Technical report: Business Types and Poverty Rates

Group C2: Luke Borowy, Tae Kosmo, Jair Vargas

Abstract

The goal of our project is to study the various factors that contribute to poverty in specific counties in the United States. Specifically we wanted to study how the poverty rate of a county is influenced by the type of businesses that are located there. We hypothesize that some types of business will increase the wealth of a county, leading to a lower poverty rate for that county compared to counties without those specific businesses, while other types of business like vice related businesses will lead to a higher rate of poverty for that county. We found that our main hypothesis was true, that some businesses were associated with increased poverty and some with decreased poverty. By pinpointing what business type has the greatest impact on poverty rates we gained a better understanding of how to potentially address these issues.

Introduction

Poverty is a complex societal problem that carries substantial consequences for individuals, families, and communities alike. The poverty rate in the United States as of 2021 is 11.6% (Census.gov). That means that over 38 million people cannot afford the basic needs for their lives. However, this rate varies greatly across different regions of the country. This highlights the need for localized approaches to poverty reduction.

Creating effective poverty reduction strategies requires understanding how different features in an area contribute to the poverty rate. Local politicians can use this information to make informed decisions about economic development initiatives to produce the greatest impact in poverty stricken areas.

While numerous factors contribute to the overall poverty rate, this paper aims to predict the poverty rate in a county based on the number of businesses of different types present in it. We will examine how poverty is affected by the rate of 9 different business types per person in each county in the United States.

Data

```
business <- read_csv("business_data_c2.csv")
set.seed(2023)
```

The dataset we selected was from the United States Environmental Protection agency. It can be located at their Environmental Dataset Gateway, at

https://edg.epa.gov/EPADataCommons/public/ORD/NHEERL/EQI/ We are specifically using the dataset of EQI data from July 2013. The dataset is complete, including data from every county in the US. There are 3140 cases. This dataset contains 227 variables categorized into 5 domains. These are Air, Water, Land, Built, and Socioeconomic. These variables are

intended to describe the environmental health of an area, as well as the human-made effects. The EPA gathered information from several other sources in order to assemble this complete report. While there are 227 variables in the data set, we are focusing only on 9 variables in the Built domain in order to explain the poverty rate. These variables are:

- **Vice related businesses**: rate_al_pn_gm_env
  - Casinos, alcohol, etc
- **Entertainment related businesses**: rate_ent_env,
- **Education related businesses**: rate_ed_env
- **Negative food related businesses**: rate_food_env_neg
  - Sell fast food, food trucks, etc
- **Positive food related businesses**: rate_food_env_pos
  - Sell healthier foods, like grocery stores, sit-down restaurants, etc
- **Healthcare related businesses**: rate_hc_env
- **Recreation related businesses**: rate_rec_env
- **Transportation related businesses**: rate_trans_env
- **Civic related businesses**: rate_civic_env
  - Government, non-profits, etc

These are all measured in (number of businesses) / (population of county).

```
business <- business %>%
  select(rate_food_env_pos, rate_food_env_neg, rate_hc_env,
                                rate_ed_env, rate_al_pn_gm_env, rate_trans_env,
                                rate_civic_env, rate_rec_env, rate_ent_env,
                                pct_pers_lt_pov, COUNTY_NAME)
skim_without_charts(business)
```

```
-- Data Summary ------------------------
                           Values
Name                       business
Number of rows             3141
Number of columns          11
-----------------------
Column type frequency:
```

```
  character                1
  numeric                  10
------------------------
Group variables            None

-- Variable type: character -----------------------------------------------------
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 COUNTY_NAME           0             1  10  17     0     1875          0

-- Variable type: numeric -------------------------------------------------------
   skim_variable     n_missing complete_rate      mean        sd         p0
 1 rate_food_env_pos         7         0.998  0.00182  0.000750  0.000192
 2 rate_food_env_neg        66         0.979  0.000830 0.000320  0.0000618
 3 rate_hc_env              33         0.989  0.00256  0.00116   0.000142
 4 rate_ed_env             125         0.960  0.000599 0.000397  0.0000473
 5 rate_al_pn_gm_env       154         0.951  0.000476 0.000398  0.0000166
 6 rate_trans_env          815         0.741  0.000131 0.000147  0.0000116
 7 rate_civic_env          838         0.733  0.000110 0.0000903 0.00000679
 8 rate_rec_env            248         0.921  0.000330 0.000329  0.0000291
 9 rate_ent_env            220         0.930  0.000428 0.000351  0.0000282
10 pct_pers_lt_pov           0         1      14.2      6.55      0
         p25        p50       p75     p100
 1 0.00142    0.00171    0.00207   0.0149
 2 0.000648   0.000791   0.000975  0.00538
 3 0.00182    0.00243    0.00318   0.0247
 4 0.000368   0.000515   0.0007    0.00392
 5 0.000233   0.000378   0.000587  0.00466
 6 0.0000686  0.000102   0.000155  0.00495
 7 0.0000571  0.0000875  0.000132  0.00106
 8 0.000181   0.000263   0.000385  0.0108
 9 0.000230   0.000364   0.000520  0.00680
10 9.5        13         17.6      56.9
```

We observed that there were a large number of counties with missing data for some of the variables. We had to drop any county that had any NA values in order to produce a regression model.

```
business <- drop_na(business)
```

This drops us down from 3140 rows to only 1878. However, we believe this is still enough to draw meaningful conclusions.

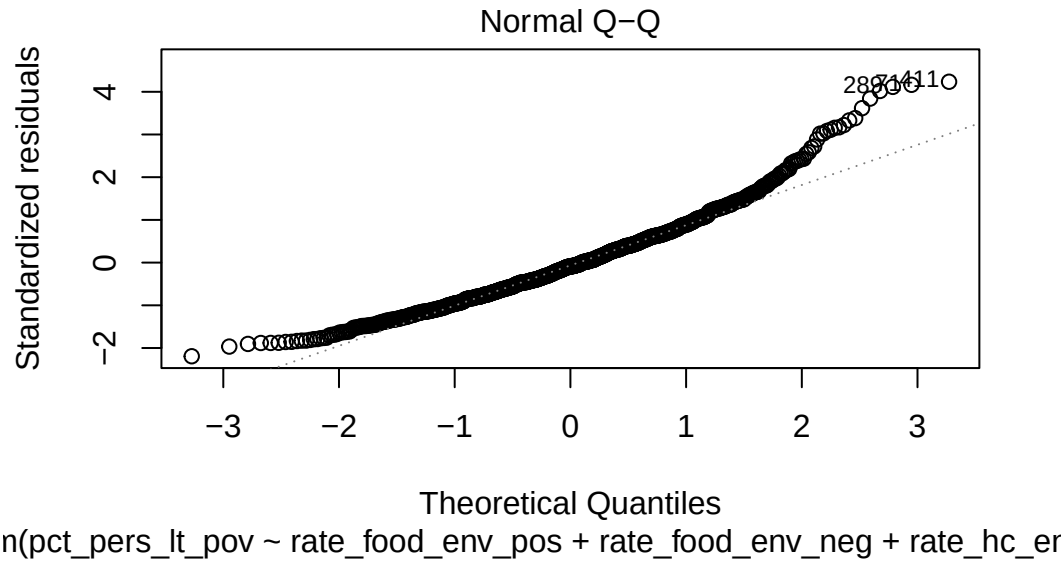
Additionally, this data set had some issues with independence because of the geographic prox-

imity of some counties. To help with this, we randomly split the data into testing and training.

```
which_train <- sample(1:1878, size = 1878/2, replace = FALSE)

training <- business %>%
  slice(which_train)
testing <- business %>%
  slice(-which_train)
```

Fitting a model directly on the raw data leads to violation of the normality condition needed for regression.

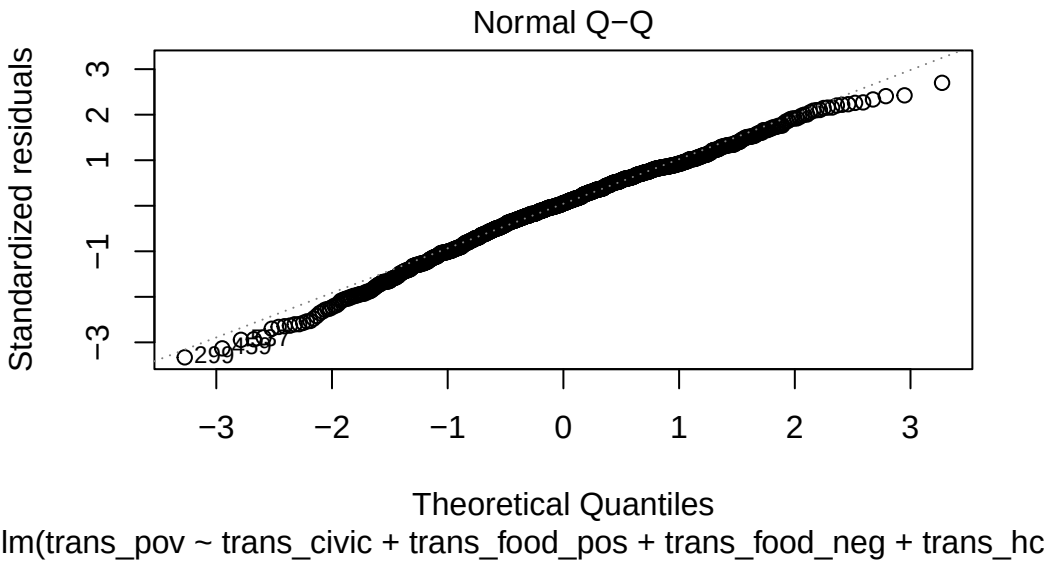
```
raw_model <- lm(pct_pers_lt_pov ~ rate_food_env_pos + rate_food_env_neg +
                  rate_hc_env + rate_ed_env+ rate_al_pn_gm_env +
                  rate_trans_env + rate_civic_env + rate_rec_env +
                  rate_ent_env, data = training)

plot(raw_model, which = 2)
```



As visible on the Normal QQ plot, the data points pull away from the line at the start and end, leading to a "U" shape. To solve this, we will take the log of the poverty rate as well as

every predictor.

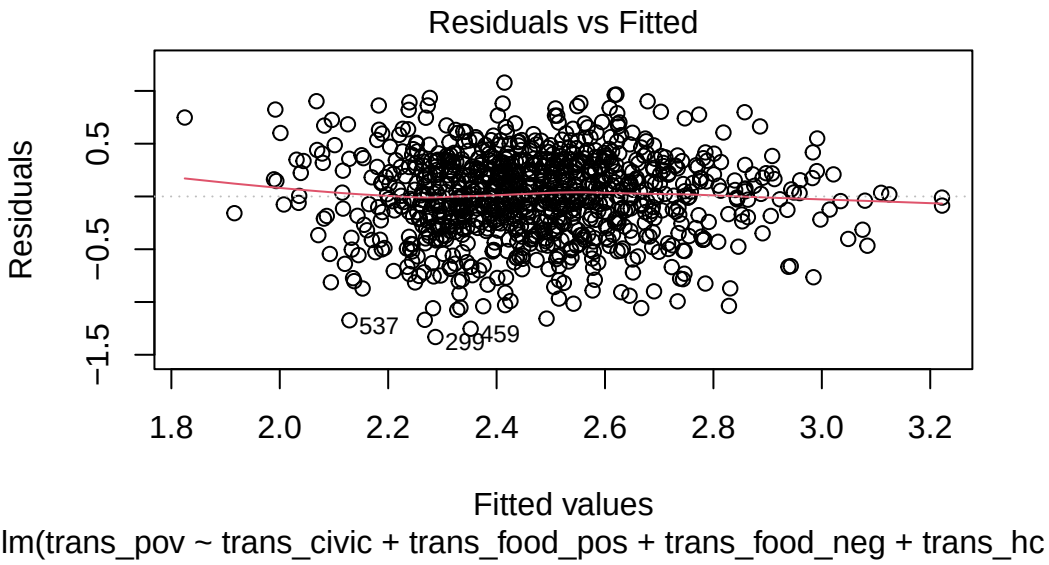
```
training <- training %>%
  mutate(trans_pov = log (pct_pers_lt_pov),
         trans_civic = log(rate_civic_env),
         trans_food_pos = log(rate_food_env_pos),
         trans_food_neg = log(rate_food_env_neg),
         trans_hc = log(rate_hc_env),
         trans_ed = log(rate_ed_env),
         trans_vice = log(rate_al_pn_gm_env),
         trans_trans = log(rate_trans_env),
         trans_rec = log(rate_rec_env),
         trans_ent = log(rate_ent_env))

testing <- testing %>%
  mutate(trans_pov = log (pct_pers_lt_pov),
         trans_civic = log(rate_civic_env),
         trans_food_pos = log(rate_food_env_pos),
         trans_food_neg = log(rate_food_env_neg),
         trans_hc = log(rate_hc_env),
         trans_ed = log(rate_ed_env),
         trans_vice = log(rate_al_pn_gm_env),
         trans_trans = log(rate_trans_env),
         trans_rec = log(rate_rec_env),
         trans_ent = log(rate_ent_env))

m1 <- lm(trans_pov ~ trans_civic + trans_food_pos + trans_food_neg +
           trans_hc + trans_ed + trans_vice + trans_trans
         + trans_rec + trans_ent, data = training)
plot(m1, which = 2)
```



This QQ plot sticks much closer to the line, and although it's not perfect it is much better than the untransformed data. There are some values pulling away at the end, but they are not very numerous compared to the rest of the data. Therefore, the normality condition is met.

To check linearity and equality of variance, we used a residuals vs fitted plot.

```
plot(m1, which = 1)
```



This shows a red line that is almost straight horizontal along the reference line. This means that the linearity condition is met. Additionally, there is not fanning out in the residuals, and the distribution above and below the line is equal. Therefore the equality of variance condition is met.

```
summary(m1)
```

```
Call:
lm(formula = trans_pov ~ trans_civic + trans_food_pos + trans_food_neg +
    trans_hc + trans_ed + trans_vice + trans_trans + trans_rec +
    trans_ent, data = training)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33163 -0.24610  0.02162  0.28090  1.07899

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.84553    0.43238   8.894  < 2e-16 ***
trans_civic     0.08449    0.02276   3.712 0.000218 ***
trans_food_pos  0.49206    0.06184   7.957 5.12e-15 ***
```

```
trans_food_neg  0.26966    0.05070   5.318 1.31e-07 ***
trans_hc       -0.14374    0.04791  -3.000 0.002768 **
trans_ed       -0.08370    0.02824  -2.963 0.003121 **
trans_vice      0.02192    0.02065   1.061 0.288786
trans_trans    -0.06494    0.02291  -2.834 0.004691 **
trans_rec      -0.09983    0.02769  -3.605 0.000329 ***
trans_ent      -0.21971    0.03004  -7.315 5.56e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4011 on 929 degrees of freedom
Multiple R-squared:  0.2049,    Adjusted R-squared:  0.1972
F-statistic:  26.6 on 9 and 929 DF,  p-value: < 2.2e-16
```

In this "kitchen sink" model, every predictor except for `trans_vice` is significant. We think that this is because of the wide range of effects casinos and bars can have in an area. Some counties with casinos are very wealthy because of them, while others have high rates of alcoholism. That may be why trans_vice is not a good predictor of poverty.

For the final model, we drop that variable, leaving 8 predictors which are all significant.

```r
m2 <- lm(trans_pov ~ trans_civic + trans_food_pos + trans_food_neg
         + trans_hc + trans_ed + trans_trans + trans_rec
         + trans_ent, data = training)
summary(m2)
```

```
Call:
lm(formula = trans_pov ~ trans_civic + trans_food_pos + trans_food_neg +
    trans_hc + trans_ed + trans_trans + trans_rec + trans_ent,
    data = training)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32859 -0.24725  0.02366  0.28140  1.07783

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.82046    0.43177   8.848  < 2e-16 ***
trans_civic     0.08712    0.02263   3.850 0.000126 ***
trans_food_pos  0.49437    0.06181   7.999 3.73e-15 ***
trans_food_neg  0.27095    0.05069   5.345 1.14e-07 ***
```

```
trans_hc       -0.14033    0.04780  -2.936 0.003412 **
trans_ed       -0.07976    0.02800  -2.848 0.004491 **
trans_trans    -0.06292    0.02283  -2.756 0.005970 **
trans_rec      -0.09396    0.02713  -3.463 0.000559 ***
trans_ent      -0.22176    0.02998  -7.398 3.09e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4012 on 930 degrees of freedom
Multiple R-squared:  0.2039,    Adjusted R-squared:  0.1971
F-statistic: 29.77 on 8 and 930 DF,  p-value: < 2.2e-16
```

We then checked to make sure that multicolinearity was not an issue. We examine the VIF values for the model, looking for anything that exceeds a VIF value of 5.

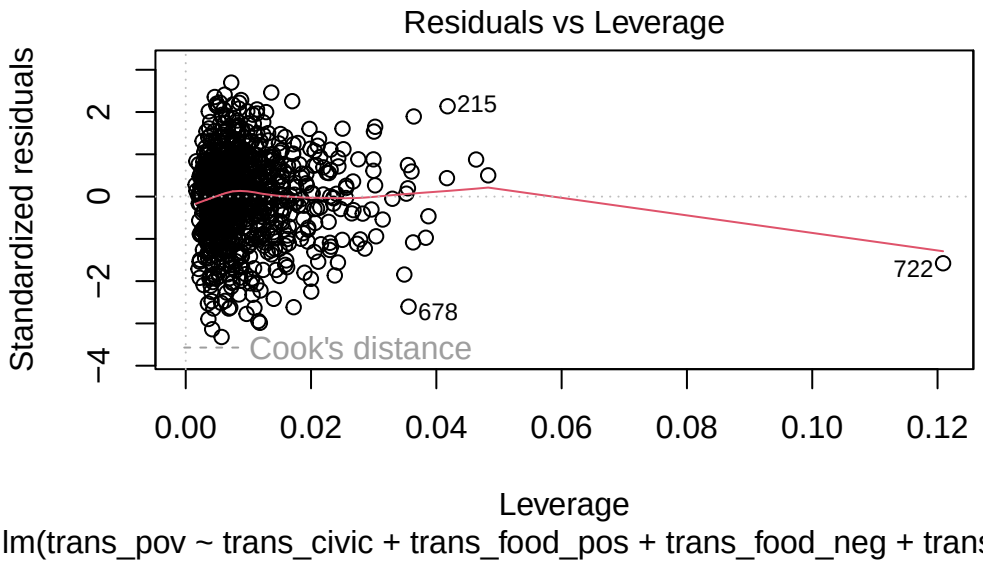
```
vif(m2)
```

```
trans_civic trans_food_pos trans_food_neg       trans_hc       trans_ed
   1.168200       1.722084       1.240231       1.726043       1.134522
trans_trans      trans_rec      trans_ent
   1.204227       1.342464       1.747042
```

Fortunately, the highest VIF value was only 1.747042, which gives evidence that this model does not exhibit multicolinearity.

We also checked to make sure there were not large outliers.

```
plot(m2, which = 5)
```



While there are some data points far from the main cluster, none of them lie outside of Cook's distance, meaning that there are not any outliers.

Results

We performed cross-validation by evaluating our model using the testing data, which was not used while fitting the model.

```
testing <- testing %>%
  mutate(yhats = predict(m2, newdata = testing))
testing <- testing %>%
  mutate(residuals = trans_pov - yhats)

testing %>%
  summarize(cor = cor(trans_pov, yhats)) %>%
  mutate(R2 = cor^2, shrinkage = summary(m1)$r.squared - R2)
```

```
# A tibble: 1 x 3
    cor    R2 shrinkage
  <dbl> <dbl>     <dbl>
1 0.438 0.192    0.0131
```

The shrinkage value is only 1.31%, leading us to believe that the model is able to accurately predict the full dataset.

```
summary(m2)
```

```
Call:
lm(formula = trans_pov ~ trans_civic + trans_food_pos + trans_food_neg +
    trans_hc + trans_ed + trans_trans + trans_rec + trans_ent,
    data = training)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32859 -0.24725  0.02366  0.28140  1.07783

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.82046    0.43177   8.848  < 2e-16 ***
trans_civic     0.08712    0.02263   3.850 0.000126 ***
trans_food_pos  0.49437    0.06181   7.999 3.73e-15 ***
trans_food_neg  0.27095    0.05069   5.345 1.14e-07 ***
trans_hc       -0.14033    0.04780  -2.936 0.003412 **
trans_ed       -0.07976    0.02800  -2.848 0.004491 **
trans_trans    -0.06292    0.02283  -2.756 0.005970 **
trans_rec      -0.09396    0.02713  -3.463 0.000559 ***
trans_ent      -0.22176    0.02998  -7.398 3.09e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4012 on 930 degrees of freedom
Multiple R-squared:  0.2039,    Adjusted R-squared:  0.1971
F-statistic: 29.77 on 8 and 930 DF,  p-value: < 2.2e-16
```

The coefficients on each of the business types varied quite a bit. 3 of the variables had a positive coefficient, meaning that having more of these types of businesses was associated with higher poverty rates. These were

- trans_civic, which is the log-transformed rate of civic businesses
- trans_food_pos, the log-transformed rate of positive food businesses
- trans_food_neg, the log-transformed rate of negative food businesses

The highest was the `trans_food_pos`. Because of the log transformations applied, this means that for every 1 percent increase in number of positive food businesses on a county, the poverty rate is expected to increase by 0.49437 percent, holding all else constant.

The rest of the business type coefficients were negative.

- trans_ed, which is the log-transformed rate of education businesses
- trans_hc, the log-transformed rate of healtcare businesses
- trans_trans, the log-transformed rate of transportation businesses
- trans_rec, the log-transformed rate of recreation businesses
- trans_ent, the log-transformed rate of entertainment businesses

The most negative coefficient was for trans_ent. This means that for every one percent increase in number of entertainment businesses in a county, the poverty rate is expected to decrease by 0.22176 percent, holding all else constant.

Conclusion

Due to the low shrinkage value on testing and training data, we believe that this model is suitable for describing the entire US. On the testing data, the model explains 19.2% of the variability in poverty rate using business types as a predictor, excluding vice businesses. We think that this is a pretty good R^2 value, considering how complex poverty is and how many other factors influence it.

The model did not find evidence that the vice business rate is a significant predictor of poverty. This was surprising, as we originally expected vice to be a good indicator of high poverty rates.

We originally set out to determine what business types contributed positively or negatively to poverty rate in an area. We find that the rate of businesses related to positive food, negative food, and nonprofit/government all contribute to a higher poverty rate in an county. On the other hand, the rate of businesses related to education, health care, transportation, recreation, and entertainment all contribute to a lower poverty rate in a county.

One major limitation of this model is that we removed 1262 counties from the dataset because they had missing values for some of the predictors. There is a possibility that these counties share some characteristics, meaning that some important information from the dataset could be lost. Another limitation is that this model cannot be applied outside of the United States because of very different economic situations and policies in other regions.

A possible use case for this model would be evaluating possible courses of actions to alleviate poverty in an area. In conclusion, this project has provided insight into the causes of poverty, and can help us address this complex social issue and advance the common good.

References

"Environmental Quality Index Results." EPA, 22 July 2013, edg.epa.gov/data/public/ORD/NHEERL/EQI/Eqi

"National Poverty in America Awareness Month: January 2023." Census.gov, Jan. 2023, www.census.gov/newsroom/stories/poverty-awareness-month.html.