Machine Learning Final Project:

Econ 178

March 21, 2025

Written for Econ 178 by:

Ashvin Chaudhary

1 Introduction

This project focuses on predicting total wealth (in US dollars) based on household data from the 1991 Survey of Income and Program Participation (SIPP). The objective is to apply various statistical methods to build and evaluate prediction models for wealth, using a range of feature variables such as income, home ownership, retirement savings (IRA, 401(k)), and other personal attributes. Methods like Ordinary Least Squares (OLS), Ridge Regression, Stepwise Selection, Lasso, and flexible linear models will be used to compare their effectiveness in predicting total wealth, ensuring that the selected models are both accurate and interpretable. The goal is to develop the best model to be used to predict total wealth on a new unseen dataset with the same predictors.

2 Exploratory Analysis

2.1 Data

The data contains 7933 observations with 17 predictors and total wealth. When making predictions, watch out for multicollinearity with education levels (and respective dummies) and home equity(=home value − mortgage).

2.2 Visualizations of Relationships

I began visualizing the relationships between various variables and total wealth. Aside from the expected relationships, such as income, I noted several key observations, which I describe below along with their graphs.

Variable	Description
`tw`	Total wealth (in US $)
`ira`	Individual retirement account (IRA) (in US $)
`e401`	1 if eligible for 401(k), 0 otherwise
`nifa`	Non-401k financial assets (in US $)
`inc`	Income (in US $)
`hmort`	Home mortgage (in US $)
`hval`	Home value (in US $)
`hequity`	Home value minus home mortgage (in US $)
`educ`	Education (in years)
`male`	1 if male, 0 otherwise
`twoearn`	1 if two earners in the household, 0 otherwise
`nohs`	1 if no high school, 0 otherwise
`hs`	1 if high school graduate, 0 otherwise
`smcol`	1 if some college, 0 otherwise
`col`	1 if college graduate, 0 otherwise
`age`	Age
`fsize`	Family size
`marr`	1 if married, 0 otherwise

Table 1: Variables and Their Descriptions

2.2.1 Family Size and Wealth

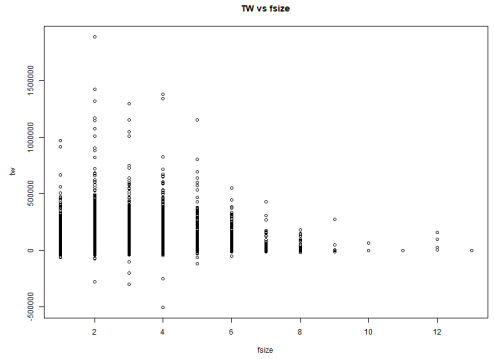


Figure 1: Total Wealth vs. Family Size

Family size exhibits a non-linear relationship with total wealth. Wealth peaks around two children and then declines. I could use a non-linear transformation in my model such as a step function.

2.2.2 Marital Status and Wealth

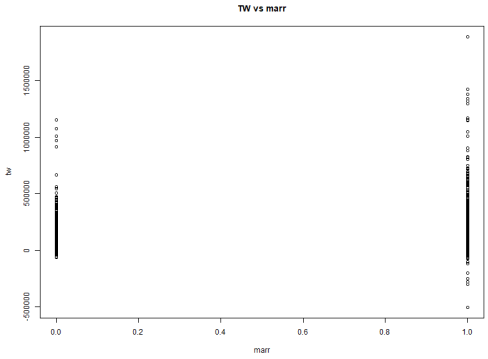
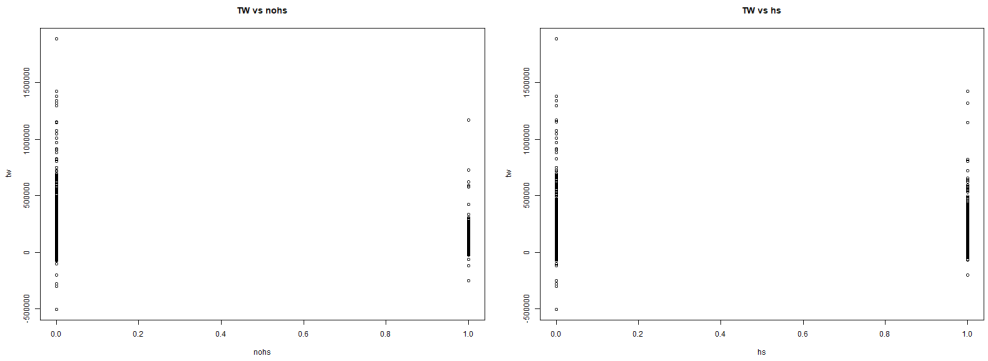


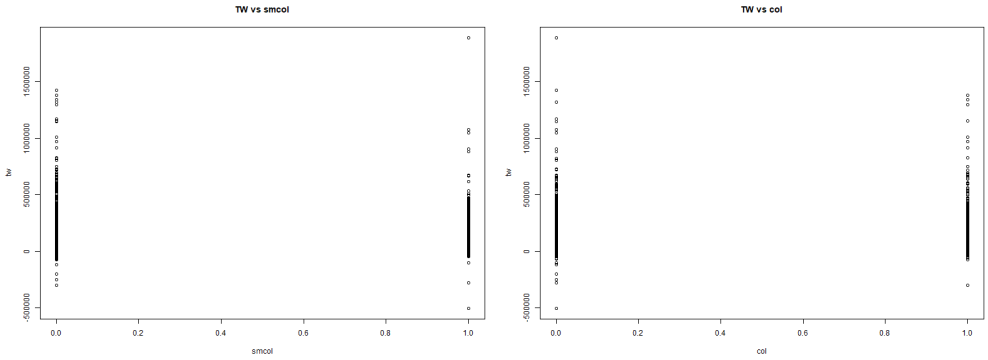
Figure 2: Total Wealth vs. Marital Status

Married individuals show more variation in total wealth but tend to have higher wealth on average. I plan to use an interaction term with other variables as relationships seems to differ if one is married or not.

2.2.3 Education Level and Wealth



(a) No High School (b) High School



(c) Some College (d) College Graduate

Figure 3: Total Wealth vs. Education Categories

The difference in wealth appears more pronounced for individuals with no high school education (nohs) compared to other education categories. When comparing some college education with other groups, the variation increases since it includes college, high school, and no high school categories in the comparison. Dummies seem to be a stronger comparison then the continuous education available.

2.2.4 Home Equity, Mortgage, and Wealth

(a) Total Wealth vs. Home Equity

(b) Home Mortgage

(c) Home Value

Figure 4: Total Wealth vs. Home Equity, Home Mortgage, and Home Value

Home equity has a much stronger linear relationship with total wealth compared to home mortgage (hmort) and home value (hval).

3 Dealing With Outliers

To improve model performance for Ridge, Lasso, and Subset Selection, I applied data cleaning techniques to handle outliers and skewed data, as these models are sensitive to extreme values. As we have not discussed this in class, I am just going to apply this to income and total wealth to avoid over complicating it.

- **IQR-Based Outlier Removal:** The Interquartile Range (IQR) method is used to detect and remove values outside the 25$ and 75$ quartile from ‘tw‘ and ‘inc‘. Learned to use box-whisker

plot in Econ5 to visualize outliers, so I thought I could apply it in this setting.

- **Wealth-Cap 1,000,000:** I used Torsha's method of Wealth-Cap at 1,000,000 as a second method to compare my IQR method too

A total of 3 observations were dropped using Method 1 (IQR Outlier Removal), and 14 observations were dropped using Method 2 (Wealth Cap at 1,000,000). Despite these removals, a substantial number of observations remain for model training.

I first created a baseline model using Ordinary Least Squares (OLS) regression with all predictors. To choose the distance from the second and third quintile to keep observations, I compared the average MSPE of all three models using k-cross validation on the training data, and landed at 15xIQR as the range since it was where the MSPE was on average the lowest. I will continue to train models on all three datasets(IQR, Torsha's-Wealth Cap, full set) and compare using cross validation to see which performs the best on the full set.

4 Ridge, Lasso, Forward/Backward Selection

4.1 Basic Linear Ridge, Lasso, Forward/Backward Selection

We began by applying Ridge, Lasso, and Forward/Backward stepwise regression models. For Ridge and Lasso, coefficient reduction played a significant role, especially as both model's coefficients were reduced with large lambda values. Comparing the Mean Squared Prediction Error (MSPE), forward stepwise regression performed the best, though it did not outperform the basic OLS model trained on the dataset after applying IQR-based outlier or the Wealth Limit. When training on a cleaned dataset, lasso performed best overall. Across cross-validation, we observed models that remove features or set them to zero perform the best, suggesting opportunities to optimize ridge regression further as our current Ridge model has all the features. Introducing interaction terms and feature selection methods could improve ridge regression, and comparisons with the base models would be insightful. It appears that the IQR cleaning method positively impacts the predictive power of the linear models so moving forward to save time I will be training on the IQR dataset and testing on the original dataset.

Table 2: Average MSPE(Original Data) for Models Trained on Different Datasets

Model	Trained on Original Data	Trained on IQR Cleaned Data	Trained on Wealth
Baseline Model	1730735289	1722469293	1726890345
Stepwise Forward	1728616000	1718798674	1724764669
Stepwise Backward	1728652755	1718798674	1724442394
Ridge Regression	1749210513	1736086915	1735930324
Lasso Regression	1728321114	1718564904	1724828668

4.2 Adding Interaction Terms

In linear models, interaction terms can be added to capture relationships between predictors that may not be apparent when considered individually. To create the interaction terms, I used the following R code:

```
interaction_terms <- combn(names(data_no_tw), 2, FUN = function(x) paste(x[1], x[2], sep=":"),
simplify = TRUE)
```

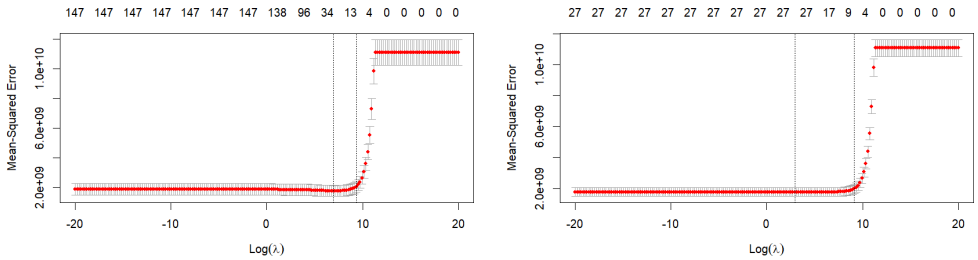
This generated a dataset containing all possible 1:1 interaction terms. Subsequently, I applied

Lasso regression for feature selection, as it can shrink coefficients to zero, thus performing variable selection. Additionally, I utilized forward and backward stepwise selection to identify the optimal combination of interaction terms and features as well. Since stepwise selection uses the AIC to compare models, I did not worry about the potential inflation of R-squared from adding more features, as AIC penalizes the inclusion of additional predictors.

Although one could refer to visualizations of the interaction terms, this method was efficient in narrowing down the most relevant combinations of interactions and features for predicting the outcome variable. Through 10-fold cross-validation, I achieved the lowest Average Mean Squared Prediction Error (MSPE) of 1,659,537,955 with Ridge Regression, using the features selected by stepwise methods. I also did Lasso Regression as well with the selected features but Ridge still performed the best.

Variable	**Coefficient**	**Variable**	**Coefficient**
(Intercept)	14293.27	inc:age	0.02
hequity	0.91	twoearn:age	−412.96
nifa	1.10	nifa:e401	0.07
ira	1.22	hequity:ira	0.00
inc	−0.73	ira:male	0.75
e401	−18526.54	e401:age	371.03
twoearn	13861.77	hequity:e401	−0.06
age	−360.66	inc:male	0.14
hmort	0.09	nifa:male	−0.08
male	−2779.73	nifa:inc	−0.00

To select the optimal value of the regularization parameter λ, I plotted λ against the Mean Squared Error (MSE). The optimal λ was found to lie within the range that was manually chosen. The corresponding plots, `Rplot01.png` and `Rplot.png`, illustrate this selection process.



(a) Lasso Features Lasso Lambda (b) Stepwise Features Ridge Lambda

Figure 5: Distribution of Lambda and MSE

4.3 Interpretiability and Performance

As we transition from basic models, like Least Squares regression, to more complex models, such as Ridge regression with interaction terms, the interpretability becomes more challenging. In the basic Least Squares regression model, the relationships between predictors and the target variable are straightforward. The coefficients are directly interpretable, with positive coefficients indicating an increase in the target variable when the predictor increases, and negative coefficients showing a decrease plus with t-tests to see if the coefficients are statistically significant(not equal zero).

Variable	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-16800.1154	8372.1175	-2.01	0.045 *
ira	1.5666	0.0590	26.56	< 0.0000000000002 ***
e401	7802.3858	1080.0309	7.22	< 0.0000000000002 ***
nifa	1.0741	0.0119	90.37	< 0.0000000000002 ***
inc	0.3347	0.0299	11.18	< 0.0000000000002 ***
hmort	-0.0196	0.0196	-1.00	0.317
hval	0.0721	0.0113	6.35	< 0.0000000000002 ***
hequity	NA	NA	NA	NA
educ	-255.3899	470.0183	-0.54	0.587
male	3493.6221	1357.3766	2.57	0.010 *
twoearn	-7447.9429	1384.8975	-5.38	< 0.00000007672 ***
nohs	-955.7204	4258.8953	-0.22	0.822
hs	-671.9751	2606.8658	-0.26	0.797
smcol	646.6415	1984.7888	0.33	0.745
col	NA	NA	NA	NA
age	308.0882	54.2376	5.68	0.00001397215 ***
fsize	-8.5144	407.5951	-0.02	0.983
marr	1725.9552	1567.9596	1.10	0.271

Table 3: Least Squares Regression Results

However, as we transition to more complex models like Ridge regression, we introduce a penalty term to shrink large coefficients. While this helps reduce overfitting, it also brings some trade-offs. The inclusion of all these interaction terms adds flexibility to the model, but it can lead to counterintuitive results. For instance, it may not make sense for income to have a negative coefficient in the base model, but show positive interaction effects with other variables like age or home equity. It is still possible to use the coefficients to calculate a predicted total wealth, but with all the interactions make it difficult to understand what type of person who are predicting without taking some time compared to a simple least squares estimate.

In Ridge regression, the coefficients are shrunk towards zero(but not zero), meaning they are no longer unbiased estimates of the "true" population values. This shrinkage causes the coefficients to appear smaller than they might be in the underlying dataset, reducing the model's interpretability. Although the coefficients are biased, this bias helps improve the predictive power of the model, as it reduces the variance by controlling the influence of less important predictors.

We use cross-validation to determine the optimal amount of bias (through the tuning of the lambda parameter) to minimize the Mean Squared Prediction Error (MSPE). By introducing bias, we improve the model's ability to generalize to new data by reducing variance(fitting to training data by shrinking coefficients) at the cost of interpretability.

In summary, while Ridge regression with interaction terms improves the model′s flexibility and accuracy by capturing complex relationships, it sacrifices some interpretability. The shrinkage of coefficients makes it harder to understand the impact of individual predictors, making the model more challenging to interpret despite its predictive benefits.

5 Flexible-Linear Models

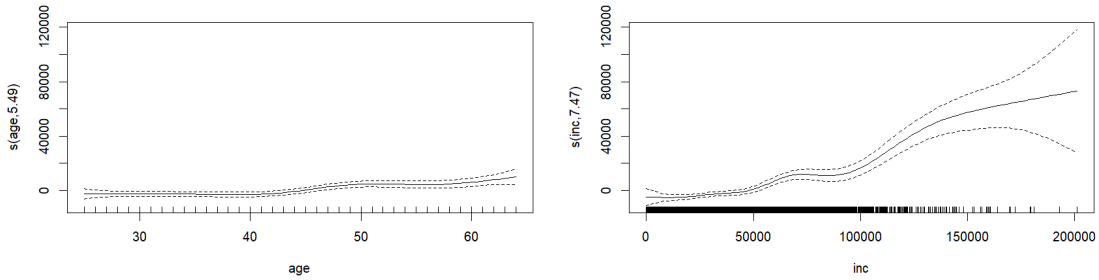
In this section, we allow each feature to undergo different non-linear transformations. Based on the scatter plots, we can determine which variables should have these transformations. For instance, variables such as home equity, mortgage, and property value exhibit clear linear relationships with total wealth, both intuitively and according to the scatter plots. Therefore, transformations are not needed for these variables in my opinion. The same applies to IRA, e401, and nifa, which intuitively make sense as they are simply additions to wealth (although they can still be used in interaction terms).

5.1 Generative Additive Models

Generalized Additive Models (GAMs) allow each feature to undergo different transformations through the use of smoothing splines which vary on predictors. To select the optimal GAM model, I created splines for variables such as education, income, age, and family size. I used an R package called `mgcv` (which we did not use in class), as it automatically selects the optimal amount of smoothness using cross validation. This is beneficial because it saves time compared to manually adjusting smoothness.

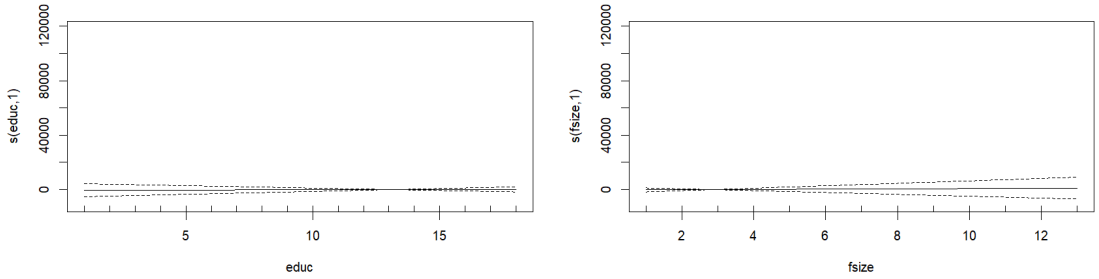
Next, similar to my approach with the linear model, I generated all possible 1:1 interaction terms between the transformed variables and the original features. I then applied lasso and stepwise (both directions) methods for feature selection. Afterward, I ran ridge regression for both models, using IQR for training and the test data for evaluation. When performing cross-validation, I encountered a significantly larger average Mean Squared Prediction Error (MSPE) compared to the linear fit. I believe this is due to overfitting, caused by the interactions and transformations.

5.1.1 How to Improve?



(a) Spline of Age

(b) Spline of Income



(c) Spline of Education

(d) Spline of Family Size

Figure 6: Visualizations of Splines for Age, Income, Education, and Family Size

When visualizing from the results of the previous GAM model, I noticed age and Income show a much bigger nonlinear relationship compared to educ and fsize when controlling for other features. This highlights one of the key advantages of GAM models: we can draw solid inferences from the visualizations while controlling for the effect of other variables. To prevent overfitting from occurring as it did here, I will use different degree polynomial regression, then add splines, and compare the models at each stage.

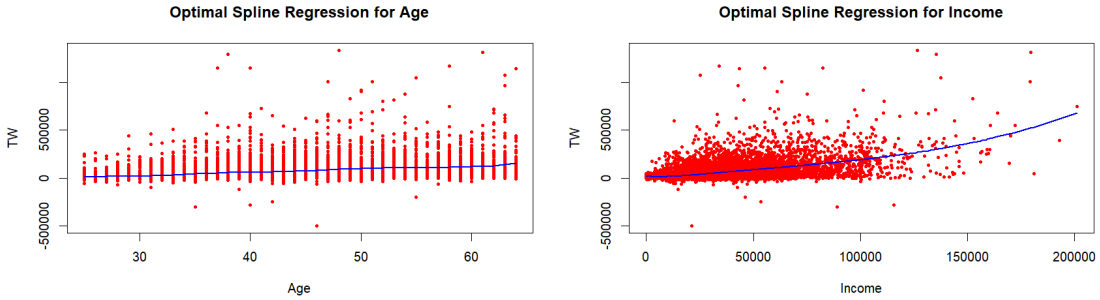
5.2 Polynomial and Spline Regression

In this section, I applied polynomial transformations of degrees 2, 3, 4, and 5 to the `age` and `income` variables. After performing k-fold cross-validation, I compared the performance of these polynomial transformations when predicting total wealth (`tw`). The results showed that while `age` did not benefit from any polynomial transformation, `income` showed improvement up to the 4th degree.

Additionally, I explored spline regression as an alternative transformation. After testing various

spline degrees, the 7-spline model improved the Mean Squared Prediction Error (MSPE) for `age`, while the 4-spline model was the best for `income`.

When comparing the four models using 10-fold cross-validation, the 4th-degree polynomial for `income` performed the best with the lowest MSPE. However, it was still not able to match the performance of the best ridge regression model, which had been selected using forward stepwise feature selection. This led me to consider incorporating the 4th-degree polynomial for `income` into my ridge regression model. However, when using Lasso for feature selection before running the ridge regression, all the polynomial features were set to zero except for the intercept. We can see also with the plot that the non linearity in age is pretty small and in income seems to occur outside where most the data points are. Maybe a more flexible model like polynomial for income was affected by the noise more nad appears more non linear in the graph.



5.3 Interpretiability and Performance

Using Generalized Additive Models (GAMs), you could not interpret at all. Although the graphs of the fitted models were informative, they did not clearly indicate the relationship between `age`, `income`, and `total wealth` compared to least squares, ridge, or lasso. We can see the relationship between `income` and `total wealth` is nonlinear, as indicated by the fitted lines, with the dotted lines representing the confidence intervals. We can looking into the smoothing splines for GAM to see where the coefficients change along the feature. As we reach extreme outliers in the data, the predicted values become less reliable. It does help us visualize nonlinear relationships, but we can't say "for every unit increase in x, y increases as well" like our other models.

For `age`, the relationship is fairly flat until around age 45, at which point the predictions begin to increase. This suggests that increasing the flexibility of the model (higher degree polynomials or

more splines) did not benefit `age` in terms of predictive power.

While the polynomial transformation improved the predictive power for `income`, there was a point where increasing the degree beyond 4 led to overfitting, as seen by the increased MSPE for higher-degree polynomials. This suggests a bias-variance tradeoff, where higher-degree transformations reduce bias but increase variance, particularly for `income` increases the MSPE rather then decrease. Despite the improvements from polynomial and spline transformations, the best predictive performance still came from the ridge regression model with forward stepwise coefficients, which used a more stable set of features selected through forward stepwise selection.

6 Conclusion

6.1 My Best Model

With the lowest average MSPE of 1,659,537,955 using k-fold cross validation, I chose my ridge regression model, with selected features from forward stepwise selection of every single interaction possible. With a lambda of 113!

$$\begin{aligned}\text{Total Wealth} = {} & 15310.062132734 + 0.887704940 \times \text{hequity} + 1.100960681 \times \text{nifa} + 1.220083300 \times \text{ira} \\ & - 0.743620224 \times \text{inc} - 19693.345496266 \times \text{e401} + 14166.766808609 \times \text{twoearn} \\ & - 380.613660672 \times \text{age} + 0.096356656 \times \text{hmort} - 3194.352652698 \times \text{male} \\ & + 0.328529043 \times \text{inc:e401} + 0.020269531 \times \text{twoearn:age} + 0.071045404 \times \text{nifa:e401} \\ & + 0.0000316 \times \text{hequity:ira} + 0.756267385 \times \text{ira:male} - 0.064043009 \times \text{e401:age} \\ & + 0.154853401 \times \text{inc:male} - 0.00000672 \times \text{hequity:hmort} + 0.000000461 \times \text{hequity:age} \\ & - 0.000000351 \times \text{hequity:nifa} + 0.00001753 \times \text{hequity:inc} - 0.059591229 \times \text{inc:twoearn} \\ & - 0.000001727 \times \text{inc:hmort}\end{aligned}$$

6.2 Thoughts on What Drives Wealth

From comparing flexible linear models to ridge regression with interaction terms, I′ve found that total wealth is relatively easy to predict once you have enough information about someone. It′s more rigid than we might think. The right combination of features can provide an accurate estimate of wealth, rather than relying on the smoother transitions of flexible linear models. This observation is based on the assumption that my R code worked as expected. It′s my interpretation from the project and my experience, and I hope you can follow my reasoning through the details in my paper.

6.3 Reflection

I′ve discovered that Machine Learning requires a lot of trial and error. There′s no single best method−each one is just an option to get closer to the true "function." We won′t know what works best until we try it out. I′ve also noticed that online forums are full of debates about the best

approach, with plenty of counterarguments to every method. More complex techniques, like flexible linear models, don′t always yield better results. I wasn′t able to optimize them as I am with ridge, lasso, and OLS, where I can clean the data and experiment with different interaction combinations. In the real world, there′s so much variation in the data and unobserved noise. Even when one models line of best fit looks better, my other model still worked the best. It′s important to keep validating with different techniques. Ultimately, all these methods are tools to help us predict, and we need to use what we know to get closer to the truth.