Implementing LASSO Regression to Examine County-Level Determinants of Mental Health

Jonathan Pipping

2024-01-18

Introduction

Few would argue that mental health represents an undeniable concern in the United States. The U.S. Substance Abuse and Mental Health Services Administration (SAMHSA) estimates that over one in five adults (59.3 million) live with a mental illness. Furthermore, an estimated 45.3% of these individuals receive treatment, leaving approximately 29 million Americans without care ("2022 National Survey on Drug Use and Health (NSDUH) Releases" 2023). In response to these staggering numbers, the U.S. Congress passed the Bipartisan Safer Communities Act (BSCA) in 2022, which included $250 million in supplemental funding to help states and territories address the mental health challenges facing their communities (Office 2023).

In the context of this recent legislation, this research aims to establish a framework for accurate mental health prediction on a county level, identify significant social, demographic, and behavioral determinants of mental health, and provide insights to state and local officials to aid their efforts in limiting the prevalence of poor mental health days within their communities.

Data

The data for this project was obtained from the University of Wisconsin's 2023 County Health Rankings data set, which is publicly available on the Harvard Dataverse (Wiemken 2023). Consisting of publicly-recorded measures and self-reported survey results, this data set includes over 70 county-level metrics such as life expectancy, median household income, and the percentage of people uninsured. All metrics are accompanied by 95% confidence intervals and associated ranks, but this project's scope relies only on point estimates.

The data was subsequently cleaned and prepared for analysis. This process involved the omission of metrics flagged as "highly unreliable" by the University of Wisconsin's research team. Additionally, metrics with high (>10%) missingness were omitted to preserve a high number of complete cases for regression. After cleaning, the final data set comprised 61 metrics from 2,636 US counties. These metrics included counties' average percentage of days people reported feeling mentally unhealthy (the outcome of interest) and 60 covariates.

Methods and Results

Linear Regression

After standardizing all variables, a multiple linear regression was fit to isolate potential causal relationships between the covariates and our outcome of interest. This model can be represented in the form $\hat{Y}_i = \beta_0 + \sum_{j=1}^{60} X_{ij}\beta_j + \epsilon_i, i \in \{1, 2, \ldots, 2,636\}$ where $\hat{Y}_i$ represents the estimated outcome (the average percentage of days people in a particular county reported feeling mentally unhealthy), the X_{ij} represent the 2,636 observed values of each of our 60 covariates, the β_j represent the coefficients that satisfy the Least-Squares solution $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, and the e_i represent the error (or residual) for each $\hat{Y}_i$ prediction. The OLS estimates for each β_j coefficient are shown in **Table 1** below.

Table 1: Linear Regression Results

Term	Estimate	Standard Error	Test Statistic	P-Value	Significant?*
perc_days_physically_unhealthy	0.9691171	0.0326973	29.6390633	0.0000000	Yes
perc_uninsured	0.2789068	0.1104900	2.5242712	0.0116535	No
perc_insufficient_sleep	0.2765008	0.0191954	14.4045400	0.0000000	Yes
perc_female	0.1557900	0.0125683	12.3954665	0.0000000	Yes
perc_adults_diabetes	0.1488152	0.0363436	4.0946748	0.0000436	Yes
perc_excessive_drinking	0.1318627	0.0136044	9.6926297	0.0000000	Yes
median_household_income	0.0812963	0.0249961	3.2523554	0.0011593	No
perc_homeowners	0.0799673	0.0174851	4.5734639	0.0000050	Yes
perc_income_required_child_care	0.0571880	0.0119841	4.7719905	0.0000019	Yes
perc_households_broadband_access	0.0557043	0.0159391	3.4948263	0.0004824	Yes
perc_adults_smoking	0.0551788	0.0289608	1.9052895	0.0568537	No
perc_food_insecure	0.0488403	0.0871471	0.5604351	0.5752315	No
perc_vaccinated	0.0481192	0.0123493	3.8965273	0.0001001	Yes
primary_care_physicians_ratio	0.0432317	0.0134279	3.2195422	0.0012999	No
perc_not_proficient_english	0.0372487	0.0188568	1.9753528	0.0483348	No
avg_daily_pm	0.0300727	0.0122395	2.4570302	0.0140746	No
perc_severe_housing_problems	0.0285048	0.0509018	0.5599952	0.5755314	No
school_funding_accuracy	0.0282255	0.0159573	1.7688166	0.0770428	No
mental_health_provider_ratio	0.0280380	0.0113155	2.4778314	0.0132820	No
perc_long_commute_alone	0.0202113	0.0141660	1.4267423	0.1537754	No
perc_completed_high_school	0.0182256	0.0223685	0.8147866	0.4152698	No
social_association_rate	0.0176326	0.0109379	1.6120570	0.1070721	No
perc_voter_turnout	0.0154377	0.0178916	0.8628506	0.3883000	No
perc_children_single_parent	0.0131822	0.0171592	0.7682261	0.4424233	No
population	0.0114884	0.0109687	1.0473841	0.2950208	No
perc_nh_opi	0.0085330	0.0099630	0.8564698	0.3918177	No
perc_low_birthweight	0.0079211	0.0158919	0.4984364	0.6182191	No
perc_with_exercise_opportunities	0.0044081	0.0131271	0.3358039	0.7370461	No
income_ratio_20_80	0.0042686	0.0134386	0.3176384	0.7507850	No
(Intercept)	0.0000000	0.0084377	0.0000000	1.0000000	No
gender_pay_gap	-0.0018220	0.0095907	-0.1899735	0.8493449	No
perc_households_severe_cost_burden	-0.0062470	0.0192447	-0.3246088	0.7455035	No
perc_census_participation	-0.0067195	0.0168903	-0.3978309	0.6907878	No
perc_rural	-0.0072396	0.0196332	-0.3687421	0.7123503	No
water_violation	-0.0076539	0.0089703	-0.8532447	0.3936030	No
other_primary_care_provider_ratio	-0.0110531	0.0117719	-0.9389358	0.3478518	No
inadequate_facilities	-0.0131680	0.0119361	-1.1032095	0.2700393	No
preventable_hospitalization_rate	-0.0144003	0.0111238	-1.2945546	0.1955900	No
perc_uninsured_children	-0.0192143	0.0287623	-0.6680374	0.5041696	No
perc_drive_alone_to_work	-0.0233262	0.0125110	-1.8644510	0.0623721	No
perc_annual_mammagram	-0.0264891	0.0132608	-1.9975461	0.0458709	No
overcrowding	-0.0272230	0.0240334	-1.1327141	0.2574398	No
severe_housing_cost_burden	-0.0391829	0.0437156	-0.8963127	0.3701696	No
perc_driving_deaths_alcohol	-0.0403303	0.0091315	-4.4166007	0.0000104	Yes
life_expectancy	-0.0472612	0.0181523	-2.6035923	0.0092782	No
perc_unemployed	-0.0479131	0.0140832	-3.4021425	0.0006788	Yes
perc_children_in_poverty	-0.0503042	0.0263198	-1.9112662	0.0560813	No
perc_adults_obesity	-0.0824102	0.0189207	-4.3555639	0.0000138	Yes
perc_under_18	-0.0858724	0.0190881	-4.4987420	0.0000071	Yes
perc_some_college	-0.1002421	0.0189363	-5.2936581	0.0000001	Yes
perc_65_over	-0.1228949	0.0197167	-6.2330478	0.0000000	Yes
perc_limited_access_healthy_foods	-0.1260822	0.0718292	-1.7553060	0.0793259	No
perc_fair_or_poor_health	-0.1820723	0.0550120	-3.3096870	0.0009469	Yes
perc_asian	-0.1877521	0.0298342	-6.2931833	0.0000000	Yes
food_environment_index	-0.2362690	0.1276886	-1.8503538	0.0643771	No
perc_physically_inactive	-0.2937239	0.0314226	-9.3475403	0.0000000	Yes
perc_uninsured_adults	-0.3554609	0.0964752	-3.6844792	0.0002339	Yes
perc_ai_an	-0.4454267	0.0656206	-6.7879123	0.0000000	Yes
perc_hispanic	-0.8904190	0.1357322	-6.5601140	0.0000000	Yes
perc_black	-0.9861326	0.1494896	-6.5966633	0.0000000	Yes
perc_non_hispanic_white	-1.2644703	0.2027857	-6.2354994	0.0000000	Yes

Note:

* At 0.001 Significance Level

Table 2: OLS, Statistically Significant Predictors

Term	Estimate	Standard Error	Test Statistic	P-Value
perc_days_physically_unhealthy	0.9691171	0.0326973	29.639063	0.0000000
perc_insufficient_sleep	0.2765008	0.0191954	14.404540	0.0000000
perc_female	0.1557900	0.0125683	12.395466	0.0000000
perc_adults_diabetes	0.1488152	0.0363436	4.094675	0.0000436
perc_excessive_drinking	0.1318627	0.0136044	9.692630	0.0000000
perc_homeowners	0.0799673	0.0174851	4.573464	0.0000050
perc_income_required_child_care	0.0571880	0.0119841	4.771991	0.0000019
perc_households_broadband_access	0.0557043	0.0159391	3.494826	0.0004824
perc_vaccinated	0.0481192	0.0123493	3.896527	0.0001001
perc_driving_deaths_alcohol	-0.0403303	0.0091315	-4.416601	0.0000104
perc_unemployed	-0.0479131	0.0140832	-3.402143	0.0006788
perc_adults_obesity	-0.0824102	0.0189207	-4.355564	0.0000138
perc_under_18	-0.0858724	0.0190881	-4.498742	0.0000071
perc_some_college	-0.1002421	0.0189363	-5.293658	0.0000001
perc_65_over	-0.1228949	0.0197167	-6.233048	0.0000000
perc_fair_or_poor_health	-0.1820723	0.0550120	-3.309687	0.0009469
perc_asian	-0.1877521	0.0298342	-6.293183	0.0000000
perc_physically_inactive	-0.2937239	0.0314226	-9.347540	0.0000000
perc_uninsured_adults	-0.3554609	0.0964752	-3.684479	0.0002339
perc_ai_an	-0.4454267	0.0656206	-6.787912	0.0000000
perc_hispanic	-0.8904190	0.1357322	-6.560114	0.0000000
perc_black	-0.9861326	0.1494896	-6.596663	0.0000000
perc_non_hispanic_white	-1.2644703	0.2027857	-6.235499	0.0000000

Note:

* At 0.001 Significance Level

Based on these results, there are 23 statistically significant predictors of the average percentage of days people reported feeling mentally unhealthy (at the $\alpha = 0.001$ confidence level). These predictors and coefficient estimates are shown in **Table 2** above. Of these covariates, the strongest positive predictor of poor mental health days is the average percentage of days people from that county reported feeling physically unhealthy. All other things equal, every one-standard-deviation increase in this variable results in an estimated increase in a county's average percentage of poor mental health days of 0.9691171 standard deviations. All other estimated coefficients can be interpreted similarly.

In addition to a relatively low MSE of 0.1877 and a high adjusted R-squared of 0.8123, preliminary residual analysis indicates a good model fit. A plot of model residuals against fitted values is shown in **Figure 1** below. Most residuals are between -0.5 and 0.5, and none have an absolute value greater than 2. Considering that our sample includes 2,636 observations, this is an encouraging result that validates our model choice.

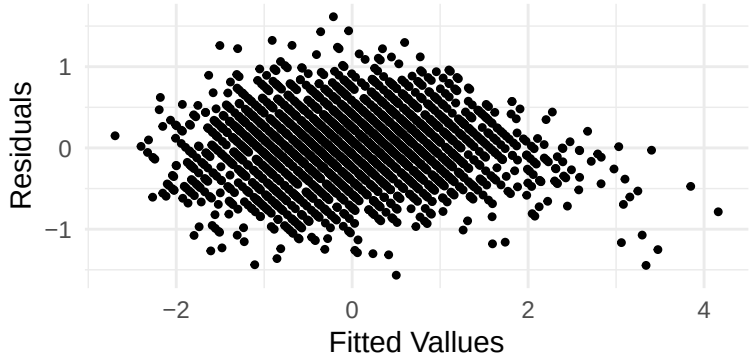

Figure 1: OLS Residuals vs Fitted Values

Table 3: Variance Inflation Factor, Top 20 Covariates

variables	vif
perc_non_hispanic_white	577.380645
perc_black	313.768552
perc_hispanic	258.674407
food_environment_index	228.924087
perc_uninsured	171.409040
perc_uninsured_adults	130.682913
perc_food_insecure	106.633366
perc_limited_access_healthy_foods	72.441778
perc_ai_an	60.459893
perc_fair_or_poor_health	42.491449
perc_severe_housing_problems	36.379253
severe_housing_cost_burden	26.832489
perc_adults_diabetes	18.545691
perc_days_physically_unhealthy	15.011048
perc_physically_inactive	13.863450
perc_asian	12.497300
perc_adults_smoking	11.776328
perc_uninsured_children	11.615412
perc_children_in_poverty	9.726445
median_household_income	8.772686

However, including all 60 covariates in this regression presents some issues, especially regarding multicollinearity and the subsequent interpretability of our estimated β_js. The Variance Inflation Factor (VIF) measures the extent to which predictors in a multiple regression model are correlated. Values between 5 and 10 generally cause mild concern, and values over 10 are potentially problematic. As **Table 3** above demonstrates, 18 of our 60 covariates have a VIF that exceeds this threshold.

Upon further inspection, it is clear that the variables with high variance inflation factors would be highly correlated with each other and, in some cases, even deterministic (ethnicity proportions summing to 1). Though this result does not compromise the model's predictive ability, accurate interpretation of its coefficients becomes nearly impossible. Therefore, we must implement a variable selection procedure to weed out surplus and correlated predictors while maintaining model integrity. With this in mind, we turn our focus to LASSO Regression.

LASSO Regression

LASSO, the "Least Absolute Shrinkage and Selection Operator," is a statistical method used for regularization and variable selection for linear models. Instead of calculating β_js using the ordinary least squares solution $\hat{\beta}_{\text{OLS}} = \underset{\beta}{\text{argmin}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2 \right\}$, LASSO introduces an additional penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$ to the optimization problem. This term is proportional to the sum of the absolute value of β_j's and induces both bias and sparsity as it pushes all β_js closer to 0. The strength of this penalty is governed by λ, which is typically tuned with k-fold cross-validation ("The Lasso" 2023).

In the context of our data set, the LASSO serves as a useful methodology for variable selection (eliminating surplus predictors with OLS coefficients close to 0 and limiting multicollinearity) while preserving the model's predictive ability and linear structure. After standardizing all variables, a LASSO regression model was fit on the data set, and β_js were calculated according to the LASSO solution $\hat{\beta}_{\text{lasso}} = \underset{\beta}{\text{argmin}} \left\{ \frac{1}{2\times 2{,}636} \sum_{i=1}^{2{,}636} (Y_i - \beta_0 - \sum_{j=1}^{60} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{60} |\beta_j| \right\}$, with λ tuned by 10-fold cross-validation. The optimal LASSO estimates (at $\lambda = 0.0043$) for each non-zero β_j coefficient are shown in **Table 4** below.

Table 4: LASSO Regression Results

Term	Estimate	Standard Error	Test Statistic	P-Value	Significant?*
perc_days_physically_unhealthy	0.9174263	0.0236629	38.7705794	0.0000000	Yes
perc_insufficient_sleep	0.2842287	0.0188927	15.0443792	0.0000000	Yes
perc_food_insecure	0.1503663	0.0250783	5.9958743	0.0000000	Yes
perc_female	0.1390360	0.0123442	11.2633045	0.0000000	Yes
perc_excessive_drinking	0.1088395	0.0130029	8.3704291	0.0000000	Yes
median_household_income	0.0800028	0.0238600	3.3530133	0.0003997	Yes
perc_non_hispanic_white	0.0685877	0.0288076	2.3808878	0.0086355	No
perc_income_required_child_care	0.0618056	0.0118494	5.2159133	0.0000001	Yes
perc_households_broadband_access	0.0543264	0.0156149	3.4791313	0.0002515	Yes
perc_vaccinated	0.0476528	0.0121505	3.9218714	0.0000439	Yes
perc_adults_smoking	0.0457616	0.0273418	1.6736855	0.0470962	No
perc_homeowners	0.0439285	0.0167506	2.6224999	0.0043644	No
mental_health_provider_ratio	0.0347891	0.0108861	3.1957494	0.0006973	Yes
avg_daily_pm	0.0345433	0.0116021	2.9773440	0.0014538	No
perc_long_commute_alone	0.0285531	0.0138025	2.0686926	0.0192875	No
primary_care_physicians_ratio	0.0274306	0.0114893	2.3874816	0.0084821	No
school_funding_accuracy	0.0250561	0.0154339	1.6234455	0.0522471	No
perc_completed_high_school	0.0198313	0.0199513	0.9939864	0.1601147	No
perc_nh_opi	0.0192969	0.0093790	2.0574641	0.0198208	No
population	0.0039549	0.0105256	0.3757395	0.3535553	No
perc_uninsured_children	0.0030142	0.0190144	0.1585227	0.4370225	No
income_ratio_20_80	0.0028580	0.0128346	0.2226795	0.4118925	No
perc_with_exercise_opportunities	0.0009233	0.0130522	0.0707425	0.4718014	No
perc_voter_turnout	0.0002561	0.0174715	0.0146559	0.4941533	No
food_environment_index	-0.0001096	0.0178931	-0.0061227	0.4975574	No
gender_pay_gap	-0.0008234	0.0095196	-0.0864918	0.4655377	No
perc_ai_an	-0.0036752	0.0144227	-0.2548173	0.3994321	No
perc_rural	-0.0039894	0.0181070	-0.2203254	0.4128089	No
preventable_hospitalization_rate	-0.0044378	0.0110612	-0.4012087	0.3441332	No
inadequate_facilities	-0.0056258	0.0095098	-0.5915833	0.2770648	No
perc_severe_housing_problems	-0.0059188	0.0163777	-0.3613919	0.3589032	No
perc_drive_alone_to_work	-0.0120388	0.0122085	-0.9861000	0.1620420	No
perc_black	-0.0128225	0.0215203	-0.5958329	0.2756434	No
perc_annual_mammagram	-0.0203372	0.0129466	-1.5708497	0.0581088	No
life_expectancy	-0.0226543	0.0175335	-1.2920603	0.0981681	No
overcrowding	-0.0229862	0.0153056	-1.5018146	0.0665725	No
perc_unemployed	-0.0231683	0.0137226	-1.6883278	0.0456742	No
perc_driving_deaths_alcohol	-0.0357845	0.0090729	-3.9441088	0.0000400	Yes
perc_uninsured_adults	-0.0416201	0.0231062	-1.8012550	0.0358313	No
perc_children_in_poverty	-0.0455739	0.0255140	-1.7862325	0.0370308	No
perc_adults_obesity	-0.0564907	0.0173800	-3.2503340	0.0005763	Yes
perc_some_college	-0.0597547	0.0185437	-3.2223739	0.0006357	Yes
perc_under_18	-0.0741654	0.0179528	-4.1311208	0.0000180	Yes
perc_65_over	-0.0777112	0.0190924	-4.0702739	0.0000235	Yes
perc_physically_inactive	-0.2802625	0.0285680	-9.8103560	0.0000000	Yes

Note:

* At 0.001 Significance Level

Table 5: LASSO, Statistically Significant Predictors

Term	Estimate	Standard Error	Test Statistic	P-Value
perc_days_physically_unhealthy	0.9174263	0.0236629	38.770579	0.0000000
perc_insufficient_sleep	0.2842287	0.0188927	15.044379	0.0000000
perc_food_insecure	0.1503663	0.0250783	5.995874	0.0000000
perc_female	0.1390360	0.0123442	11.263305	0.0000000
perc_excessive_drinking	0.1088395	0.0130029	8.370429	0.0000000
median_household_income	0.0800028	0.0238600	3.353013	0.0003997
perc_income_required_child_care	0.0618056	0.0118494	5.215913	0.0000001
perc_households_broadband_access	0.0543264	0.0156149	3.479131	0.0002515
perc_vaccinated	0.0476528	0.0121505	3.921871	0.0000439
mental_health_provider_ratio	0.0347891	0.0108861	3.195749	0.0006973
perc_driving_deaths_alcohol	-0.0357845	0.0090729	-3.944109	0.0000400
perc_adults_obesity	-0.0564907	0.0173800	-3.250334	0.0005763
perc_some_college	-0.0597547	0.0185437	-3.222374	0.0006357
perc_under_18	-0.0741654	0.0179528	-4.131121	0.0000180
perc_65_over	-0.0777112	0.0190924	-4.070274	0.0000235
perc_physically_inactive	-0.2802625	0.0285680	-9.810356	0.0000000

Note:
* At 0.001 Significance Level

The results in **Table 4** only include 45 coefficients, indicating that the LASSO shrunk the coefficients of 15 weakly-relevant covariates to 0. Additionally, only 16 variables were deemed statistically significant predictors of our outcome (at $\alpha = 0.001$), and each of their coefficients is slightly closer to 0 than their corresponding OLS estimate. Unsurprisingly, the average percentage of days spent physically unhealthy remains the clearest predictor of the average percentage of days spent mentally unhealthy. Meanwhile, no ethnicity percentages were deemed statistically significant (OLS had 5). These results, displayed in **Table 5** above, showcase the power of LASSO for shrinkage and variable selection, even in data sets riddled with multicollinearity. Additionally, **Figure 2** provides a visualization of how the strength of these properties varies across different values of λ.

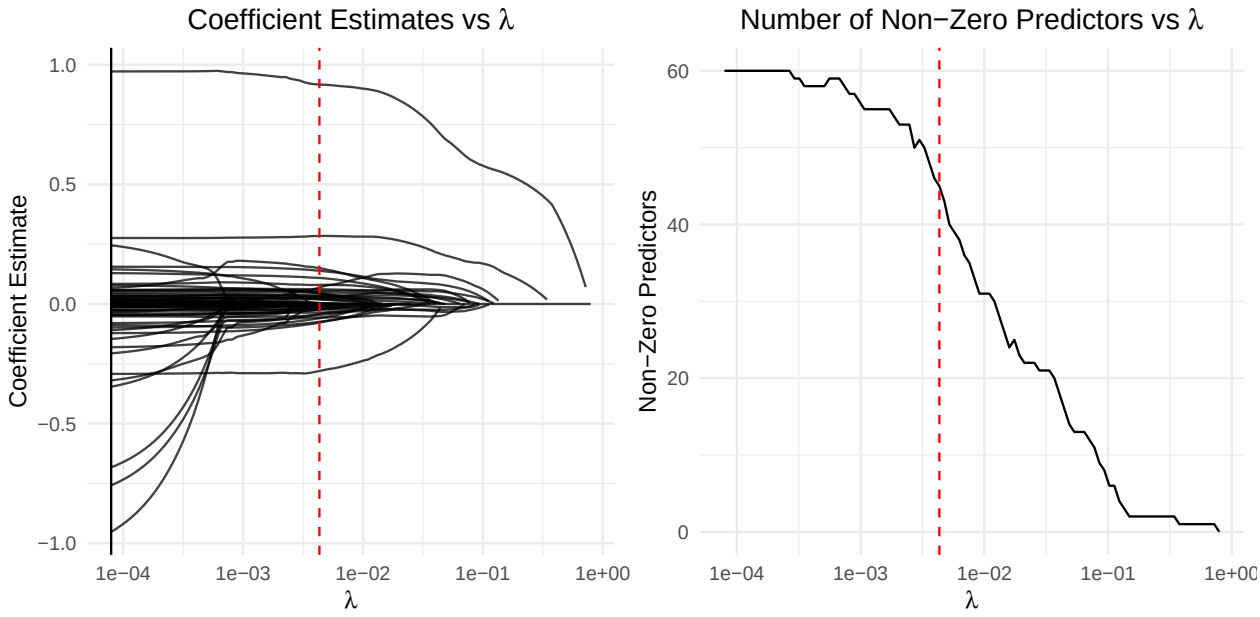

Figure 2: LASSO Visualizations

Despite eliminating 15 variables through shrinkage, the LASSO only increased the MSE by 0.0115 compared to our linear regression model. Additionally, this algorithm does not induce outliers, as evidenced by **Figure 3** below. This unique ability to preserve the predictive ability of a linear regression while making its covariates more independent and its coefficients more interpretable sets the LASSO apart as a viable statistical method for econometric analysis.

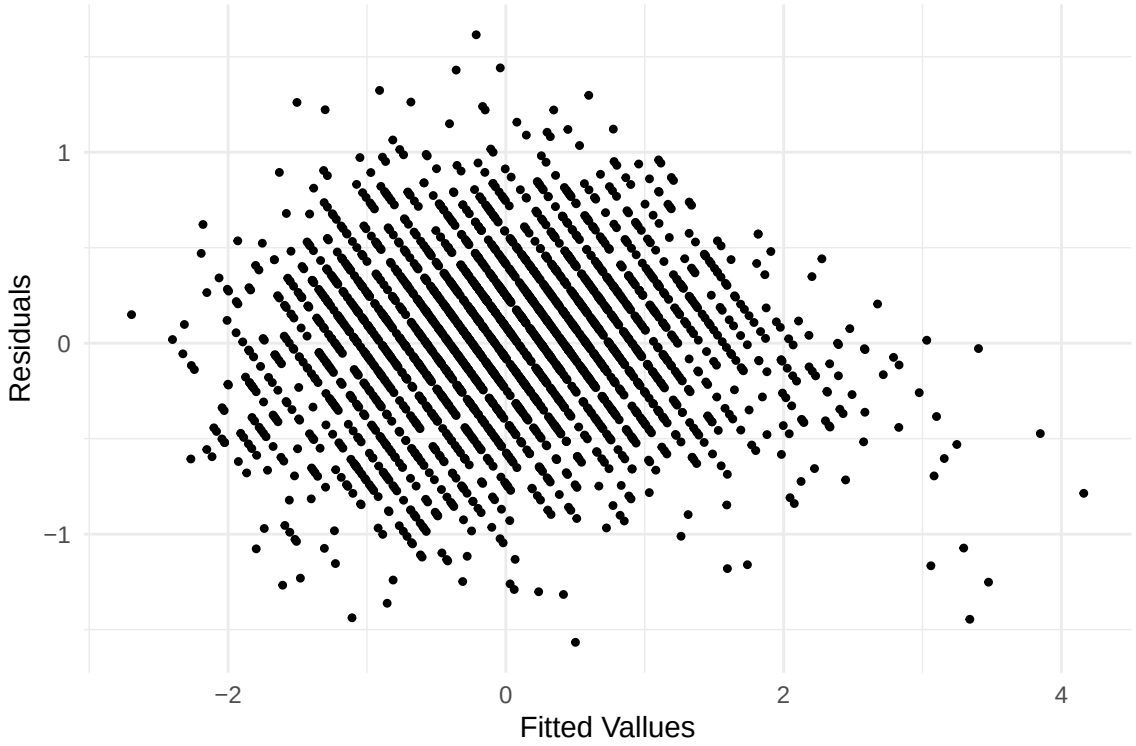

Figure 3: LASSO Residuals vs Fitted Values

Turning our attention to the problem at hand, it is clear that physical health is highly connected to mental health: all else held equal, a one-standard-deviation increase in a county's average percentage of days reported as physically unhealthy results in a 0.9174263 SD increase in its average percentage of days reported as mentally unhealthy. This finding supports the biopsychosocial model of health, which emphasizes the connectivity of biological, social, and psychological wellness (Megan 2021). The two next-most significant positive predictors (the percentage of adults who report less than 7 hours of sleep per night and the percentage of adults who suffer food insecurity) also directly support this framework, together indicating that addressing mental health concerns on a county level necessitates investment in both mental health treatment and support for impoverished and overworked communities that enable them to live healthy lives.

However, the significant negative coefficient associated with the percentage of physically inactive adults is puzzling, as it starkly contrasts the marginal association between these variables. This indicates either the presence of a confounding variable or of continued multicollinearity in the LASSO regression setting. Additionally, the significant positive coefficients associated with median household income and broadband internet access indicate that this study may suffer from biased sampling or self-reporting bias.

Finally, the significance of demographic variables, such as the percentage of females in a county and the percentage of (non) working-age people (18-65) in predicting poor mental health, may indicate under-reporting in specific demographics. However, it could also guide government officials to focus their resources and efforts on helping these groups of people.

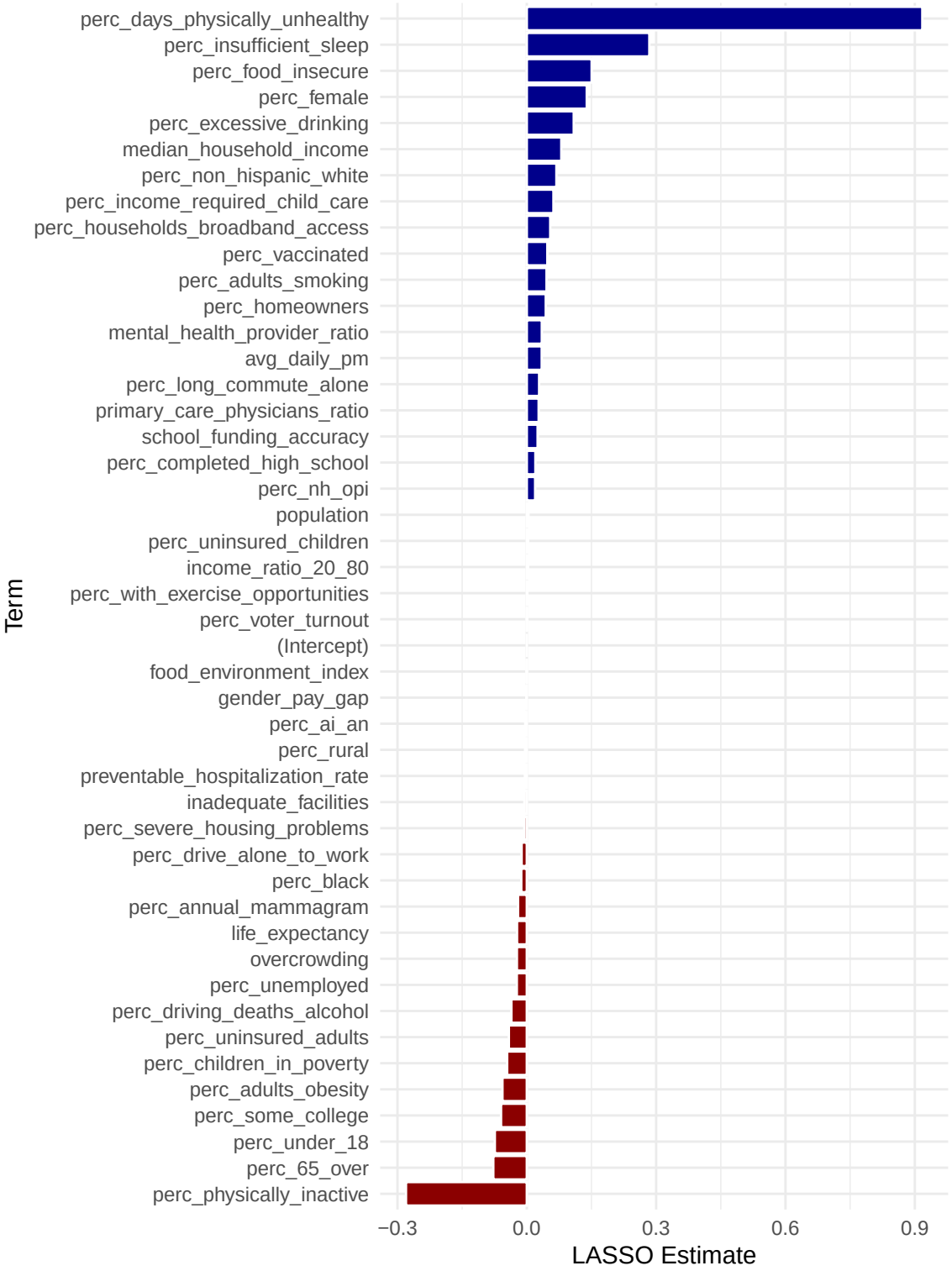

Figure 4: LASSO Coefficient Plot

Conclusions

Though least-squares regression remains an essential econometric method, the interpretation of its coefficients is easily complicated by multicollinearity in higher-dimensional settings. We argue the viability of LASSO regression as a solution to this problem based on its properties, which preserve OLS' inferential and predictive properties while reducing multicollinearity and increasing the interpretability of its coefficients. Finally, we interpret the optimal LASSO coefficients and provide policy recommendations to help state and local government officials address their communities' mental health challenges. Future work includes investigating methods to de-bias LASSO coefficients for inference when the optimal λ is large and developing variance inflation factor methodology to quantify the effect of LASSO regression on multicollinearity more accurately.

References

"2022 National Survey on Drug Use and Health (NSDUH) Releases." 2023. 2023. https://www.samhsa.gov/data/release/2022-national-survey-drug-use-and-health-nsduh-releases.

Megan. 2021. "Three Aspects of Health and Healing: The Biopsychosocial Model in Medicine." 2021. https://surgery.wustl.edu/three-aspects-of-health-and-healing-the-biopsychosocial-model/.

Office, HHS Press. 2023. 2023. https://www.hhs.gov/about/news/2023/08/23/biden-harris-administration-awards-more-than-64-million-grants-fund-mental-health-services-awareness-training-united-states-territories.html.

"The Lasso." 2023. 2023. https://online.stat.psu.edu/stat508/lesson/5/5.4.

Wiemken, Timothy. 2023. "United States County Level County Health Rankings." 2023. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FGSVXG&version=1.0.