Paycheck Pathways

Unveiling the Key Factors Shaping Early Career Earnings

Carter Hall[1] Connor McNeill[1] Miles Woollacott[1]

[1]Department of Statistics, North Carolina State University

April 29, 2024

Outline

Background

- **Motivation:** Recent research has shown that where students attended college was a key factor into how much they made post-graduation
- But this research did not delve into what was associated with this discrepancy

Background

- **Motivation:** Recent research has shown that where students attended college was a key factor into how much they made post-graduation
- But this research did not delve into what was associated with this discrepancy
- The natural question emerges:
  What factors lead to differences in salary after graduation?

Why is this important?

- Allows us to see if this discrepancy is related to academics and/or socioeconomics
- Idea of the "Cycle of Poverty"
- Other factors may be university-specific

The Dataset

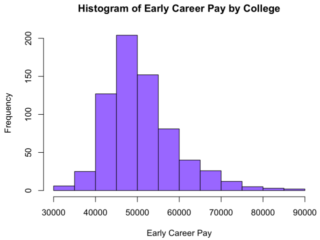
The data comes from the US Department of Education with 58 variables.

- University information: acceptance rate, average SAT score, region/locale
- Student Body information: diversity and domestic/international, socioeconomic status, % STEM and "Better World"
- Tuition/Financial Aid information: tuition revenue/cost of attendance, Pell Grants, etc.

Distribution of Early Career Pay

Our response variable is early career pay, measured as an average across the alumni student body of each college.

1. Early career pay appears to follow some right-skewed and positive distribution, which indicates we need to transform our response variable, or fit a model with a positive response



Response & Predictor Transformations

Our response variable is early career pay, measured as an average across the alumni student body of each college.

1. Early career pay appears to follow some right-skewed and positive distribution, which indicates we need to transform our response variable, or fit a model with a positive response
2. As such, a log transformation will be considered for early career pay for our linear models
3. Predictor transformations include:
   - Admission rate: inverse transformation
   - Total enrollment: log transformation
   - % Domestic students: quartic transformation

Linear Dependencies & Multicollinearity

- A few variables were linear combinations of one another – this caused linear dependencies to occur
- Some multicollinearity between some of the strongly correlated data:
  - → Multiple variables related to tuition (in-state, out-of-state, total costs, etc.)
  - → Certain diversity factors
  - → Economic factors such as median household income and poverty rate
- These predictors were dropped from consideration in all of our models in order to meet assumptions

Outline

Stepwise Selection

Motivation: Determine a minimal subset of predictors that accurately predict early career pay with ease of interpretability.

1. We kept the same variable transformations as in OLS
2. Applied 10-fold CV to further reduce the risk of overfitting

Motivation: Determine a minimal subset of predictors that accurately predict early career pay with ease of interpretability.

1. We kept the same variable transformations as in OLS
2. Applied 10-fold CV to further reduce the risk of overfitting
3. Reduced number of terms from 53 to 8 (including intercept)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.5680	0.1780	59.37	< 0.0001
Avg SAT Score	-1.564e-04	3.861e-05	-4.05	0.0001
% Students in STEM	0.3241	0.0205	15.79	< 0.0001
Tuition Revenue per Student	3.310e-06	4.807e-07	6.89	< 0.0001
Avg Faculty Salary	2.256e-05	1.707e-06	13.21	< 0.0001
% Students with Pell Grants	-0.2935	0.0268	-10.95	< 0.0001
% Domestic students	0.5419	0.2727	1.99	0.0474
$(\%\ \text{Domestic})^4$	-0.3730	0.1068	-3.49	0.0005

LASSO

Motivation: Determine a minimal subset of predictors that accurately predict early career pay.

1. Retained all variable transformations
2. $\lambda \approx 0.006270$ was selected via 10-fold CV

Motivation: Determine a minimal subset of predictors that accurately predict early career pay.

1. Retained all variable transformations
2. $\lambda \approx 0.006270$ was selected via 10-fold CV

Variable	Coefficient
(Intercept)	10.7034
% World Better	0.0391
% Students in STEM	0.2263
Located in Rural Town	-0.0110
Tuition Revenue per Student	1.769e-06
Avg Faculty Salary	1.761e-05
% Students on Pell Grants	-0.1957
Graduation Rate	0.0826
% Households with Graduate Degree	0.2248
(% Domestic Students)^4	-0.1143
% Students identifying as Female	-0.0930
% Students identifying as Asian	0.0979

Denotes a variable that was also in Stepwise Selection

Partial Least Squares

Motivation: Perform dimension reduction.

1. Retained all variable transformations
2. 3 components were chosen via 10-fold CV

Motivation: Perform dimension reduction.

1. Retained all variable transformations
2. 3 components were chosen via 10-fold CV

	Component 1	Component 2	Component 3
% Students in STEM*	0.0095	0.0264	0.0369
log(total enrollment)	0.0046	0.0148	0.0127
Tuition Revenue per Student*	0.0103	0.0088	0.0145
Avg Faculty Salary*	0.0128	0.0201	0.0233
% Students on Pell Grants*	-0.0097	-0.0174	-0.0232
Graduation Rate*	0.0119	0.0137	0.0165
% Households with Graduate Degree*	0.0111	0.0093	0.0104
% Students identifying as Female*	-0.0061	-0.0202	-0.0217
% Students identifying as Asian*	0.0095	0.0128	0.0147

Predictors were included if there was a loading α that satisfied $|\alpha| > 0.015$ for any of the three components.
Denotes a predictor that also appeared in the Stepwise model.
* Denotes a predictor that also appeared in the LASSO model.

Outline

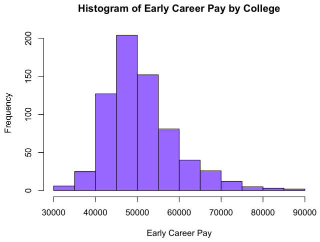
Why consider a Gamma GLM?

Note that the previous methods required a log-transformation of our response variable. It's natural to consider a Generalized Linear Model:

Why consider a Gamma GLM?

Note that the previous methods required a log-transformation of our response variable. It's natural to consider a Generalized Linear Model:

1. Financial data often follows a Gamma distribution - supports our earlier remarks about early career pay



Why consider a Gamma GLM?

Note that the previous methods required a log-transformation of our response variable. It's natural to consider a Generalized Linear Model:

1. Financial data often follows a Gamma distribution - supports our earlier remarks about early career pay
2. There are benefits of working with a model that does not require further transformations.

About the Gamma GLM

- A slightly different parameterization is used - with the shape parameter ν and then scale parameter $= \frac{\nu}{\mu}$
- A log-link $\log(\mu)$ was used (instead of the canonical link)
- When variance is small, the Gamma GLM with log-link performs rather similar to a Gaussian linear model with a log-transformed response.

Predictors Included

We utilized the predictors that were screened from the stepwise regression model.

There is not variable selection or dimension reduction built-in.

Our Results

- The coefficients and standard errors were nearly exactly the same as those of the stepwise model w/ log transformation
- This is because the dispersion parameter - $1/\hat{\nu}$ - was small, and with large ν the Gamma distribution can be approximated by Normal

Our Results

- The coefficients and standard errors were nearly exactly the same as those of the stepwise model w/ log transformation
- This is because the dispersion parameter - $1/\hat{\nu}$ - was small, and with large ν the Gamma distribution can be approximated by Normal
- But...the Analysis of Deviance test failed to reject the null model.

Outline

BART: **B**ayesian **A**dditive **R**egression **T**rees

Notable questions answered in this presentation:

- What led us to consider a regression tree model?
  1. Non-parametric
  2. Capable of capturing nonlinear relationships
- Why *Bayesian* [additive] regression trees?

BART: **B**ayesian **A**dditive **R**egression **T**rees

Notable questions answered in this presentation:

- What led us to consider a regression tree model?
    1. Non-parametric
    2. Capable of capturing nonlinear relationships
    3. Trees are *weak learners*.
- Why *Bayesian* [additive] regression trees?
    1. Each tree intended to address *different* aspects of the prediction problem.
    2. No need for 'greedy growing' of each tree and subsequent pruning, as in CART models – see Ročková and Saha, 2018. Instead, a **prior** is used to combat overfitting.

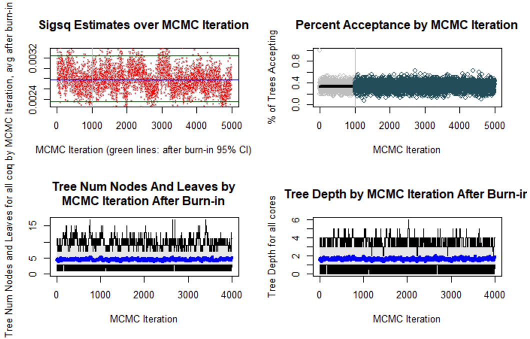
$$Y = \sum_{j=1}^{m} g(x; \underbrace{T_j, M_j}_{(2)}) + \epsilon \qquad \epsilon \sim N(0, \underbrace{\sigma^2}_{(1)})$$

1. **Variance of error term**: Errors are $\epsilon \sim N(0, \sigma^2)$ for mathematical tractability.
   - *Prior is inverse chi-squared* with scaling factor determined by hyperparameters on the *center* and *shape* of the distribution. (ν, q)
2. **Pairs** of (T_j, M_j) – T_j are binary regression trees that split the range of predictors into subsets; M_j are parameters of [terminal] nodes.
   - Prior includes factor of prior on $M_j | T_j$, affected by **depth** of a node and assigning high probability mass to the interval $(y_{\min}, y_{\max})$. (k)

Cross-validation chose hyperparameters $k = 5, \nu = 3, q = 0.90, m = 40$.

$m = 40$ was highest considered value – runs with higher m led to intractability when visualizing trees.

Other notable parameters: 1000 iterations for burn-in; 4000 iterations after burn-in concluded.



Two R packages – BARTMAN (BART **M**odel **An**alysis) and BARTMACHINE (running BART).

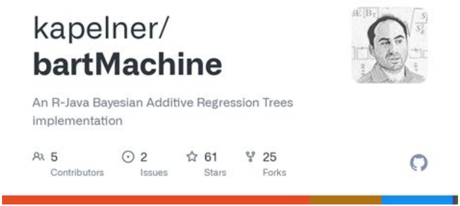


Figure: Profile of bartMachine package repository

Figure: "Bartman," alternate persona of Bart Simpson

To *run BART* is not a computational challenge on a laptop (tech-lab computers did not have Java :( ); to *visualize* the $\approx$ 1.6 million trees, however, took $\approx$ 5 hours and all but 4 MB of 15.8 GB of available RAM.

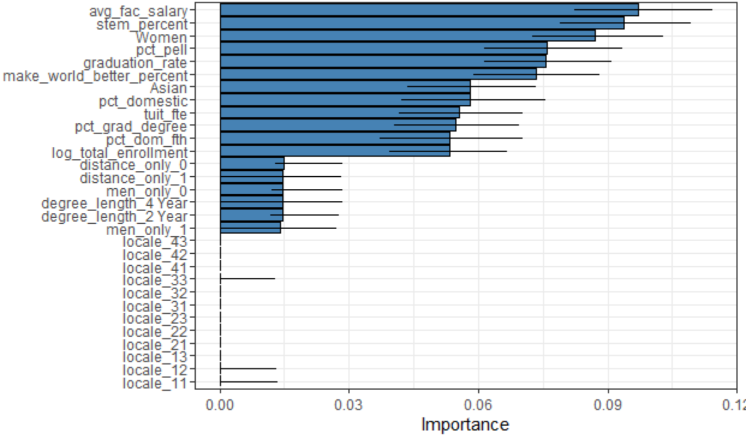


Figure: Variable Importance for BART

Outline

	OLS	Stepwise	LASSO	PLS	GLM	BART
RMSE (Train)	0.0557	0.0639	0.0634	0.0605	0.0639	**0.0465**
RMSE (Test)	0.0760	0.0775	0.0775	0.0726	0.0774	**0.0655**
Difference	0.0203	0.0136	0.0141	**0.0121**	0.0135	0.0190
Num. Terms	53	8	15	**3***	8	–

* number of components retained

Table: Summary of various comparison methods for our models. Note the errors are presented on the log-scale.

Answering our Research Question

What factors lead to differences in salary after graduation?

Important for both PLS and BART	Is only important in BART
% of Students in STEM	% Make World Better
Average Faculty Salary	% Domestic Students
log(total enrollment)	(% Domestic Students)^4
Tuition Revenue per Student	Distance Only
% of Students on Pell Grants	Men Only
Graduation Rate	Locale
% Students identifying as Female	Degree Length
% Students identifying as Asian	
% Households with Graduate Degree	

Predictors in red had negative relationship with early career pay (PLS).
Note: We can only determine correlation, not causation

Questions?

Any and all questions are welcome!
If you are curious, our paper can be obtained by the QR code below: