Investigating the Relationship Between Mental Health and Economic Inequality

Team 6: Mark Febrizio, Shumel Siraj, Alex Thiersch, and Xuan Zou

The George Washington University

DATS 6101: Introduction to Data Science

Final Project

May 1, 2022

Introduction

Original Research Topic

SMART Question:

- *What is the relationship between mental health and economic inequality within the United States from 2016 to 2021?*

- Null Hypothesis: There is no relationship between between mental health and economic inequality that varies by geography.

- Alternative Hypothesis: There is a relationship between between mental health and economic inequality that varies by geography.

Background Information:

- Many scholars have studied the relationship between income inequality and human health, less research has focused on income inequality and people's mental health.

- An initial literature review suggests that the relationship between income inequality and mental health has not been conclusively answered (Tibber et al. 2022). Many articles use different measures and study different geographies.

- Therefore, this issue is worthy of further study

New Research Topic

New SMART Questions:

- *Is the relationship between mental health and income inequality across the U.S. counties robust when including other economic variables?*
- *Does the relationship between mental health and income inequality differ across regions of the U.S.?*
- *Does the relationship between mental health and income inequality on the way mental health is measured?*

Background Information:

- The following SMART questions were motivated by our original EDA.
- Previous EDA revealed that the relationship between mental health and income inequality has statistically significant correlations and could differ across regions.

Description of the Dataset

Dataset:

- Source: Robert Wood Johnson Foundation (RWJF)
- County Health Rankings & Roadmaps (2016 – 2021)
  - *Consists of economic and health data for all states at the county level in the United States*

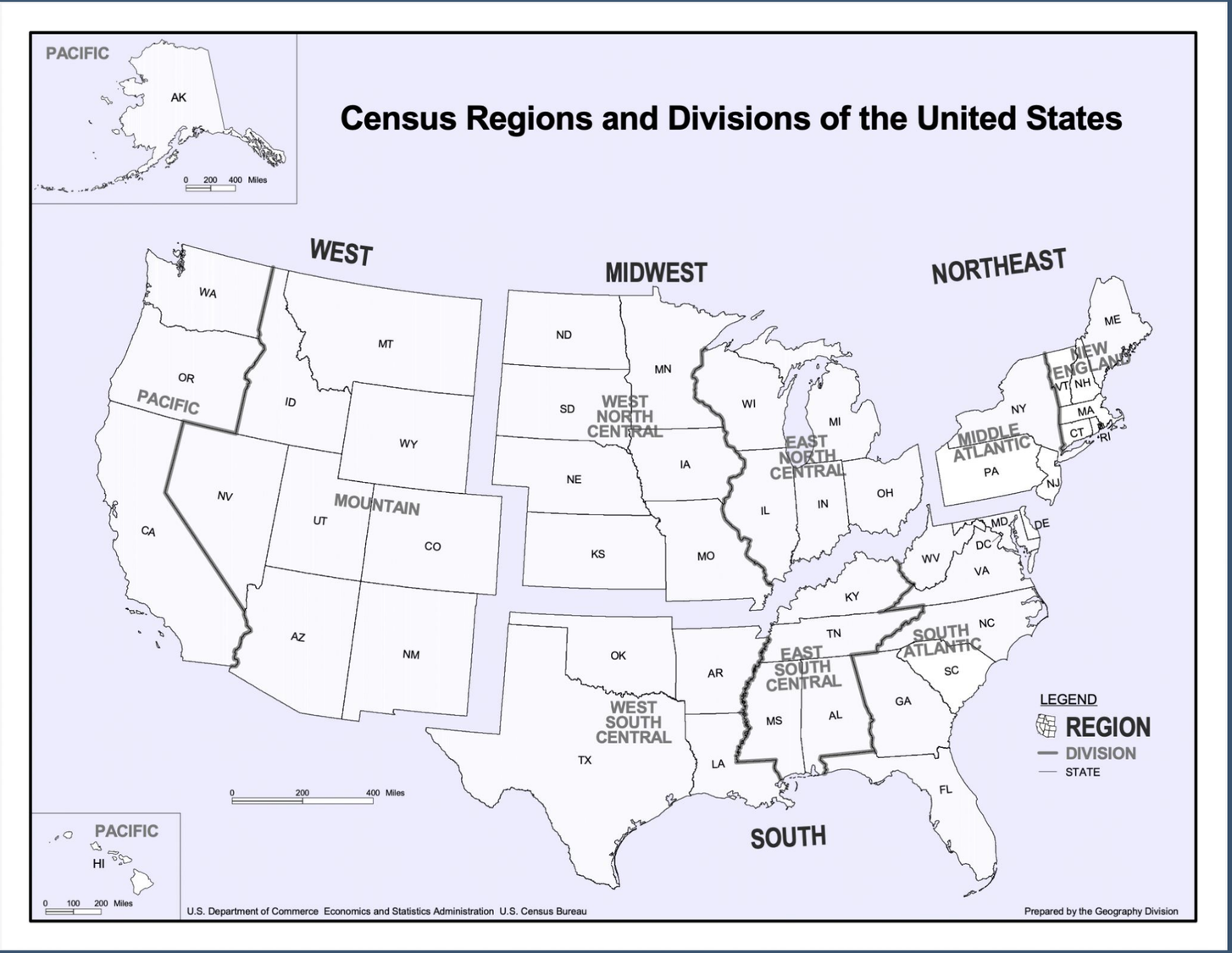
Variables:

- Independent Variables:
  - **inequality** = *Household Income Ratio*
  - **median_inc** = *Median Household Income*
  - **hs_grad** = *Percentage of high school graduates*
  - **college** = *Percentage of adults with some college education*
  - **unempl** = *Percentage of unemployment*
  - **child_poverty** = *Percentage of child poverty*
  - **single_parent** = *Percentage of single parent households*
  - **severe_housing** = *Percentage of households with severe housing problems*
  - **food_index** = *Food environment index*
  - **mh_providers** = *Mental health provider rate*
  - **pop_provider_ratio** = *Mental health provider ratio*
  - **region** = *Midwest, Northeast, South, and West*
  - **year** = *2016 - 2020*
- Dependent Variables:
  - **mental_health_days** = *Number of mentally unhealthy days*
  - **mental_distress_rate** = *Percentage frequent mental distress*

Regional Categories

Four Regional Categories:

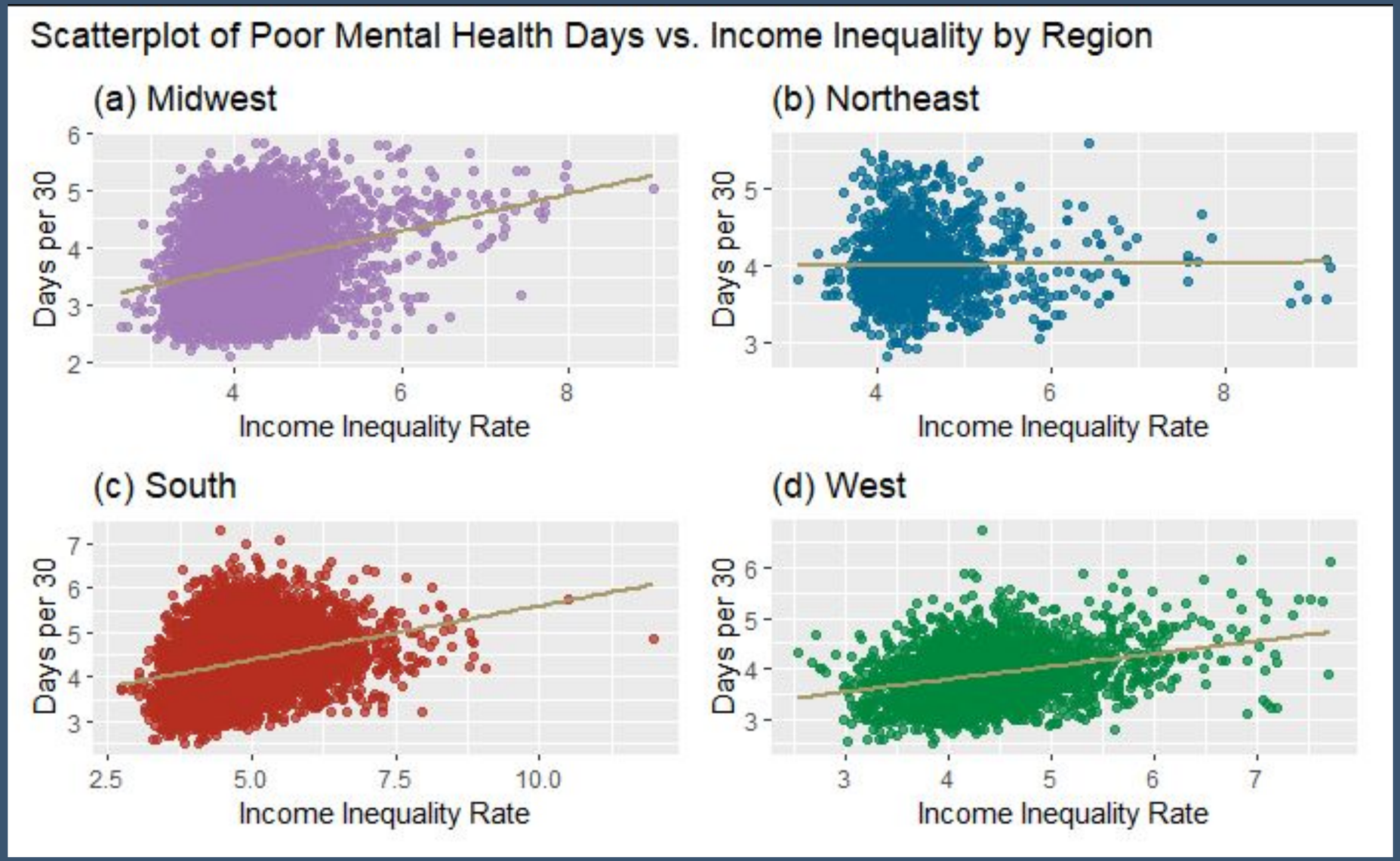
- Midwest
- Northeast
- South
- West



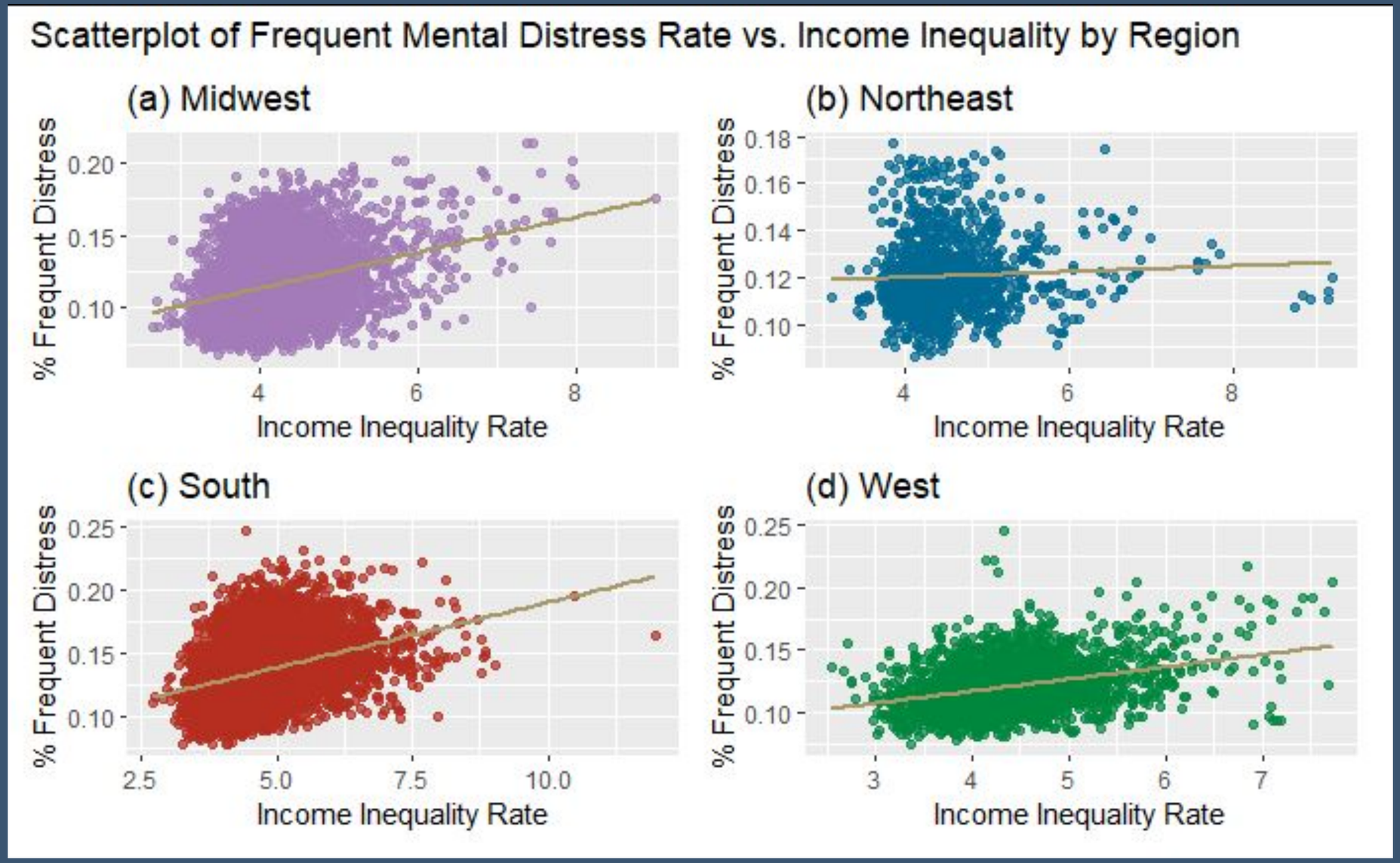
Source: U.S. Census Bureau

Summary of EDA

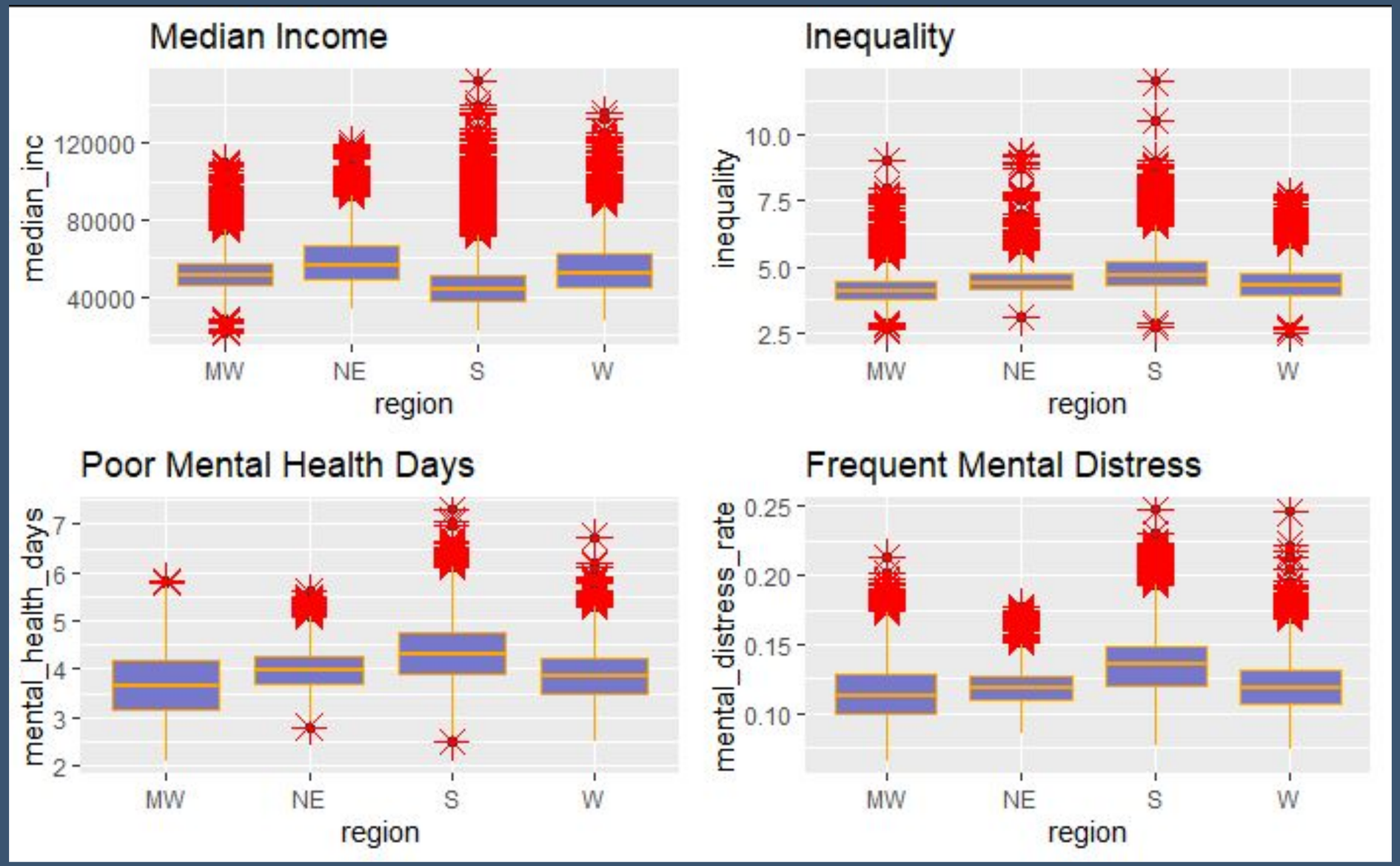
Scatterplots: Poor Mental Health Days



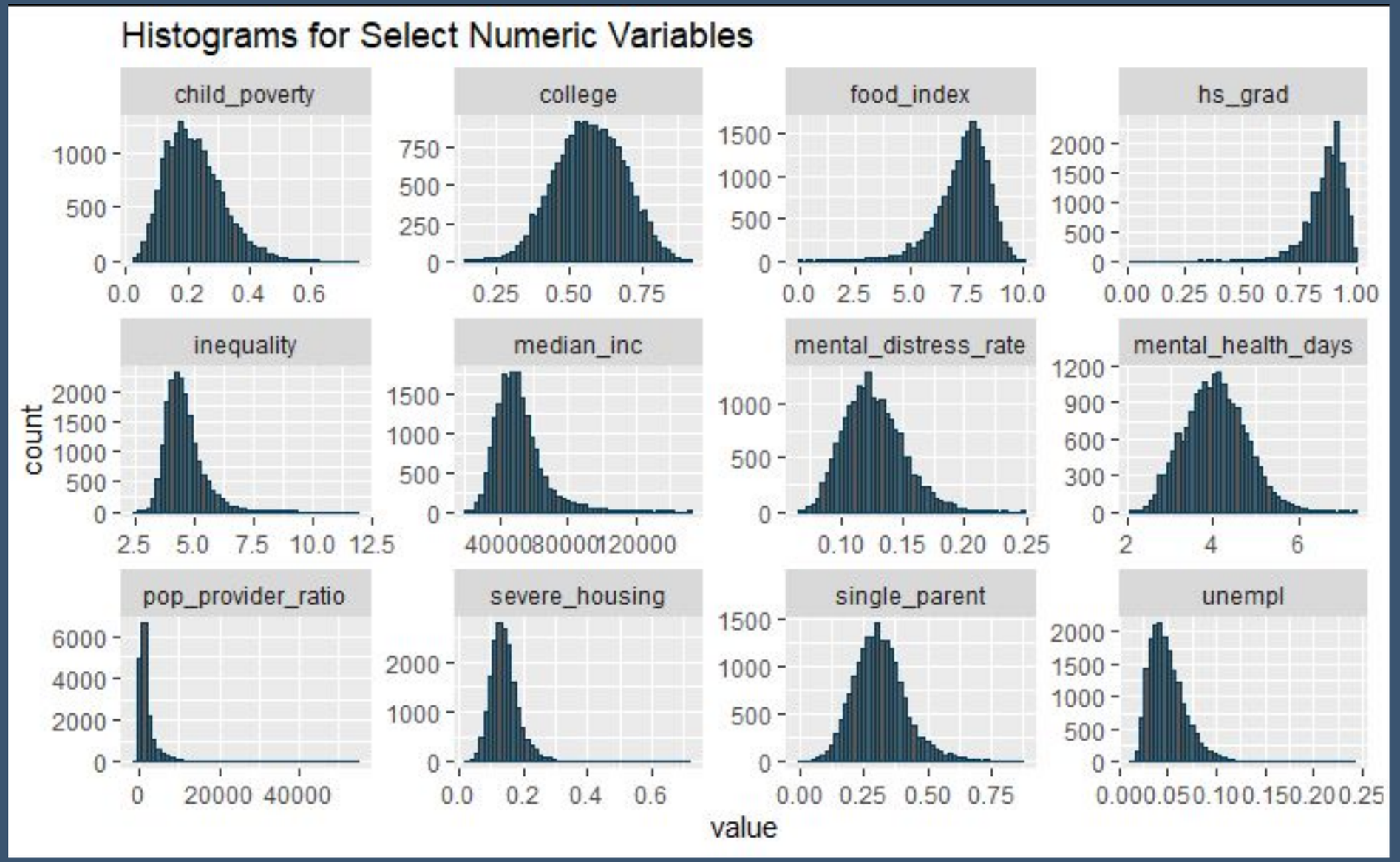
Scatterplots: Frequent Mental Distress Rate



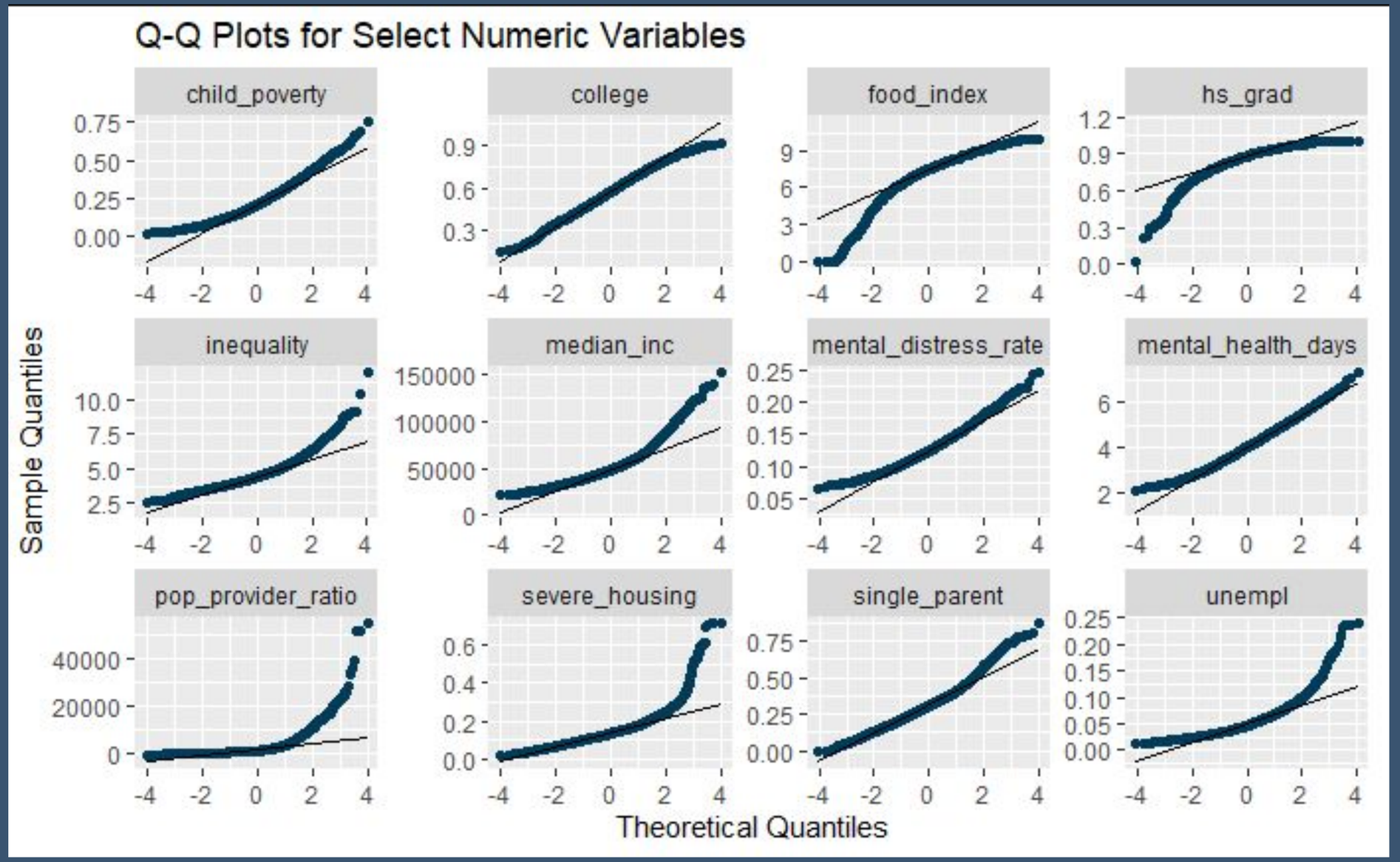
Boxplots


Median Income
median_inc
120000
80000
40000
MW
NE
S
W
region
Inequality
inequality
10.0
7.5
5.0
2.5
MW
NE
S
W
region
Poor Mental Health Days
mental_health_days
7
6
5
4
3
2
MW
NE
S
W
region
Frequent Mental Distress
mental_distress_rate
0.25
0.20
0.15
0.10
MW
NE
S
W
region

Histograms

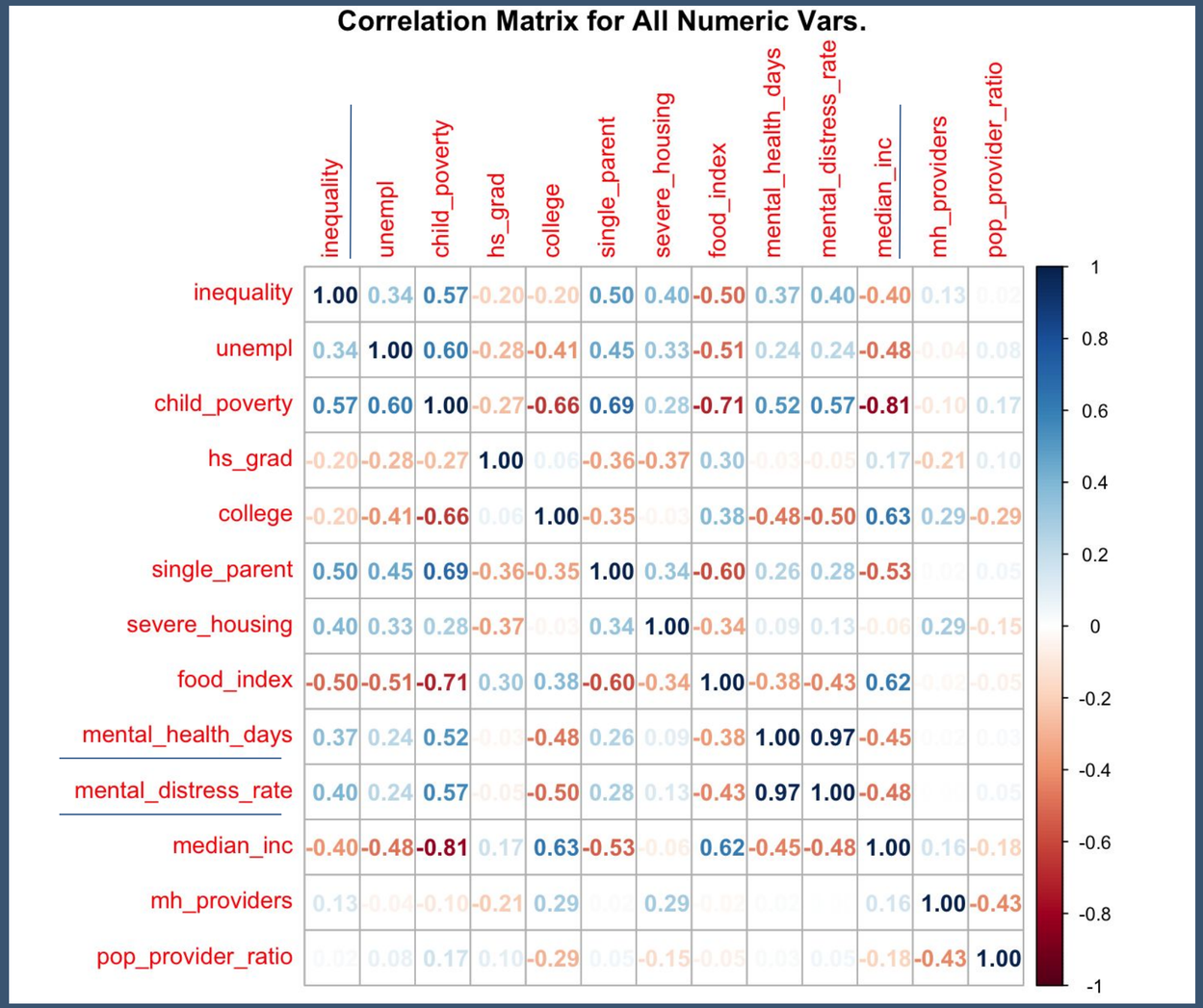

Histograms for Select Numeric Variables
child_poverty
college
food_index
hs_grad
inequality
median_inc
mental_distress_rate
mental_health_days
pop_provider_ratio
severe_housing
single_parent
unempl
count
value

Normal Q-Q Plots



Correlation Matrix: All Numeric Variables

- Correlation coefficient between *mental_health_days* and *inequality* is 0.37.
- Correlation coefficient between *mental_distress_rate* and *inequality* is 0.40
- Correlation coefficient between *mental_health_days* and *median_inc* is -0.45
- Correlation coefficient between *mental_distress_rate* and *median_inc* is -0.48



Model Testing

Model Testing/Development Roadmap

Models Used:

- Linear Regression
- Linear Regression Feature Selection
- Regression Trees
- Random Forests

Linear Models

Linear Models: Poor Mental Health Days

Model 1: base specification with region factor var

Model 2: adds year factor var

Model 3: adds economic var

Model 4: adds demographic var

	1	2	3	4
(Intercept)	2.68 ***	2.32 ***	3.94 ***	5.86 ***
	(0.03)	(0.03)	(0.08)	(0.32)
inequality	0.25 ***	0.25 ***	0.05 ***	0.07 ***
	(0.01)	(0.01)	(0.01)	(0.01)
Northeast	0.21 ***	0.21 ***	0.28 ***	0.25 ***
	(0.02)	(0.02)	(0.01)	(0.01)
South	0.47 ***	0.47 ***	0.27 ***	0.32 ***
	(0.01)	(0.01)	(0.01)	(0.01)
West	0.12 ***	0.12 ***	0.06 ***	0.19 ***
	(0.01)	(0.01)	(0.01)	(0.01)
college			-0.90 ***	-1.64 ***
			(0.04)	(0.04)
hs_grad			0.27 ***	0.15 ***
			(0.05)	(0.04)
unempl			6.49 ***	5.19 ***
			(0.23)	(0.21)
child_poverty			0.76 ***	1.34 ***
			(0.09)	(0.08)
single_parent			-0.13 *	0.39 ***
			(0.05)	(0.05)
severe_housing			-0.38 ***	0.39 ***
			(0.09)	(0.09)
food_index			-0.02 ***	-0.04 ***
			(0.00)	(0.00)
median_inc			-0.00 ***	-0.00 ***
			(0.00)	(0.00)
pop_provider_ratio			-0.00 ***	-0.00 ***
			(0.00)	(0.00)
pop				0.00 ***
				(0.00)
Num.obs	18465	18465	15741	15728
R2	0.23	0.45	0.67	0.74
adj.R2	0.23	0.45	0.67	0.74

*** p < 0.001; ** p < 0.01; * p < 0.05.

Linear Models: Poor Mental Health Days

Base specification for regional models (Model 1):

- Midwest: mental_health_days = 2.676 + 0.247 * inequality
- Northeast: mental_health_days = 2.676 + 0.212 + 0.247 * inequality
  - new intercept: 2.888
- South: mental_health_days = 2.676 + 0.472 + 0.247 * inequality
  - new intercept: 3.148
- West: mental_health_days = 2.676 + 0.121 + 0.247 * inequality
  - new intercept: 2.797

Linear Models: Poor Mental Health Days

Model 1: base specification with region factor var

Model 5: considers VIFs

Model 6: adds interaction terms

Model 7: adds interaction terms to Model 5

	1	5	6	7
(Intercept)	2.68 ***	2.39 ***	2.36 ***	2.13 ***
	(0.03)	(0.09)	(0.06)	(0.10)
inequality	0.25 ***	0.12 ***	0.32 ***	0.18 ***
	(0.01)	(0.01)	(0.01)	(0.01)
Northeast	0.21 ***	0.18 ***	1.61 ***	0.45 ***
	(0.02)	(0.01)	(0.13)	(0.08)
South	0.47 ***	0.32 ***	0.80 ***	0.73 ***
	(0.01)	(0.01)	(0.07)	(0.05)
West	0.12 ***	0.14 ***	0.41 ***	0.19 **
	(0.01)	(0.01)	(0.10)	(0.06)
inequality:NE			-0.31 ***	-0.06 ***
			(0.03)	(0.02)
inequality:SO			-0.08 ***	-0.09 ***
			(0.02)	(0.01)
inequality:WE			-0.07 **	-0.01
			(0.02)	(0.01)
college		-2.19 ***		-2.19 ***
		(0.04)		(0.04)
hs_grad		-0.04		-0.03
		(0.04)		(0.04)
unempl		6.37 ***		6.39 ***
		(0.21)		(0.21)
single_parent		0.92 ***		0.92 ***
		(0.05)		(0.05)
severe_housing		0.47 ***		0.44 ***
		(0.09)		(0.09)
food_index		-0.10 ***		-0.10 ***
		(0.00)		(0.00)
Num.obs	18465	15728	18465	15728
R2	0.23	0.72	0.23	0.72
adj.R2	0.23	0.72	0.23	0.72

*** p < 0.001; ** p < 0.01; * p < 0.05.

Linear Models: Poor Mental Health Days

Regional models with controls and interaction terms (Model 7):

- Midwest: mental_health_days = 2.13 + 0.182 * inequality + Vc
- Northeast: mental_health_days = 2.13 + 0.448 + (0.182 - 0.0616) * inequality + Vc
  - new intercept: 2.578
  - inequality coef: 0.1204
- South: mental_health_days = 2.13 + 0.734 + (0.182 - 0.0944) * inequality + Vc
  - new intercept: 2.864
  - inequality coef: 0.0876
- West: mental_health_days = 2.13 + 0.186 + (0.182 - 0.0121) * inequality + Vc
  - new intercept: 2.316
  - inequality coef: 0.1699

Linear Models: Frequent Mental Distress Rates

Model 1: base specification with region factor var

Model 2: adds year factor var

Model 3: adds economic var

Model 4: adds demographic var

	1	2	3	4
(Intercept)	0.07 ***	0.06 ***	0.12 ***	0.16 ***
	(0.00)	(0.00)	(0.00)	(0.01)
inequality	0.01 ***	0.01 ***	0.00 ***	0.00 ***
	(0.00)	(0.00)	(0.00)	(0.00)
Northeast	0.00	0.00 *	0.00 ***	0.00 ***
	(0.00)	(0.00)	(0.00)	(0.00)
South	0.01 ***	0.01 ***	0.01 ***	0.01 ***
	(0.00)	(0.00)	(0.00)	(0.00)
West	0.00 ***	0.00 ***	-0.00	0.00 ***
	(0.00)	(0.00)	(0.00)	(0.00)
college			-0.03 ***	-0.05 ***
			(0.00)	(0.00)
hs_grad			0.01 ***	0.01 ***
			(0.00)	(0.00)
unempl			0.18 ***	0.14 ***
			(0.01)	(0.01)
child_poverty			0.04 ***	0.05 ***
			(0.00)	(0.00)
single_parent			-0.00 **	0.01 ***
			(0.00)	(0.00)
severe_housing			0.01 *	0.02 ***
			(0.00)	(0.00)
food_index			-0.00 ***	-0.00 ***
			(0.00)	(0.00)
median_inc			-0.00 ***	-0.00 ***
			(0.00)	(0.00)
pop_provider_ratio			-0.00 ***	-0.00 ***
			(0.00)	(0.00)
pop				0.00 ***
				(0.00)
Num.obs	18465	18465	15741	15728
R2	0.24	0.52	0.77	0.82
adj.R2	0.24	0.52	0.77	0.82

*** p < 0.001; ** p < 0.01; * p < 0.05.

Linear Models: Frequent Mental Distress Rates

Regional models (base specification):

- Midwest: mental_distress_rate = 0.073 + 0.010 * inequality
- Northeast: mental_distress_rate = 0.073 + 0.001 + 0.010 * inequality
- South: mental_distress_rate = 0.073 + 0.014 + 0.010 * inequality
- West: mental_distress_rate = 0.073 + 0.002 + 0.010 * inequality

Linear Models: Frequent Mental Distress Rates

Model 1: base specification with region factor var

Model 5: considers VIFs

Model 6: adds interaction terms

Model 7: adds interaction terms to Model 5

	1	5	6	7
(Intercept)	0.07 ***	0.06 ***	0.06 ***	0.04 ***
	(0.00)	(0.00)	(0.00)	(0.00)
inequality	0.01 ***	0.00 ***	0.01 ***	0.01 ***
	(0.00)	(0.00)	(0.00)	(0.00)
Northeast	0.00	0.00 ***	0.05 ***	0.02 ***
	(0.00)	(0.00)	(0.00)	(0.00)
South	0.01 ***	0.00 ***	0.02 ***	0.02 ***
	(0.00)	(0.00)	(0.00)	(0.00)
West	0.00 ***	-0.00 *	0.01 ***	0.01 ***
	(0.00)	(0.00)	(0.00)	(0.00)
inequality:NE			-0.01 ***	-0.00 ***
			(0.00)	(0.00)
inequality:SO			-0.00 ***	-0.00 ***
			(0.00)	(0.00)
inequality:WE			-0.00 ***	-0.00 ***
			(0.00)	(0.00)
college		-0.07 ***		-0.07 ***
		(0.00)		(0.00)
hs_grad		0.00 **		0.00 **
		(0.00)		(0.00)
unempl		0.19 ***		0.19 ***
		(0.01)		(0.01)
single_parent		0.00 **		0.00 **
		(0.00)		(0.00)
severe_housing		-0.01 **		-0.01 **
		(0.00)		(0.00)
food_index		-0.00 ***		-0.00 ***
		(0.00)		(0.00)
Num.obs	18465	15728	18465	15728
R2	0.24	0.77	0.25	0.77
adj.R2	0.24	0.76	0.25	0.77

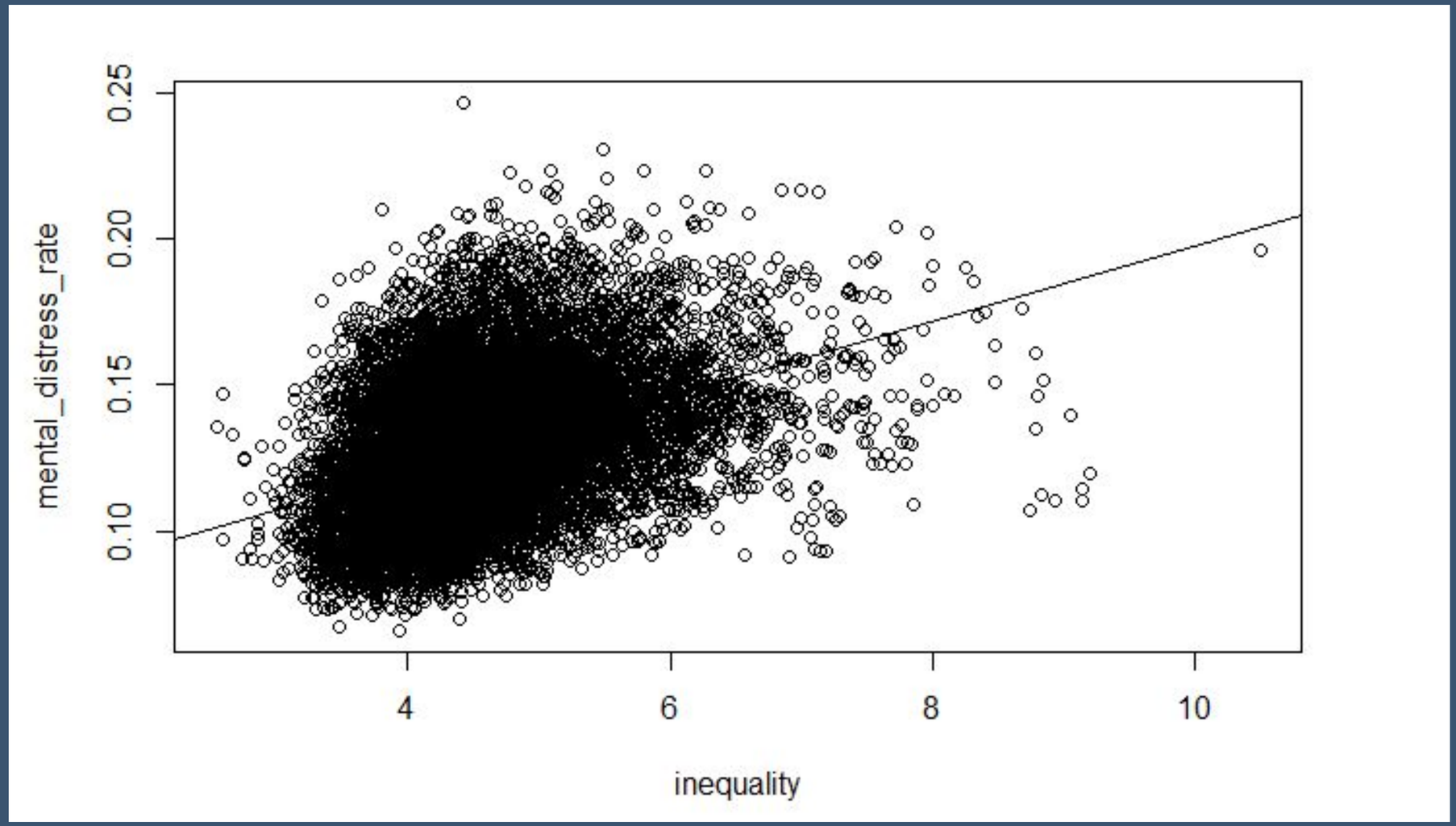
*** p < 0.001; ** p < 0.01; * p < 0.05.

Linear Models: Frequent Mental Distress Rates

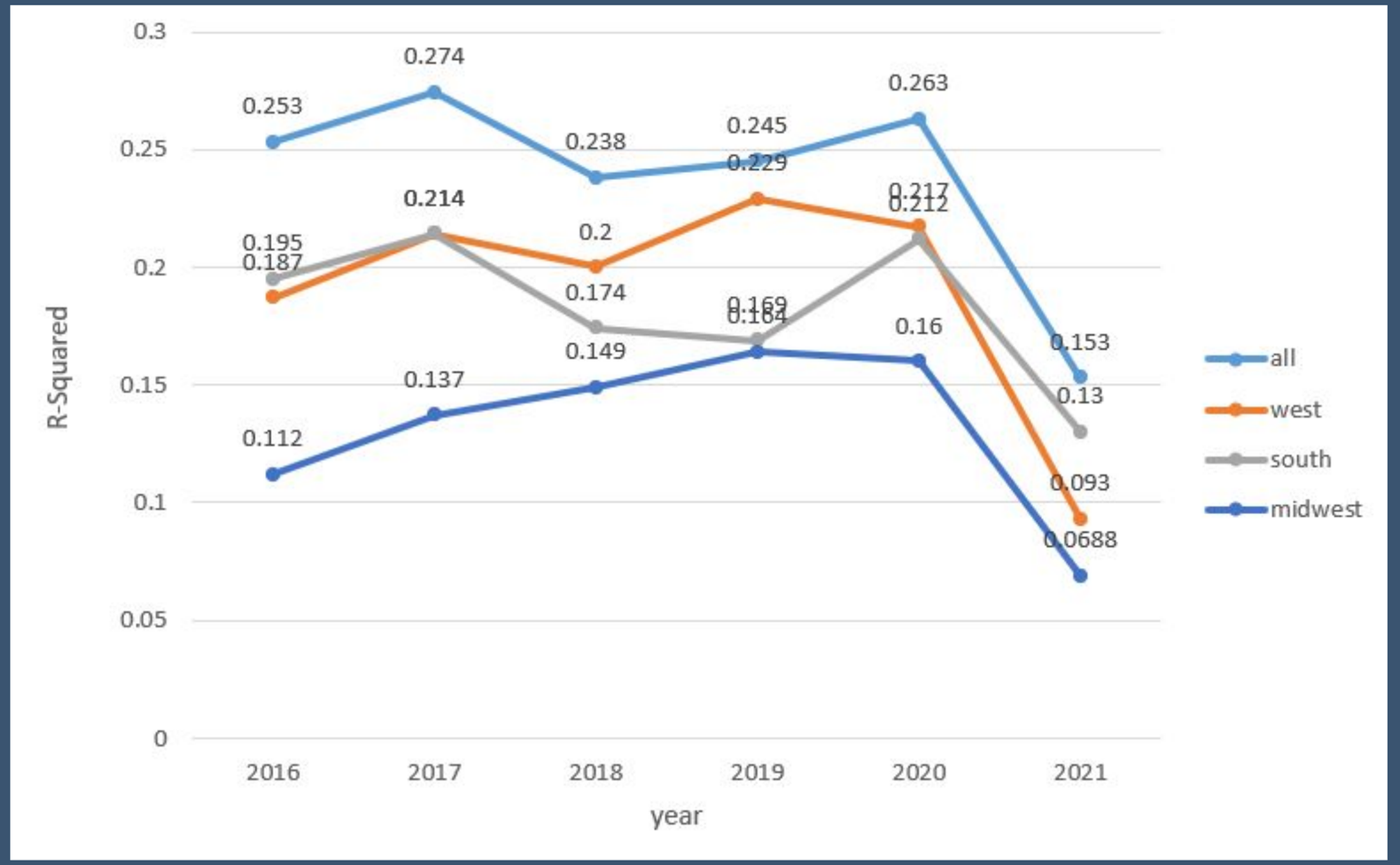
Regional models with controls and interaction terms:

- **Midwest: mental_distress_rate = 0.0445 + 0.00696 * inequality + Vc**
- **Northeast: mental_distress_rate = 0.0445 + 0.0165 + (0.00696 - 0.00348) * inequality + Vc**
- **South: mental_distress_rate = 0.0445 + 0.0198 + (0.00696 - 0.00350) * inequality + Vc**
- **West: mental_distress_rate = 0.0445 + 0.0119 + (0.00696 - 0.00300) * inequality + Vc**

Linear Models: Frequent Mental Distress Rates



Linear Models: Frequent Mental Distress Rates



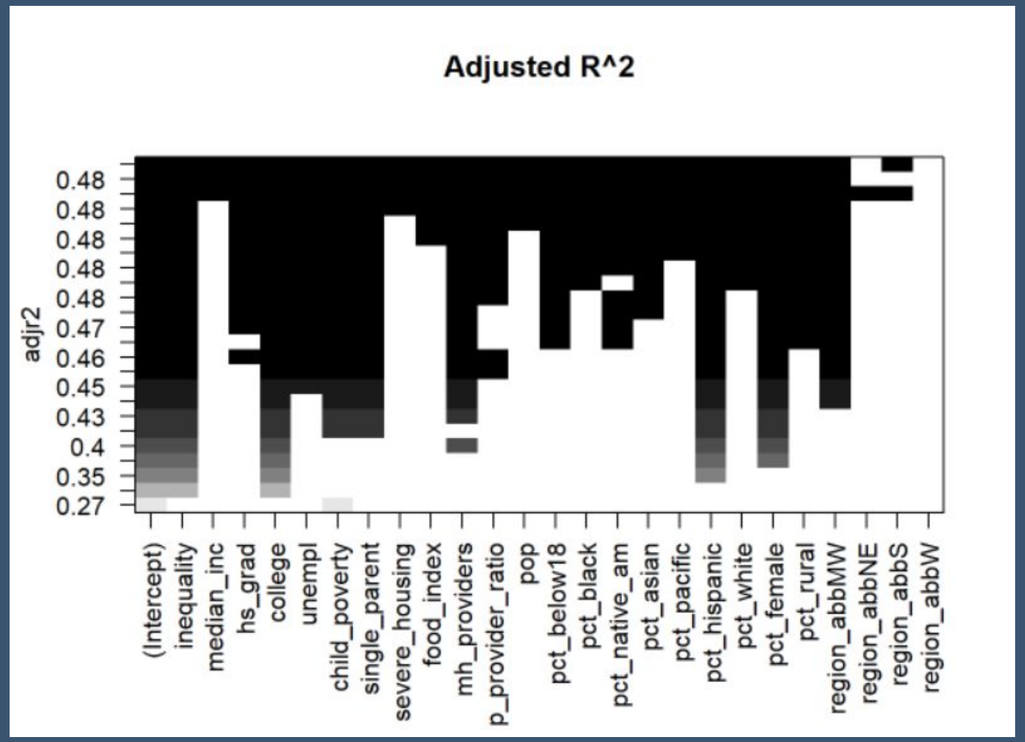
Linear Models: Feature Selection

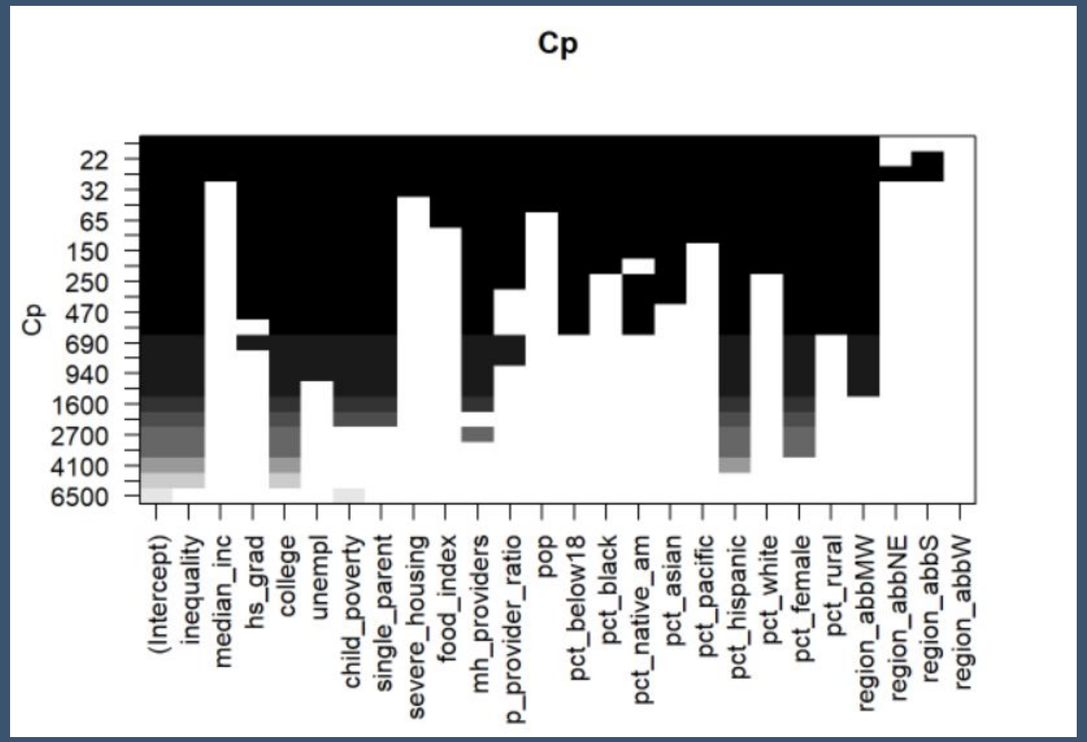
Linear Models: Poor Mental Health Days

Linear model with all the variables for taking mental health days as target variable and excluding the other target variable i.e mental distress rate.

```
##
## Call:
## lm(formula = mental_health_days ~ . - mental_distress_rate, data = rankedfs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3919 -0.3232 -0.0201  0.3089  2.2005
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.14e+00   4.51e-01   15.84  < 2e-16 ***
## inequality          1.49e-01   7.50e-03   19.87  < 2e-16 ***
## median_inc          2.26e-06   6.38e-07    3.54  0.00040 ***
## hs_grad             6.10e-01   5.95e-02   10.26  < 2e-16 ***
## college            -2.68e+00   5.87e-02  -45.71  < 2e-16 ***
## unempl             -4.32e+00   2.67e-01  -16.16  < 2e-16 ***
## child_poverty       2.88e+00   1.16e-01   24.73  < 2e-16 ***
## single_parent      -1.71e+00   6.77e-02  -25.28  < 2e-16 ***
## severe_housing     -4.23e-01   1.26e-01   -3.35  0.00081 ***
## food_index         -3.71e-02   6.10e-03   -6.08  1.2e-09 ***
## mh_providers        3.62e+01   3.09e+00   11.71  < 2e-16 ***
## pop_provider_ratio -1.56e-05   1.55e-06  -10.10  < 2e-16 ***
## pop                 6.98e-08   1.30e-08    5.36  8.4e-08 ***
## pct_below18        -1.99e+00   1.52e-01  -13.13  < 2e-16 ***
## pct_black          -5.04e+00   4.46e-01  -11.30  < 2e-16 ***
## pct_native_am      -4.20e+00   4.68e-01   -8.98  < 2e-16 ***
## pct_asian          -7.42e+00   5.12e-01  -14.50  < 2e-16 ***
## pct_pacific        -9.93e+00   1.49e+00   -6.67  2.6e-11 ***
## pct_hispanic       -5.86e+00   4.31e-01  -13.59  < 2e-16 ***
## pct_white          -5.08e+00   4.49e-01  -11.32  < 2e-16 ***
## pct_female          6.28e+00   2.26e-01   27.83  < 2e-16 ***
## pct_rural          -2.19e-01   2.02e-02  -10.86  < 2e-16 ***
## region_abbNE        1.59e-01   1.68e-02    9.44  < 2e-16 ***
## region_abbS         1.76e-01   1.15e-02   15.30  < 2e-16 ***
## region_abbW         1.58e-01   1.55e-02   10.16  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 15703 degrees of freedom
##   (2742 observations deleted due to missingness)
## Multiple R-squared:  0.485,  Adjusted R-squared:  0.485
## F-statistic:  617 on 24 and 15703 DF,  p-value: <2e-16
```

Feature Selection: Poor Mental Health Days





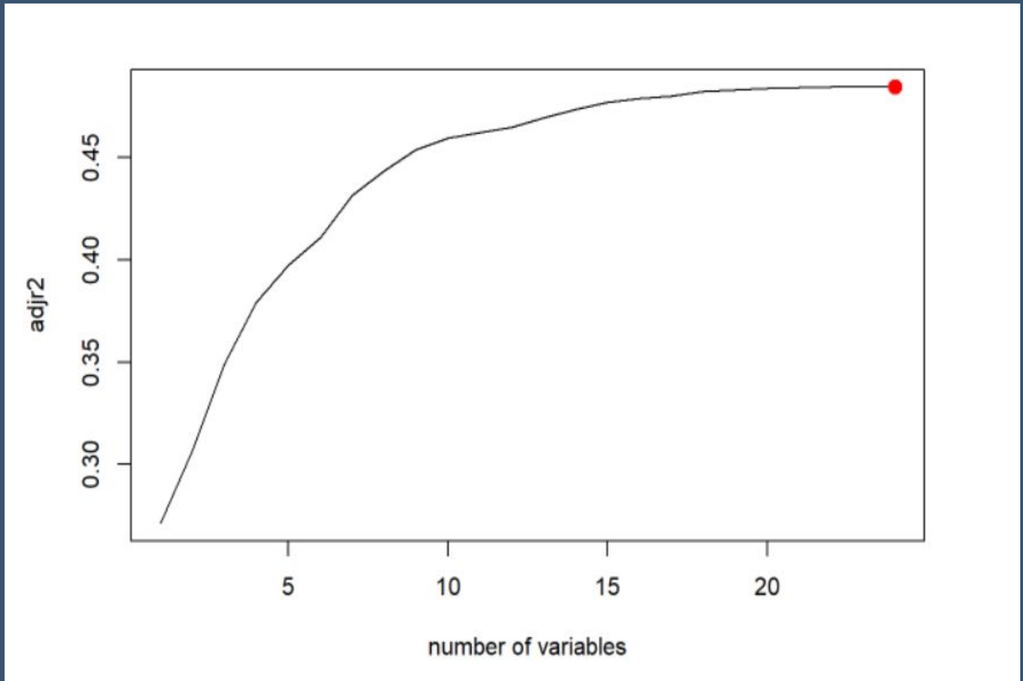
Linear Models: Poor Mental Health Days

```
   (Intercept)         inequality         median_inc            hs_grad 
          TRUE               TRUE              FALSE              FALSE 
     college              unempl      child_poverty       single_parent 
          TRUE               TRUE               TRUE               TRUE 
severe_housing         food_index      mh_providers pop_provider_ratio 
         FALSE              FALSE               TRUE               TRUE 
         pop       pct_below18          pct_black      pct_native_am 
       FALSE              FALSE              FALSE              FALSE 
  pct_asian       pct_pacific      pct_hispanic          pct_white 
       FALSE              FALSE               TRUE              FALSE 
 pct_female          pct_rural      region_abbMW      region_abbNE 
        TRUE              FALSE               TRUE              FALSE 
region_abbS       region_abbW 
      FALSE              FALSE 
```

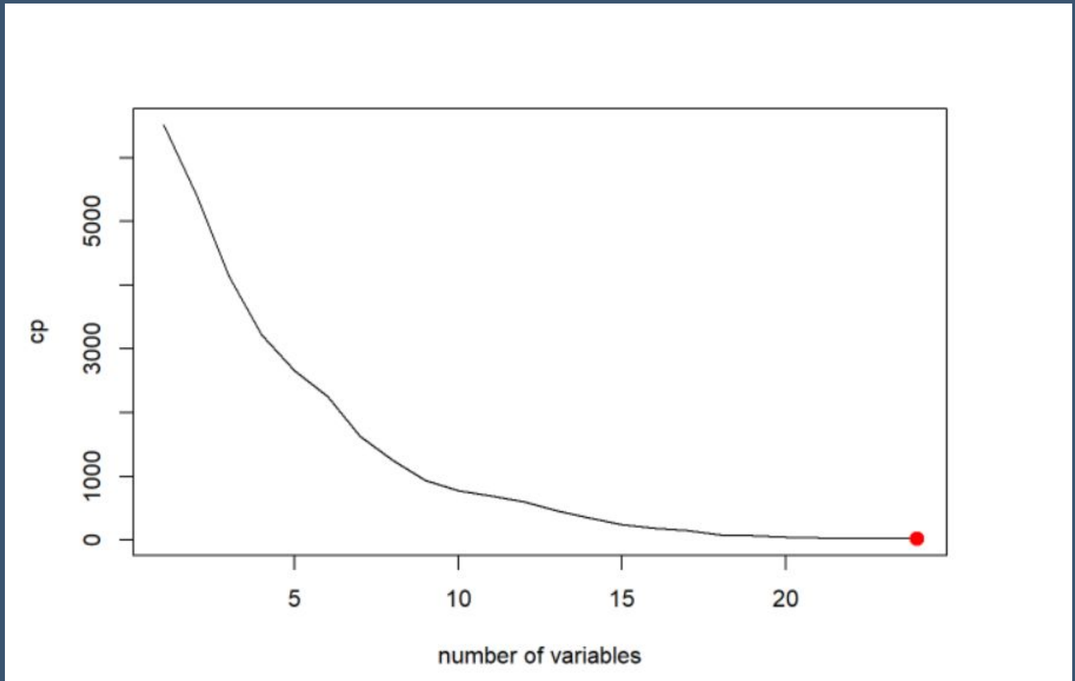
```
Call:
lm(formula = mental_health_days ~ child_poverty + inequality + 
    unempl + college + mh_providers + pop_provider_ratio + region_abb + 
    single_parent + pct_hispanic + pct_female, data = rankedfs)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9949 -0.3438 -0.0244  0.3178  2.1867 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         2.12e+00   1.02e-01   20.72   <2e-16 ***
child_poverty       2.73e+00   9.27e-02   29.40   <2e-16 ***
inequality          1.51e-01   7.28e-03   20.79   <2e-16 ***
unempl             -4.09e+00   2.64e-01  -15.50   <2e-16 ***
college            -2.47e+00   5.47e-02  -45.05   <2e-16 ***
mh_providers        5.09e+01   3.00e+00   16.98   <2e-16 ***
pop_provider_ratio -2.05e-05   1.55e-06  -13.25   <2e-16 ***
region_abbNE        1.62e-01   1.61e-02   10.02   <2e-16 ***
region_abbS         2.18e-01   1.05e-02   20.70   <2e-16 ***
region_abbW         1.33e-01   1.45e-02    9.17   <2e-16 ***
single_parent      -1.57e+00   5.65e-02  -27.81   <2e-16 ***
pct_hispanic       -1.08e+00   3.18e-02  -33.96   <2e-16 ***
pct_female          5.45e+00   2.16e-01   25.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.49 on 15715 degrees of freedom
Multiple R-squared:  0.461,     Adjusted R-squared:  0.461 
F-statistic: 1.12e+03 on 12 and 15715 DF,  p-value: <2e-16
```

Linear Models: Frequent Mental Distress Rates

The model with 24 variables has the highest adjusted R^2



The model with 24 variables has the lowest CP.



Feature Selection: Poor Mental Health Days

ANOVA Comparison between lm1 and lm2.

lm1 <- lm(rankedfs,formula=mental_health_days ~. -mental_distress_rate)

lm2 <- lm(rankedfs,formula=mental_health_days ~ child_poverty + inequality + unempl + college + mh_providers + pop_provider_ratio + unempl + single_parent + pct_hispanic + pct_female)

ANOVA comparison between two linear models for mental health days : lm1 and lm2

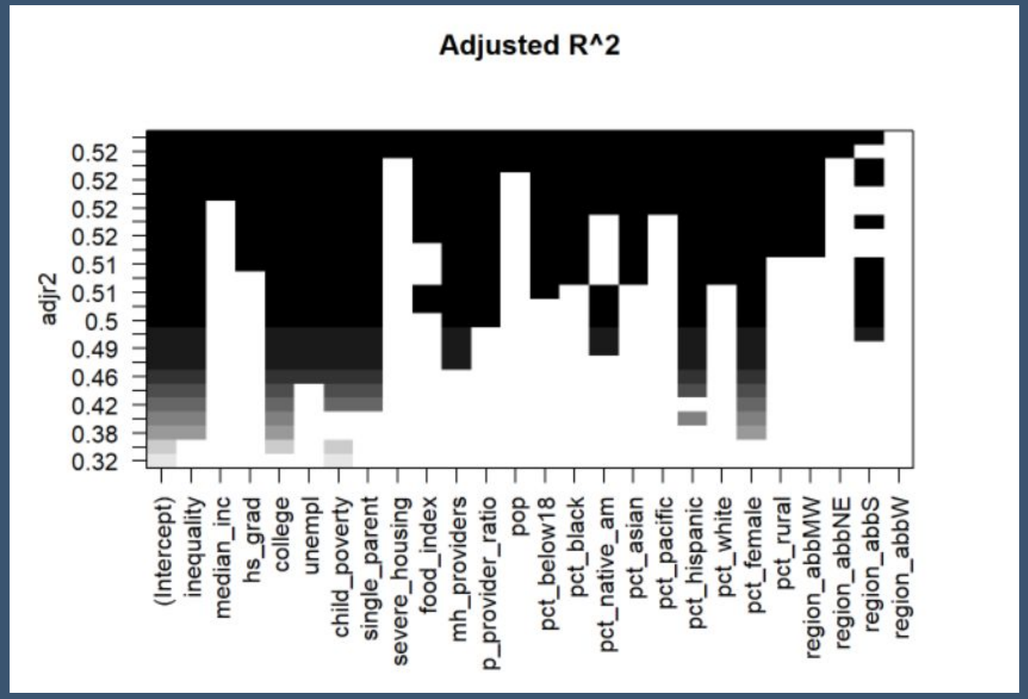
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
15703	3608	NA	NA	NA	NA
15715	3777	-12	-169	61.4	0

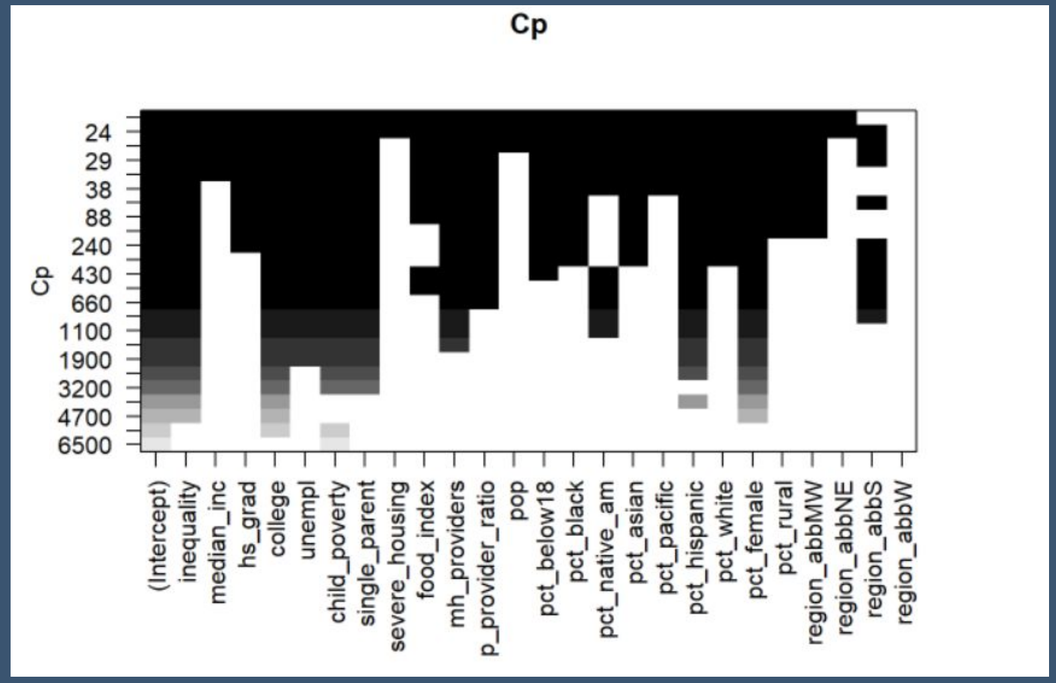
Linear Models: Frequent Mental Distress Rates

Linear model with all the variables for taking mental distress rate as target variable and excluding the other target variable i.e mental health days.

```
##
## Call:
## lm(formula = mental_distress_rate ~ . - mental_health_days, data = rankedfs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08588 -0.01023 -0.00131  0.00901  0.07827
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.11e-01   1.50e-02   14.06  < 2e-16 ***
## inequality          5.66e-03   2.50e-04   22.69  < 2e-16 ***
## median_inc          6.03e-08   2.12e-08    2.84   0.0045 **
## hs_grad             2.09e-02   1.98e-03   10.56  < 2e-16 ***
## college            -8.58e-02   1.95e-03  -43.94  < 2e-16 ***
## unempl             -2.07e-01   8.88e-03  -23.33  < 2e-16 ***
## child_poverty       1.11e-01   3.87e-03   28.67  < 2e-16 ***
## single_parent      -7.13e-02   2.25e-03  -31.66  < 2e-16 ***
## severe_housing     -6.85e-03   4.20e-03   -1.63   0.1031
## food_index         -1.67e-03   2.03e-04   -8.23  < 2e-16 ***
## mh_providers        1.11e+00   1.03e-01   10.75  < 2e-16 ***
## pop_provider_ratio -5.44e-07   5.15e-08  -10.57  < 2e-16 ***
## pop                 1.14e-09   4.33e-10    2.64   0.0083 **
## pct_below18        -6.11e-02   5.05e-03  -12.08  < 2e-16 ***
## pct_black          -1.36e-01   1.48e-02   -9.14  < 2e-16 ***
## pct_native_am      -9.82e-02   1.56e-02   -6.31  2.9e-10 ***
## pct_asian          -2.33e-01   1.70e-02  -13.69  < 2e-16 ***
## pct_pacific        -2.49e-01   4.95e-02   -5.04  4.8e-07 ***
## pct_hispanic       -1.67e-01   1.43e-02  -11.64  < 2e-16 ***
## pct_white          -1.51e-01   1.49e-02  -10.12  < 2e-16 ***
## pct_female          2.06e-01   7.51e-03   27.50  < 2e-16 ***
## pct_rural          -4.51e-03   6.71e-04   -6.72  1.9e-11 ***
## region_abbNE        1.69e-03   5.60e-04    3.02   0.0025 **
## region_abbS         3.24e-03   3.83e-04    8.47  < 2e-16 ***
## region_abbW         2.59e-03   5.16e-04    5.03  5.0e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0159 on 15703 degrees of freedom
## Multiple R-squared:  0.522,  Adjusted R-squared:  0.521
## F-statistic:  714 on 24 and 15703 DF,  p-value: <2e-16
```

Feature Selection: Frequent Mental Distress Rates





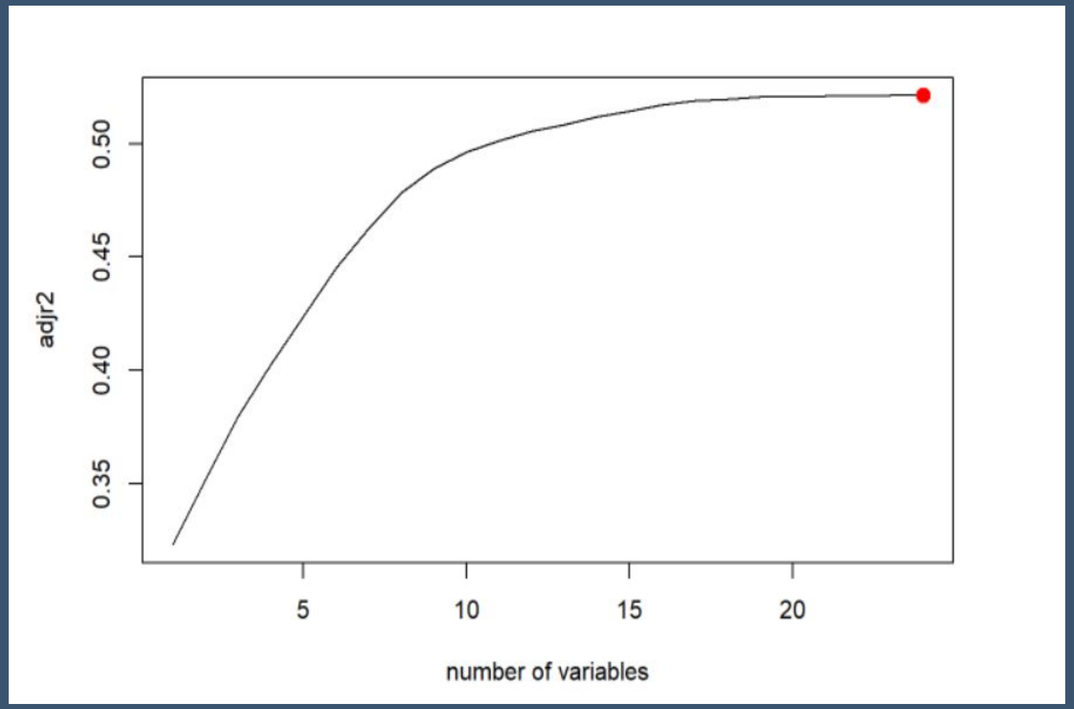
Linear Models: Frequent Mental Distress Rates

(Intercept)	inequality	median_inc	hs_grad	college	unempl	child_poverty	single_parent
TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
severe_housing	food_index	mh_providers	pop_provider_ratio	pop	pct_below18	pct_black	pct_native_am
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
pct_asian	pct_pacific	pct_hispanic	pct_white	pct_female	pct_rural	region_abbMW	region_abbNE
FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
region_abbS	region_abbW						
TRUE	FALSE						

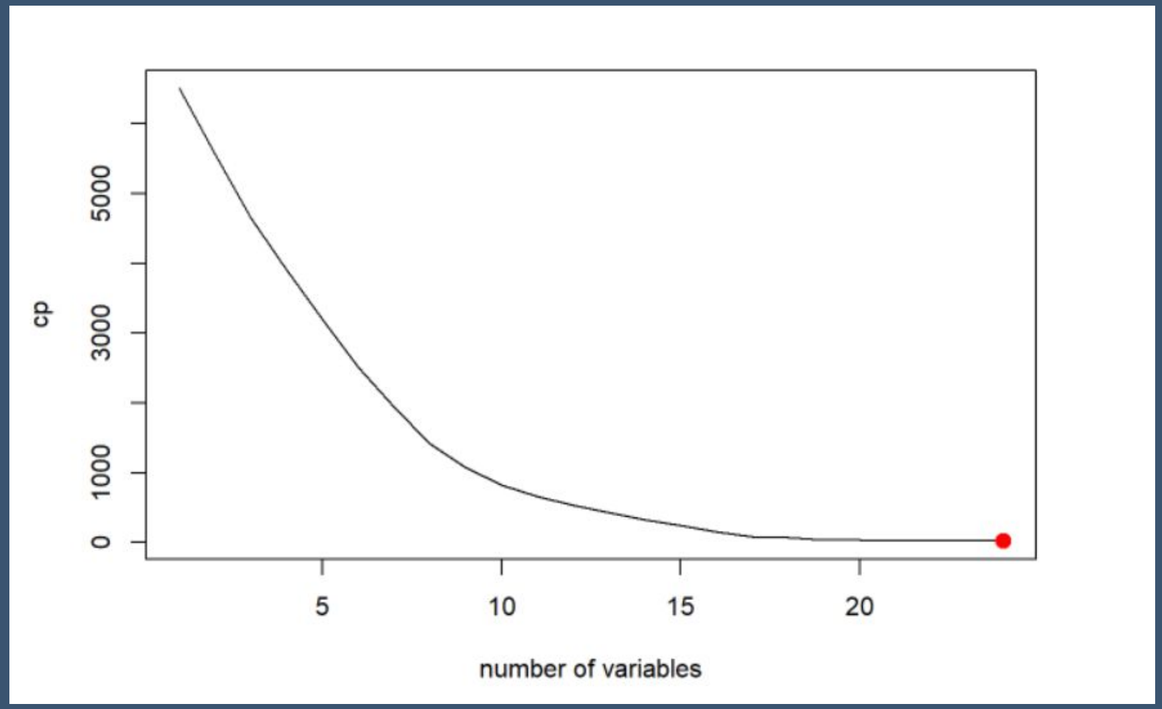
```
##
## Call:
## lm(formula = mental_distress_rate ~ inequality + college + mh_providers +
##     pct_female + child_poverty + unempl + single_parent + region_abb +
##     pct_hispanic + pct_native_am, data = rankedfs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06968 -0.01070 -0.00149  0.00914  0.07660
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.053433   0.003393   15.75  < 2e-16 ***
## inequality     0.005854   0.000243   24.10  < 2e-16 ***
## college       -0.075564   0.001817  -41.58  < 2e-16 ***
## mh_providers   1.739600   0.095220   18.27  < 2e-16 ***
## pct_female     0.179805   0.007196   24.99  < 2e-16 ***
## child_poverty  0.116025   0.003090   37.54  < 2e-16 ***
## unempl        -0.210209   0.008833  -23.80  < 2e-16 ***
## single_parent -0.057428   0.001887  -30.43  < 2e-16 ***
## region_abbNE   0.002034   0.000539    3.77  0.00016 ***
## region_abbS    0.005652   0.000351   16.10  < 2e-16 ***
## region_abbW    0.001965   0.000487    4.04  5.4e-05 ***
## pct_hispanic  -0.029361   0.001060  -27.70  < 2e-16 ***
## pct_native_am  0.041526   0.002153   19.28  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0163 on 15715 degrees of freedom
## Multiple R-squared:  0.498,  Adjusted R-squared:  0.497
## F-statistic: 1.3e+03 on 12 and 15715 DF,  p-value: <2e-16
```

Feature Selection: Frequent Mental Distress Rates

The model with 24 variables has the highest adjusted R^2



The model with 24 variables has the lowest CP.



Feature Selection: Frequent Mental Distress Rates

ANOVA Comparison between lm3 and lm4.

lm3 <- lm(rankedfs,formula=mental_distress_rate ~. -mental_health_days)

lm4 <- lm(rankedfs,formula=mental_distress_rate ~ inequality + college + mh_providers + pct_female + child_poverty + unempl + single_parent + region_abb + pct_hispanic + pct_native_am)

ANOVA comparison between two linear models for mental distress rate : lm3 and lm4

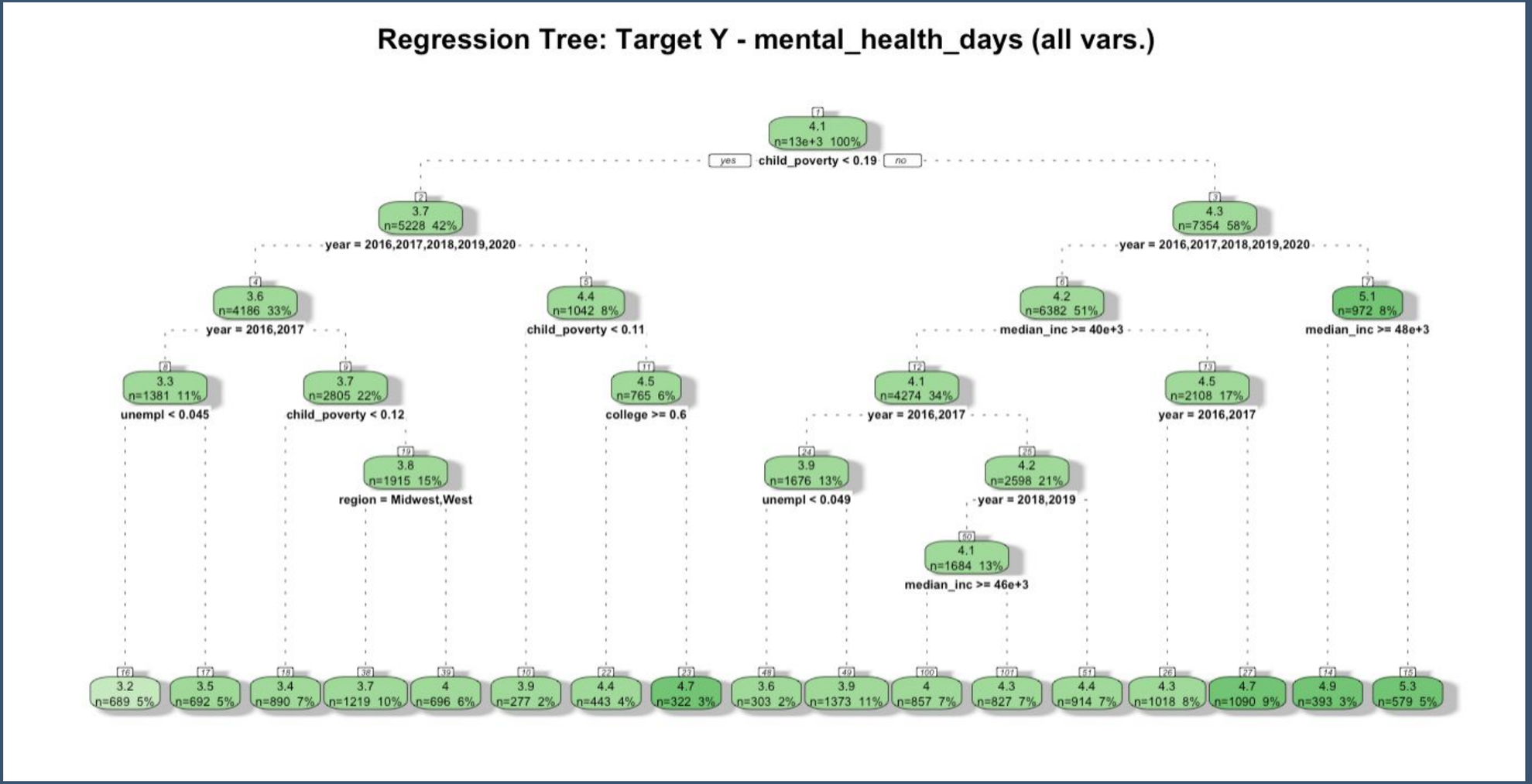
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
15703	3.99	NA	NA	NA	NA
15715	4.20	-12	-0.204	66.7	0

Regression Trees

All Models used an 80:20 Train-Test Split

Regression Trees: mental_health_days

- The tree has 17 leaf nodes



Regression Trees: mental_health_days

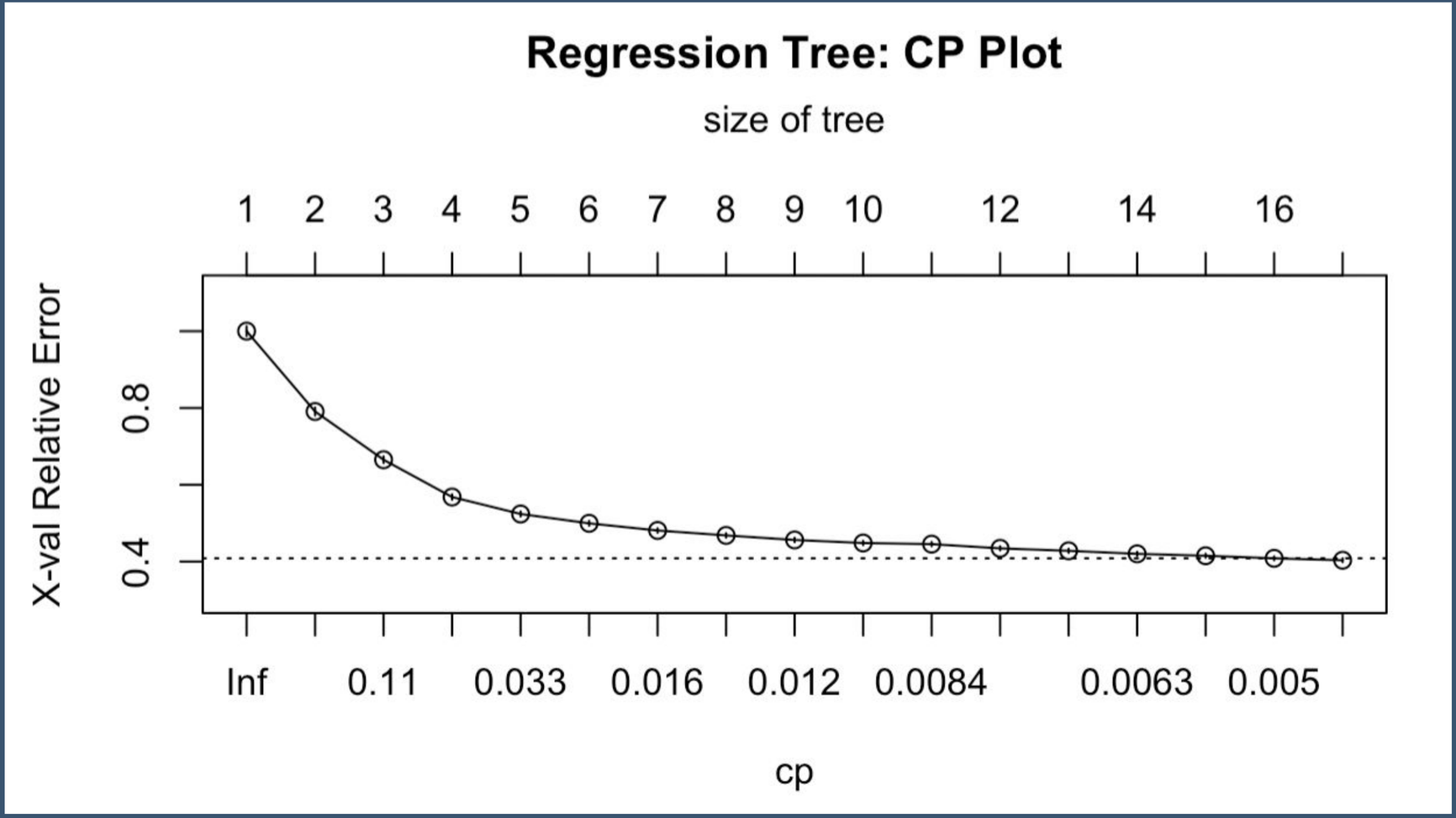
- Variable Importance Table

Variable Importance: Target Y - mental_health_days (all vars.)

	Feature Importance
year	1586.83
child_poverty	1502.05
median_inc	1173.68
food_index	615.87
unempl	588.94
college	583.42
single_parent	545.39
inequality	54.16
region	52.43
mh_providers	14.08
hs_grad	13.31
pop_provider_ratio	12.39
severe_housing	7.11

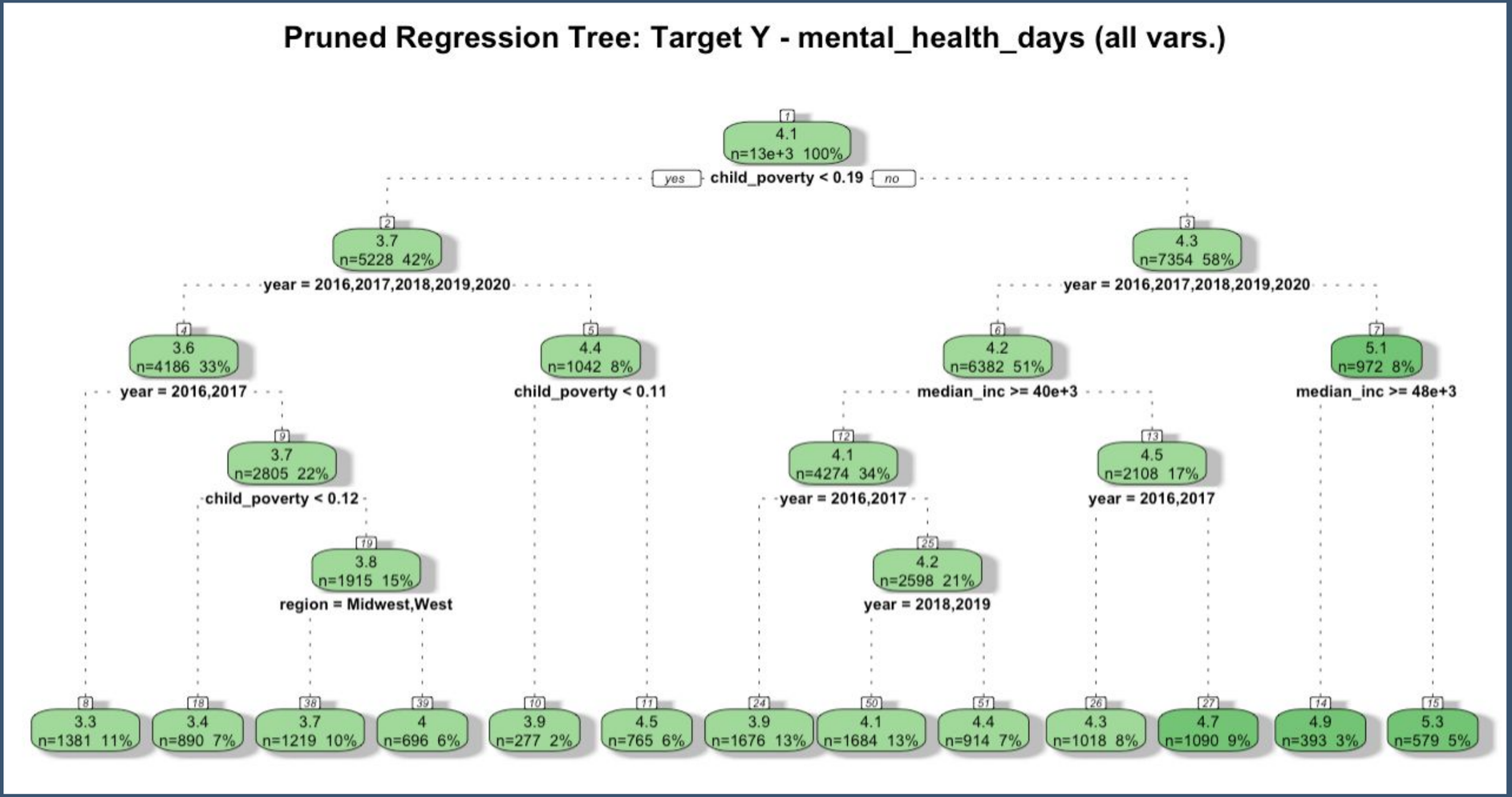
Regression Trees: mental_health_days

- Max Depth = 8
- cp = 0.004



Pruned Regression Trees: mental_health_days

- The tree has 13 leaf nodes



Pruned Regression Trees: mental_health_days

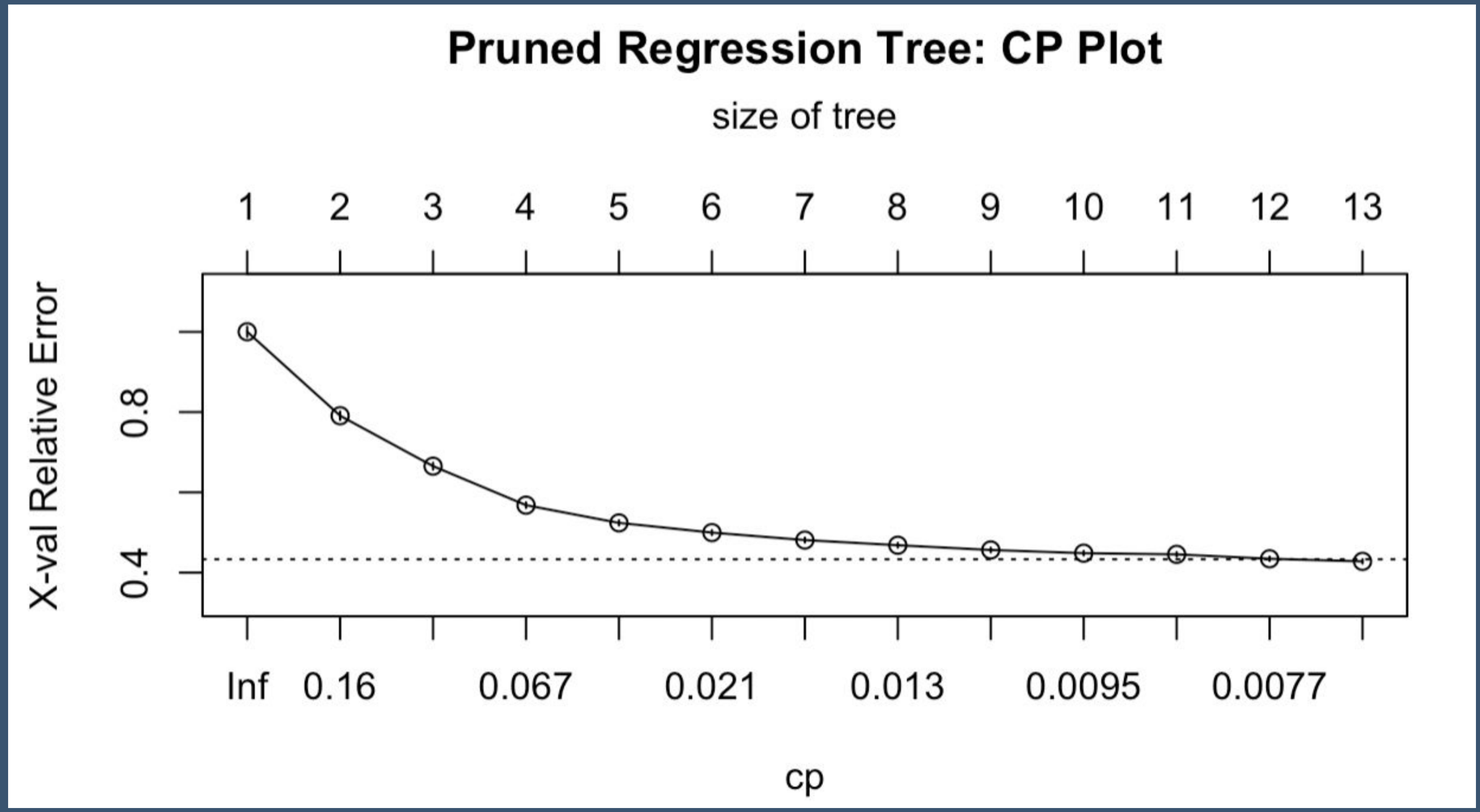
- Variable Importance Table

Pruned Variable Importance: Target Y - mental_health_days (all vars.)

	Feature Importance
year	1579.616
child_poverty	1479.220
median_inc	1143.893
food_index	611.472
college	544.814
single_parent	536.570
unempl	516.602
inequality	53.924
region	52.431
hs_grad	11.096
mh_providers	3.515
pop_provider_ratio	1.827
severe_housing	0.868

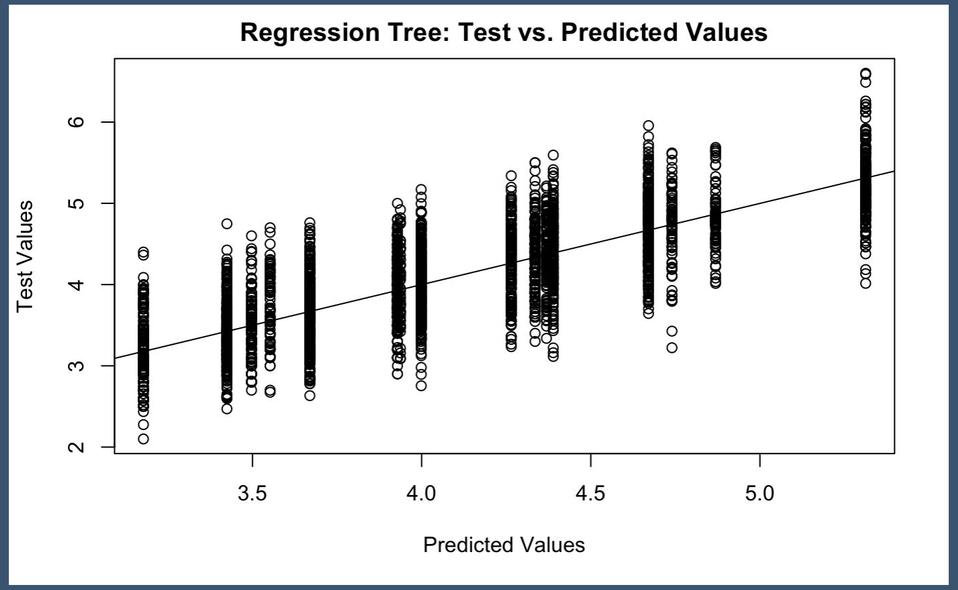
Pruned Regression Trees: mental_health_days

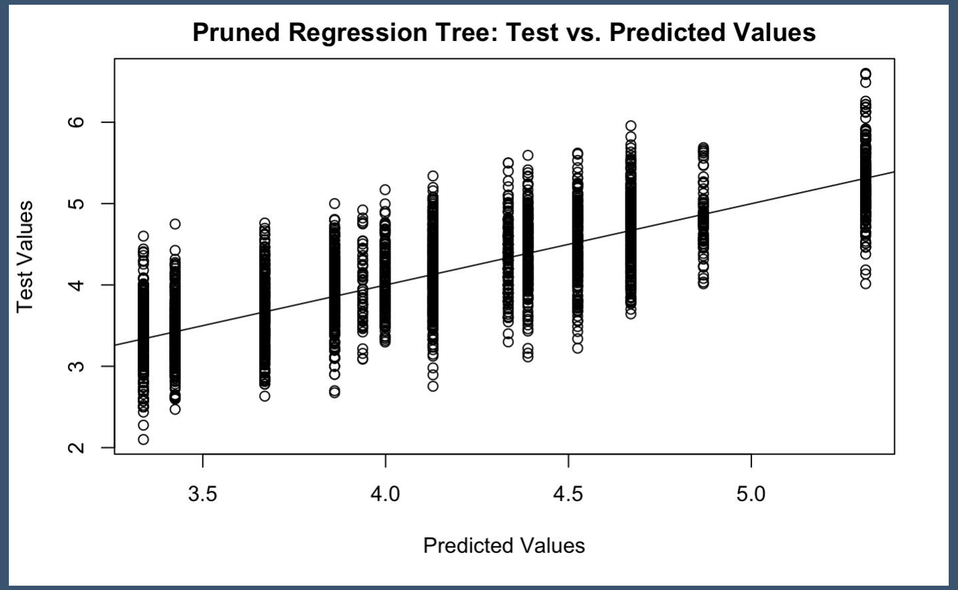
- cp = 0.0067



Reg. Tree and Prune Tree RMSE: mental_health_days

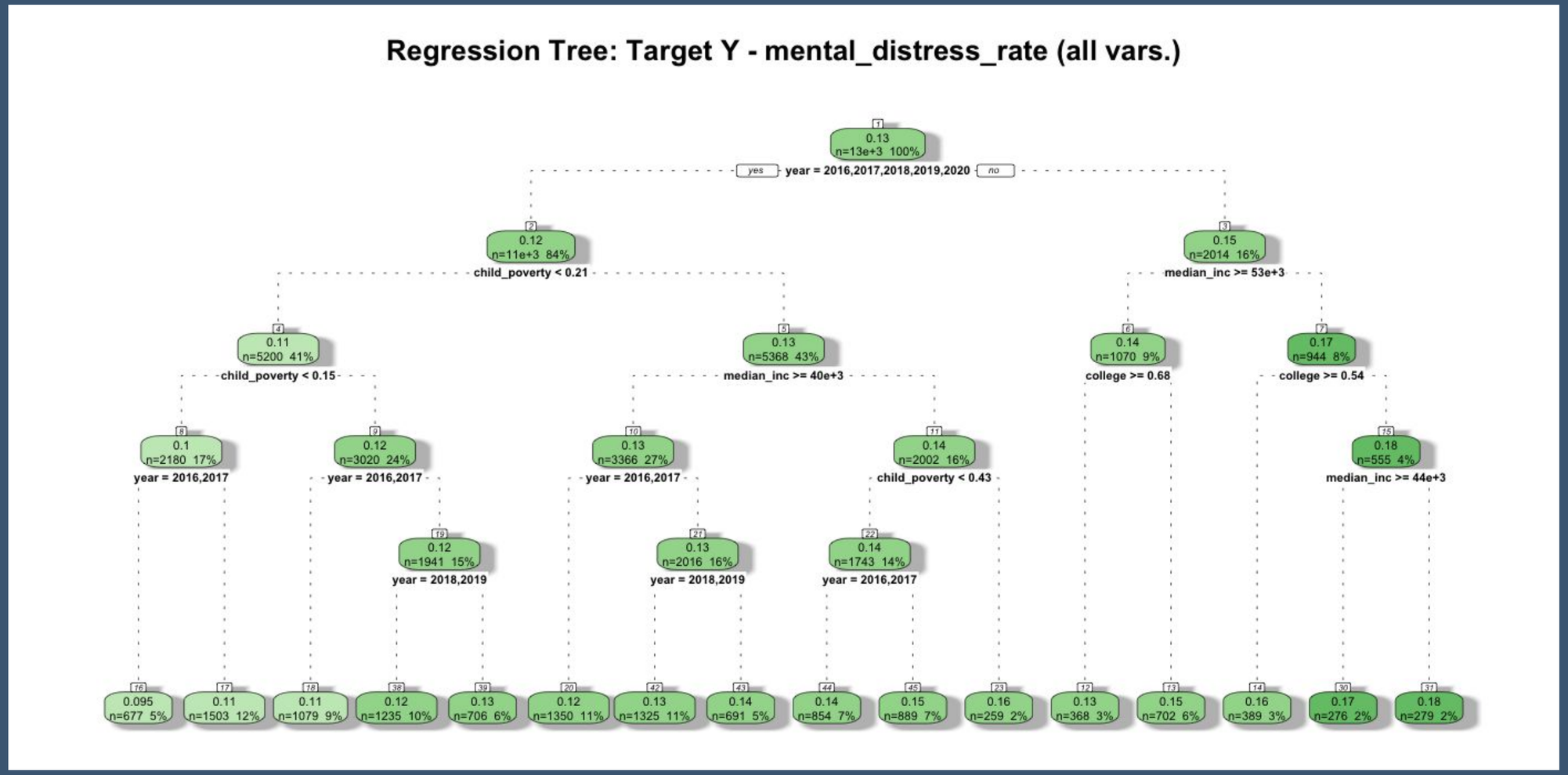
- Reg. Tree RMSE: 0.427
- Pruned Reg. Tree RMSE: 0.435





Regression Trees: mental_distress_rate

- The tree has 17 leaf nodes



Regression Trees: mental_distress_rate

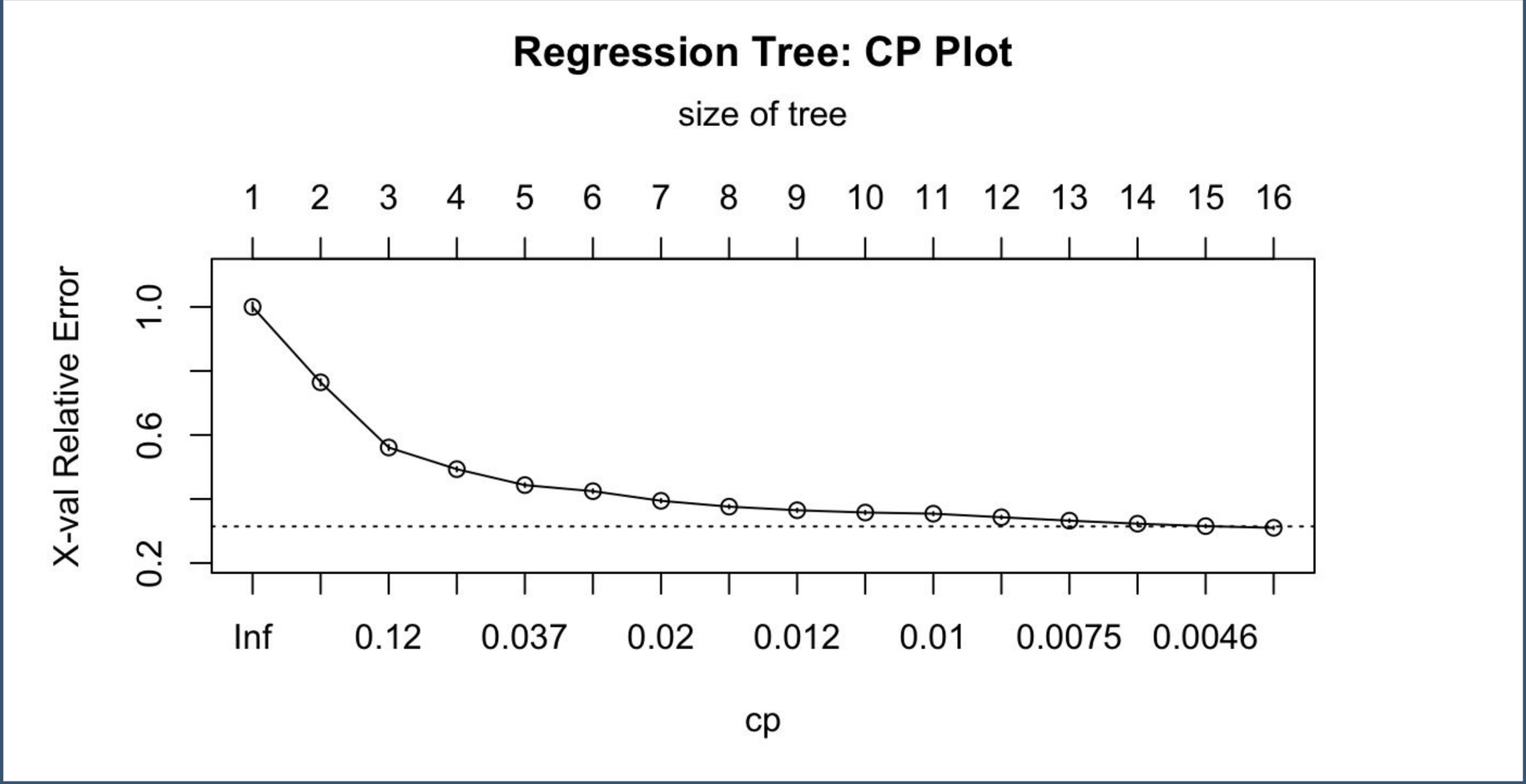
- Variable Importance Table

Variable Importance: Target Y - mental_distress_rate (all vars.)

	Feature Importance
child_poverty	2.210
year	2.038
median_inc	1.987
food_index	1.117
college	1.111
single_parent	0.995
unempl	0.803
inequality	0.243
mh_providers	0.036
pop_provider_ratio	0.035
hs_grad	0.015
region	0.012
severe_housing	0.002

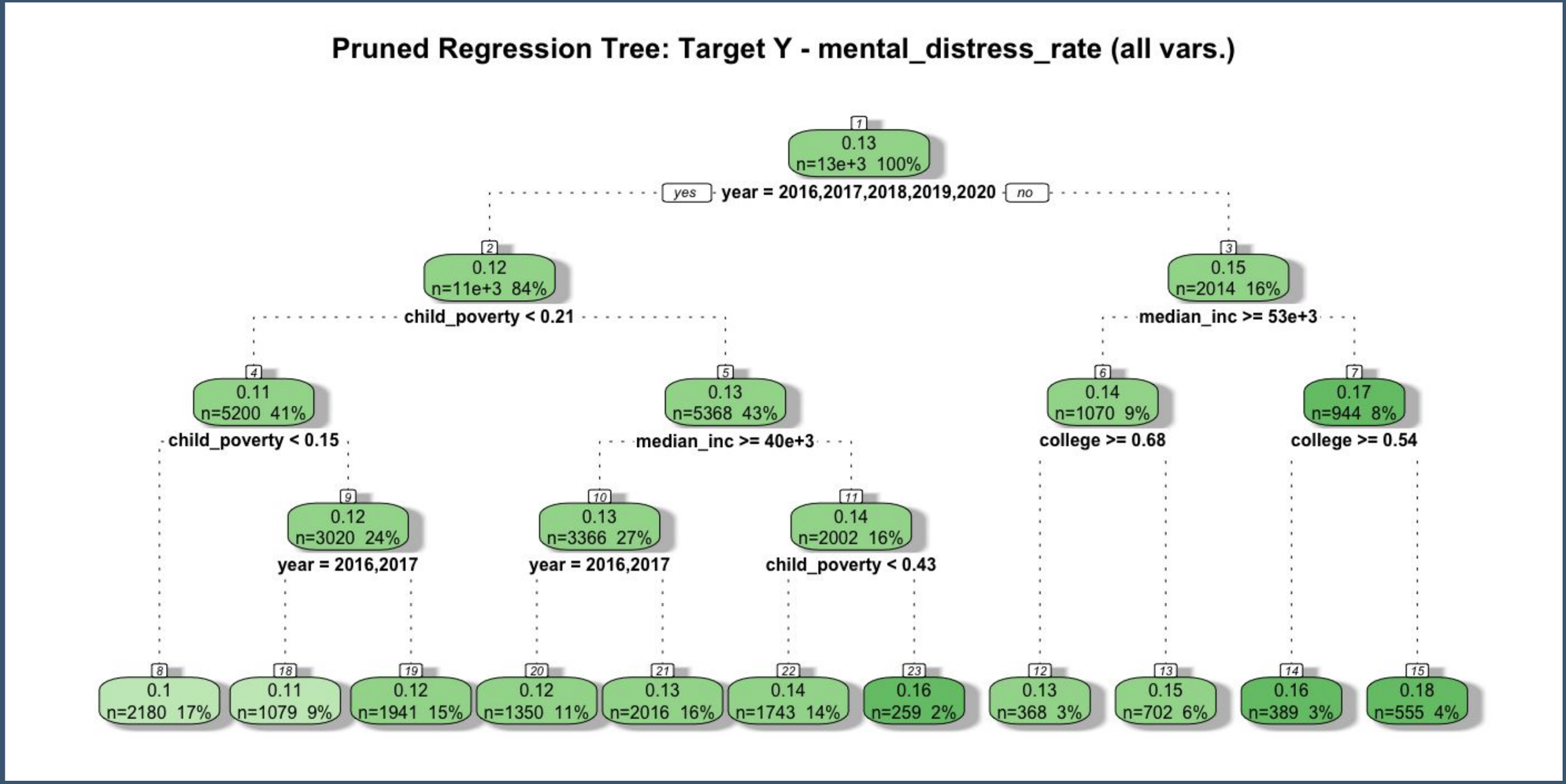
Regression Trees: mental_distress_rate

- Max Depth = 8
- cp = 0.004



Pruned Regression Trees: mental_distress_rate

- The tree has 13 leaf nodes



Pruned Regression Trees: mental_distress_rate

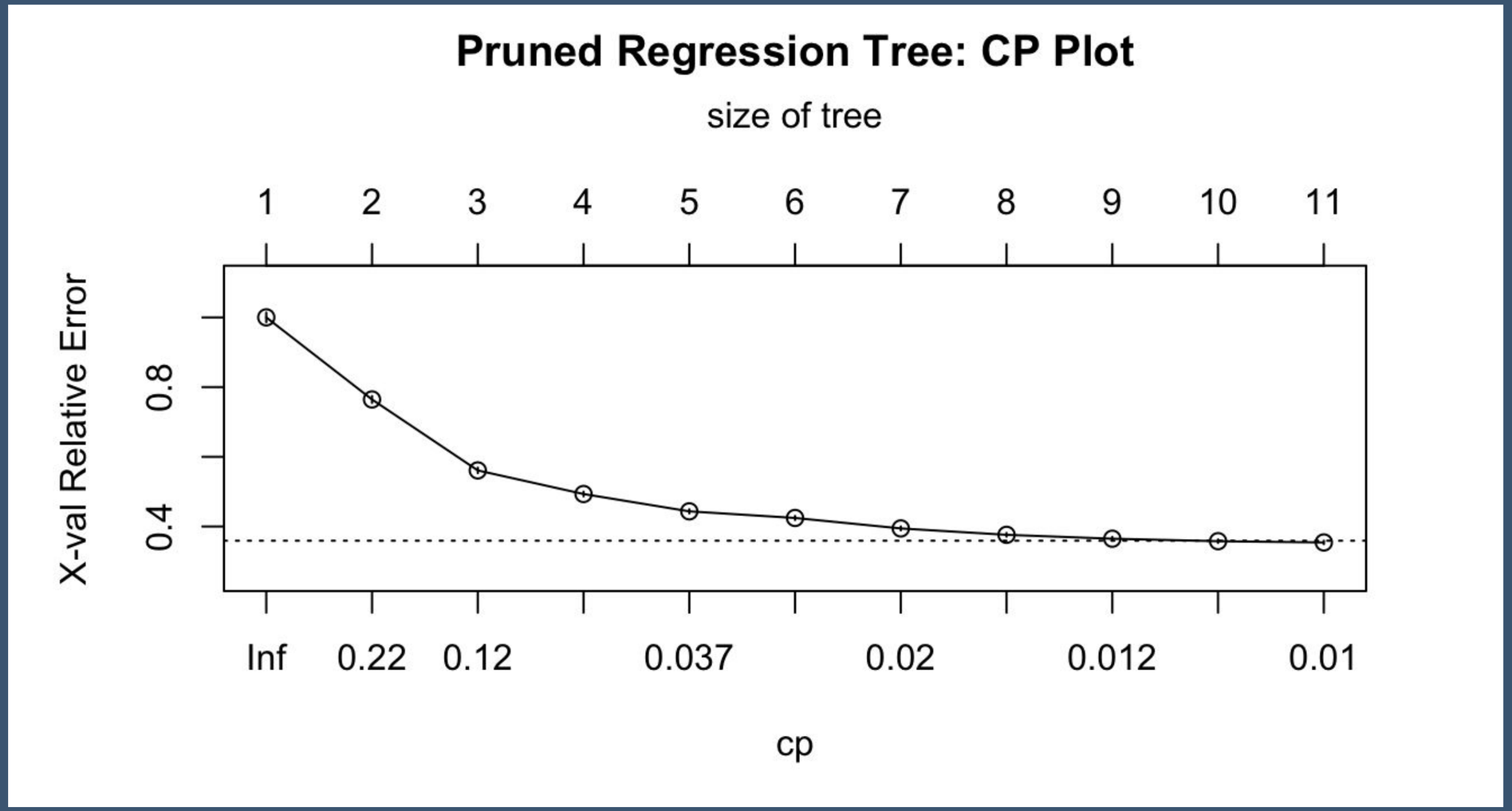
- Variable Importance Table

Pruned Variable Importance: Target Y - mental_distress_rate (all vars.)

	Feature Importance
child_poverty	2.193
median_inc	1.960
year	1.837
college	1.111
food_index	1.098
single_parent	0.981
unempl	0.751
inequality	0.227
mh_providers	0.035
pop_provider_ratio	0.034
region	0.012
hs_grad	0.006
severe_housing	0.001

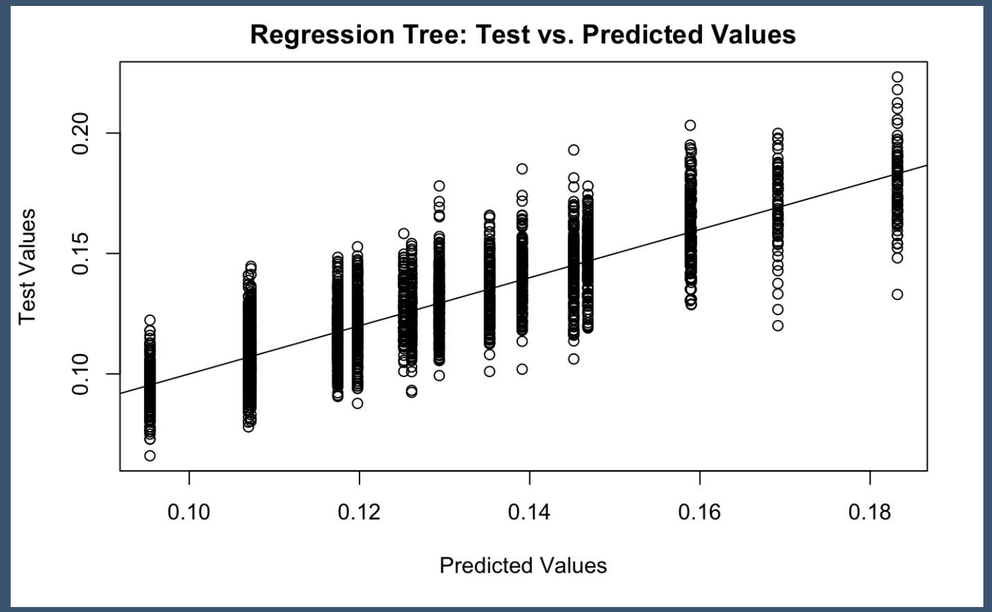
Pruned Regression Trees: mental_distress_rate

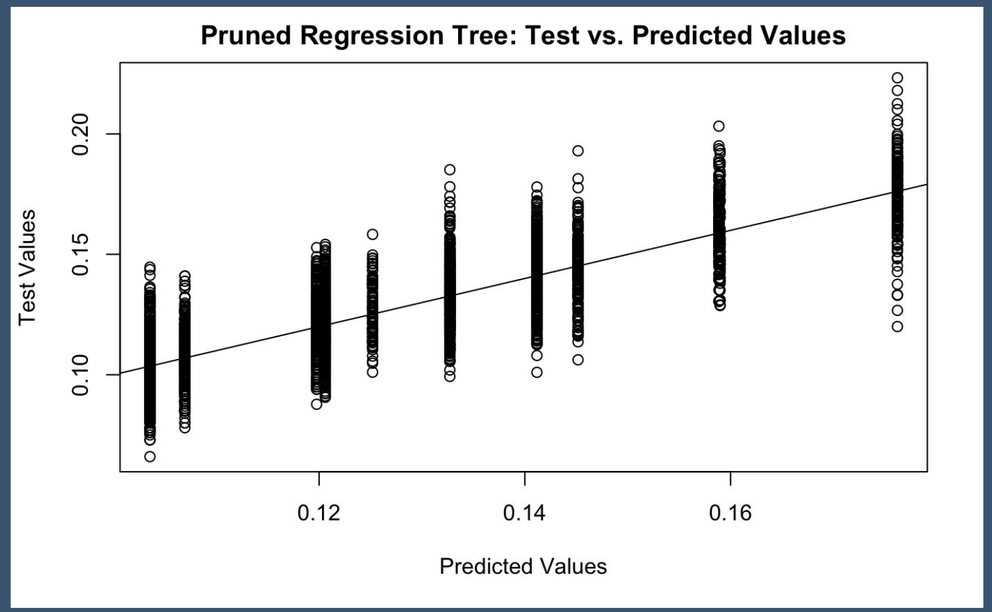
- cp = 0.01



Reg. Tree and Prune Tree RMSE: mental_distress_rate

- Reg. Tree RMSE: 0.0126
- Pruned Reg. Tree RMSE: 0.0134



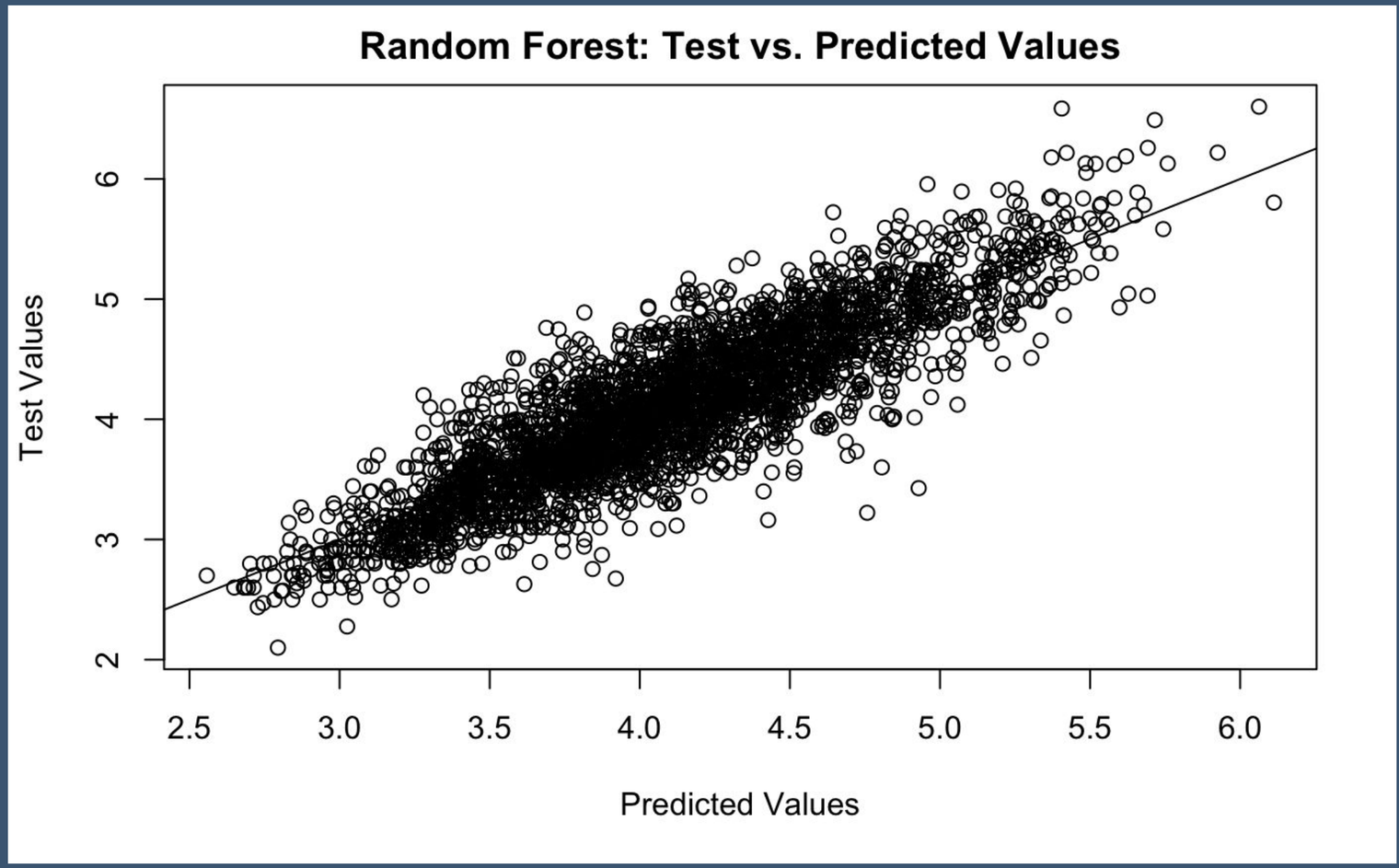


Random Forests

All Models used an 80:20 Train-Test Split

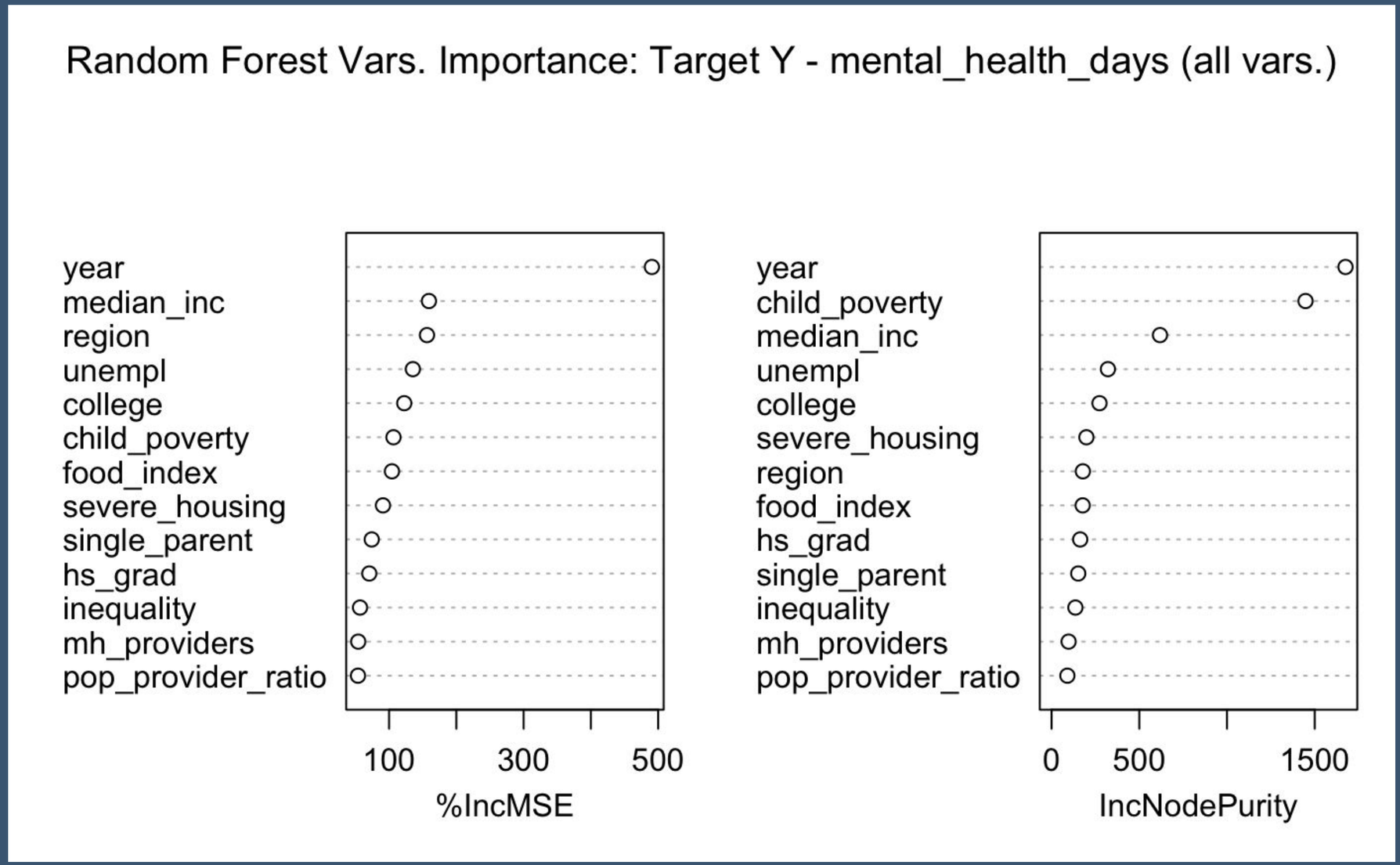
Random Forest: mental_health_days

- RMSE: 0.327



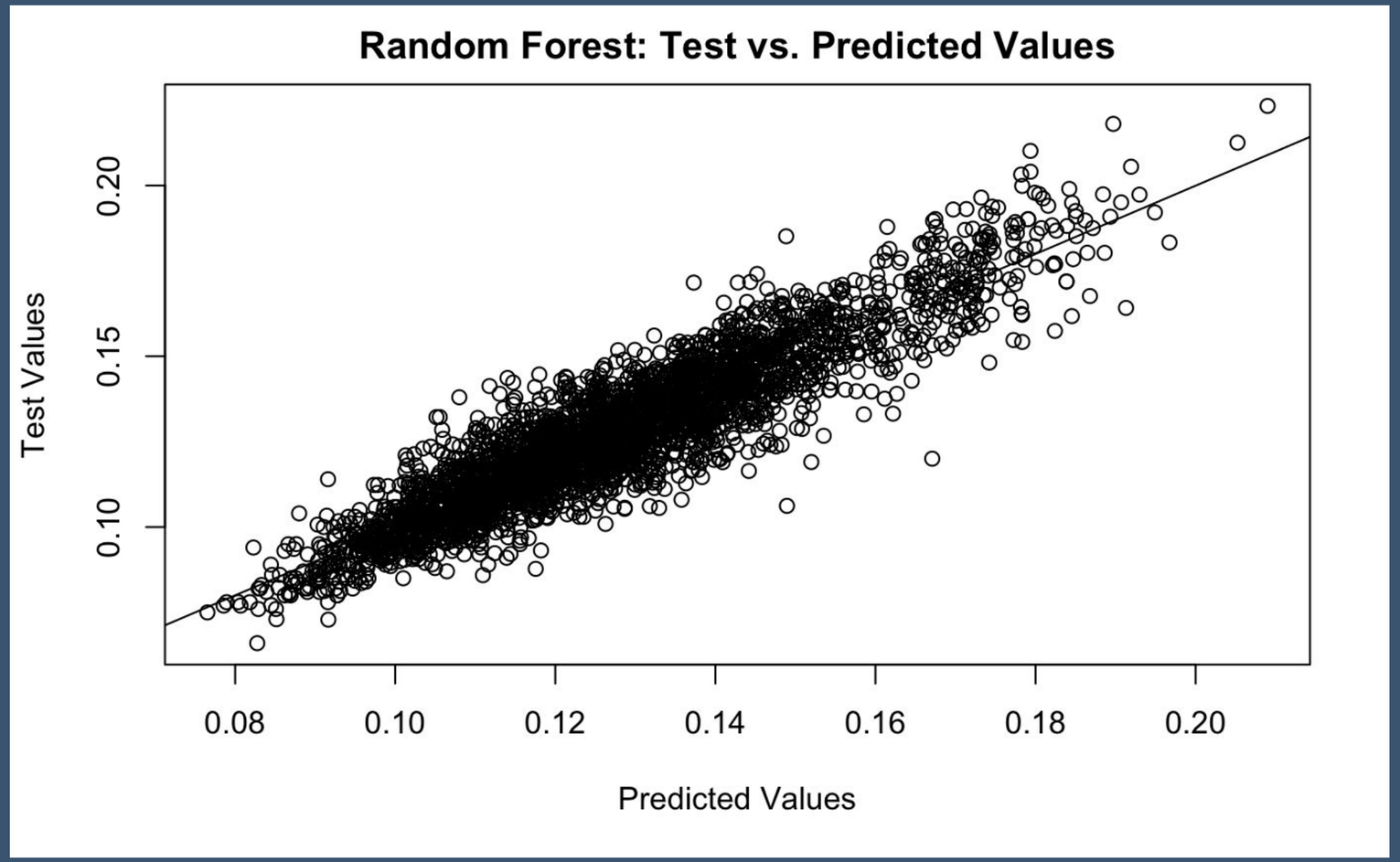
Random Forest: mental_health_days

- Variable Importance Tables



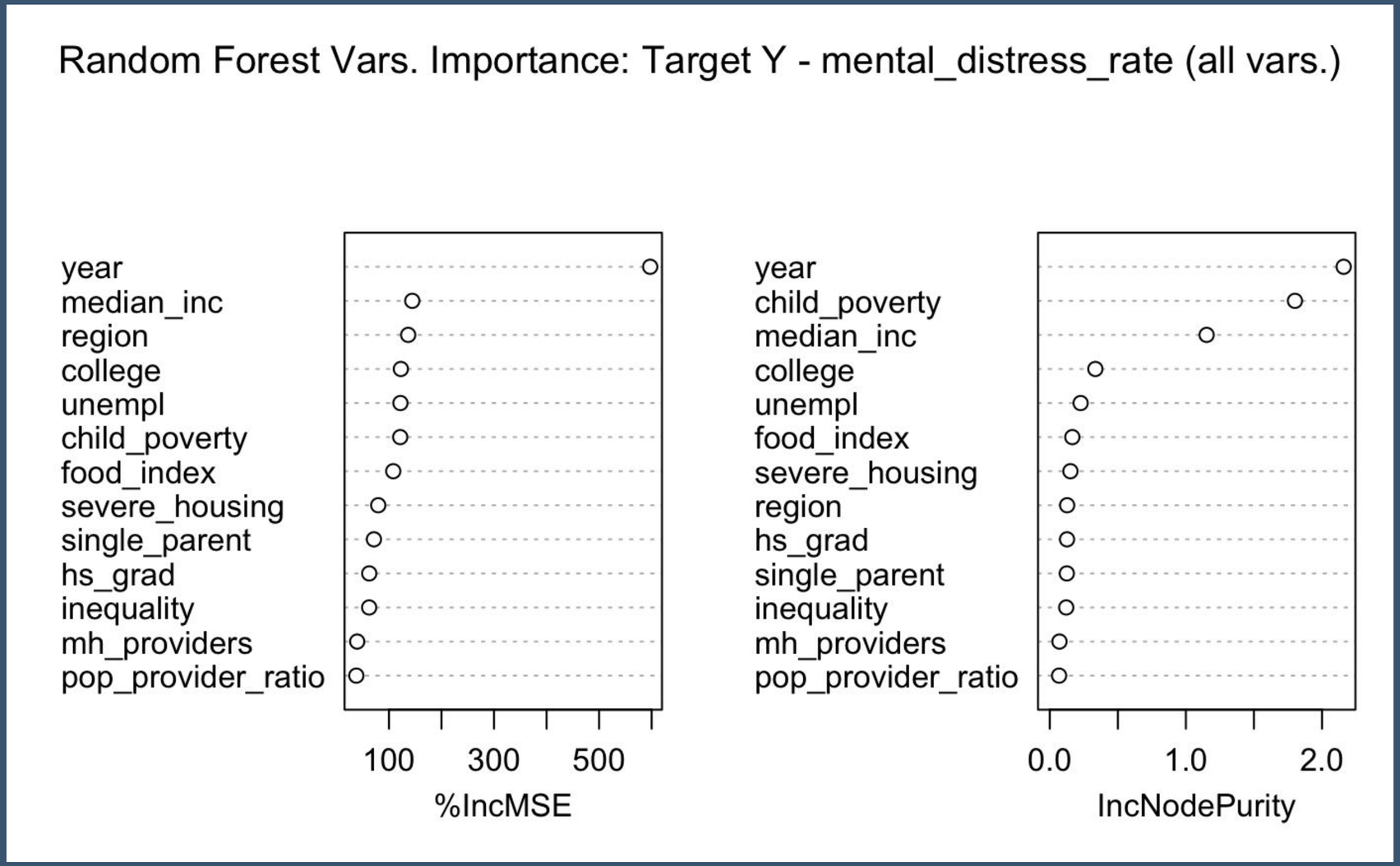
Random Forest: mental_distress_rate

- RMSE: 0.00915



Random Forest: mental_distress_rate

- Variable Importance Tables



RMSE Comparisons for All Models

RMSE Comparisons - mental_health_days

Models	RMSE
Linear Model	0.390
Reg. Tree	0.427
Pruned Tree	0.435
Random Forest	0.327

RMSE Comparisons - mental_distress_rate

Models	RMSE
Linear Model	0.0110
Reg. Tree	0.0126
Pruned Tree	0.0134
Random Forest	0.0091

Conclusion

Conclusion

- EDA and hypothesis tests from Midterm project suggested that the relationship between mental health and income inequality differs by U.S. region.

- Linear modeling indicates this relationship is statistically significant and strength of association varies by region. Both measures exhibit similar relationship but association with mental health days is more pronounced.
  - Midwest has the strongest association
  - South has weakest association

- Regression trees for both target mental health variables indicated that the top 3 most important variables to *year*, *%child poverty*, and *median income*.

- Random forests displayed more conflicting results on variable importance.
  - %IncMSE: *year* and *median income.*
  - IncNodePurity: *year* and *child poverty.*

- Overall, RMSE values indicate that the random forests are the best predicting models for both target mental health variables due to their low RMSE values.

Limitations

- The target variables rely on self-reported mental health data.
  - This data cannot be validated with medical records.
  - Further research could be done using a binary mental health variable.
  - Prevalence changes as mental health becomes more socially acceptable.

- There is a strong time influence, as indicated by *year* being a top important variable.
  - This makes it difficult to determine if economic inequality actually has a substantial impact on mental health or if they are just correlated over time.
  - Next time use a dataset with no time variable.

- Regression trees could be reduced in size to give better general predictability.

- Although the random forests had the lowest RMSE their ability to predict accurately is likely curtailed by the *year* variable.
  - Would ***not*** use these models for predicting or determining causality.
  - Models are likely unreliable.

References

Bechtel, Lucy, Grace Lordan, and DS Prasada Rao. "Income inequality and mental health—empirical evidence from Australia." *Health economics* 21 (2012): 4-17.

Kelley, Jonathan, and Mariah DR Evans. "Societal Inequality and individual subjective well-being: Results from 68 societies and over 200,000 individuals, 1981–2008." *Social science research* 62 (2017): 1-23.

Layte, Richard. "The association between income inequality and mental health: testing status anxiety, social capital, and neo-materialist explanations." *European Sociological Review* 28.4 (2012): 498-511.

Matthew, Pravin, and Donka Mirtcheva Brodersen. "Income inequality and health outcomes in the United States: An empirical analysis." *The Social Science Journal* 55.4 (2018): 432-442.

Pickett, Kate E., and Richard G. Wilkinson. "Income inequality and health: a causal review." *Social science & medicine* 128 (2015): 316-326.

Ribeiro, Wagner Silva, et al. "Income inequality and mental illness-related morbidity and resilience: a systematic review and meta-analysis." *The Lancet Psychiatry* 4.7 (2017): 554-562.

Robert Wood Johnson Foundation. "2021 County Health Rankings National Data." *County Health Rankings & Roadmaps.* https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation. Accessed: March 13, 2022.

Sommet, Nicolas, Davide Morselli, and Dario Spini. "Income inequality affects the psychological health of only the people facing scarcity." *Psychological Science* 29.12 (2018): 1911-1921.

Tibber, Marc S., et al. "The association between income inequality and adult mental health at the subnational level—a systematic review." *Social psychiatry and psychiatric epidemiology* (2021): 1-24.

U.S. Census Bureau. "Census Regions and Divisions of the United States." *Economics and Statistics Administration.*

Zimmerman, Frederick J., and Janice F. Bell. "Income inequality and physical and mental health: testing associations consistent with proposed causal pathways." *Journal of Epidemiology & Community Health* 60.6 (2006): 513-521.

Appendix

Literature Review

Bechtel et al. (2012): determined that mental health is only adversely affected in the context of income inequality to a very small degree.

Kelley & Evans (2017): determined that there was no strong relationship between income inequality and subjective well-being.

Layte (2012): determined that there is an empirical consensus on the existence of an effect regarding income as a determinant of mental and physical health outcomes, but there is differing views on the explanation and causes of the effect

Matthew & Brodersen (2018): found a significant negative association between mental health and inequality.

Picket & Wilkinson (2015): determined that there is a strong causal link between income inequality and population, but a connection to mental health is less clear.

Ribeiro et al. (2017): found significant small effects that income inequality affects mental health.

Sommet et al. (2018): determined that income inequality affects the psychological health of individuals that are economically vulnerable.

Tibber et al. (2022): found that area-level income inequality is associated with poorer mental health.

Zimmerman & Bell (2006): found that the effect of income inequality on health may work through the differences in social capital among different racial groups.

Descriptive Statistics

Dataset:

- RWJF County Health Rankings & Roadmaps (2016 – 2021)
  - *Consists of economic and health data for all states at the county level in the United States*.

Descriptive Statistics of Dataset

	Var Num.	Count	Mean	Std. Dev.	Median	Trimmed Mean	MAD	Minimum	Maximum	Range	Skewness	Kurtosis	S.E.
mental_health_days	1	18469	4.04e+00	6.92e-01	4.02e+00	4.03e+00	7.06e-01	2.100	7.29e+00	5.19e+00	0.223	-0.113	0.005
mental_distress_rate	2	18469	1.26e-01	2.40e-02	1.24e-01	1.25e-01	2.40e-02	0.066	2.47e-01	1.81e-01	0.532	0.301	0.000
inequality	3	18465	4.52e+00	7.33e-01	4.41e+00	4.45e+00	6.28e-01	2.543	1.20e+01	9.43e+00	1.254	3.431	0.005
median_inc	4	18469	5.08e+04	1.36e+04	4.87e+04	4.93e+04	1.10e+04	21658.000	1.52e+05	1.30e+05	1.424	3.680	99.952
hs_grad	5	16563	8.70e-01	7.90e-02	8.83e-01	8.78e-01	6.70e-02	0.025	1.00e+00	9.75e-01	-1.592	5.592	0.001
college	6	18469	5.71e-01	1.16e-01	5.72e-01	5.72e-01	1.22e-01	0.152	9.11e-01	7.59e-01	-0.088	-0.263	0.001
unempl	7	18469	5.00e-02	2.00e-02	4.60e-02	4.80e-02	1.70e-02	0.012	2.40e-01	2.28e-01	1.690	6.491	0.000
child_poverty	8	18469	2.21e-01	9.10e-02	2.09e-01	2.14e-01	9.00e-02	0.024	7.47e-01	7.23e-01	0.695	0.536	0.001
single_parent	9	18469	3.14e-01	1.06e-01	3.06e-01	3.08e-01	9.60e-02	0.000	8.72e-01	8.72e-01	0.696	1.250	0.001
severe_housing	10	18470	1.42e-01	4.70e-02	1.37e-01	1.39e-01	3.70e-02	0.022	7.13e-01	6.91e-01	2.086	14.165	0.000
food_index	11	18394	7.32e+00	1.18e+00	7.50e+00	7.44e+00	1.04e+00	0.000	1.00e+01	1.00e+01	-1.385	3.715	0.009
mh_providers	12	17028	1.00e-03	2.00e-03	1.00e-03	1.00e-03	1.00e-03	0.000	2.40e-02	2.40e-02	3.457	22.562	0.000
pop_provider_ratio	13	17028	2.00e+03	2.84e+03	9.90e+02	1.38e+03	8.97e+02	-957.000	5.49e+04	5.58e+04	4.291	34.700	21.797

Descriptive Statistics - Std. Dev. & Variance

- The small standard deviations and variances for each variable suggest that the data for each variable are relatively close to the mean.
- In other words, the data for most variables has low spread. Median income seems to be an exception.

Descriptive Statistics of Dataset - Std. Dev. & Variance

	Var Num.	Std. Dev.	Variance
mental_health_days	1	6.92e-01	4.78e-01
mental_distress_rate	2	2.40e-02	1.00e-03
inequality	3	7.33e-01	5.38e-01
median_inc	4	1.36e+04	1.85e+08
hs_grad	5	7.90e-02	6.00e-03
college	6	1.16e-01	1.30e-02
unempl	7	2.00e-02	0.00e+00
child_poverty	8	9.10e-02	8.00e-03
single_parent	9	1.06e-01	1.10e-02
severe_housing	10	4.70e-02	2.00e-03
food_index	11	1.18e+00	1.40e+00
mh_providers	12	2.00e-03	0.00e+00
pop_provider_ratio	13	2.84e+03	8.09e+06

Regional Histogram Plots