County-level socioeconomic outcomes of Oil and Gas Extraction

Abstract

The oil industry is a key constituent of the greater U.S. economy. As a large-scale employer, source of exports, and as an industry closely tied to national security, the oil business is deeply interwoven into the fabric of the United States. In this paper, we investigate the effect of oil extraction on socioeconomic outcomes in local communities. In particular, we use techniques of causal inference to deduce relationships between oil dependence and specific socioeconomic outcomes. We draw potential causal links between income specialization in oil/gas, and characteristics such as increased per-capita income volatility, increased wealth, and poor health outcomes. We further go on to analyze oil industry involvement on education, and public opinion. We find significant correlations between oil industry presence, political affiliation, and climate-related public opinion as determined through a national survey. Overall, this report reinforces the notion that the oil industry is associated with, and may potentially cause, specific adverse and beneficial effects to communities in which a high degree of extraction takes place.

Keywords: Causal Inference, Statistical Modeling, Oil & Gas extraction, Correlation One

Table of contents

Executive summary

US Oil Industry

The oil and gas industry holds an integral, albeit controversial, place within the US economy. The US oil industry may be traced to 1859 when Edwin Drake struck oil in Titusville, Pennsylvania (Drake Well Museum and Park). Although Drake was interested in refining petroleum into Kerosene, the United States soon developed a demand for oil to fuel its burgeoning industry and the rise of automobiles. By 1920, the United States produced over one million barrels of oil per day (U.S. Energy Information Administration). By 1970, this figure had increased ten-fold to over ten million barrels of oil per day (U.S. Energy Information Administration). The growth of the US oil industry has been facilitated by the expansion of corporations into different oil fields across the United States and in coastal waters. To date, the United States is responsible for about 17% of all oil produced in the world (Cleveland et al.). Today, the US oil industry is responsible for roughly 12.3 million jobs, $1.6 trillion in tax revenue (US Department of Energy), and nearly 8% of the US GDP (American Petroleum Institute). In 2017, the average salary of a non-entry level or service station employee within the oil and gas industry was over $100,000 (American Petroleum Institute). Despite its clear contribution to the strength of the US economy, the oil industry has received extensive criticism, largely due to its perceived negative environmental influence and attempts to undermine or otherwise escape regulation.

Resource Curse

The “resource curse” is a well-documented phenomenon wherein countries that become reliant on a particular natural resource (oil, copper, gold, cobalt, etc.) tend to have poorer markers of social and economic progress [1]. In his survey of studies relating to the resource curse, Frankel makes note of the pervasiveness of a phenomenon known as “Dutch Disease.” He describes it as “...possibly unpleasant side

effects of a boom in oil or other mineral and agricultural commodities." In the case of the United States, the 21st century has brought a natural resource revolution of its own—one pertaining to shale oil. With the advent of hydraulic fracturing, the U.S. has gained access to oil reserves far in excess of its previous capacity. The result of which is that formerly rural, agricultural communities now have access to new economic opportunities pertaining to oil extraction. While this confers potential for economic growth, Frankel notes that "Many African countries such as Angola, Nigeria, Sudan, and the Congo are rich in oil, diamonds, or other minerals, and yet their peoples continue to experience low per capita income and low quality of life…" (2). That is, the existence of natural resources does not necessarily confer positive social and economic outcomes to a community dependent on them. We investigate whether this claim holds true in the U.S. in the aftermath of the shale oil revolution.

Longitudinal Study

Experimental Design

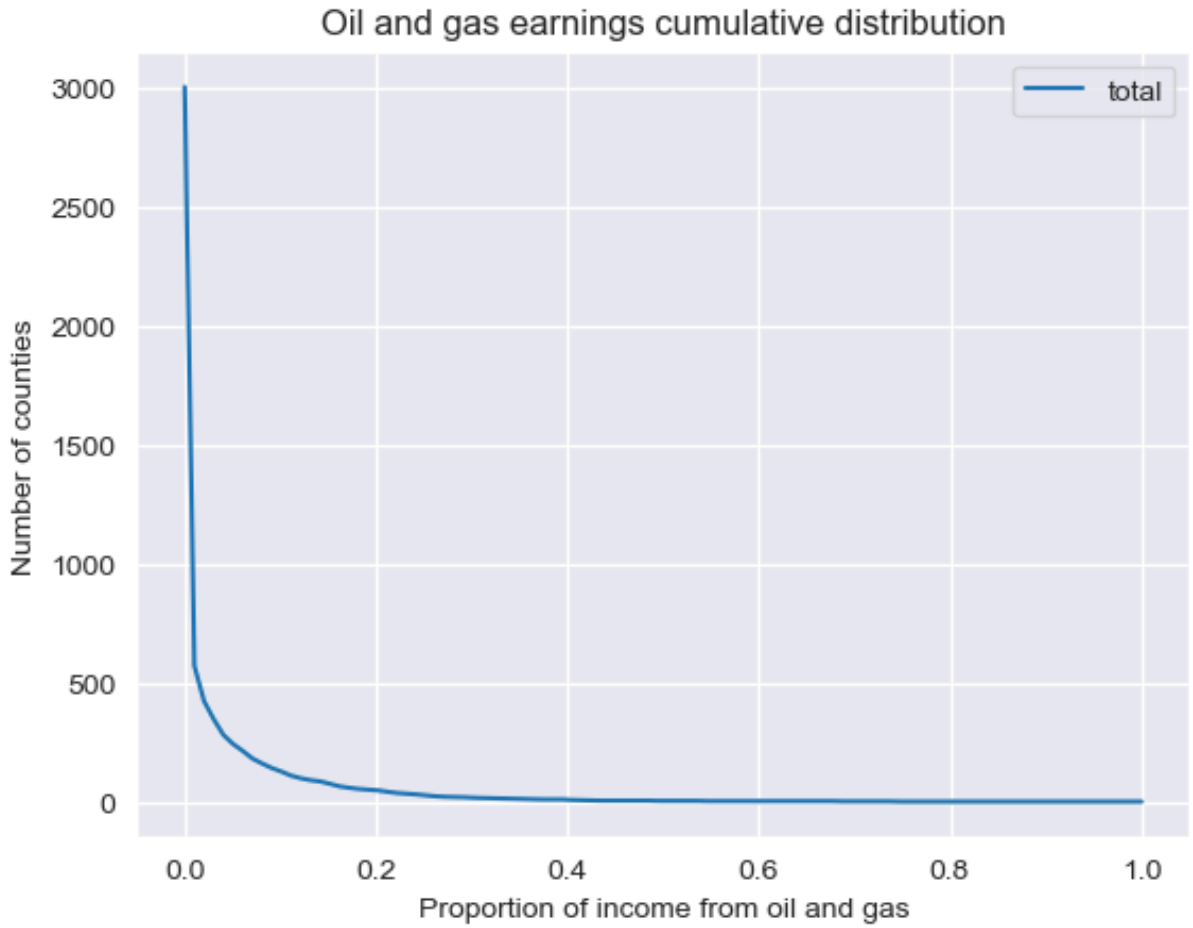


Figure 1: Cumulative distribution of counties with a proportion of income from oil and gas

The first thing that we needed to determine was which counties in the U.S. are particularly dependent on oil and gas as a substantial portion of their economic activity. To this end, we created a feature for `% income from oil and gas` based on the BEA dataset as a proxy for economic dependence on oil. We then computed a cumulative distribution over this quantity (Fig 1). Notice that the vast majority of counties

have none of their income coming from oil and gas extraction. As soon as this number becomes positive, we drop to around 500 candidate communities.

To account for small variance in measurements we establish a cutoff percentage where the contribution of oil and gas extraction becomes significant. In our modeling, we use a cutoff of 5%. The reasoning behind this is both qualitative and quantitative. From a sanity-checking perspective, this cutoff is reasonable since 5% is a substantial portion of income for any community (that also must rely on shopkeepers, teachers, policemen, etc.) We also observe, as we see later on, that communities which exhibit income proportions beyond this threshold share similar idiosyncrasies in economic characteristics. Counties that are otherwise similar but that don't reach this cutoff generally do not exhibit these behaviors. In figure 2 we can see the distribution of such counties across the United States.

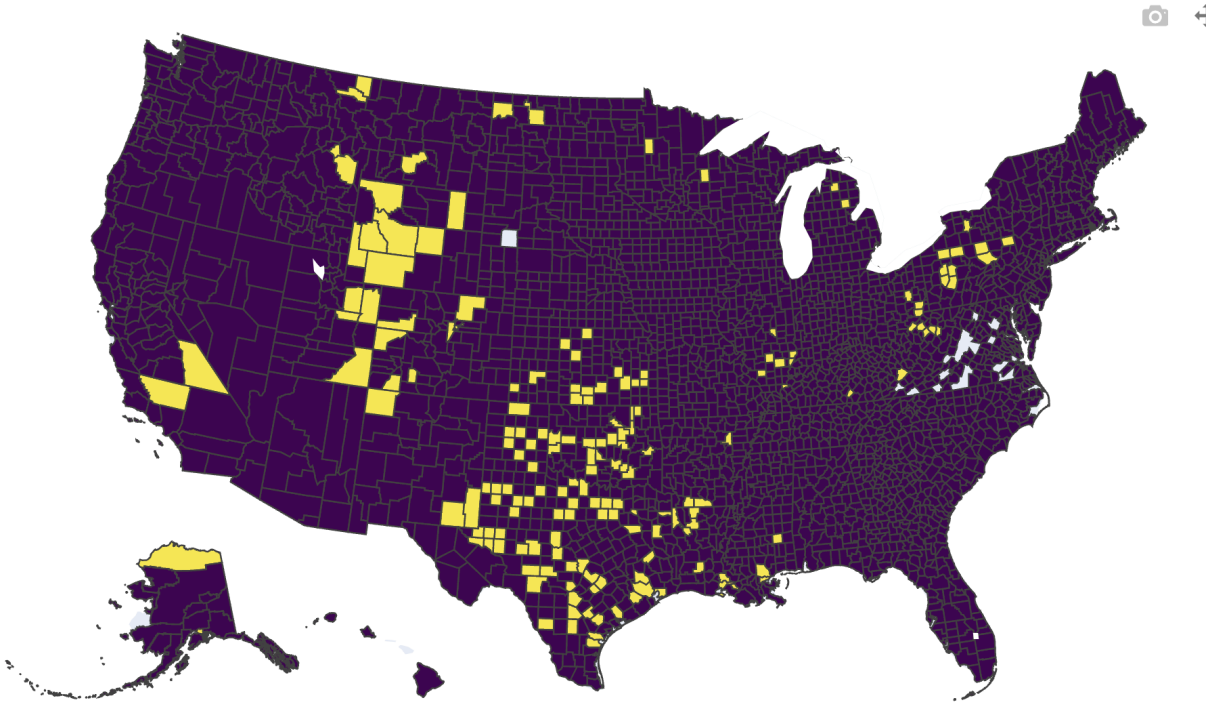


Figure 2: Counties that are reliant on oil by our metric

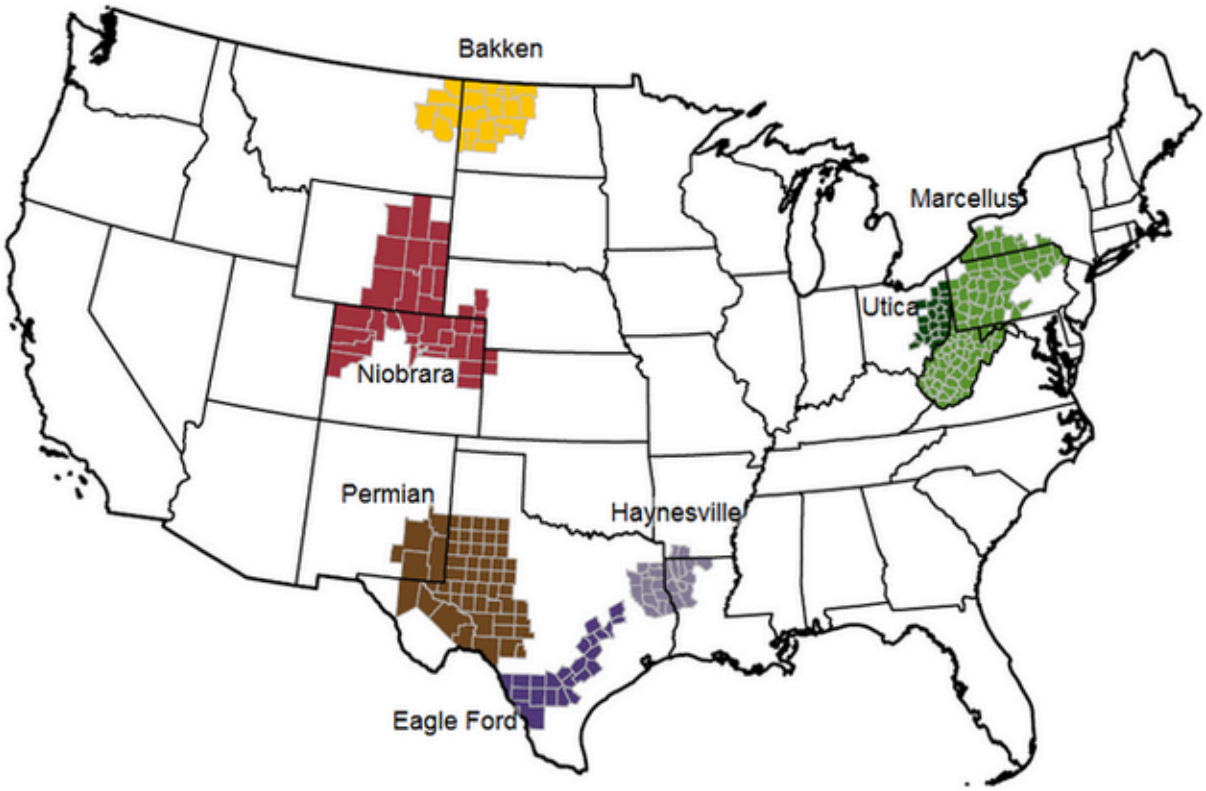


Figure 3: Major shale oil plays in the U.S.

Notice that comparing Fig 2 and Fig 3, we see that many of the counties we singled out coincide with geographic regions where shale oil is currently being extracted. The presence of the other counties likely just reflect non-shale oil development (regular drilling). This gives some qualitative merit to our choice of counties. Notably, most of the counties in question tend to be rural with very low populations. These are the counties that tend to experience the erratic patterns associated with income specialization.

One important observation that we made early on is that the direct economic effects of oil and gas extraction on counties is **large** but often **transitory**. Figure 4 shows us this effect qualitatively. We can see that if a county crosses the threshold, it often does so briefly and with a **large magnitude**. These rapid changes likely have downstream effects on social and economic outcomes, but cannot be predicted by just

using the feature as part of some regression model. For our investigation, we were faced with a few choices for how to robustly model the effects of these changes. Since our goal is to establish causal relationships between the dependence on oil extraction and particular social and economic outcomes, we create synthetic control and treatment groups to simulate a controlled trial.

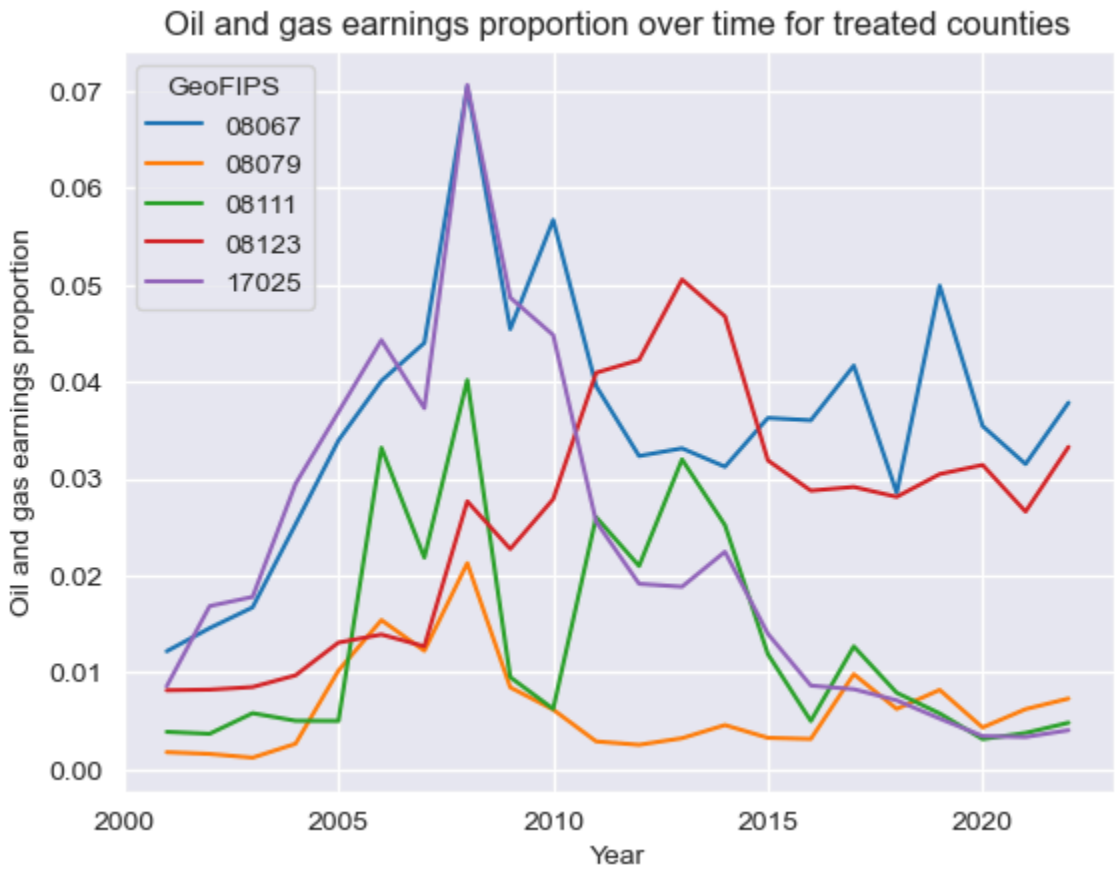


Figure 4: Proportion of income from oil and gas in affected counties

The treatment group consists of counties with **little to no dependence on oil** prior to 2010, and **some level of dependence** beyond 2010 according to our cutoff criterion. There were many reasons why we designed our control group in this way. The first is that we wanted to distinguish between proximal and long-term effects. By asserting that all counties in both groups are identical for some period of time before observing the effects of the “intervention” (oil and gas extraction), we can be sure that

the effects we observe obey the same chronology. That is, we won't conflate long-term observations from a county that became reliant on the 1980 oil boom with short-term observations from a county that benefits from shale oil discovered in 2015. Another reason is that the ACS dataset began tracking selected social characteristics in 2010. Thus, the cutoff gives us a convenient starting point to investigate the divergences between our treatment and control groups.

Our control group has the same size as the treatment group, with the additional stipulation that they never become reliant on oil at any point during the 2001-2023 period. To isolate the effects of shale, the goal is to select a control group that most closely resembles the economic characteristics of the treatment group prior to 2010. We use cluster analysis to select this subset of counties, and then qualitatively verify that the analysis is sound. Indeed, the set of states represented by the control and treatment groups are similar, and both groups tend to contain similar characteristics such as population, per capita income, education rates, and so forth.

To confirm this, we performed a simple hypothesis test on the per capita incomes of the control group and treatment group in 2010. That is, we have

$$\begin{aligned} \mu_1 &= \text{Mean per capita income of control set} \\ \mu_2 &= \text{Mean per capita income of treatment set} \\ H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

$$P(H_0) = 0.7737$$

Thus, we fail to reject the null hypothesis. This is good news since we had hoped that the two datasets would share similar characteristics, With this experimental setup in place, we begin investigating some hypotheses

Per capita income volatility

One well-researched aspect of natural research extraction is the causation of boom-bust cycles in local communities. The reasoning behind this is that often labor demand spikes during the construction phase of large-scale resource extraction operations. This has proximal effects of increasing housing prices, gentrification, and strain on public resources. Ruddell et. al. in their report on “Youth crime in North Dakota boom communities" contend that

> “...more workers are required for the construction of the extraction facilities as well as modes of distribution such as pipelines, and these communities are unprepared for the rapid population increase that follows. Community populations, however, eventually decrease and stabilize after the transition to the production phase, and local gov- ernments—including the justice system—are often able to match the demands for services.”

That is to say, natural resource extraction may lead to economic growth within a community, but that growth may be unstable and short-lived. Most of these effects are observed on a large scale since these studies are primarily conducted in developing countries. Our team found that it would be interesting to investigate if these effects held locally. That is, we hypothesize that **per capita income in the treated group is more volatile than per capita income in the control group.**

First, we see if we can observe this effect qualitatively. To this end, we plot the per-capita incomes of subsets of the treatment and control groups

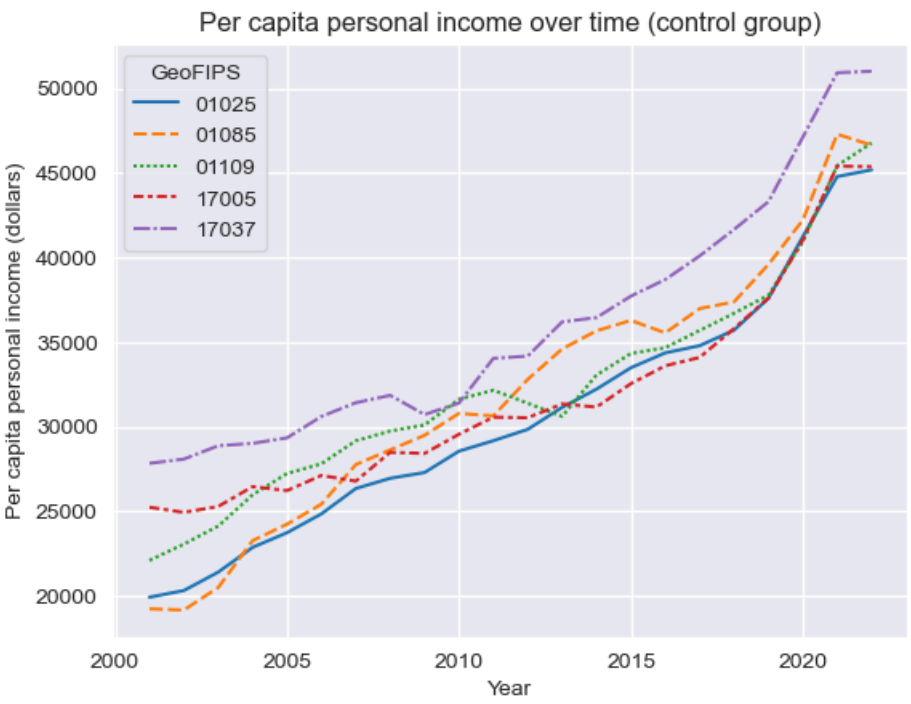


Figure 5: Per capita personal income (control group)

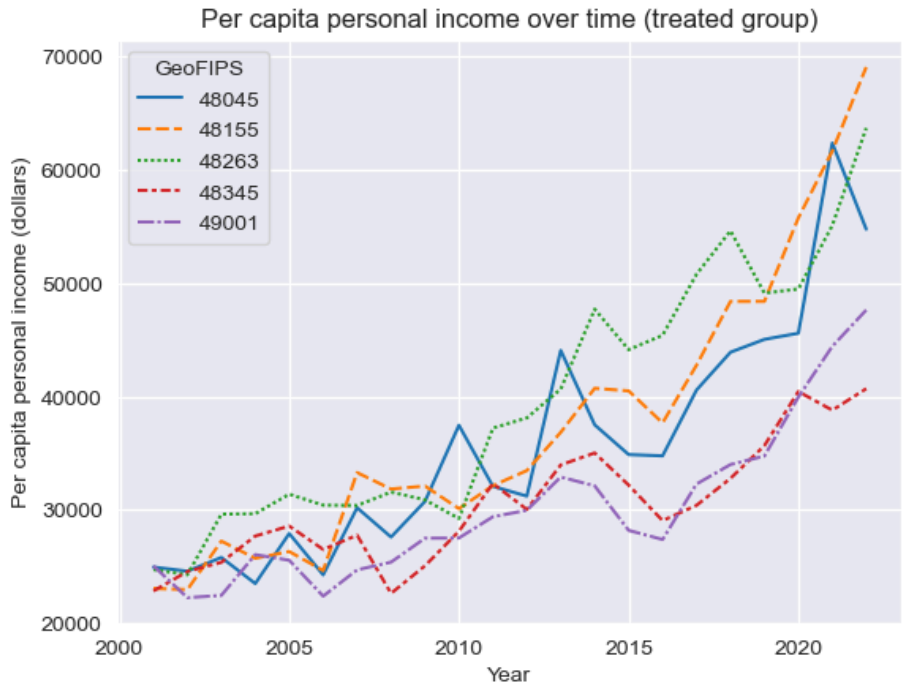


Figure 6: Per capita personal income (treated group)

From the plot, it is plausible that the volatility of the two income curves is different. With some temporary satisfaction from these qualitative findings, we set out to test this relationship quantitatively. What this amounts to is determining the difference in mean volatility between the treatment and control time-series.

The first thing to make note of is that to test the volatility of these time series, we should ensure that they are stationary. In both of these cases, it is clear that they are not. This is primarily because both incomes experience consistent positive growth independent of their membership in either the control or treatment group. We can attribute this to factors such as inflation, and some baseline level of economic growth. To account for this, we take **second-order differences** for both time series. For first order differences, we isolate variance around some constant number (representing baseline growth). For instance, after a first order difference our time-series may look like random noise around the value 5,000. Taking the difference once more gives us “pure” variability in net personal income.

We test the stationarity of the resulting second order differences using the Augmented Dickey-Fuller test. Doing so yields the following results:

Statistic	Treated Value	Control Value
count	40	40
mean p. value	0.321	0.329

In both cases we **reject the null hypothesis that the time-series has a unit root**. This gives us some confidence to assert that both series are nonstationary. Then, we decide on which method to use to test the volatility of both time-series. The conventional approach is to just compare variances, however this can be problematic if certain assumptions are not met. In particular, if the time-series are heteroskedastic then we may increase the risk of type II error. I.e. erroneously failing to reject the null hypothesis that the time series have identical volatilities. To avoid this, we run a

Breusch-Pagan test on the time series to determine whether they are heteroskedastic. The results of this test are tabulated below.

lm	lm_pvalue	fvalue	f_pvalue
0.5934	0.4411	0.5504	0.4677

The main thing to note is the p-value of 0.4411. This is above our significance level of 0.1, and therefore we **fail to reject the null hypothesis that the data are heteroskedastic**. Given this fact, we can more confidently use total variance as a measure of volatility for both sets of time series. More formally,

$$\sigma_1 = \text{Volatility of control set}$$
$$\sigma_2 = \text{Volatility of treated set}$$
$$H_0 : \sigma_1 \geq \sigma_2$$
$$H_1 : \sigma_1 < \sigma_2$$

$$P(H_0) = 0.0133$$

Note the result that p = 0.01 which is below our significance level of 0.05. This allows us to reject the null hypothesis, and claim that the treated group has a higher volatility in per capita personal income than the control group

Change in Net Personal Income

We first aim to test the plausibility of the resource curse hypothesis by analyzing the relationship between historical reliance on the oil and gas industry on long-term economic outcomes. We adopt our methodology from a 2014 paper titled *Long-term effects of income specialization in oil and gas extraction: The U.S. West, 1980–2011* (Haggerty et al.).

We first limit our dataset to the top 200 counties by average percentage of oil and gas income from 2001-2005. We note many of these counties lie in the rural, oil-rich areas mapped in Figure 3. To quantify “reliance on the oil and gas industry”, we construct two distinct features for each county. The first feature, “avg_pct_5” gives the average percentage of income from oil and gas from 2001-2005. The second feature, “oil_var_5”, gives the variance in the percentage of income from oil and gas extraction from 2001-2005. Together, these features are intended to represent the degree of reliance on the oil industry, as well as presence of “boom/bust” patterns commonly exhibited within oil-rich regions. To represent long-term economic outcomes, we construct a third feature, “pc_growth”, roughly corresponding to the percent change in per-capita income from 2001-2022. We first analyze the relationship between our two independent variables, “avg_pct_5” and “oil_var_5” against the target variable, “pc_growth”.

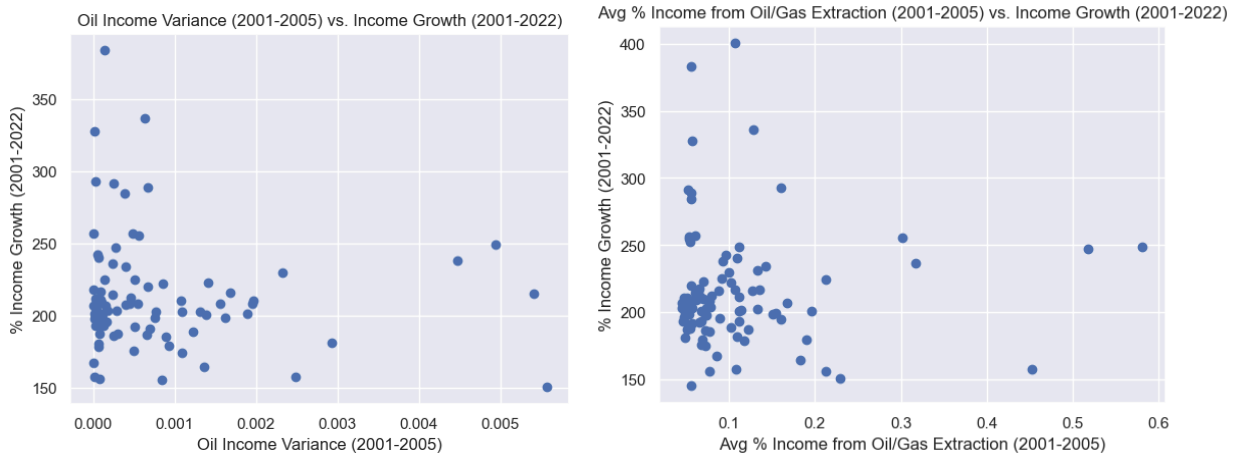


Figure 7: Oil income variance plots

At first glance, there is no distinctly visible non-zero linear relationship between the target variable and each of the constructed features. We also note heteroskedasticity in the above plots as demonstrated by the below plots of the model residuals.

Figure 8: OLS residuals plot

Drawing from the same paper cited above, we apply a general estimating equations (GEE) model to test the hypothesis that there is a significant relationship between our features and the target variable. We choose to implement a GEE model as it does not assume normally-distributed residuals, a linear relationship, and allows for

the specification of custom covariance structures to account for correlated input features.

We obtain the following result table:

	Coeff	Std. Err.	t	$P > \|t\|$	[0.025	0.975]
Intercept	$2.35e+04$	978.973	24.003	0.000	$2.16e+04$	$2.54e+04$
oil_var_5	$-9.795e+04$	$5.62e+05$	-0.174	0.862	$-1.2e+06$	$1e+06$
avg_pct_5	$2.431e+04$	$1.07e+04$	2.271	0.023	3325.149	$4.53e+04$

We note a statistically significant relationship, positive relationship between the avg_pct_5 variable and pc_growth at a 5% significance level. That is, the GEE model would suggest that among oil-reliant counties, a higher initial reliance on oil is associated with positive future income growth. These results would not support the “resource curse” hypothesis within the past 20 years within highly oil-dependent counties within the United States.

Public health outcomes

The next thing that we investigated were public health outcomes. Most studies on the resource curse make note of the fact that oil and gas extraction is usually correlated with poorer health outcomes for the surrounding community. The causal factors behind this could be the pollutants released into the air and into groundwater that ultimately makes its way into the surrounding population’s food and tap water. We were curious whether such effects could be observed in the U.S. in communities that have recently experienced booms in oil and gas extraction.

To investigate this, we procured and cleaned the IHME public health dataset. In particular, we track the life-expectancy of individuals in the 25-30 year as a proxy for the

overall health outcomes of a particular community. We do this for a few reasons—the primary one being that young adults in this age range are most likely to have jobs relating to oil and gas extraction. The data ranges from 2000 to 2019, and is on a per-county basis. We also know that estimates for this group are more likely to be accurate since there are generally more young people than people who are significantly older.

Recall that our goal is to establish causal relationships between that treatment status of a county and its public health outcomes. To this end, we employ a Difference in Differences (DiD) model, which is a statistical technique used to estimate the causal effect of a treatment by comparing the changes in outcomes over time between a treatment group and a control group. This method is particularly useful since it can attempt to establish causality in an observational dataset such as ours. DiD leverages the longitudinal aspect of the data, allowing us to control for unobserved confounding variables that are constant over time and specific to each group.

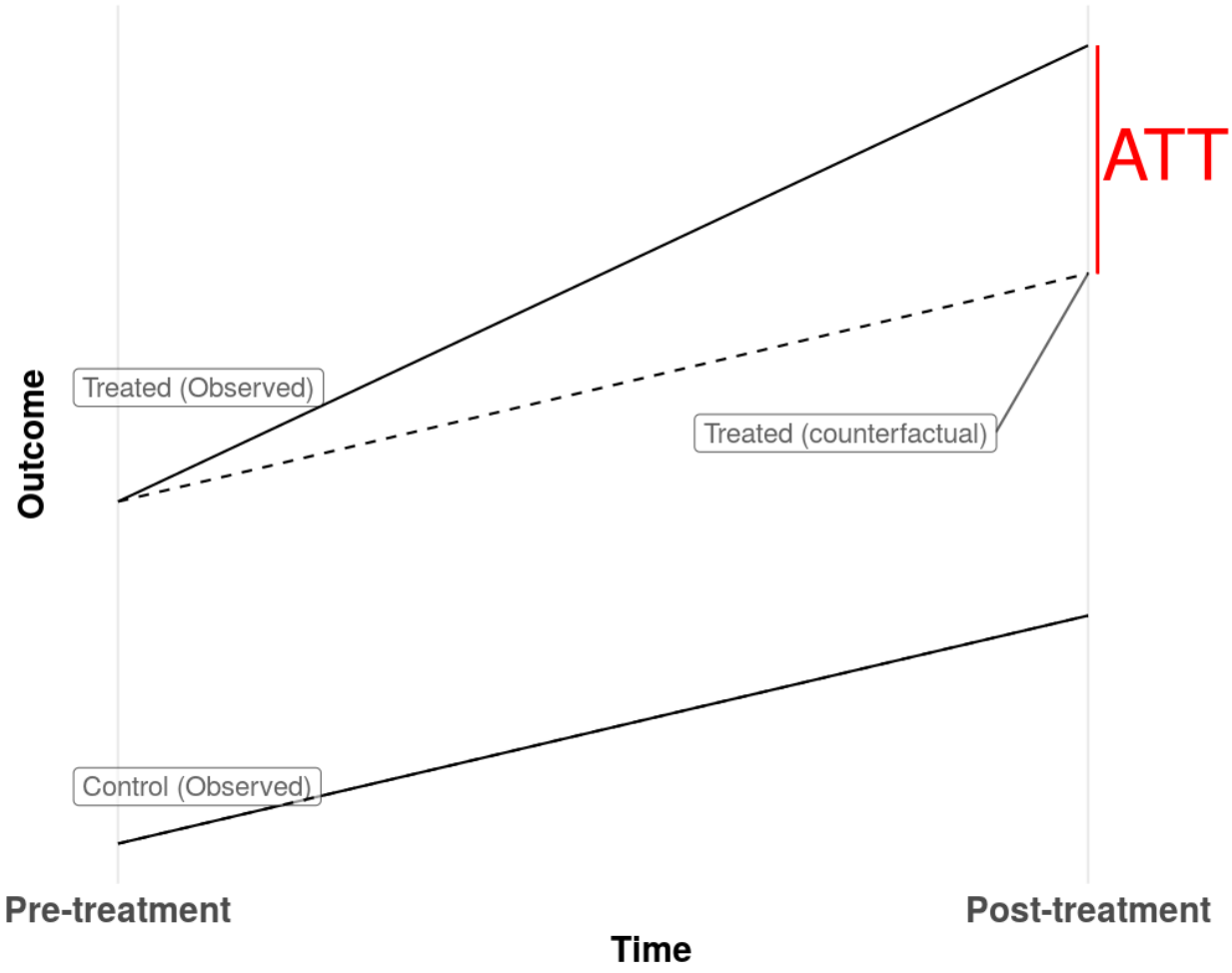


Figure 9: DiD in theory

The core idea behind the DiD approach is to observe the average outcome in both the treatment and control groups before and after the intervention. The 'difference in differences' is the change in the outcome variable for the treatment group minus the change for the control group. This calculation helps to net out any time trends that affect both groups similarly, isolating the impact of the intervention. If the treatment had no effect, we would expect the differences over time between the two groups to be the same; any significant deviation from this can be attributed to the treatment effect.

However, how do we determine when a treatment effect starts? The obvious choice would be when the percentage of income from oil and gas exceeds our

predefined cutoff. The problem with this approach is that, as seen earlier, this percentage income tends to be “spiky.” Thus, there are generally periods of time where this percentage income far exceeds the cutoff, and beyond this there are periods where the income is less than the cutoff. The main insight is to realize that we still want to consider such periods in the treatment group. That is, any period past a “spike” in percentage income from oil and gas extraction should be considered as in treatment, since the county is now actively experiencing the effects of oil and gas extraction.

Running OLS, we get the follow regression results ($p < 0.05$):

	Coeff	Std. Err.	t	$P > \|t\|$	[0.025	0.975]
Intercept	294.7563	48.951	6.021	0.000	198.656	390.856
treatment_post	−0.5126	0.240	−2.132	0.033	−0.985	−0.041
year	−0.1226	0.024	−5.009	0.000	−0.171	−0.075
population	-3.9×10^{-6}	6.72×10^{-7}	−5.829	0.000	-5.24×10^{-6}	-2.6×10^{-6}
income	0.0002	9.79×10^{-6}	18.077	0.000	0.000	0.000

Notice that our treatment status `treatment_post` has a statistically significant nonzero value. This may indicate that there is some relationship between oil and gas extraction and poorer health outcomes in these counties. Notably, time seems to also have a negative correlation with health outcomes, which indicates that these counties are getting unhealthier *anyways* over time. Thus, we may have also discovered an interesting yet unsettling aspect of rural america. Notice that while these coefficients have statistical significance, it is difficult to see which ones are *more* significant than others since they have different scales. To this end, we normalize our features to retrieve the beta-star coefficients for the regression However, before making conclusions about the results from the regression, it is prudent to test the assumptions behind this. First, we test for outliers in our dataset by creating a leverage plot from the regression.

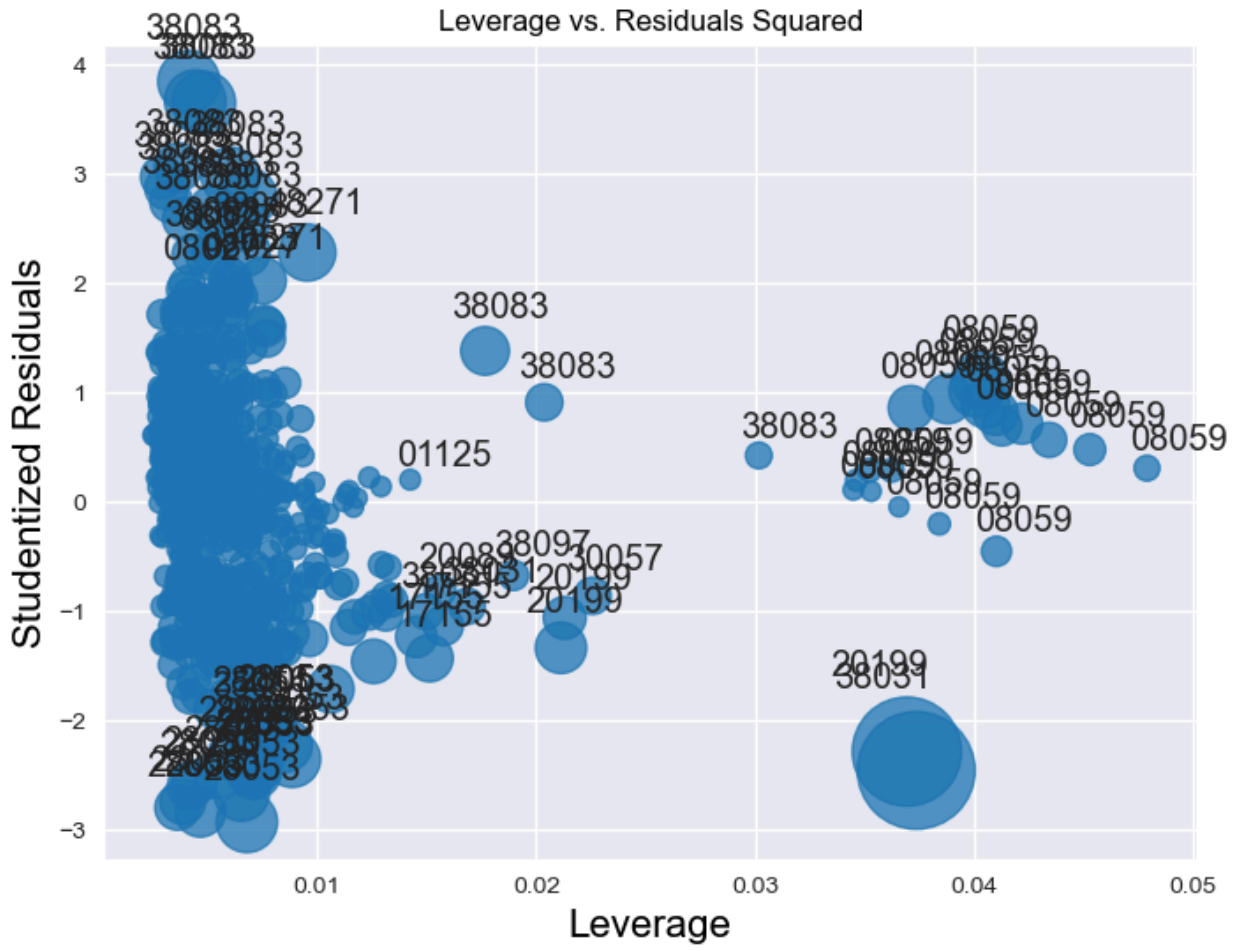


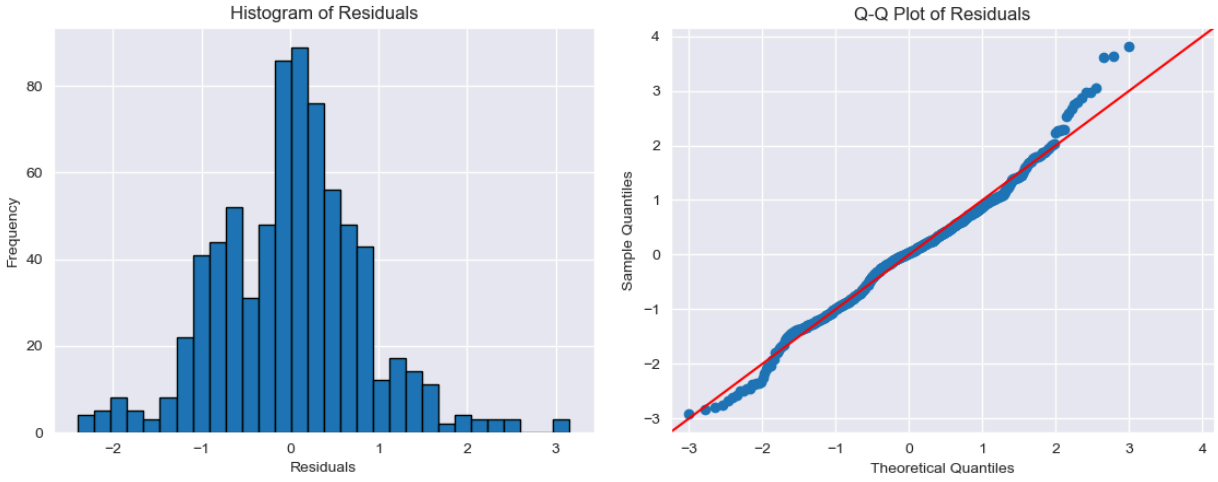
Figure 10: OLS leverage plot

Notice that there are data points which have both substantial leverage and high residuals. Thus, we can conclude that there are outliers in the dataset that could severely increase the variance of our regression. Thus, we consider a variant of OLS that is more robust to these outliers. Specifically, we choose OLS with a Hubert Loss function, which is partially quadratic and partially linear. This way, the model is more robust to extreme outliers while still minimizing the L2 loss for data points near the line of regression.

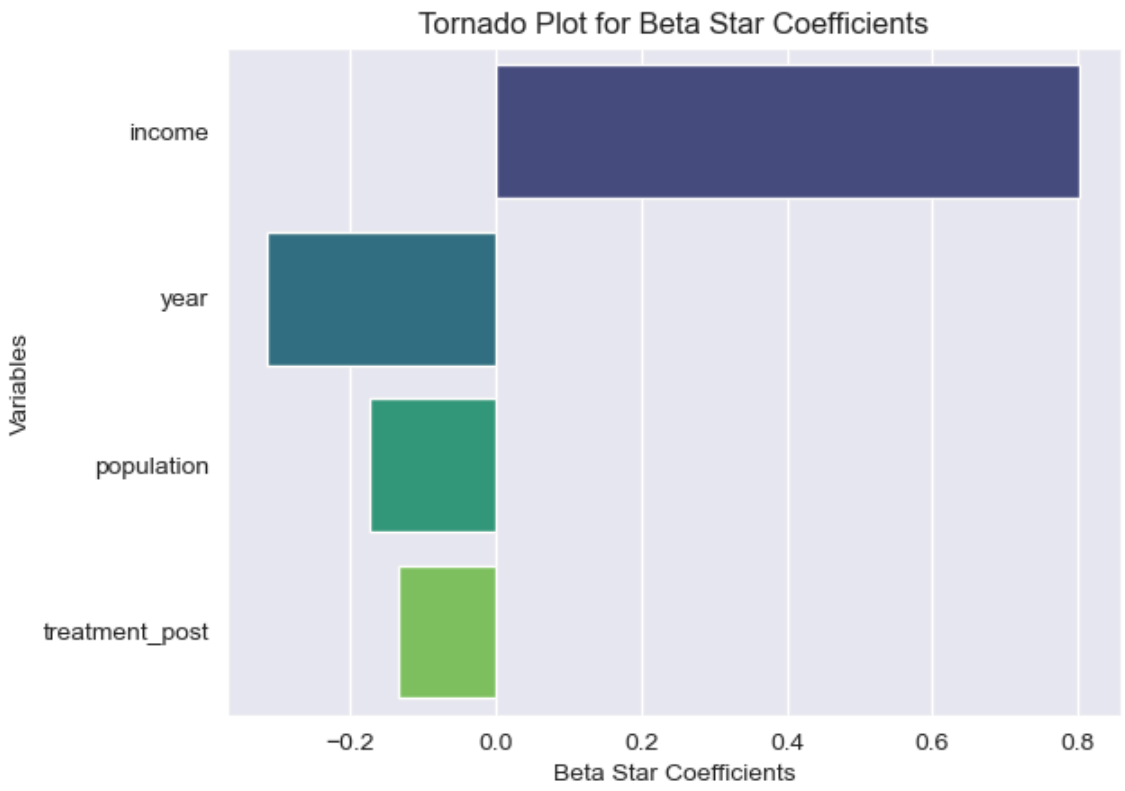
Now, performing robust least squares we get

	Coeff	Std. Err.	t	$P > \|t\|$	[0.025	0.975]
const	0.0060	0.029	0.207	0.836	−0.051	0.063
year	−0.3139	0.063	−4.950	0.000	−0.438	−0.190
income	0.8028	0.043	18.593	0.000	0.718	0.887
population	−0.1748	0.030	−5.746	0.000	−0.234	−0.115
treatment_post	−0.1331	0.057	−2.349	0.019	−0.244	−0.022

Finally, let's make sure that the residuals are normally distributed. We can observe this by plotting the residuals themselves and by creating a Q-Q plot



Notice that the residuals lie close to the line in the Q-Q plot, which indicates that the residuals are likely normally distributed. We could in theory verify this statistically with the Kolmogorov-Smirnov test, but qualitatively we feel that these graphs are compelling enough evidence for this assumption to hold true. With this, we can finally create a tornado plot of beta star coefficients for the regression:



Notice that income seems to be an outsized factor in health outcomes, which aligns with our expectations. As observed earlier, the passage of time, increased population, and `treatment_post` (dummy treatment variable) have negative correlations with health outcomes in decreasing order. Thus, we can say that the oil and gas industry does seem to have adverse effects on the health of a community, but **the effect seems to be small**.

Education outcomes

One of the biggest conclusions from research on the resource curse is that educational outcomes tend to be poorer in communities with high natural resource income specialization. This is explained by the fact that tertiary education is often not required for the work involved in mining/extraction, and thus education is not promoted

in these communities. In Haggerty et. al.'s paper "Long-term effects of income specialization in oil and gas extraction: The U.S. West, 1980–2011," they note that,

> "For counties with high participation during the 1980–82 boom, per capita income over the period 1980–2011 decreases with longer above average income from oil and gas. The magnitude of this relationship is substantial, decreasing per capita income by as much as $7000 for a county with high participation in the boom and long-term specialization (greater than 10 years) versus a hypothetical identical county with only one year of specialization in oil and gas." (Haggerty 193)

Thus, there is evidence that this hypothesis is true for counties that experienced economic growth during the oil boom in the 1980s. Our team investigated whether this trend holds in recent decades. We hypothesized that education outcomes would be poorer in counties wherein oil production played a major role in economic growth.

To test this, we used the column for % of population with bachelor's degree or higher in the ACS dataset as a proxy for educational attainment in these counties. Unfortunately, this ACS data has only been collected on a county-level scale since 2010. Thus, we are unable to replicate the DiD analysis as before since there is not enough data to establish parallel trends. Thus, we abandon the longitudinal aspect of this portion, and run a simple hypothesis test on educational outcomes in 2010 in contrast to 2022.

First, we run a hypothesis test where we see if the treated group and control group have similar educational outcomes in 2010. We get $P(H0) = 0.836$ which is well above our significance level. Thus, we fail to reject the null hypothesis that the two means are the same. Since we are now confident that there is little difference between the means, we calculate Δ%education for each subsequent year. That is, the change in the percentage of people with a bachelor's degree or higher relative to 2010. We then run a hypothesis test on whether the two are statistically different in 2022.

To our surprise, we not only were unable to reject the null hypothesis, but we actually observed the opposite result to what we initially expected. It turns out that counties that rely on oil and gas as a source of income seem to have superior education outcomes in the long term

Also notice that by the graph, there seems to be some merit to the claim that proximally, the effect of oil drilling is to **decrease** the share of the population with advanced degrees. However, in the long term, these communities tend to attract or produce **more** people that receive advanced degrees. There are a few plausible explanations for this phenomenon. Perhaps the presence of oil-related jobs requiring skilled labor encourages students in these communities to pursue higher education. Perhaps in the aftermath of COVID-19, more petroleum engineers wanted to go to the countryside and live in smaller communities? It's almost impossible to say much about this without further analysis of confounding variables, but this is nonetheless an intriguing finding.

Oil Industry and Public Opinion

In addition to economic effects, the local presence of the oil industry may be a powerful driver of public opinion, especially on highly politicized issues. This effect is potentially to be expected. If the oil industry plays a major role within a local economy, residents may be more likely to hold positive opinions towards the oil industry/a specific local company. This effect is likely visible across other industries as well. However, the influence of public opinion within highly oil-dependent counties is of particular interest given the often polarizing nature of the oil industry. In wake of the global climate crisis, lawmakers have attempted to pass regulations intended to reduce carbon emissions. For decades, the oil industry was known to publicly downplay the threat of global warming (Taylor and Cassidy), and actively waged misinformation campaigns to thread doubt regarding humanity's role in the changing climate. Given the United States system of representation, public opinion on issues involving climate, regulation, and

corporate policy may drive legislative action via the electoral process, lobbying, or otherwise civil discourse. We seek to better understand the relationship of the presence of the oil industry on US public opinion.

Data and Methods

We collect public opinion data on issues relating to the climate from the Yale Climate Opinions Dataset for 2020 (Marlon et al.). The public opinions dataset consists of aggregated binary survey responses to a set of climate-related statements (e.g., “Global warming is happening”). The data gives an estimated percent of individuals who responded positively to each question for each county in the dataset. We collect 2020 presidential election data from the MIT Election Data and Science Lab (MIT Election Data and Science Lab). We collect oil industry income data from the BEA dataset referenced previously (Bureau of Economic Analysis).

We seek to test the hypothesis that the presence of the oil industry, as quantified by the level of income attributable to oil and gas extraction (avg_3), is correlated to public opinion on climate-related issues. We seek to address multiple potential confounding variables in this analysis by considering political orientation as an independent variable. We represent political orientation by the percentage of voters who voted Republican in the 2020 general election.

We evaluate our hypothesis via linear regression on the independent variables and the set of response variables in the public opinion dataset. We limit our analysis to the top 150 counties according to the avg_3 variable (given the small proportional oil industry representation across most US counties) At the 5% significance level, we find there to be a significant relationship between oil industry involvement and survey response to the following questions:

Statement	avg_3 p-value	Republic Vote % p-value	avg_3 coefficient	Republic Vote % coefficient
Estimated percentage who think global warming will harm them personally a moderate amount/a great deal	0.0084	1.17E-34	9.93	-24.98
Estimated percentage who think global warming will harm them personally not at all/only a little	0.0155	1.06E-29	-10.83	26.504
Estimated percentage who think their local officials should be doing more/much more to address global warming	0.0276	2.94E-59	5.91	-27.88
Estimated percentage who are not very/not at all worried about global warming	0.0278	2.97E-52	-8.14	34.28
Estimated percentage who are somewhat/very worried about global warming	0.0281	1.49E-52	8.08	-34.30

From the above table, we see that there is a significant relationship between county-level responses to several climate-related questions, even when considering the influence of political partisanship. We note that the questions most closely correlated with the avg_3 independent variable tend to be those that involve attitude towards the risk and level of personal threat presented by climate change. Across the above questions, respondents were less likely to be worried about global warming and less likely to perceive a personal threat. Normalizing for how Republican the respondents identified as, we found a statistically significant relationship between responses from

people residing in counties with heavy oil involvement and correspondingly more positive views on average towards oil. While we cannot establish a causal relationship from this data alone, we believe this data may suggest that the presence of the oil industry may correlate with public opinion on a local level.

Conclusion and Discussion

The findings in this report indicate that in some ways, aspects of the resource curse hypothesis still apply to rural oil-reliant communities in the United States today. However, in other surprising ways, we see that these counties diverge from expectations. We find a possible causal link between increased reliance on oil and gas extraction, and more volatile income streams, greater wealth, and poorer health outcomes. The education dataset was interesting to investigate, but ultimately did not contain data over a long enough time horizon to form causal inferences.

Another of the potential areas of improvement in this study pertains to the experimental design. While the cutoff criterion works quite well in determining treatment and control groups, the discretionary nature of the choice of cutoff is ultimately a source of statistical error. If given the opportunity to restart, we would use *synthetic control trials* over DiD to establish stronger potential causal relationships since these models are more statistically robust, and take away the element of choice when determining the cutoff. Furthermore, there are some assumptions that are difficult to test. For instance, one of the assumptions with DiD is that of parallel trends—i.e. that the dependent variable exhibits the same trends in both the treatment and control groups prior to the intervention. We verified this qualitatively, however there is no well-agreed upon test to rigorously demonstrate this quality. Furthermore, multicollinearity was likely present in our regression models, and may slightly influence some of the confidence intervals/parameter estimates from the DiD regression.

Ultimately, our team is quite pleased with both the positive and null results we have explored throughout this journey, as well as the beauty of statistical modeling. We sincerely hope that the judges will appreciate this work as well.

Works Cited

American Petroleum Institute. *OIL & NATURAL GAS: SUPPORTING THE ECONOMY, CREATING JOBS, DRIVING AMERICA FORWARD*. 2018, https://www.api.org/-/media/files/policy/taxes/dm2018-086_api_fair_share_onepager_fin3.pdf. Accessed 8 April 2024.

Bureau of Economic Analysis. *CAINC5N Personal income by major component and earnings by NAICS industry 1*. 16 November 2023. *Bureau of Economic Analysis*, Bureau of Economic Analysis, https://apps.bea.gov/itable/?ReqID=70&step=1&_gl=1*1n8vqqr*_ga*ODk4NzI5MTkxLjE3MTI1OTI3ODY.*_ga_J4698JNNFT*MTcxMjU5Mjc4NS4xLjEuMTcxMjU5MjgxMi4zMy4wLjA.#eyJhcHBpZCI6NzAsInN0ZXBzIjpbMSwyOSwyNSwzMSwyNiwyNywzMF0sImRhdGEiOltbIlRhYmxlSWQiLCIzMiJdLFsiTWFqb3JfQXJl. Accessed 8 April 2024.

Cleveland, Cutler, et al. “The history of oil production in the United States.” *Visualizing Energy*, 11 September 2023, https://visualizingenergy.org/the-history-of-oil-production-in-the-united-states/. Accessed 8 April 2024.

Drake Well Museum and Park. “Site History.” *Drake Well Museum*, 2024, https://www.drakewell.org/about-us/site-history. Accessed 8 April 2024.

Haggerty, Julia, et al. “Long-term effects of income specialization in oil and gas extraction: The U.S. West, 1980–2011.” *Energy Economics*, vol. 45, 2014, pp. 186-195. *ScienceDirect*, https://www.sciencedirect.com/science/article/pii/S0140988314001534?via%3Dihub. Accessed 8 April 2024.

Institute for Health Metrics and Evaluation (IHME). *United States Mortality Rates by Causes of Death and Life Expectancy by County, Race, and Ethnicity 2000-2019*. 2023. *GHDx*,

https://ghdx.healthdata.org/record/ihme-data/united-states-causes-death-life-expectancy-by-county-race-ethnicity-2000-2019.

Marlon, Jennifer, et al. “Yale Climate Change Opinion Maps (YCOM).” *GitHub*, https://github.com/yaleschooloftheenvironment/Yale-Climate-Change-Opinion-Maps. Accessed 8 April 2024.

MIT Election Data and Science Lab. “County Presidential Election Returns 2000-2020.” *Harvard Dataverse*, 2021, https://doi.org/10.7910/DVN/VOQCHQ. Accessed 8 April 2024.

PricewaterhouseCoopers International Limited. *Economic Impacts of the Oil and Natural Gas Industry on the US Economy in 2011*. July 2013. *American Petroleum Institute*, https://www.api.org/~/media/files/policy/jobs/economic_impacts_ong_2011.pdf. Accessed 8 April 2024.

Shaffer, Brenda, and Taleh Ziyadov, editors. *Beyond the Resource Curse*. University of Pennsylvania Press, Incorporated, 2011. Accessed 8 April 2024.

Sorokin, Leonid. “Exploring the Relationship between Crude Oil Prices and Renewable Energy Production: Evidence from the USA.” *Energies*, vol. 16, no. 11, 2023, p. 4306. *Energies*, https://www.mdpi.com/1996-1073/16/11/4306. Accessed 8 April 2024.

Taylor, Charlotte, and Cecilia Cassidy. “Defense, Denial, and Disinformation: Uncovering the Oil Industry's Early Knowledge of Climate Change - Common Home.” *Common Home | Georgetown University*, 25 October 2023, https://commonhome.georgetown.edu/issues/summer-2023/defense-denial-and-disinformation-uncovering-the-oil-industrys-early-knowledge-of-climate-change/. Accessed 8 April 2024.

US Department of Energy. *THE ECONOMIC BENEFITS OF OIL & GAS*. https://www.energy.gov/articles/economic-impact-oil-and-gas. Accessed 8 April 2024.

U.S. Energy Information Administration. “U.S. Field Production of Crude Oil (Thousand Barrels per Day).” *EIA*, 29 March 2024,

https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=pet&s=mcrfpus2&f=m. Accessed 8 April 2024.

Appendices

Datasets

The primary datasets we worked with were from the Bureau of Economic Activity (Personal income by major component and earnings by NAICS industry) (Bureau of Economic Analysis), U.S. Census Bureau (5 year selected social characteristics), and the IHME Global Health Data Exchange (United States Mortality Rates by Causes of Death and Life Expectancy by County, Race, and Ethnicity 2000-2019)

The BEA dataset consists of economic data pertaining to each county. The column schema is as follows

GeoFIPS	County Identifier (STR)
GeoName	County Name (STR)
IndustryClassification	Industry code (INT)
Description	Description of value (STR)
Unit	Unit (STR)
2001	Value in year 2001 (OPTIONAL[FLOAT])
…	
2022	Value In year 2022 (OPTIONAL[FLOAT])

The ACS dataset consists of 5-year averages of particular social characteristics of a community in a given year. This dataset has many columns, and during our analysis we only used a subset of them. Here is a subset of the column schema

GEO_ID	County Identifier (STR)
…	
DP02_0067E	# people with a bachelor’s degree or higher (FLOAT)
DP02_0067M	Margin of error (FLOAT)
DP02_0067PE	% population with a bachelor’s degree or higher (FLOAT)
DP02_0067PM	% margin of error (FLOAT)
DP02_0068E	# people with veteran status (FLOAT)
…	

The IHME dataset contains various factors relating to health and mortality. Once again, this dataset contains many features which we do not ultimately use. Here is the schema:

measure_id	Measurement identifier (INT)
measure_name	Name of measurement (STR)
location_id	Location identifier (INT)
location_name	Name of location (STR)
fips	FIPS code (STR)
race_id	Race identifier (INT)
race_name	Name of race (STR)

sex_id	Sex identifier (INT)
sex_name	Name of sex (STR)
age_group	Age group identifier (INT)
age_name	Age group description (STR)
year	Year (INT)
metric_id	Metric identifier (INT)
metric_name	Name of metric (STR)
val	Corresponding value (FLOAT)
upper	Upper confidence bound (FLOAT)
lower	Lower confidence bound (FLOAT)