College Athletic Success with Machine Learning: Ridge Regression with Bootstrapping

Jacob Lara, Daniel Pate, Gerard Tetegan, and Lane Whitehead

May 2024

1 Abstract

The objective of this study is to replicate the findings of Anderson's research, which explored the causal relationships between football success and various metrics such as college donations, applicant rates, and academic reputation. He employed propensity score matching combined with weighted linear regression to address endogeneity concerns. Significant levels of multicollinearity was revealed during the replication, performed in Python, which led to the adoption of Ridge Regression to mitigate the limitations associated with multicollinearity in linear regression, while maintaining the model's while maintaining the model's interpretive advantages in assessing how independent variables linearly affect dependent variables. Furthermore, bootstrapping was integrated with Ridge Regression to estimate standard errors. Although ridge regression usually exhibits higher bias compared to linear regression, it ensures lower variance.

2 Introduction

College athletics, notably football, are widely recognized for their considerable influence on university finances and student applications, with the prevailing belief that athletic success translates into significant institutional benefits, including increased donations, higher student application rates, and an enhanced academic reputation. Anderson's research used propensity score matching (PSM) to establish causal relationships between college football team victories and various institutional outcomes, revealing a positive correlation with metrics such as donations and applicant numbers.

The aim of this study is to reassess these findings by employing machine learning techniques, such as ridge regression, within a Sequential Treatment Effects (STE) model framework. This approach ensures the robustness and reproducibility of the results. While traditional PSM methods effectively balance covariates between treated and control groups, they often overlook multicollinearity among covariates, which can destabilize estimation procedures. To address this issue, ridge regression has been integrated into the Sequential Treatment Estimation (STE) process in this replication study.

Unlike ordinary least squares (OLS) regression, Ridge regression does not provide direct standard error estimates for coefficient estimates. Consequently, bootstrapping serves as an alternative approach.

By combining these methods within the STE framework, this study not only replicates but also significantly enhances the methodological rigor of previous research efforts.

3 Literature Review

Ridge regression, a widely used technique in statistical modeling, offers a solution to multicollinearity and overfitting by introducing a penalty term to the traditional least squares estimation. This method has garnered significant attention in various fields due to its ability to handle high-dimensional data and improve prediction accuracy. McDonald (2009) delves into the theoretical aspects of Ridge regression, emphasizing its role in mitigating multicollinearity issues in regression analysis. By introducing a penalty term, Ridge regression offers a trade-off between bias and variance, making it particularly useful in scenarios where predictor variables are highly correlated. Concurrently, bootstrapping, a resampling technique, has gained prominence for its robustness in estimating the sampling distribution of a statistic by repeatedly sampling with replacement from the observed data.

Bootstrapping is a resampling technique introduced by Efron in the late 1970s, offering a non-parametric approach to estimating the sampling distribution of a statistic. Unlike traditional parametric methods, bootstrapping does not rely on assumptions about the underlying distribution of the data, making it particularly useful in situations where the data are non-normal or the sample size is small. By repeatedly sampling with replacement from the observed data, bootstrapping generates a large number of pseudo-samples, allowing for the estimation of standard errors, confidence intervals, and hypothesis testing.

Young (1994) delves into the intricacies of bootstrapping and its wide-ranging applications in statistical inference. Young highlights the versatility of bootstrapping across various statistical problems, including parameter estimation, model selection, and hypothesis testing. Through empirical examples and simulations, the paper illustrates the efficacy of bootstrapping in providing reliable estimates and addressing the limitations of classical inferential methods.

Jawara (2020) utilized bootstrapping techniques in conjunction with Propensity Score Matching (PSM) analysis to examine the relationship between access to savings and household welfare in The Gambia. By resampling from the observed data and employing PSM to balance covariates, the study provided robust estimates of the causal effect of savings access on household welfare outcomes.

In recent years, researchers have explored the integration of Ridge regression with bootstrapping techniques to enhance the stability and reliability of predictive models. This amalgamation has shown promise in addressing the challenges posed by complex datasets with interrelated predictors. Gonzalez Salas Duhne et al. (2022) applied Ridge regression in their study on predicting early dropout in online versus face-to-face guided self-help interventions. Their research highlights the efficacy of Ridge regression in handling high-dimensional data and improving the predictive performance of dropout rates in different intervention modalities.

While studies have independently demonstrated the effectiveness of Ridge regression and bootstrapping with PSM analysis in addressing specific statistical challenges, limited research has explored their combined application. Our paper hopes to bridge this gap by furthering research from Anderson on the effects of college athletic success.

4 Results

The following tables present the replicated results from Anderson's study, starting with Tables 1 and 2, which replicate the Standard Treatment Effect (STE) analysis of Table 2 from Anderson's study with least weighted squares and ridge regression respectively.

4.1 Standard Treatment Effect Analysis

Table 1: Table 2 STE Replication

Outcome	Coefficient	SE	N
Alumni Athletic Operating Donations	-44.8	59.2	477
Alumni Nonathletic Operating Donations	16.1	89.6	477
Total Alumni Donations	291.3	341.8	1094
Alumni Giving Rate	0.0000	0.0007	1106
Academic Reputation	0.003	0.002	600
Applicants	-16.0	60.1	526
Acceptance Rate	0.003	0.002	926
First-Time Out-of-State Enrollment	2.9	3.8	941
First-Time In-State Enrollment	-4.1	8.1	941
25th Percentile SAT	1.4	0.8	427

Table 2: Table 2 Ridge Regression with Bootstrapping

Outcome	Coefficient	SE	Lower CI	Upper CI	N
Alumni Athletic Operating Donations	-43.9	51.4	-147.0	59.3	477
Alumni Nonathletic Operating Donations	16.5	116.5	-211.0	244.0	477
Total Alumni Donations	292.1	248.5	-171.3	755.3	1094
Alumni Giving Rate	0.0000	0.0010	-0.0019	0.0020	1106
Academic Reputation	0.003	0.004	-0.006	0.012	600
Applicants	-16.2	140.2	-292.7	260.2	526
Acceptance Rate	0.003	0.003	-0.003	0.009	926
First-Time Out-of-State Enrollment	2.8	8.6	-14.7	20.4	941
First-Time In-State Enrollment	-4.2	10.7	-25.0	16.6	941
25th Percentile SAT	1.4	1.2	-1.0	3.8	427

Following the initial analysis, we also replicated Table 3 as well as included a ridge regression afterwards.

Table 3: Table 3 STE Replication

Outcome	Coefficient	SE	N
Alumni Athletic Operating Donations	191.2	65.0	616
Alumni Nonathletic Operating Donations	-137.4	96.1	616
Total Alumni Donations	267.4	267.1	1258
Alumni Giving Rate	0.0002	0.0007	1287
Academic Reputation	0.003	0.002	650
Applicants	81.1	60.4	528
Acceptance Rate	-0.003	0.002	979
First-Time Out-of-State Enrollment	1.6	5.0	962
First-Time In-State Enrollment	12.6	6.4	962
25th Percentile SAT	0.8	0.7	426

Table 4: Table 3 Ridge Regression with Bootstrapping

Outcome	Coefficient	SE	Lower CI	Upper CI	N
Alumni Athletic Operating Donations	190.5	34.6	122.1	258.9	616
Alumni Nonathletic Operating Donations	-138.2	117.6	-368.9	92.5	616
Total Alumni Donations	269.8	215.1	-152.7	692.3	1258
Alumni Giving Rate	0.0002	0.0009	-0.0015	0.0019	1287
Academic Reputation	0.003	0.002	-0.001	0.007	650
Applicants	80.4	121.0	-156.2	317.0	528
Acceptance Rate	-0.003	0.002	-0.007	0.001	979
First-Time Out-of-State Enrollment	1.6	7.2	-12.5	15.7	962
First-Time In-State Enrollment	12.7	10.8	-8.6	34.0	962
25th Percentile SAT	0.8	1.7	-2.5	4.1	426

5 Discussion

The tables in our study—Tables 1 and 3—were recalculated to adjust standard errors to reflect the heterogeneity of variances across observations, utilizing weighted least squares estimation. This adjustment was crucial in ensuring that our estimates could robustly handle the diverse variance structures within the data, a common challenge in educational and donation outcome data. Meanwhile, Tables 2 and 4 were designed to enhance the robustness of our estimates further and address potential multicollinearity issues through the application of ridge regression and bootstrapping. This approach not only reinforces the stability of the parameter estimates but also provides a more reliable inference by enhancing the precision of standard error calculations. As discussed in McDonald (2009), the incorporation of a penalty term in ridge regression helps manage the bias-variance trade-off, enhancing prediction accuracy in high-dimensional data settings. This approach is echoed in our application of ridge regression to the multicollinearity evident in the variable 'season wins', particularly in our handling of the years 2001-2003 and 2005-2009, and reintegrating 2004 into the analysis in Table 6. All the outcome variables is the impact of season wins on the outcome variables. This report examines Anderson's Table 3, which is pivotal in understanding the relationship between season wins and various educational and donation outcomes using a Sequential Treatment Effects (STE) model.

All the outcome variables are the coefficients for the impact of season wins on each outcome variable. To understand how the weighted least squares were improved upon by ridge regression, we will examine one of the variables; Alumni Athletic Operating Donations. The regression to calculate the coefficient of Alumni Athletic Operating Donations is in Tables 5 and 6. In the replication of the original findings, several year variables were excluded being the interaction of years 1987-2000 and 2004 due to the multicollinearity issue. Table 6 introduces the ridge regression to increase the predictability of the season wins on the alumni operating donations.

A key aspect of our replication involves the examination of standard errors associated with these estimates, which provides insights into the precision of our regression results. Notably, the standard errors for the alumni athletic operating donations and total alumni donations are smaller in our replication than in the original study. This reduction in standard errors suggests that the ridge regression method, known for its ability to handle multicollinearity effectively, has enhanced the reliability of these estimates. The smaller standard errors indicate a higher level of statistical confidence in the reported effects, affirming the strength of the relationship between season wins and increases in specific types of donations.

Table 5 includes the years 2001-2003 and 2005-2009 as the other years contributed to high multicollinearity. The ridge regression results in Table 6 includes the year 2004, meaning that the original regression omitting 2004 was likely due to specific data characteristics or anomalies in that year which initially suggested potential distortions in the analysis. The model's ability to integrate the year 2004 without significant losses in precision or increases in standard errors validates the effectiveness of ridge regression in handling datasets with potential multicollinearity among temporal variables.

The coefficient for the variable 'seasonwins_m' is the alumni operating donations coefficient in the final regression. The original study reported a significant increase in alumni athletic donations per additional win in Table 3, quantified at $191,200. The ridge regression confirms this positive effect, with a similarly robust increase of $190,500 in Table 4.

Table 5: WLS for Alumni Athletic Operating Donations

Variable	**Coefficient**	**Std. Error**	**P score**
const	3.326e+05	2.12e+06	0.876
seasonwins_m	**191.2**	**65.0**	**0.003**
lag3_seasonwins	2.149e+05	6.92e+04	0.002
lag_seasongames	-1.359e+05	2.32e+05	0.559
lag3_seasongames	3.873e+04	1.28e+05	0.762
_Iyear_2001	-3.646e+04	3.15e+05	0.908
_Iyear_2002	-5.159e+04	3.67e+05	0.888
_Iyear_2003	4.232e+05	8.99e+05	0.638
_Iyear_2005	-3.565e+05	4.93e+05	0.470
_Iyear_2006	-6.387e+04	4.34e+05	0.883
_Iyear_2007	9.588e+05	5.42e+05	0.077
_Iyear_2008	2.712e+05	2.41e+05	0.260
_Iyear_2009	1.374e+05	3.35e+05	0.681
Adjusted R^2		0.048	

Table 6: Ridge Regression for Alumni Athletic Operating Donations

Variable	**Coefficient**	**Std. Error**	**p-value**
Intercept	399092.3	–	–
seasonwins_m	**190.5**	**34.6**	**0.000**
lag3_seasonwins	214370.562	43180.916	0.000
lag_seasongames	-122689.484	103817.392	0.270
lag3_seasongames	32803.102	122014.501	0.824
_Iyear_1987	0.000	0.000	1.000
_Iyear_1988	0.000	0.000	1.000
_Iyear_1989	0.000	0.000	1.000
_Iyear_1990	0.000	0.000	1.000
_Iyear_1991	0.000	0.000	1.000
_Iyear_1992	0.000	0.000	1.000
_Iyear_1993	0.000	0.000	1.000
_Iyear_1994	0.000	0.000	1.000
_Iyear_1995	0.000	0.000	1.000
_Iyear_1996	0.000	0.000	1.000
_Iyear_1997	0.000	0.000	1.000
_Iyear_1998	0.000	0.000	1.000
_Iyear_1999	0.000	0.000	1.000
_Iyear_2000	0.000	0.000	1.000
_Iyear_2001	-164488.094	138186.967	0.785
_Iyear_2002	-178484.375	182898.737	0.385
_Iyear_2003	261398.188	215360.562	0.308
_Iyear_2004	-145783.453	153337.392	0.755
_Iyear_2005	-477957.688	301788.082	0.556
_Iyear_2006	-190721.016	136860.169	0.328
_Iyear_2007	782654.375	185669.401	0.000
_Iyear_2008	117812.453	156693.337	0.453
_Iyear_2009	-4427.238	195086.718	0.977
Adjusted R^2	0.0291007		

6 Conclusion

This study revisited Anderson's (2017) examination of the causal effects of college football success on institutional outcomes using advanced statistical techniques. Our application of ridge regression and bootstrapping methods not only confirmed Anderson's original findings but also introduced a higher degree of robustness and reliability into the analysis.

The implementation of ridge regression effectively controlled for multicollinearity, which was a limitation in the original study. This approach ensured that the coefficients remained stable across different models, providing confidence in the reproducibility and accuracy of our results. Notably, the ridge regression model demonstrated a consistent impact of season wins on alumni athletic donations, which aligns with Anderson's findings but with improved precision and lower standard errors.

Furthermore, the use of bootstrapping allowed for a thorough examination of the variability and reliability of the estimated effects. This technique enhanced our understanding of the confidence intervals around the estimated coefficients, affirming the robustness of the causal links between college football success and its effects on donations and applicant numbers.

Our study contributes to the body of knowledge by demonstrating the utility of combining ridge regression with bootstrapping in econometric analyses. This approach is particularly valuable in scenarios where traditional models face challenges with multicollinearity and variance instability. The findings suggest that university administrators and policymakers should consider the broader implications of athletic success, not only for its potential revenue impacts but also for strategic planning in admissions and funding.

In conclusion, by enhancing the methodological rigor of the analysis, this study provides a more definitive understanding of the relationship between college athletic success and key university outcomes. Future research could expand upon these methods to explore other variables or settings where similar econometric challenges exist, potentially offering new insights into the strategic benefits of athletic programs in educational institutions.

References

Capur, Mira. (2006). Bootstrap estimation of standard error of ridge estimates. *UNLV Retrospective Theses & Dissertations*. 2044. http://dx.doi.org/10.25669/l41i-my4c

Gonzalez Salas Duhne, P., Delgadillo, J., & Lutz, W. (2022). Predicting early dropout in online versus face-to-face guided self-help: A machine learning approach. *Behaviour Research and Therapy*, 159, 104200. https://doi.org/10.1016/j.brat.2022.104200

Jawara, H. (2020). Access to savings and household welfare evidence from a household survey in The Gambia. *African Development Review*, 32(2), 138–149. https://doi.org/10.1111/1467-8268.12423

McDonald, G. C. (2009). Ridge regression. *WIREs Computational Statistics*, 1, 93–100. DOI: 10.1002/wics.14

Revan Özkale, M., & Altuner, H. (2023). Bootstrap confidence interval of ridge regression in linear regression model: A comparative study via a simulation study. *Communications in Statistics-Theory and Methods*, 52(20), 7405-7441. https://doi.org/10.1080/03610926.2022.2045024

Yu, Z., Mao, S., & Lin, Q. (2022). Has China's carbon emissions trading pilot policy improved agricultural green total factor productivity?. *Agriculture*, 12(9), 1444. https://doi.org/10.3390/agriculture12091444