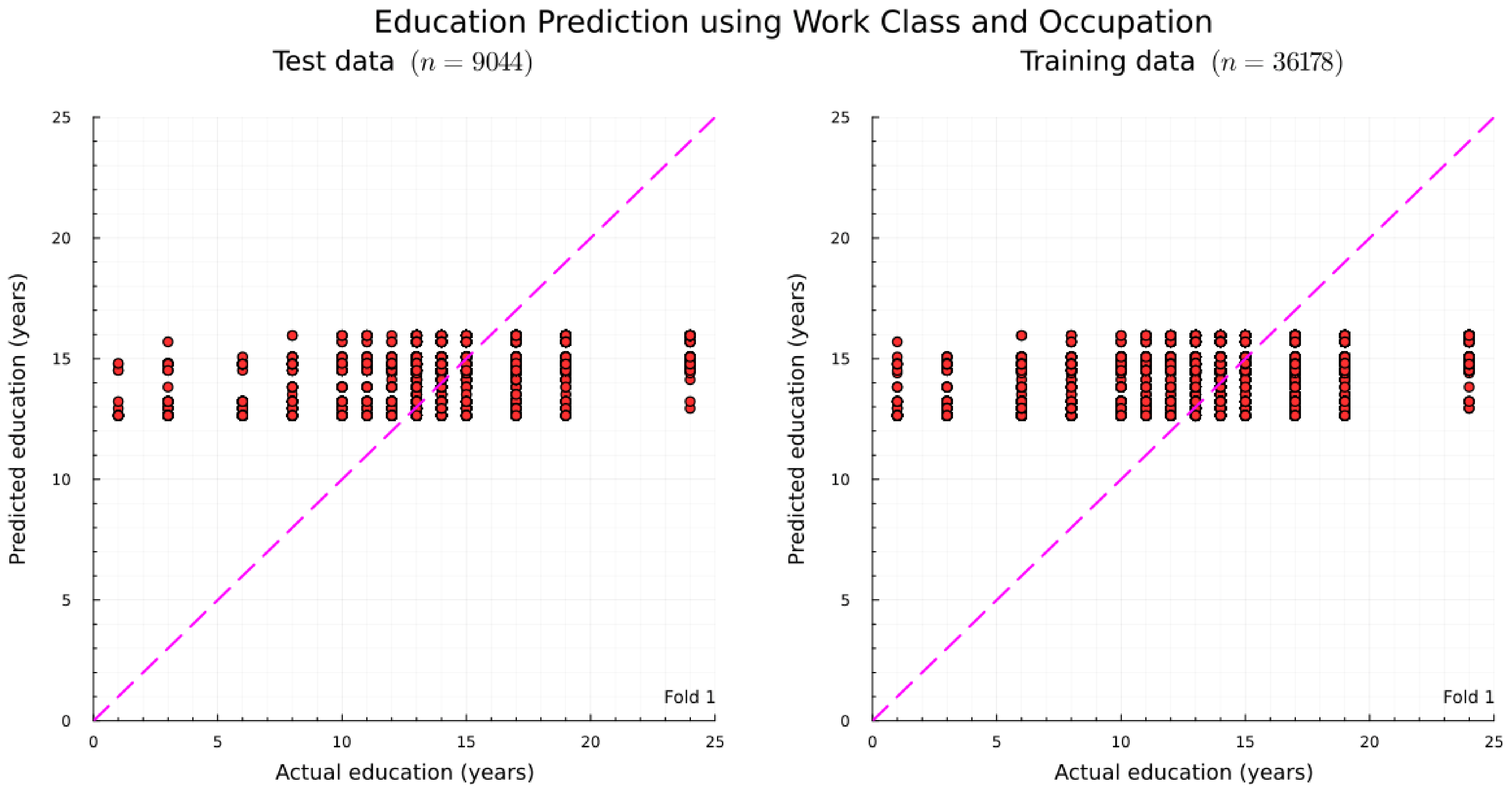
Cross-Validation

Model 1: Cross-Validation



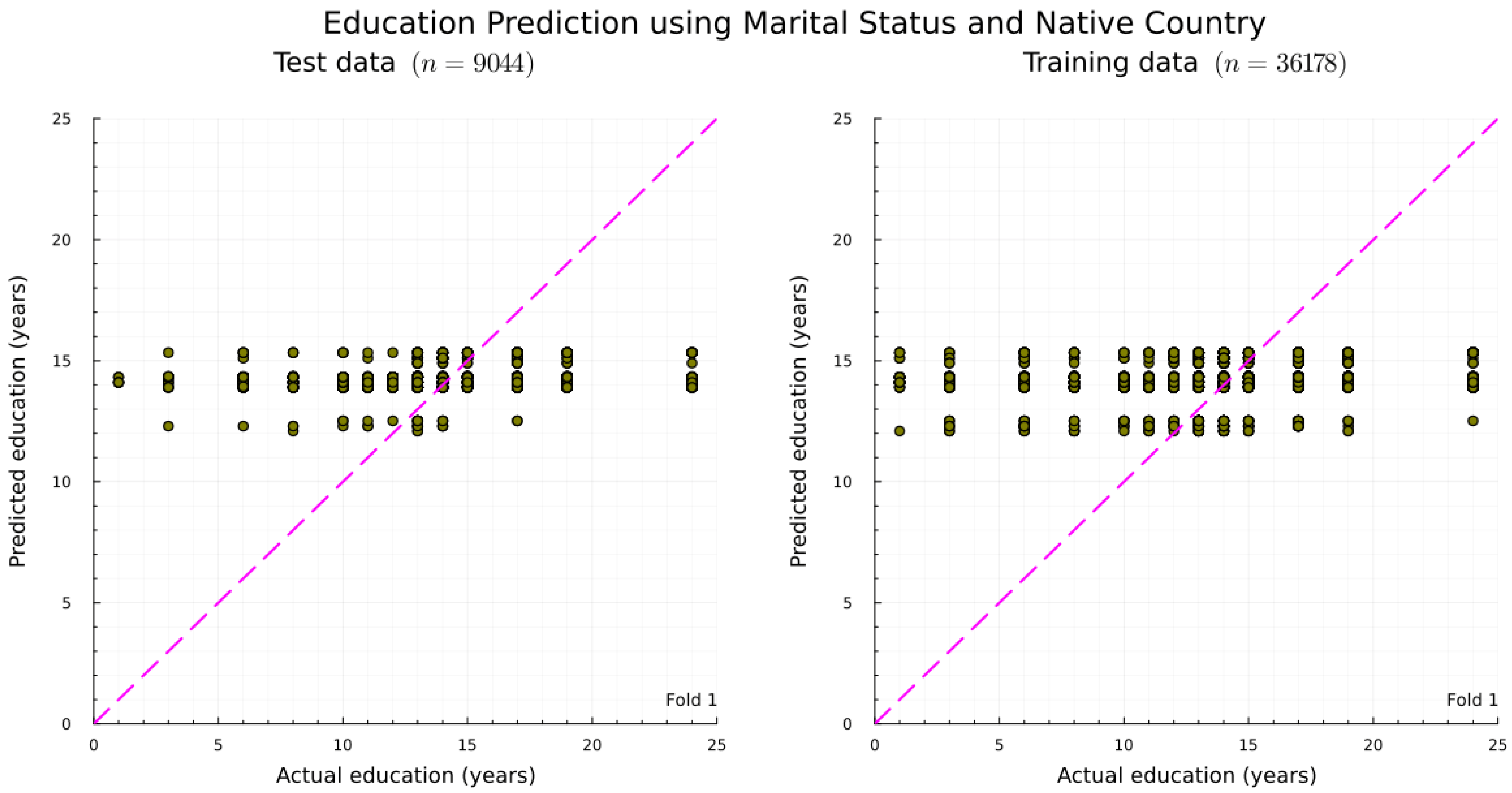
Training RMS	**2.67745**	**2.68094**	**2.67959**	**2.6779**	**2.66435**
Testing RMS	**2.67061**	**2.65658**	**2.66205**	**2.66884**	**2.72266**

Model 2: Cross-Validation



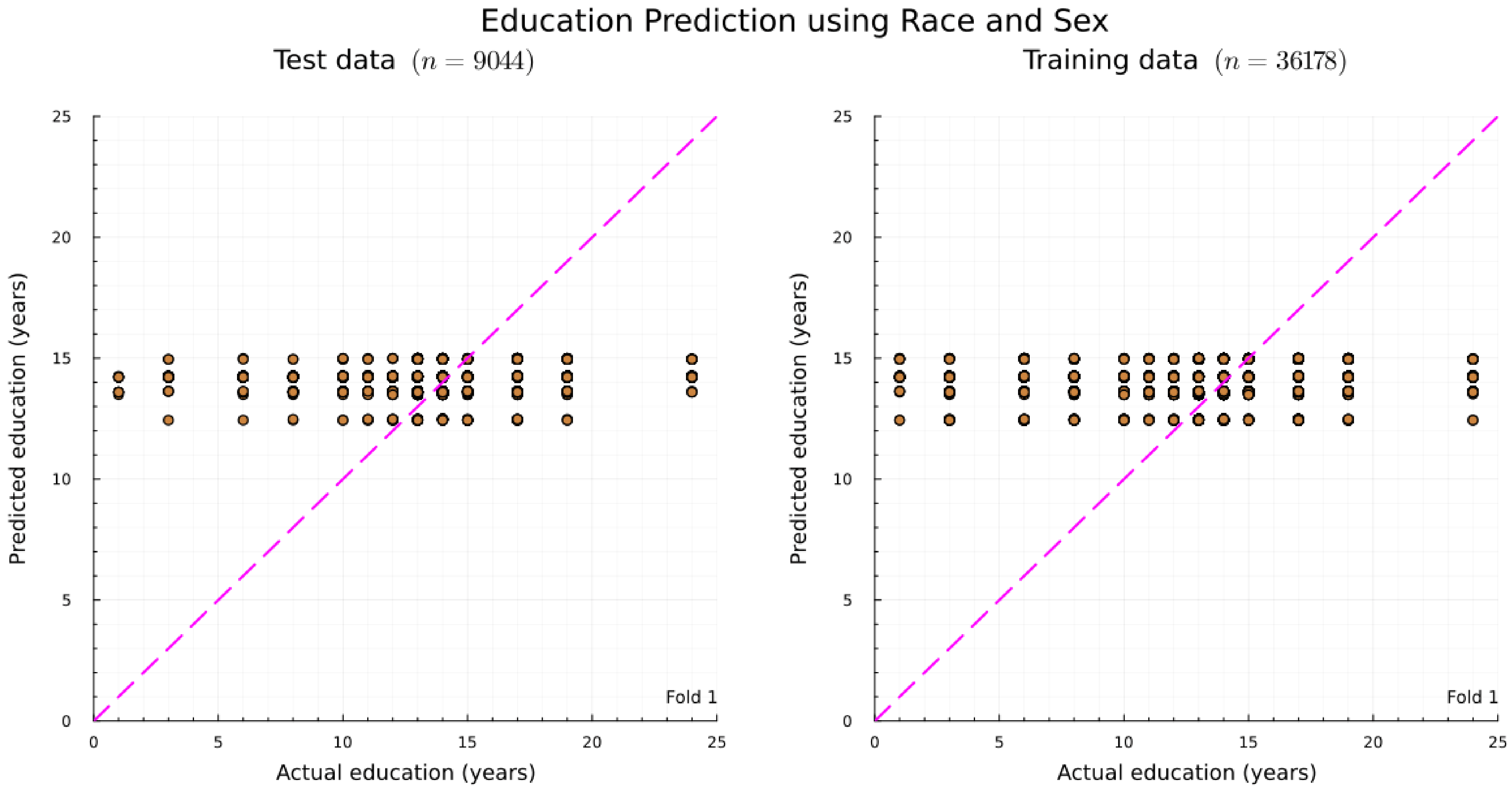
Training RMS	2.50425	2.50096	2.51185	2.50032	2.50428
Testing RMS	2.5053	2.51863	2.47474	2.52122	2.50516

Model 3: Cross-Validation



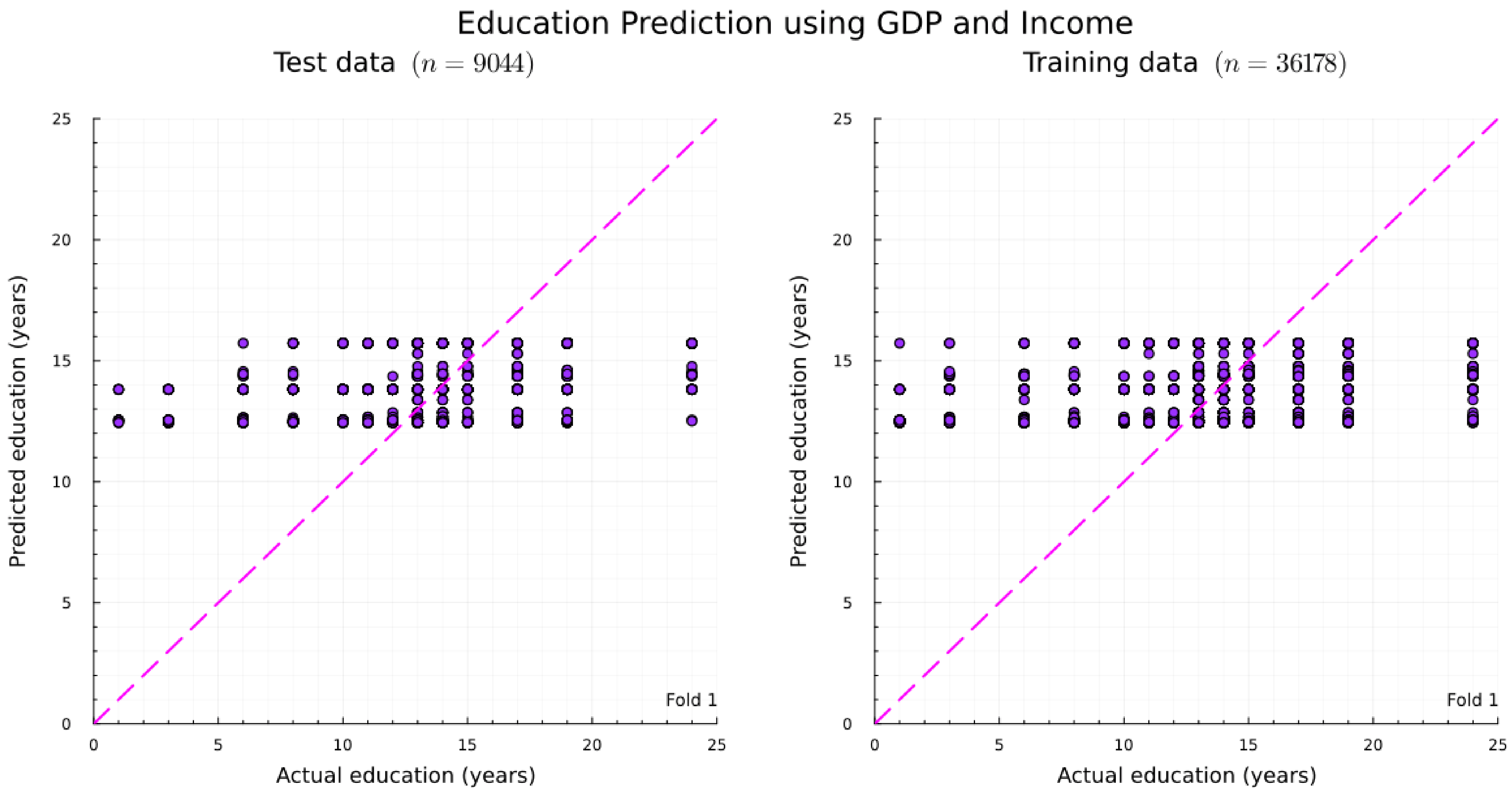
Training RMS	2.69031	2.69045	2.68677	2.67987	2.6775
Testing RMS	2.66411	2.66357	2.67825	2.7057	2.71531

Model 4: Cross-Validation



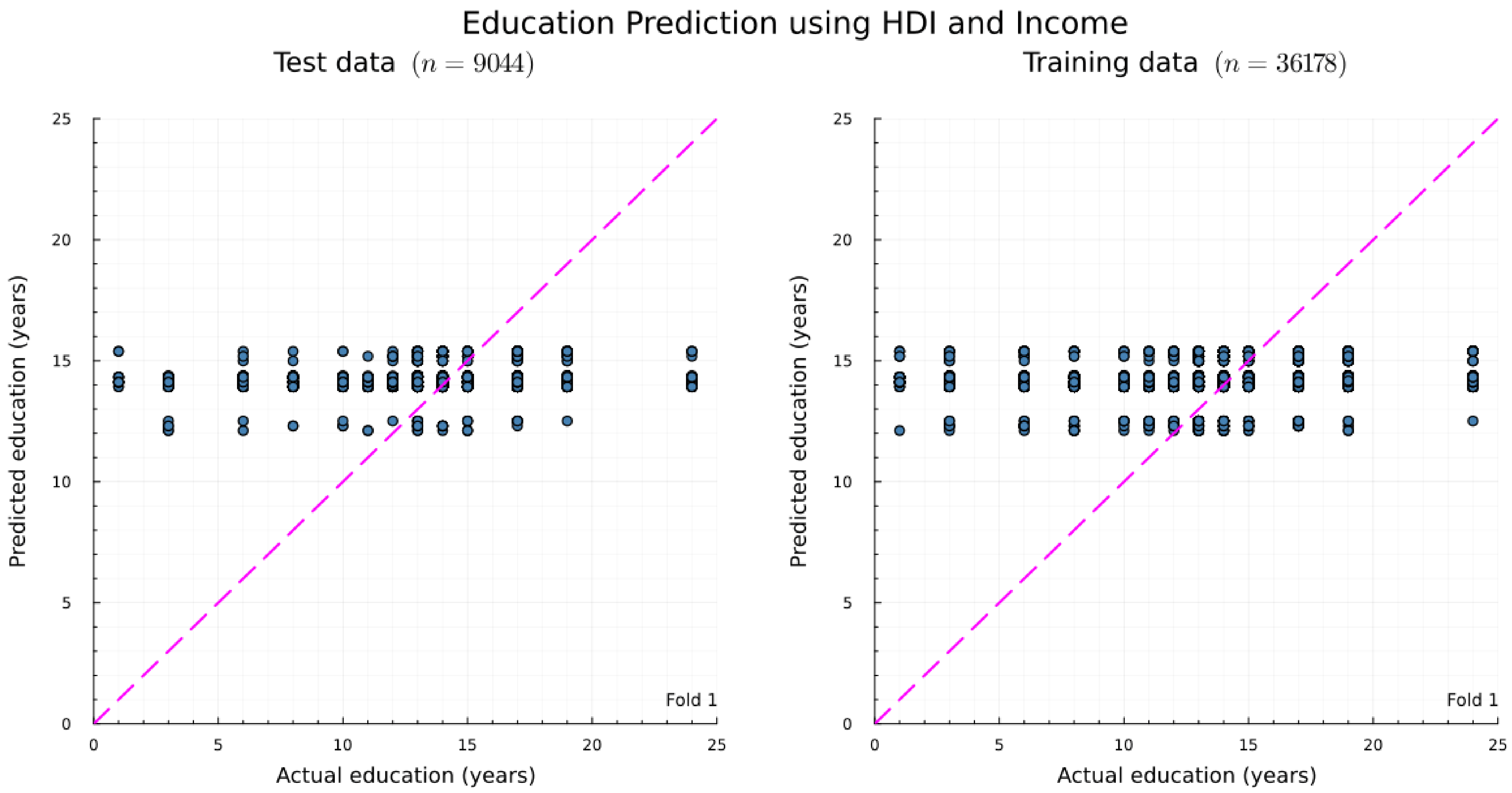
Training RMS	2.69404	2.68734	2.68132	2.68417	2.67377
Testing RMS	2.64442	2.67151	2.69568	2.68441	2.7255

Model 5: Cross-Validation



Training RMS	**2.53674**	**2.54578**	**2.53636**	**2.52809**	**2.55153**
Testing RMS	**2.55167**	**2.51548**	**2.55322**	**2.58578**	**2.4921**

Model 6: Cross-Validation



Training RMS	2.68376	2.68492	2.68126	2.68909	2.68584
Testing RMS	2.69063	2.68577	2.70024	2.66901	2.68212

Conclusions: Discovery Models

- Nothing highly significant
  - “On average” 2-3 years off
- Cross-Validation showed consistency
- Lowest RMS Error: Work Class and Occupation
- Highest RMS Error: Marital Status and Native Country

Predicting Years of Education

Categories

Uncontrollable (Model 7)

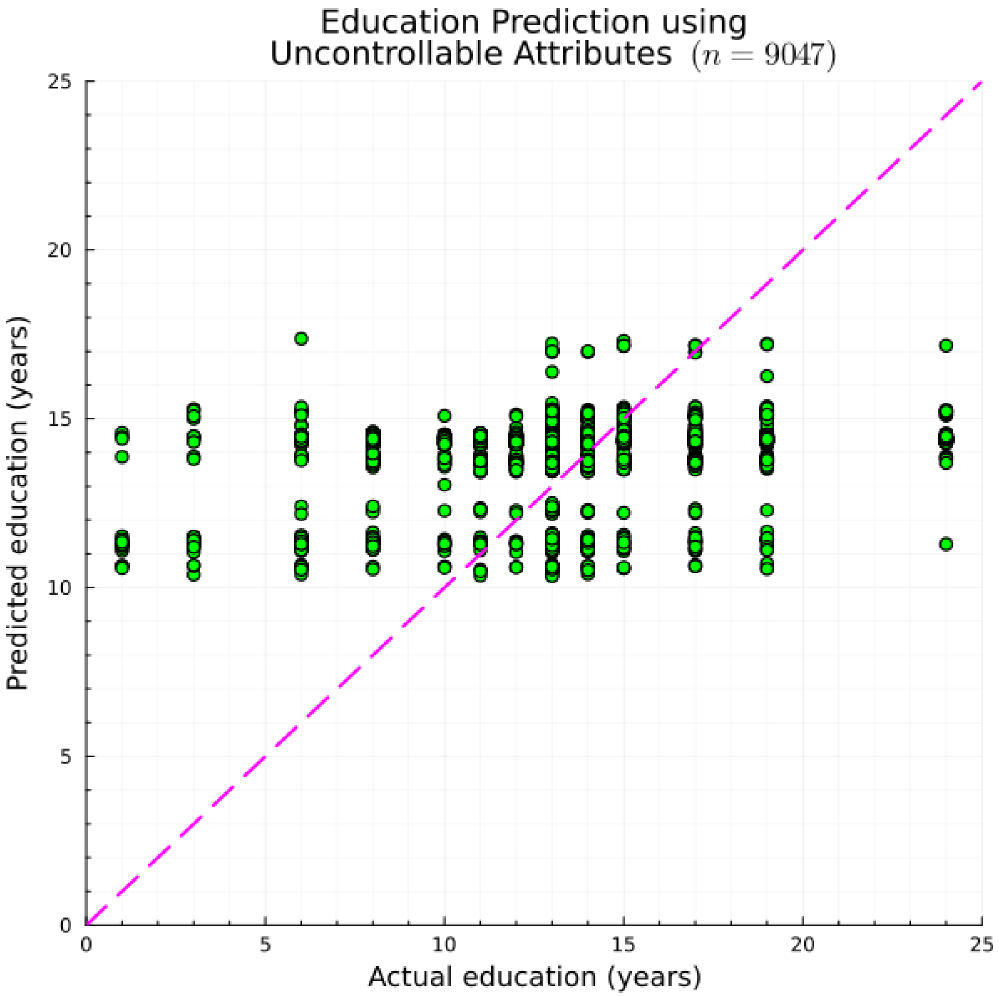
- Native Country
- Gross Domestic Product (GDP)
  - For Native Country
- Race
- Sex
- Age

Controllable (Model 8)

- Work Class
- Occupation
- Marital Status
- Income
- Hours Worked Per Week

Initial Least Squares

Model 7: Least Squares

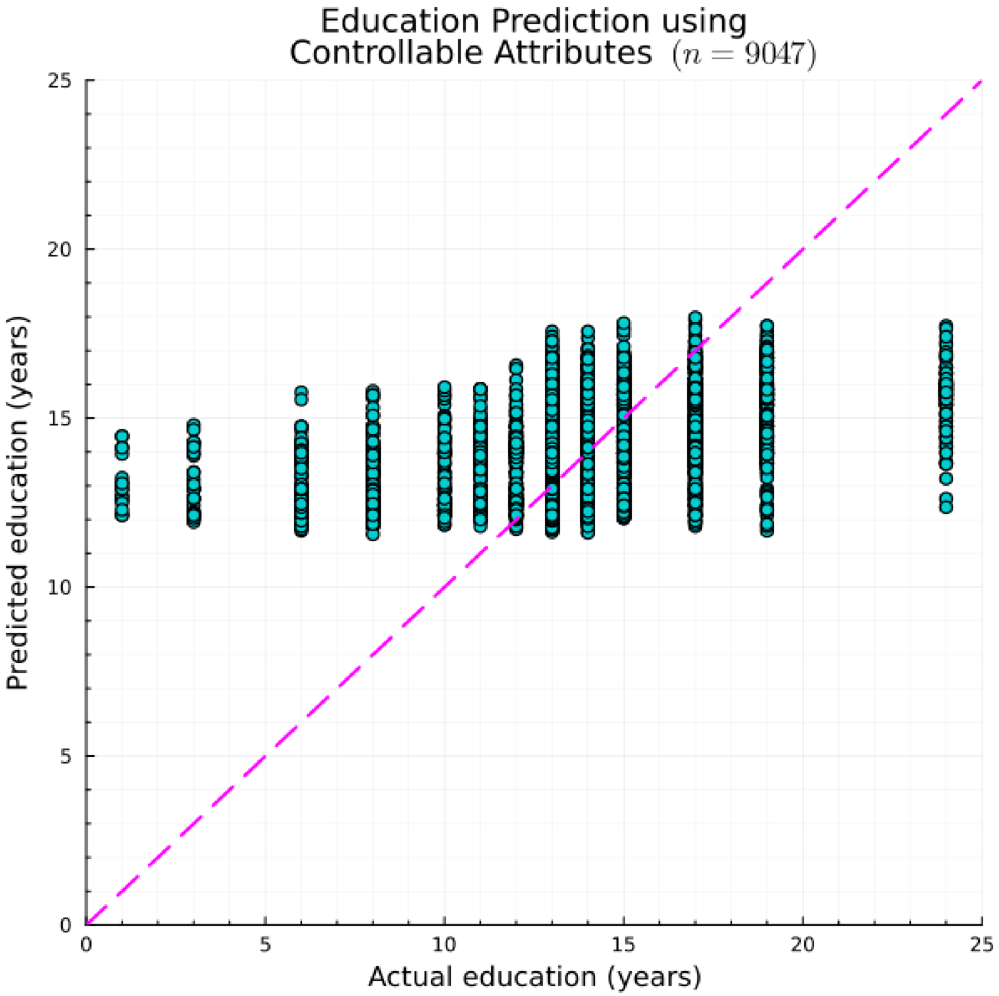


X-Hat Values:

Attribute		Value
Y-intercept		13.042
Native Country:	North America	-2.72529
	South America	-1.54074
	Asia	0.915771
Native GDP (in $billions)		0.000450635
Race:	White	0.531306
	Asian/Pacific Islander	0.877301
	Native American	-0.165588
	Black	-0.0172389
Sex:	Female	0.02444
Age:		0.00543944

	Training RMS	Testing RMS
Uncontrollable	2.6137331343608308	2.6037659893776155

Model 8: Least Squares



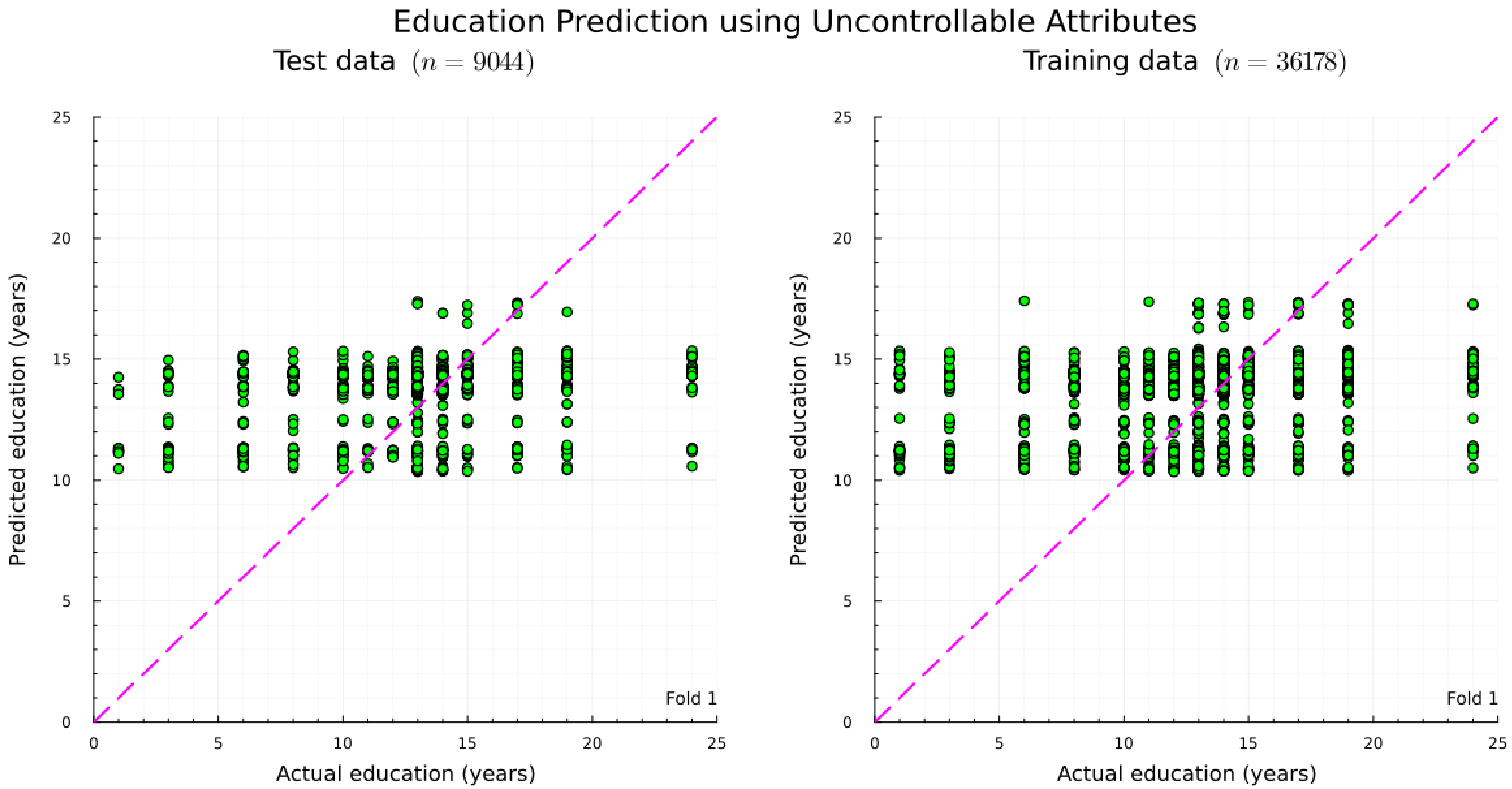
X-Hat Values:

Attribute		Value
Y-intercept		13.66
Work Class:	Private	0.0269376
	Self-employed	0.206359
	Government	1.12609
Occupation:	Engineering	0.35586
	Business	1.44653
	Technical	1.68272
	Non-degree	-0.123143
Marital Status:	Single	0.449082
	Married	-0.325552
Income		-1.71803
Hours Worked per Week		0.0151695

	Training RMS	Testing RMS
Controllable	2.398386831440762	2.417192522529947

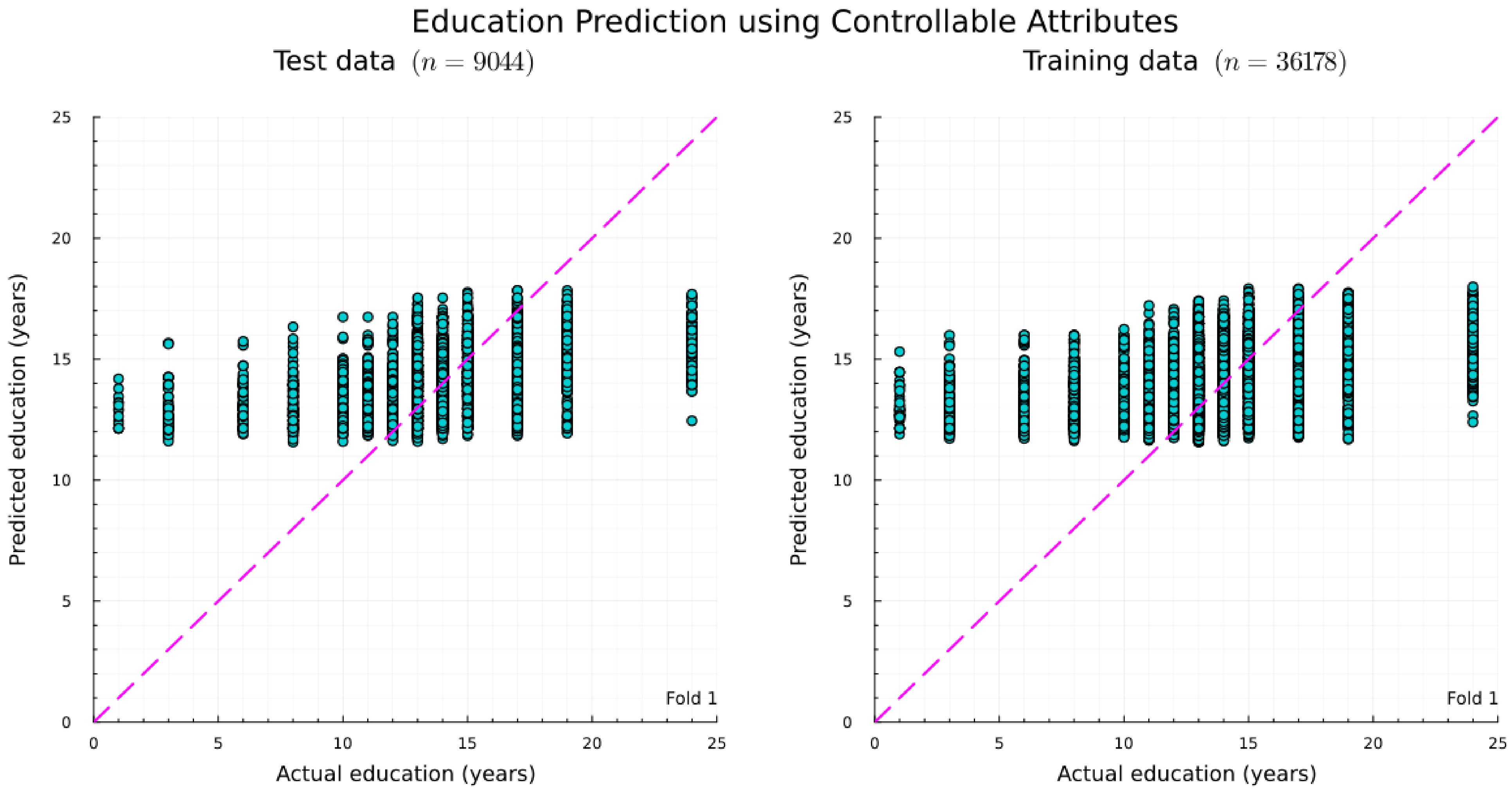
Cross-Validation

Model 7: Cross-Validation



Training RMS	2.61364	2.61541	2.62344	2.59976	2.60386
Testing RMS	2.6025	2.59521	2.56259	2.65769	2.64146

Model 8: Cross-Validation



Training RMS	**2.38612**	**2.41069**	**2.40545**	**2.39829**	**2.40919**
Testing RMS	**2.46485**	**2.36761**	**2.38852**	**2.41742**	**2.37347**

Conclusions: Predicting Years of Education

- Slightly more significant
- Cross-Validation showed consistency
- Observations
  - Controllable had a higher accuracy
    - ~2.4 years error < ~2.6 years error
  - More slope (better weighting)

Predicting Human Development Index (HDI)

Categories

Uncontrollable (Model 9)

- Native Country
- Gross Domestic Product (GDP)
  - For Native Country
- Race
- Sex
- Age

Controllable (Model 10)

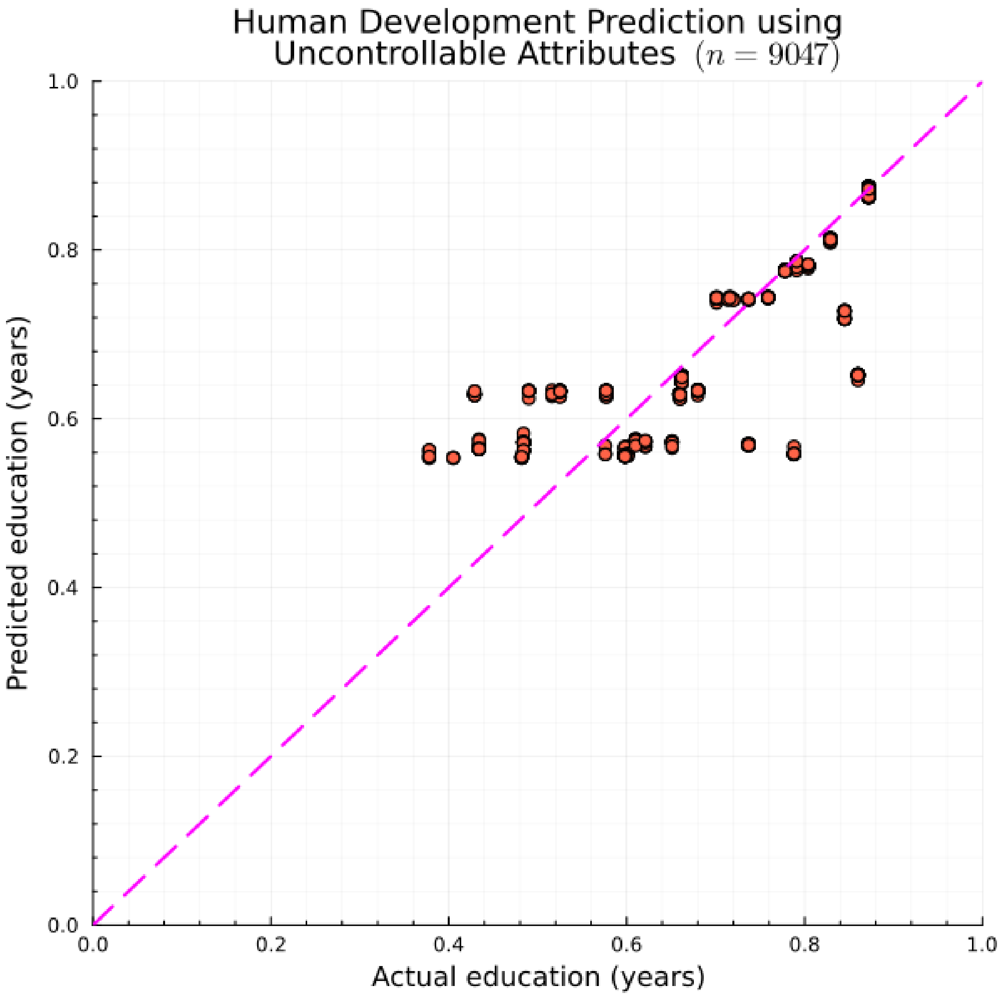
- Work Class
- Occupation
- Marital Status
- Income
- Hours Worked Per Week

Initial Least Squares

Note: HDI is measured on a scale from 0 (worst) to 1 (best)

Best Model Award

Model 9: Least Squares

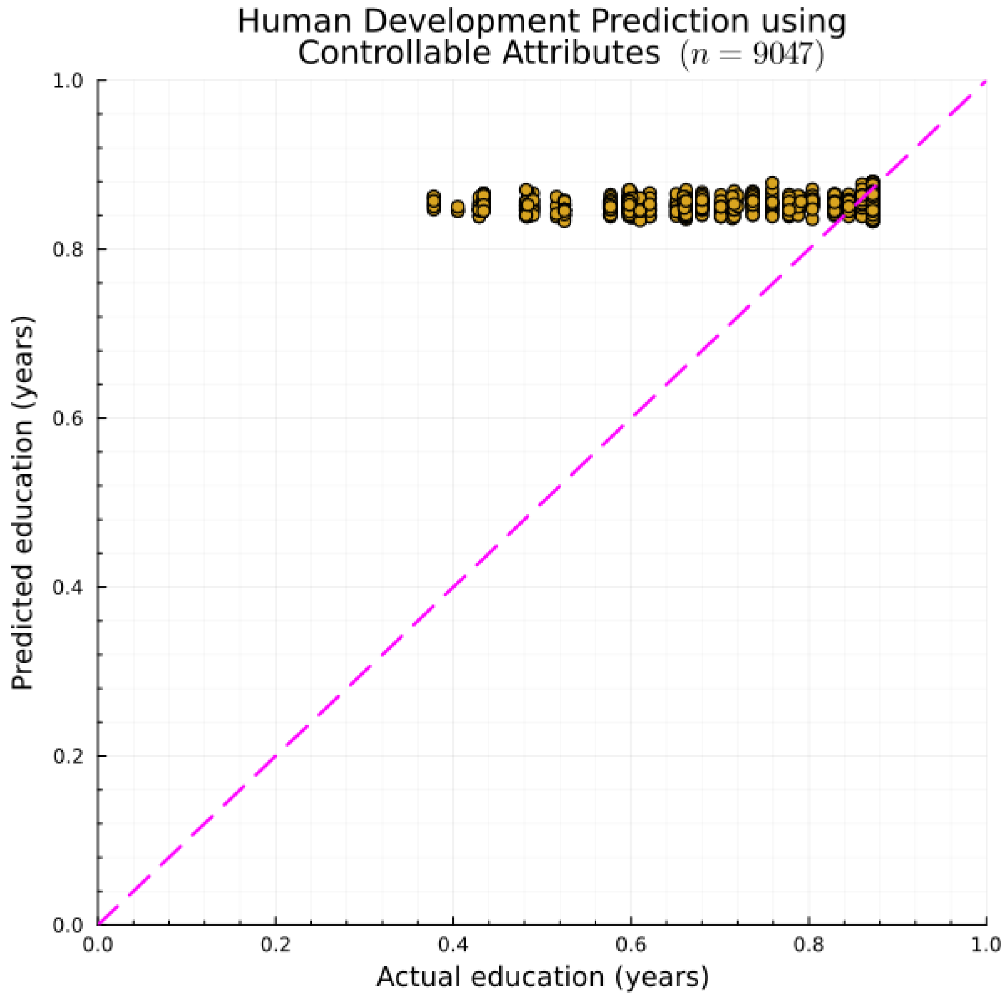


X-Hat Values:

Attribute		Value
Y-intercept		0.734605
Native Country:	North America	-0.109331
	South America	-0.162632
	Asia	-0.176781
Native GDP (in $billions)		0.0000333779
Race:	White	0.00261042
	Asian/Pacific Islander	-0.0068521
	Native American	0.00351155
	Black	-0.00155952
Sex:	Female	0.000928655
Age:		0.0000355595

	Training RMS	Testing RMS
Uncontrollable	0.024364083939679447	0.024681003356696894

Model 10: Least Squares



X-Hat Values:

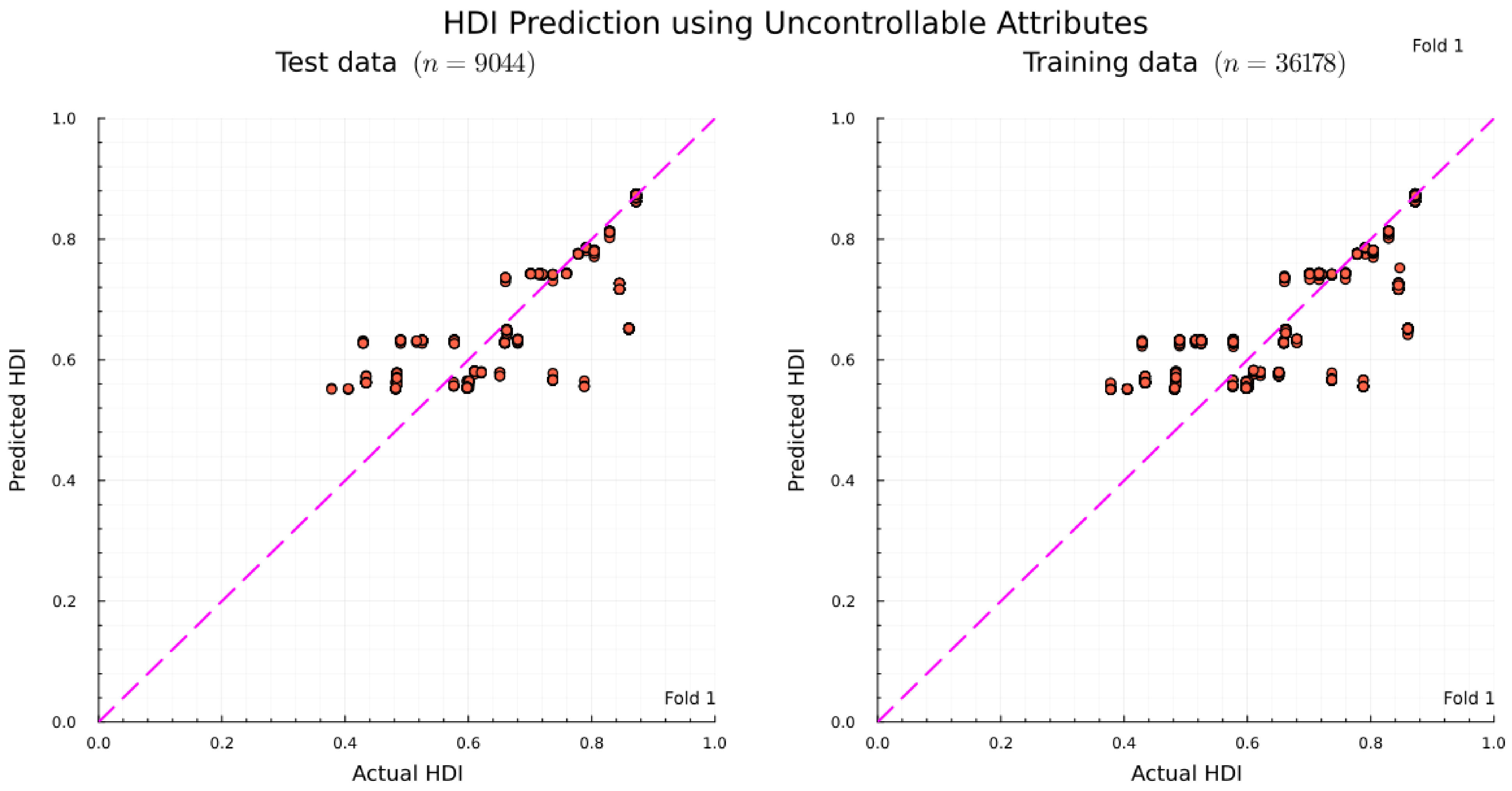
Attribute		Value
Y-intercept		0.861265
Work Class:	Private	-0.00947663
	Self-employed	-0.00360414
	Government	-0.00121374
Occupation:	Engineering	0.00499996
	Business	0.015298
	Technical	0.00959016
	Non-degree	0.00565312
Marital Status:	Single	-0.005171
	Married	-0.0111575
Income		-0.00746344
Hours Worked per Week		0.0000600894

	Training RMS	Testing RMS
Controllable	0.07015984984327628	0.06857875224288963

Cross-Validation

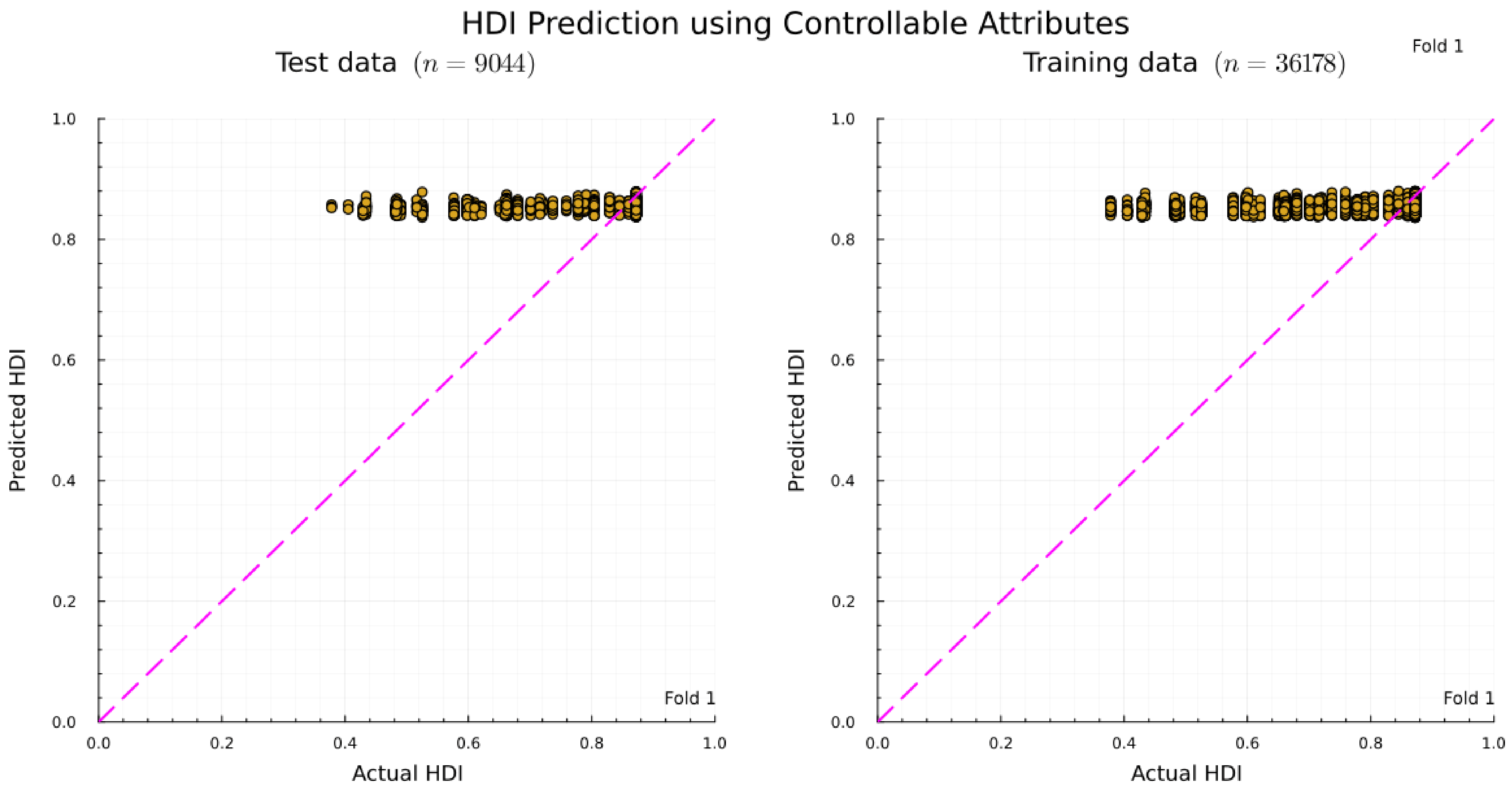
Note: HDI is measured on a scale from 0 (worst) to 1 (best)

Model 9: Cross-Validation



Training RMS	0.0245069	0.0245038	0.0240429	0.0245159	0.0245856
Testing RMS	0.0241563	0.024161	0.0259486	0.0241315	0.0238383

Model 10: Cross-Validation



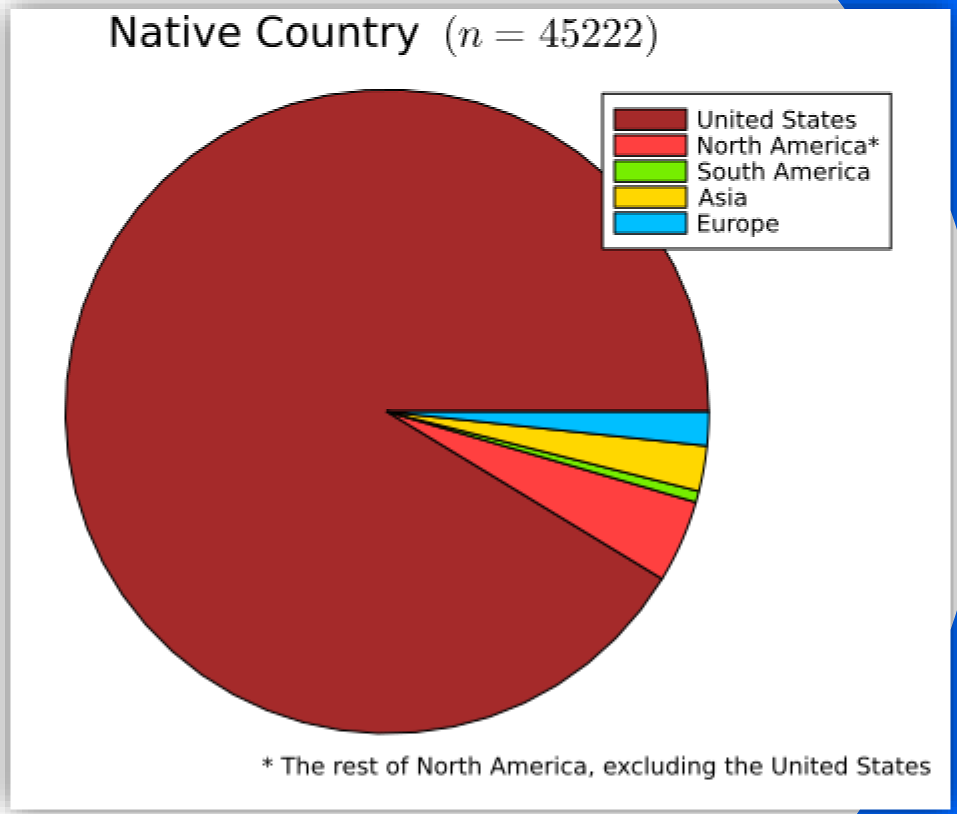
Training RMS	0.0692897	0.0698806	0.0701464	0.0698257	0.0700713
Testing RMS	0.0720302	0.0697094	0.0686374	0.069935	0.0689464

Conclusions: Predicting HDI

- Significant results
- Cross-Validation showed consistency
- Observations
  - Model 9 has a reasonable slope (~2% error)
  - Model 10 has no slope (~7% error)
  - Both predict within 10% error

Results

- Model 9 was significant
- Refining attributes was important
- Diversity of United States makes prediction hard
  - See chart (over 91% United States)
- Skewed attributes make prediction hard
  - Income (75% under $50k)
  - Education (mostly high school grads)
  - Race (mostly white)



Future Work

- What if we had the heritage of individuals instead of native country?
    - How does culture impact years of education?
    - Can we predict where someone is from?
- Can we generalize models to 2020 census data?
- What if we had more attributes?
    - Number of kids, parental education, state of residence…
    - Raw GDP vs. Per Capita GDP with better location data
- What if we had more inclusive/detailed attributes?
    - Native country, race, occupation to name a few!

Thank you!

Benton Stacy

(bmstacy7127@eagle.fgcu.edu)

Katarya Johnson-Williams

(kajohnsonwilliam3168@eagle.fgcu.edu)

References

Becker, Barry. *Adult*. UCI Machine Learning Repository, 1996. *DOI.org (Datacite)*, https://doi.org/10.24432/C5XW20.

GDP - Gross Domestic Product 1994 | Countryeconomy.Com. https://countryeconomy.com/gdp?year=1994. Accessed 17 Apr. 2023.

Home · Plots. https://docs.juliaplots.org/stable/. Accessed 24 Apr. 2023.

Introduction · DataFrames.Jl. https://dataframes.juliadata.org/stable/. Accessed 24 Apr. 2023.

Lemon, Chet, et al. *Predicting If Income Exceeds $50,000 per Year Based on 1994 US Census Data with Simple Classification Techniques*. https://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/048.pdf.

References (cont.)

Nations, United. *Documentation and Downloads*. United Nations. *hdr.undp.org*, https://hdr.undp.org/data-center/documentation-and-downloads. Accessed 17 Apr. 2023.

Nations, United. *Human Development Index*. United Nations. *hdr.undp.org*, https://hdr.undp.org/data-center/human-development-index. Accessed 23 Apr. 2023.

Our Censuses. United States Census Bureau, https://www.census.gov/programs-surveys/censuses.html. Accessed 17 Apr. 2023.

Subramanian, S. V., et al. “Multilevel Perspectives on Modeling Census Data.” *Environment and Planning A: Economy and Space*, vol. 33, no. 3, Mar. 2001, pp. 399–417. *DOI.org (Crossref)*, https://doi.org/10.1068/a3357.