1 Data Science Final Project

Data Science for Business

AUTHOR
Mahsa Abdollahi Mirzanagh, 400355424

PUBLISHED
July 22, 2024

1 Abstract

This report presents a replication of a previous study that investigates the influence of neighborhood poverty on various socio-economic outcomes with R programming language. Utilizing publicly available U.S. data, the analysis focuses on multiple geographic scales, including commuting zones, school districts, and census tracts, to explore how poverty rates affect indicators such as employment rates, life expectancy, upward mobility, and academic achievement.

The replication involves a detailed examination of these relationships, highlighting significant associations between higher poverty rates and poorer outcomes across all metrics. Specifically, increased poverty is shown to correlate with lower adult employment rates, reduced life expectancy, diminished upward mobility for children, and lower levels of academic achievement. The replication confirms the original study's results and offers further insights into the effects of neighborhood poverty on individual and family well-being, emphasizing the critical role of residential environments in shaping socio-economic outcomes.

2 Introduction

"Social scientists have long hypothesized that living in a disadvantaged area directly affects the outcomes of adults and life courses of children. Descriptive research has supported this idea by showing that individuals living in high-poverty areas fare worse both contemporaneously and over the long-run in terms of important outcomes such as education, criminal involvement, health, and earnings (Wilson 1987; Jencks and Mayer 1990; Brooks-Gunn et al.1993; Sampson, Morenoff, and Gannon-Rowley 2002; Sharkey and Faber 2014)." (Chyn and F.Katz, 2021, P.197)

In *"Neighborhoods Matter: Assessing the Evidence for Place Effects"*, Eric Chyn and Lawrence F.Katz explore the relationship between neighborhood poverty and various socio-economic outcomes, including employment rates, life expectancy, upward mobility, and academic achievement by Using a dataset comprising these indicators at different geographic levels. This report reproduce the original analyses and visualizations done by Chyn and F.Katz , specifically focusing on bin scatter plots and linear regression models using the statistical programming language R.

Our replication confirms the original study's findings that neighborhood poverty significantly correlates with adverse socio-economic outcomes. However, we critique the use of bin scatter plots for potentially oversimplifying the data , which could lead to misleading interpretations. Instead, we advocate for a more detailed visualization approach that retains individual data points to better capture the underlying variability.

This section utilizes publicly available U.S. data to replicate the analysis of how place of residence influences various outcomes. Following the methodology of previous studies on neighborhood effects, the analysis focuses

on several geographical units. The largest units considered are commuting zones, which are groups of counties based on commuting patterns from the 1990 Census, approximating local labor markets. There are 741 commuting zones in the U.S. Additionally, more detailed levels such as over 12,000 school districts and approximately 72,000 census tracts are examined, with census tracts typically encompassing a few thousand residents and closely resembling what is commonly known as a “neighborhood.”

To classify these geographic areas by economic opportunity, the poverty rate from the 2000 Decennial Census is used. Poverty rates are a widely used measure in neighborhood effects research (Sampson and Sharkey 2008) and can be broadly interpreted as a summary index of the bundle of characteristics associated with a neighborhood (Kling, Liebman, and Katz 2007).

Figure 1 illustrates a strong association between an area’s poverty rate and various outcomes for adults and children. Each panel plots averages based on grouping commuting zones (in panels A, B, and C) or school districts (in panel D) into one of 20 “bins” by poverty rate. The results in panel A show that a one percentage point increase in the poverty rate in a commuting zone is associated with a 0.8 percentage point decline in the adult employment rate using data from the 2000 U.S. Census. Panel B shows that adult health, as measured by life expectancy at age 40, also decreases with the poverty rate. Life expectancy is measured using data from Chetty et al. (2016a) and is based on mortality records from the Social Security Administration.

The results in Figure 1 also show that upward mobility and academic achievement of children both decrease with the poverty rate. The measure of upward mobility is the mean household income (measured at ages 31–37) for children who grew up in each commuting zone and were born to low-income parents (those at the 25th percentile of the income distribution) from the Opportunity Atlas (Chetty et al. 2020a). The measure of achievement is based on the mean of standardized test scores for school districts from the Stanford Education Data Archive. Panels C and D of Figure 1 show that a one percentage point increase in the poverty rate is associated with declines of $371 in a child’s expected adult income and 0.025 standard deviations in academic achievement, respectively. All the relationships in Figure 1 are statistically significant at the 1 percent level, as indicated in the regression results reported in columns 1–4 of Table 1.

2.0.0.1 Figure 1 Description

This figure provides binned scatter plots of the relationship between the poverty rate and the following measures of average resident outcomes: employment rates, life expectancy, upward mobility, and test scores. The unit of analysis in panels A, B, and C is a commuting zone. In panel D, the unit of analysis is a school district.

2.0.0.2 Table 1 Description

Table 1 reports estimates from a regression model where the dependent variable is a measure of adult or child outcomes (specified in each column header) in a geographic area. Geographic areas are commuting zones (CZ), school districts, or Census tracts. The independent variable of interest is a location-specific measure of the poverty rate (the fraction of residents living below the poverty line). Columns 1, 2, 3, 5, and 6 use poverty rates from the 2000 Decennial Census. Column 4 uses poverty rates averaged over 2007–2016 from the American Community Survey (the combined files for the years 2007-2011 and 2012–2016). The dependent variables in columns 1 and 5 are measures from the 2000 Decennial Census. Column 2 uses the life expectancy measure from Chetty et al. (2016a,b) based on mortality data from Social Security Administration death records. Columns 3 and 6 use the “Upward Mobility” measure from the Opportunity Atlas (Chetty et al. 2020a,b), which is the mean later-life

household income rank (measured at ages 31–37) for children whose parents were at the twenty-fifth percentile of the national income distribution. Column 4 uses the test-based achievement measure from the Stanford Education Data Archive (SEDA), which is an estimate of mean test scores on a cohort-standardized scale. The test score means are constructed using data from the National Assessment of Educational Progress (NAEP), as detailed in Fahle et al. (2019). Standard errors are clustered at the county level in columns 5 and 6. The correlations between the poverty rate and outcomes are not simply due to broad differences across metropolitan areas. Columns 5 and 6 of Table 1 present correlations between poverty rates and resident outcomes at the census-tract level using data on all U.S. tracts and specifications that control for county fixed effects. These within-county results generate estimates similar to those observed in the commuting-zone level analysis.

3 for github implementation:

if (!require(pacman)) install.packages("pacman") pacman::p_load(usethis) create_github_token()

4 Loading the Initial libraries and data

For starting fresh without any lingering variables or data frames from previous work, it is useful to remove all objects from the current workspace.

To ensure our environment is properly set up, we begin by verifying the installation of essential packages. The first line checks whether the `pacman` package is already installed. If it is not, it installs `pacman`. This is crucial as `pacman` simplifies the process of loading and managing other packages.

`ggplot2` is fundamental for creating advanced and customizable graphics, whereas `haven` allows us to seamlessly import data type '.dta' from the files attached to the original paper.

```
rm(list = ls())

# Install and load necessary packages
if (!require(pacman)) install.packages("pacman")
```

```
Loading required package: pacman
```

```
#Loading required package: pacman
pacman::p_load(ggplot2, haven)
```

This line of code sets the working directory to the specified path on the computer.

```
#setting the working directory
setwd('C:/OLD Asus/Cdoc/Hs-Fresenius/PDFs/Third Semester/Data Science/Project/142621-V1/
```

To analyze various socioeconomic outcomes at different levels of aggregation, we need to load the relevant datasets. Each dataset corresponds to different figures and tables that illustrate these outcomes. We use the

`read_dta` function from the `haven` package to load Stata files into R.

```
#loading the data sets

# first figure (employment rate(CZ Level))
cz_adult_emp_pov <- read_dta("Replication_Data\\cz_adult_emp_pov.dta")

# second figure (life expectancy (CZ Level))
cz_life_expect_pov <- read_dta("Replication_Data\\cz_life_expect_pov.dta")

# third figure (Upward mobility(CZ Level))
cz_upward_mobility_pov <- read_dta("Replication_Data\\cz_upward_mobility_pov.dta")

# fourth figure (mean test score(CZ Level))
district_test_scores_pov <- read_dta("Replication_Data\\district_test_scores_pov.dta")

#loading Table1 related data

# fifth dataset(tract_adult_emp)
tract_adult_emp_pov <- read_dta("Replication_Data\\tract_adult_emp_pov.dta")

#sixth dataset(tract_upward_mobility)
tract_upward_mobility_pov <- read_dta("Replication_Data\\tract_upward_mobility_pov.dta")

#seventh datasets(district_test_scores)
district_test_scores_pov <- read_dta("Replication_Data\\district_test_scores_pov.dta")
```

5 Analysis

5.1 Figure 1: Association Between Adult and Child Outcomes and Neighborhood Poverty

Figure 1 illustrates a strong association between an area's poverty rate and various outcomes for both adults and children. The data is grouped into 20 "bins" based on poverty rate, with different panels representing different outcomes and geographic levels.

5.1.0.1 **Panel A. Adult employment rate 2000**

Panel A shows the relationship between poverty rates and adult employment rates at the level of commuting zones. The data indicates that for each one percentage point increase in a commuting zone's poverty rate, there is a corresponding 0.8 percentage point decrease in adult employment rates. This data is derived from the 2000 US Census. In this section, we will analyze the adult employment rate at the Commuting Zone (CZ) level.

5.1.0.1.1 **Loading and Inspecting the Data**:

```
#Panel A. Adult employment rate (2000)
```

```
# Inspecting the first few rows of the data
head(cz_adult_emp_pov)
```

```
# A tibble: 6 × 4
     cz czname         emp2000 pov_rate
  <dbl> <chr>            <dbl>    <dbl>
1   100 Johnson City     0.558     14.0
2   200 Morristown       0.589     14.4
3   301 Middlesborough   0.443     26.8
4   302 Knoxville        0.589     13.0
5   401 Winston-Salem    0.629     10.6
6   402 Martinsville     0.565     13.1
```

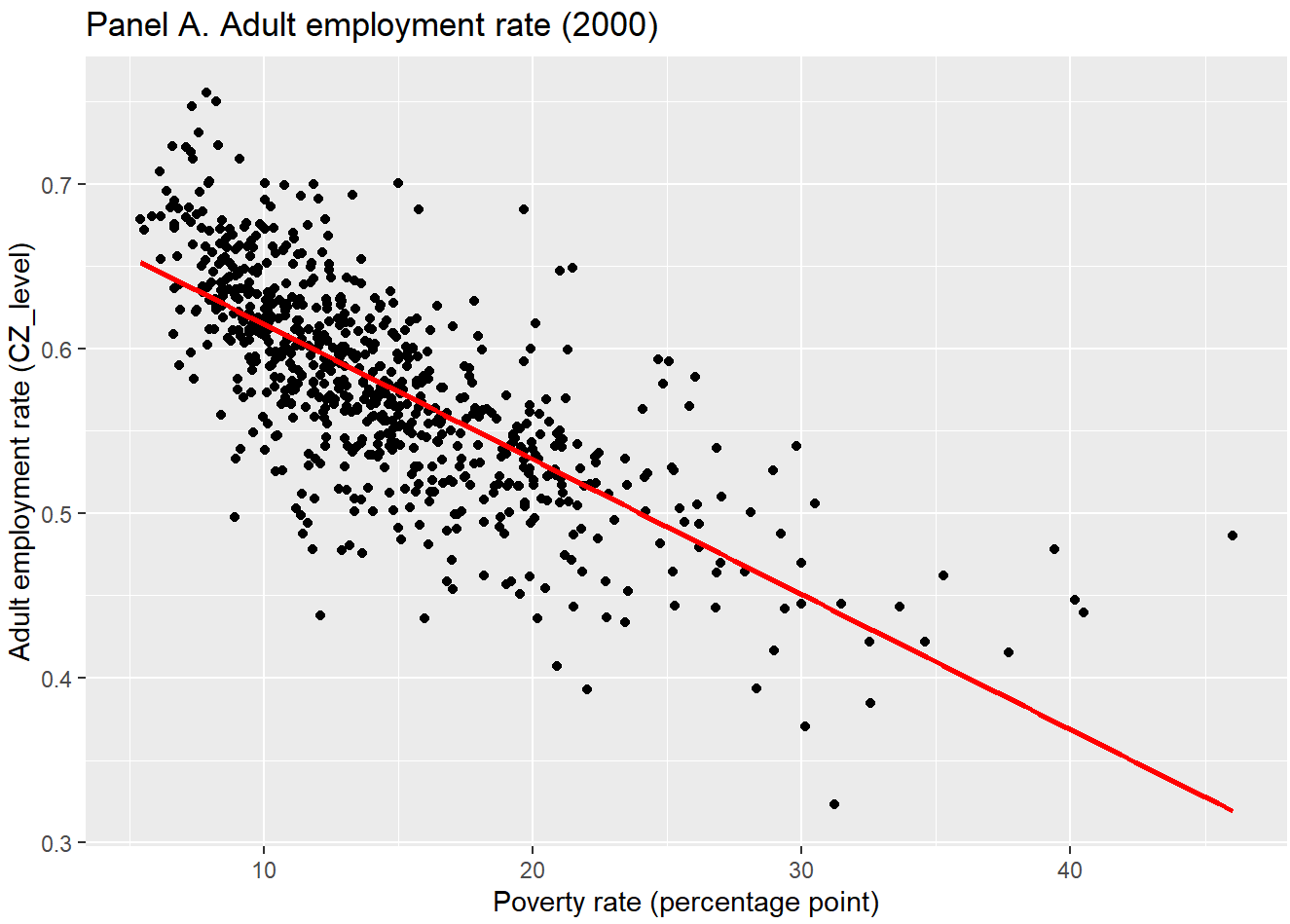
```
# Fitting a linear regression model
model<-lm(emp2000 ~ pov_rate , data = cz_adult_emp_pov)
show(model)
```

```
Call:
lm(formula = emp2000 ~ pov_rate, data = cz_adult_emp_pov)

Coefficients:
(Intercept)     pov_rate
   0.696980    -0.008208
```

5.1.0.1.2 **Standard Scatter Plot with Linear Regression**:

ggplot(): Creates a scatter plot of emp2000 against pov_rate.

```
#Normal plot
library("ggplot2")

# Creating the scatter plot
ggplot(cz_adult_emp_pov, aes(x = pov_rate, y = emp2000)) +
  geom_point() +
  stat_smooth(formula = y ~ x, method = "lm", se = FALSE, colour = "red", linetype = 1) +
  labs(x = "Poverty rate (percentage point)", y = "Adult employment rate (CZ_level)" ,
       title = "Panel A. Adult employment rate (2000)")
```



5.1.0.1.3 **Binned Scatter Plot**:

As also discussed by Cattaneo et al. (2024),to generate a binned scatter plot, I used the `binsreg` function from the `binsreg` package. This function divides the data into 20 bins (`nbins = 20`) based on the poverty rate (`pov_rate`) in the `cz_adult_emp_pov` dataset. The `binsreg` function fits a polynomial regression of degree 1 (`polyreg = 1`) to each bin, without displaying confidence intervals (`ci = FALSE`).

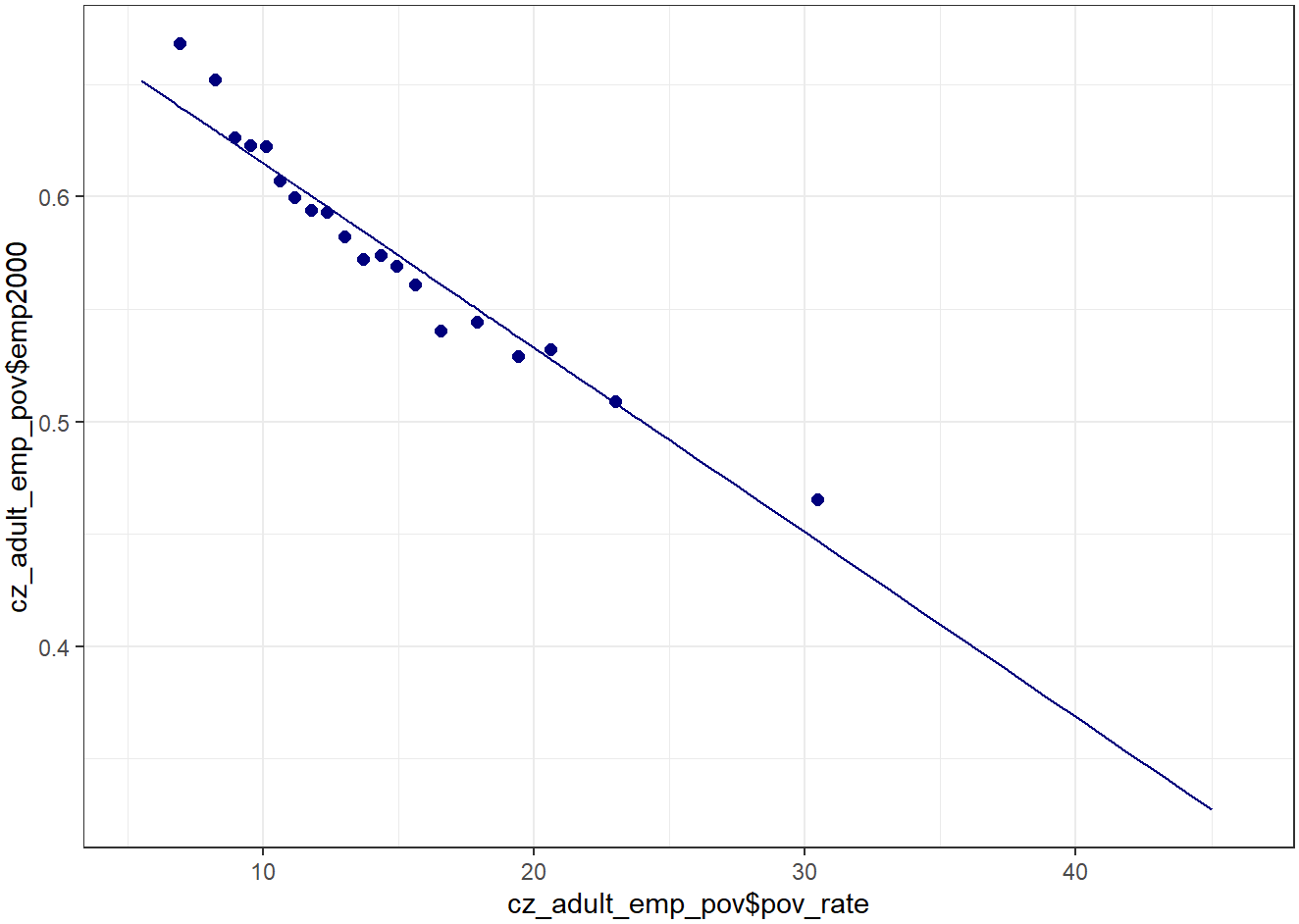
This approach allows for a segmented view of how adult employment rates (`emp2000`) vary with changes in poverty rates (`pov_rate`), providing insights into potential nonlinear relationships that may exist within the data.

To customize the appearance of the binned scatter plot, I applied the `theme_minimal()` function for a clean visual style. Initially, I attempted to change the color of the regression line using `geom_smooth(method = "lm", se = FALSE, color = "red")`. However, despite specifying red as the color, the plot did not reflect this change. To address this, I manually added a separate `geom_smooth()` layer with the desired color adjustment.

This discrepancy likely arises from the internal plot construction of the `binsreg` package. Sometimes, specialized plotting functions such as those in `binsreg` may impose limitations on directly modifying or adding layers using standard `ggplot2` functions.

```
#binned scatter plot
if (!require(pacman)) install.packages("pacman")
pacman::p_load(ggplot2, binsreg, haven, dplyr)

library(binsreg)
```

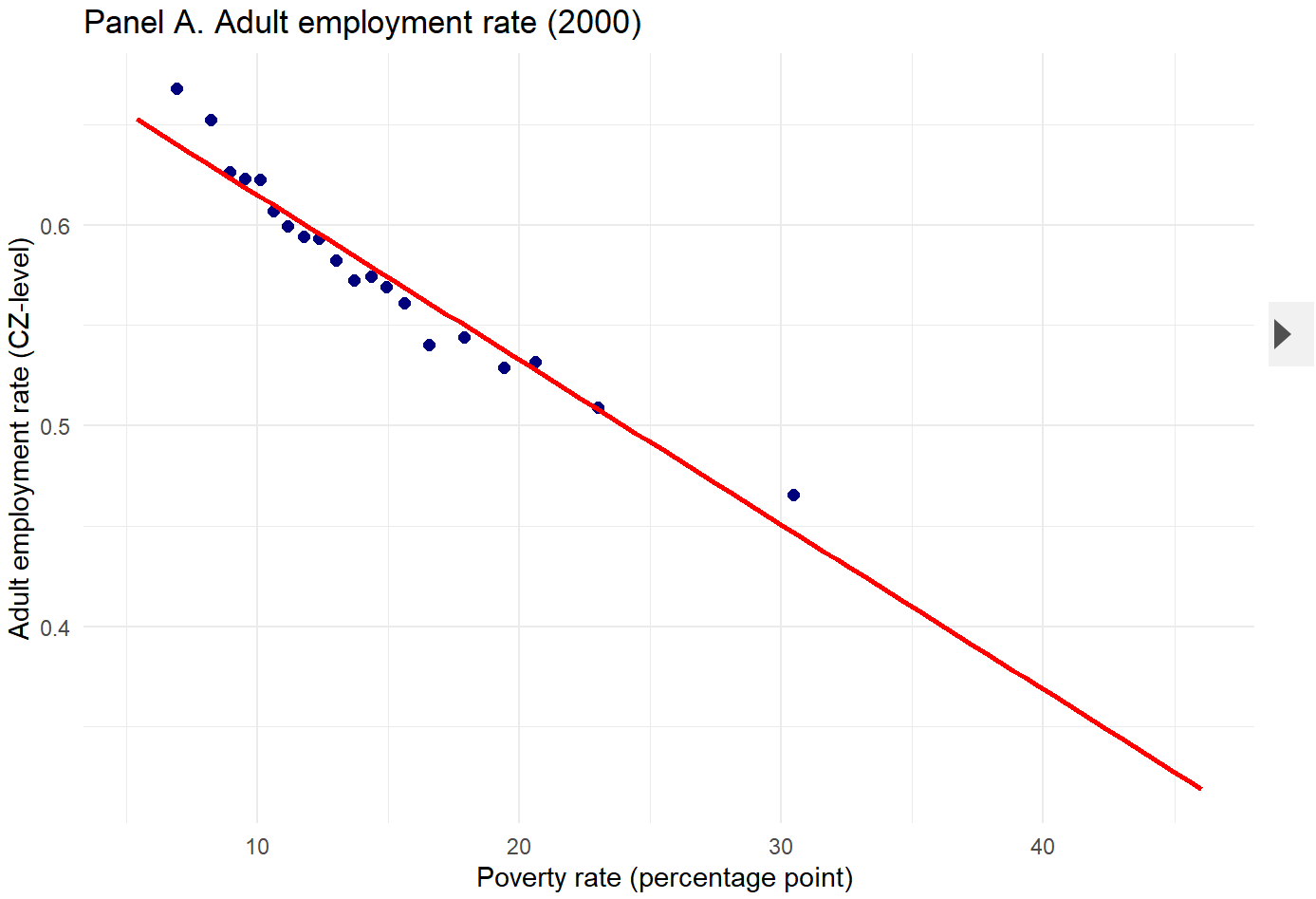
```
library(ggplot2)

# Creating the binned scatter plot
binsreg_result <- binsreg(
  y = cz_adult_emp_pov$emp2000,
  x = cz_adult_emp_pov$pov_rate,
  nbins = 20,
  polyreg = 1,
  ci = FALSE
)
```




```
# Customizing the plot using ggplot2
binsreg_plot <- binsreg_result$bins_plot +
  labs(
    x = "Poverty rate (percentage point) ",
    y = "Adult employment rate (CZ-level)",
    title = "Panel A. Adult employment rate (2000)"
  ) +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE, color = "red")  # Change the line color to red

#so i add it manually
binsreg_plot <- binsreg_plot +
```

```
  geom_smooth(data = cz_adult_emp_pov, aes(x = pov_rate, y = emp2000), method = "lm", se

# Printing the customized plot
print(binsreg_plot)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



The `ntile` function from `dplyr` is used to divide the `pov_rate` variable into 20 quantiles (`n = 20`), and each observation is assigned to a corresponding bin (`bin`).

`summarise` calculates the mean values of `emp2000` (employment rate) and `pov_rate` (poverty rate) for each bin, stored in `new_cz_adult_emp_pov`.

`ggplot` creates a scatter plot (`geom_point`) where `pov_rate_mean` is plotted on the x-axis and `emp2000_mean` on the y-axis.

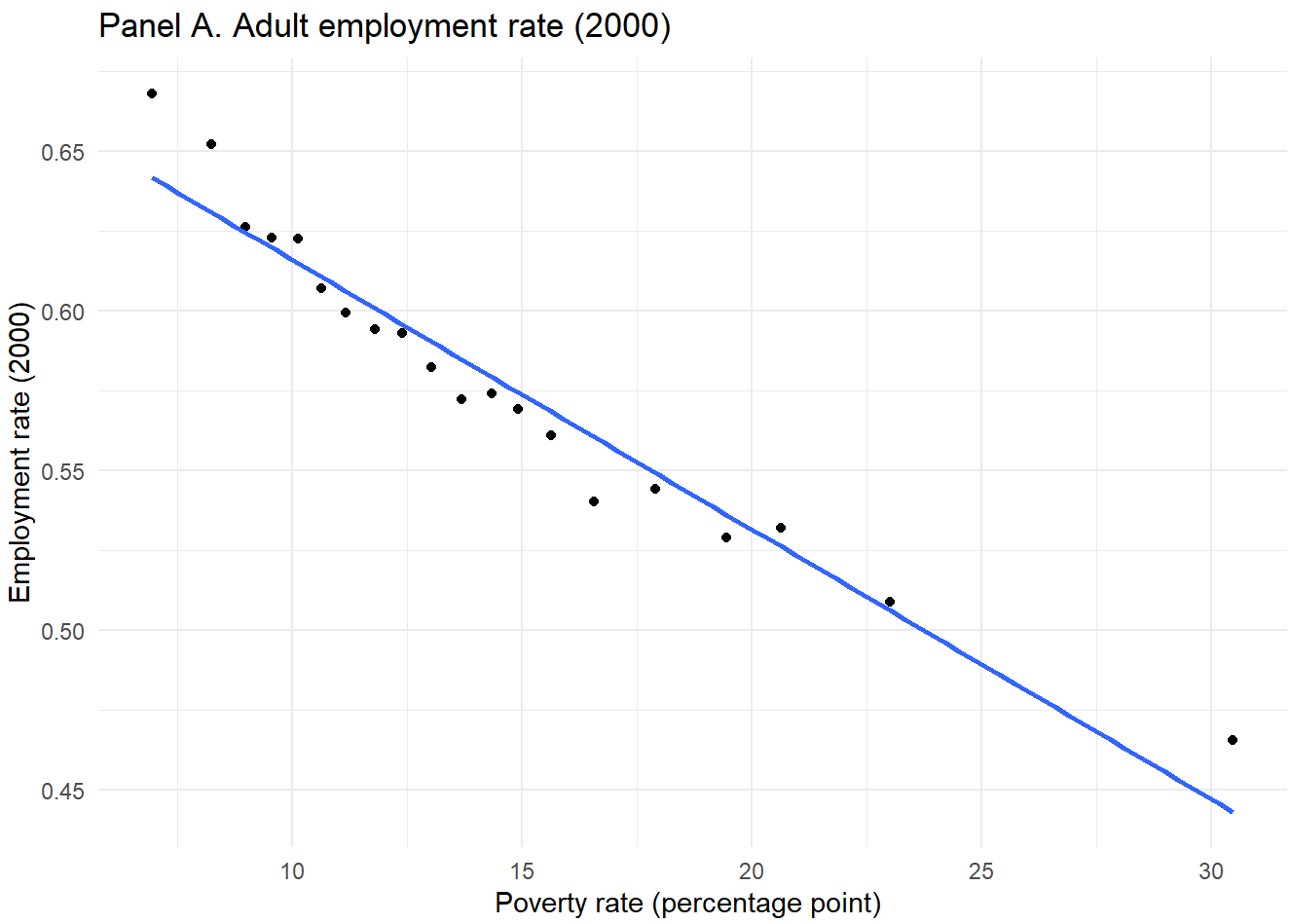
`geom_smooth(method = "lm", se = FALSE)` adds a linear regression line (`method = "lm"`) to show the overall trend between poverty rate and employment rate without displaying standard error (`se = FALSE`).

`labs` sets the axis labels and plot title.

`theme_minimal()` adjusts the plot theme for a cleaner appearance.

```
#The method for creating binned scatter plots is well-documented (Lost Stats, n.d.).

library(dplyr)

# Create 20 quantiles using y and assign observations to bins
cz_adult_emp_pov = cz_adult_emp_pov %>% mutate(bin = ntile(pov_rate, n=20))

# Calculate mean values of employment rate (emp2000) and poverty rate (pov_rate) for eac
new_cz_adult_emp_pov = cz_adult_emp_pov %>% group_by(bin) %>% summarise(emp2000_mean = m

# Create a scatter plot with a linear regression line
ggplot(new_cz_adult_emp_pov, aes(x = pov_rate_mean, y = emp2000_mean)) +
  geom_point()+
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Poverty rate (percentage point) ", y = "Employment rate (2000)",
       title = "Panel A. Adult employment rate (2000)") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Each panel follows a similar structure:

- Fit a linear model (`lm()`) relating a variable to `pov_rate` .
- Create a normal plot (`ggplot()` + `geom_point()` + `stat_smooth()`).
- Create a binned scatter plot (`binsreg()` + `geom_smooth()`).

5.1.0.2 **Panel B. Life expectancy**

Panel B highlights the correlation between poverty rates and life expectancy at age 40 for adults. It demonstrates that higher poverty rates are linked to lower life expectancy. The life expectancy data comes from Chetty et al. (2016a) and is based on mortality records from the Social Security Administration.

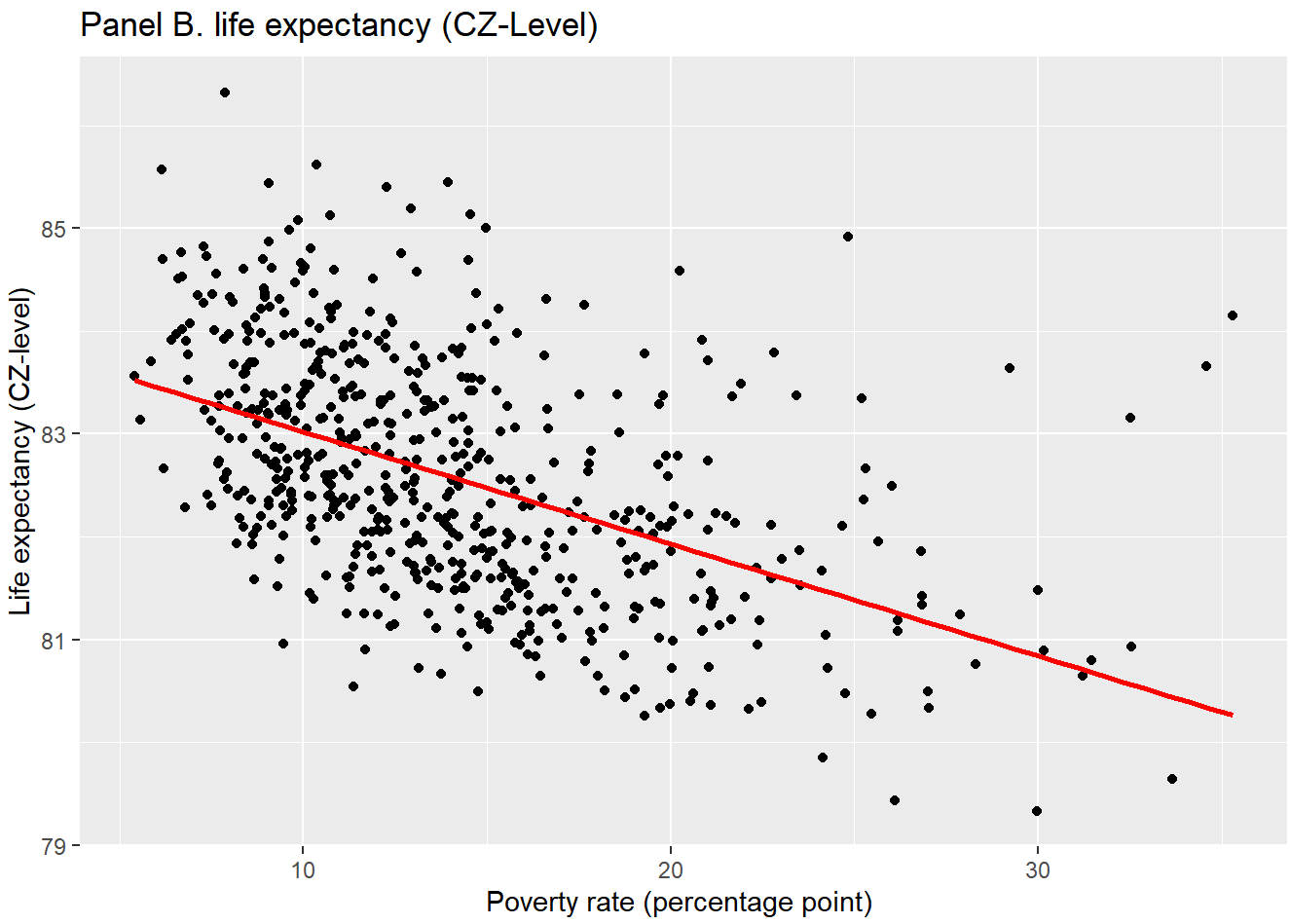
```
# Panel B. life expectancy (CZ-Level)

# Loading and inspecting the data
head(cz_life_expect_pov)
```

```
# A tibble: 6 × 4
     cz czname         le_agg pov_rate
  <dbl> <chr>           <dbl>    <dbl>
1   100 Johnson City     82.2     14.0
2   200 Morristown       81.5     14.4
3   301 Middlesborough   81.9     26.8
4   302 Knoxville        82.5     13.0
5   401 Winston-Salem    82.6     10.6
6   402 Martinsville     82.0     13.1
```

```
# Fitting a linear regression model
model<-lm(le_agg ~ pov_rate , data = cz_life_expect_pov)
show(model)
```

```
Call:
lm(formula = le_agg ~ pov_rate, data = cz_life_expect_pov)

Coefficients:
(Intercept)     pov_rate
    84.1113      -0.1091
```

```
# Normal plot with ggplot2
ggplot(cz_life_expect_pov, aes(x = pov_rate, y = le_agg)) +
  geom_point() +
  stat_smooth(formula = y ~ x, method = "lm", se = FALSE, colour = "red") +
  labs(x = "Poverty rate (percentage point) ", y = "Life expectancy (CZ-level)",title =
```




```
# Create quantiles based on poverty rate (pov_rate) and assign observations to bins
cz_life_expect_pov = cz_life_expect_pov %>% mutate(bin = ntile(pov_rate, n=20))

# Calculate mean values of life expectancy (le_agg) and poverty rate (pov_rate) for each
new_cz_life_expect_pov = cz_life_expect_pov %>% group_by(bin) %>% summarise(le_agg_mean =

# Create the binned scatter plot with a linear regression line
ggplot(new_cz_life_expect_pov, aes(x = pov_rate_mean, y = le_agg_mean)) +
  geom_point()+
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Poverty rate (percentage point) ", y = "Life expectancy (CZ-level)",title =
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

5.1.0.3 **Panel C. Intergenerational mobility for low-income children**

Panel C focuses on the impact of poverty rates on children's upward mobility, measured by the mean household income at ages 31-37 for children who grew up in each commuting zone and were born to low-income parents (at the 25th percentile of the income distribution). According to data from the Opportunity Atlas (Chetty et al. 2020a), a one percentage point increase in the poverty rate is associated with a $371 decrease in the expected adult income of these children.

```
# Panel C. Intergenerational mobility for low-income children

# Loading and inspecting the data
head(cz_upward_mobility_pov)
```

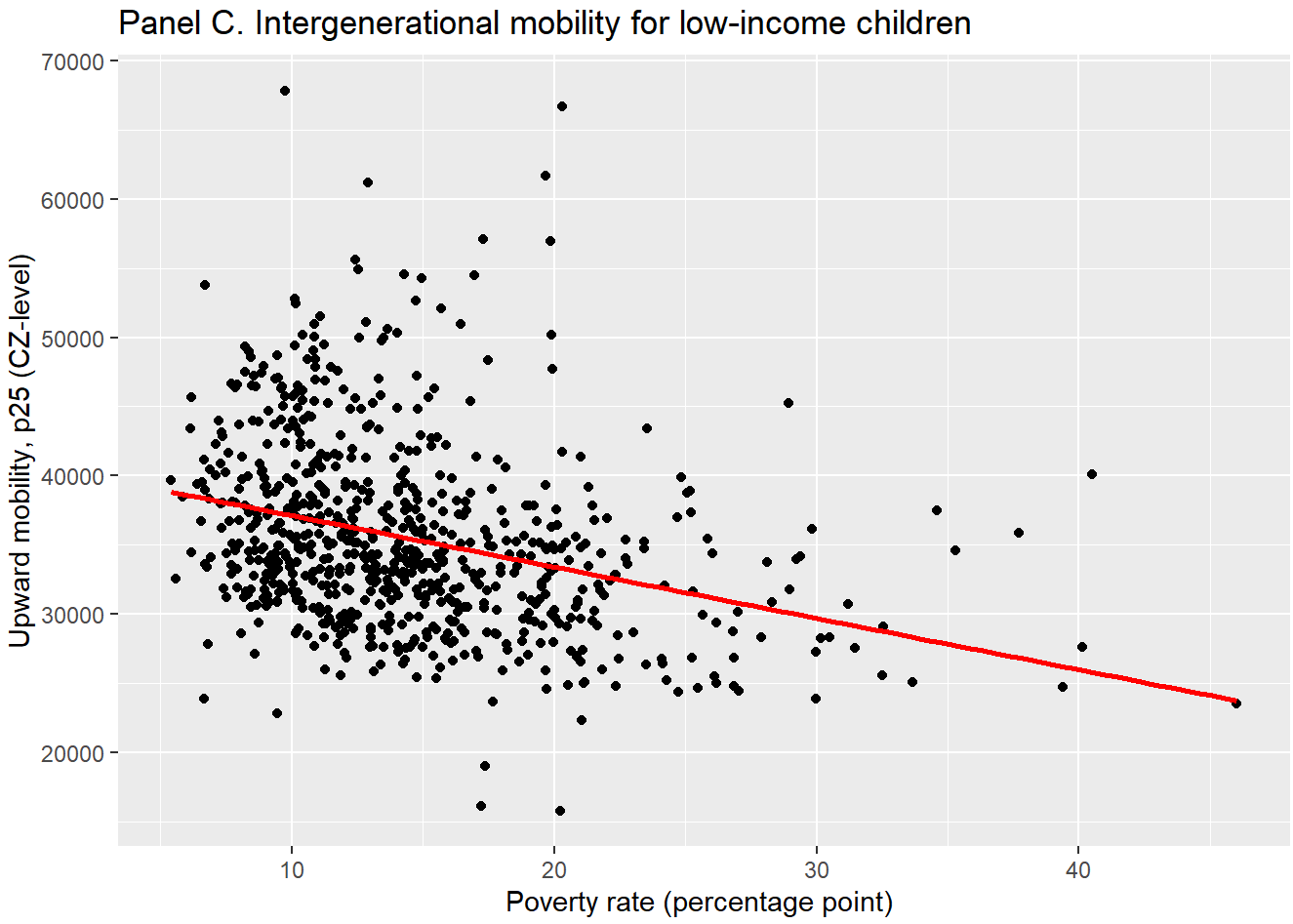
```
# A tibble: 6 × 4
     cz czname         kfr_pooled_p25_cz pov_rate
  <dbl> <chr>                      <dbl>    <dbl>
1   100 Johnson City              27755.     14.0
2   200 Morristown                27514.     14.4
3   301 Middlesborough            28746.     26.8
4   302 Knoxville                 29040.     13.0
5   401 Winston-Salem             28443.     10.6
6   402 Martinsville              27650.     13.1
```

```r
# Fitting a linear regression model
model<-lm(kfr_pooled_p25_cz ~ pov_rate , data = cz_upward_mobility_pov)
show(model)
```

```
Call:
lm(formula = kfr_pooled_p25_cz ~ pov_rate, data = cz_upward_mobility_pov)

Coefficients:
(Intercept)     pov_rate
    40840.0       -371.5
```

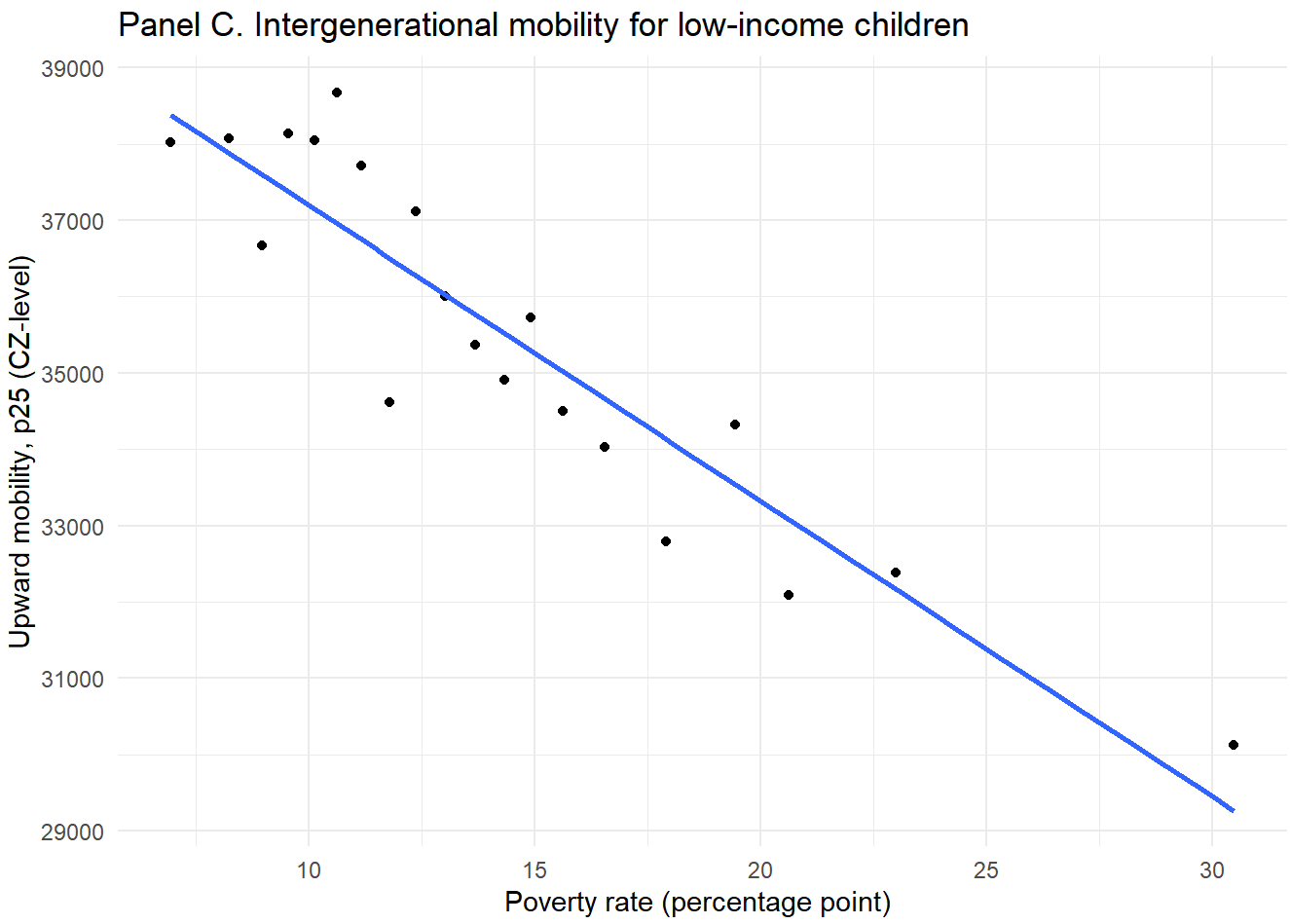
```r
# Normal plot with ggplot2
ggplot(cz_upward_mobility_pov, aes(x = pov_rate, y = kfr_pooled_p25_cz)) +
  geom_point() +
  stat_smooth(formula = y ~ x, method = "lm", se = FALSE, colour = "red", linetype = 1) +
  labs(x = "Poverty rate (percentage point) ", y = "Upward mobility, p25 (CZ-level)", ti
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```




```
# Create quantiles based on poverty rate (pov_rate) and assign observations to bins
cz_upward_mobility_pov = cz_upward_mobility_pov %>% mutate(bin = ntile(pov_rate, n=20))

# Calculate mean values of upward mobility (kfr_pooled_p25_cz) and poverty rate (pov_rat
new_cz_upward_mobility_pov = cz_upward_mobility_pov %>% group_by(bin) %>% summarise(kfr_

# Create the binned scatter plot with a linear regression line
ggplot(new_cz_upward_mobility_pov, aes(x = pov_rate_mean, y = kfr_pooled_p25_cz_mean)) +
  geom_point()+
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Poverty rate (percentage point) ", y = "Upward mobility, p25 (CZ-level)", ti
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 1 row containing non-finite outside the scale range (`stat_smooth()`).
Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



5.1.0.4 **Panel D. Academic achievement**

Panel D examines the effect of poverty rates on academic achievement for children, using mean standardized test scores for school districts from the Stanford Education Data Archive. The results show that a one percentage point increase in the poverty rate corresponds with a 0.025 standard deviation decrease in academic achievement.

```
# Panel D. Academic achievement

# Loading and inspecting the data
head(district_test_scores_pov)
```

```
# A tibble: 6 × 3
  leaidC  mn_avg_ol pov_rate
  <chr>       <dbl>    <dbl>
1 0100002    NA        NA
2 0100005    -0.291    39.4
3 0100006    -0.187    26.0
4 0100007     0.229     8.60
5 0100008     0.475     7.11
6 0100009    NA        NA
```

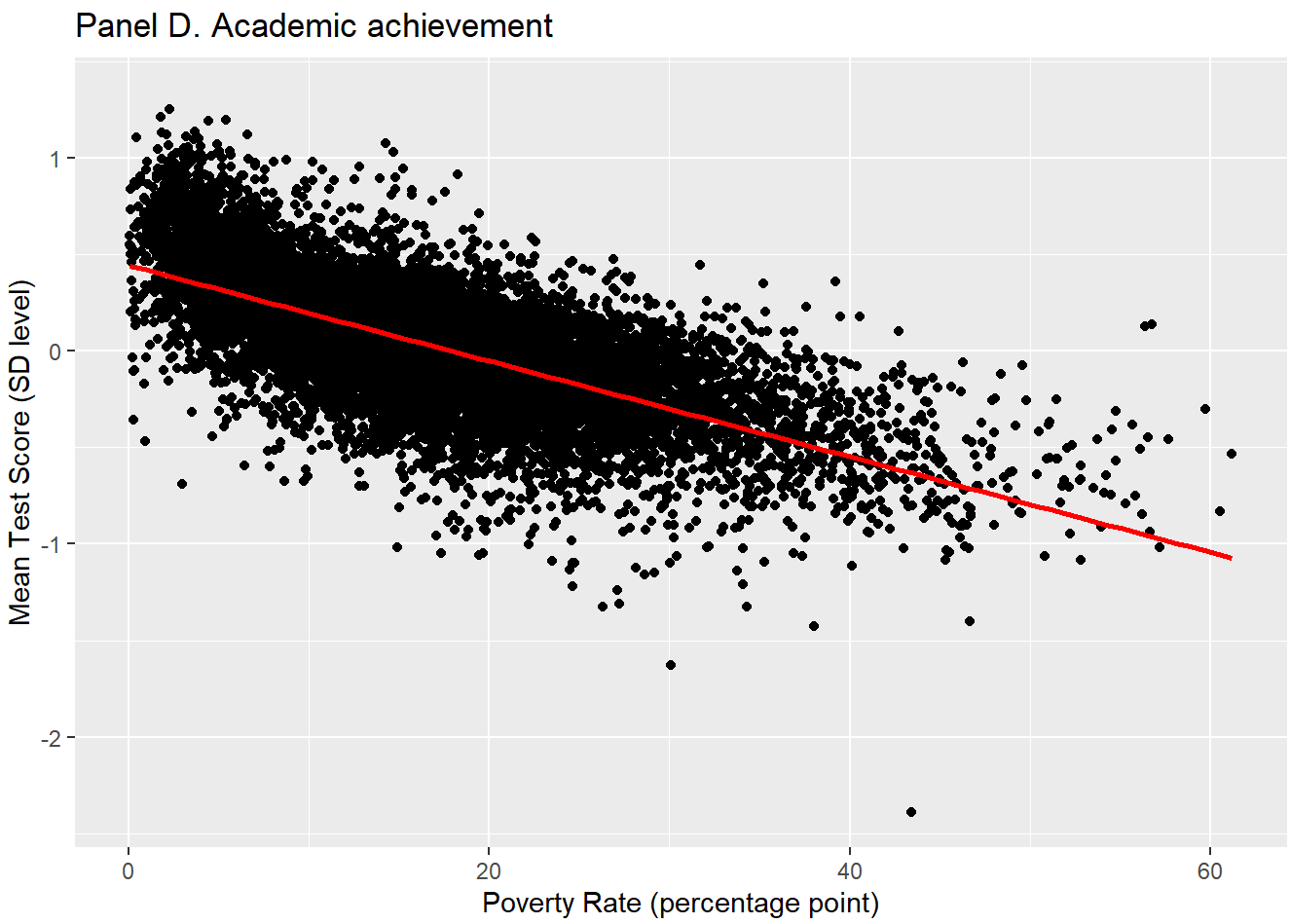
```
# Fitting a linear regression model
model<-lm(mn_avg_ol ~ pov_rate , data = district_test_scores_pov)
```

```
show(model)
```

```
Call:
lm(formula = mn_avg_ol ~ pov_rate, data = district_test_scores_pov)

Coefficients:
(Intercept)     pov_rate
     0.4451      -0.0248
```

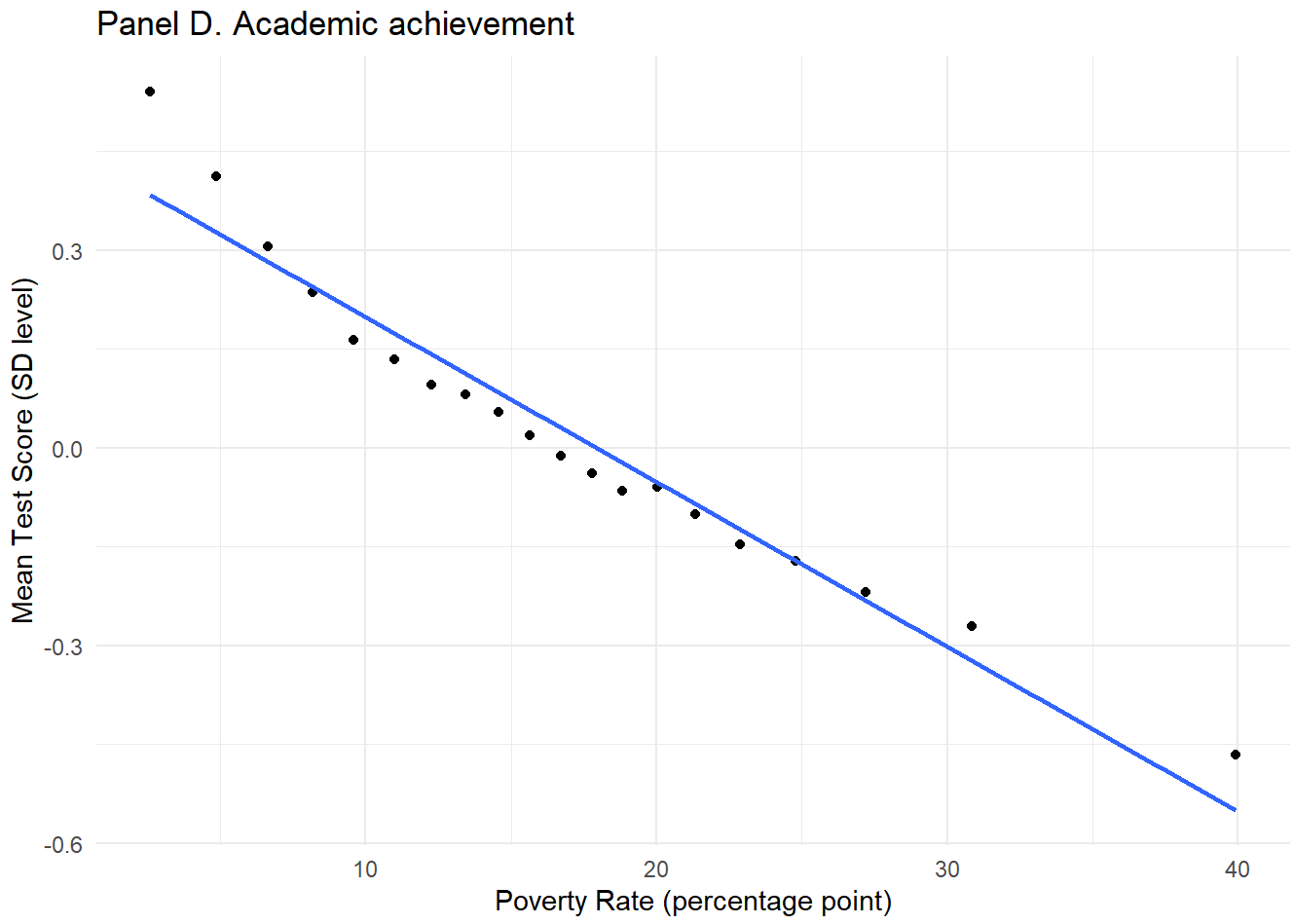
```
# Normal plot with ggplot2
ggplot(district_test_scores_pov, aes(x = pov_rate, y = mn_avg_ol)) +
  geom_point() +
  stat_smooth(formula = y ~ x, method = "lm", se = FALSE, colour = "red", linetype = 1) +
  labs(x = "Poverty Rate (percentage point)", y = "Mean Test Score (SD level)", title =
```

```
Warning: Removed 559 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 559 rows containing missing values or values outside the scale range
(`geom_point()`).
```



```
# Create quantiles based on poverty rate (pov_rate) and assign observations to bins
district_test_scores_pov = district_test_scores_pov %>% mutate(bin = ntile(pov_rate, n=2

# Calculate mean values of mean test score (mn_avg_ol) and poverty rate (pov_rate) for e
new_district_test_scores_pov = district_test_scores_pov %>% group_by(bin) %>% summarise(

# Create the binned scatter plot with a linear regression line
ggplot(new_district_test_scores_pov, aes(x = pov_rate_mean, y = mn_avg_ol_mean)) +
  geom_point()+
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Poverty Rate (percentage point)", y = "Mean Test Score (SD level)", title =
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



Overall, Figure 1 underscores the significant negative impact that higher poverty rates have on crucial outcomes for both adults and children, highlighting the importance of addressing poverty to improve employment, health, upward mobility, and educational performance.

5.2 Table1

5.3 Associations between Adult and Child Outcomes and Neighborhood Poverty

5.3.0.1 Tract Adult Employment

This code segment effectively visualizes how tract adult employment (measured by `emp2000`) varies with changes in poverty rates (`pov_rate`), both in a continuous scatter plot and a segmented binned scatter plot for deeper analysis.

```
# fifth dataset
#Panel E .tract adult employment

# Loading and inspecting the data
head(tract_adult_emp_pov)
```

```
# A tibble: 6 × 6
  tract county state emp2000 fips  pov_rate
  <dbl>  <dbl> <dbl>   <dbl> <chr>    <dbl>
1 20100      1     1   0.567 01001    12.7
2 20200      1     1   0.493 01001    22.7
3 20300      1     1   0.579 01001     7.66
4 20400      1     1   0.597 01001     4.55
5 20500      1     1   0.661 01001     3.68
6 20600      1     1   0.643 01001    15.2
```

```
# Fitting a linear regression model
model<-lm(emp2000 ~ pov_rate , data = tract_adult_emp_pov)
show(model)
```

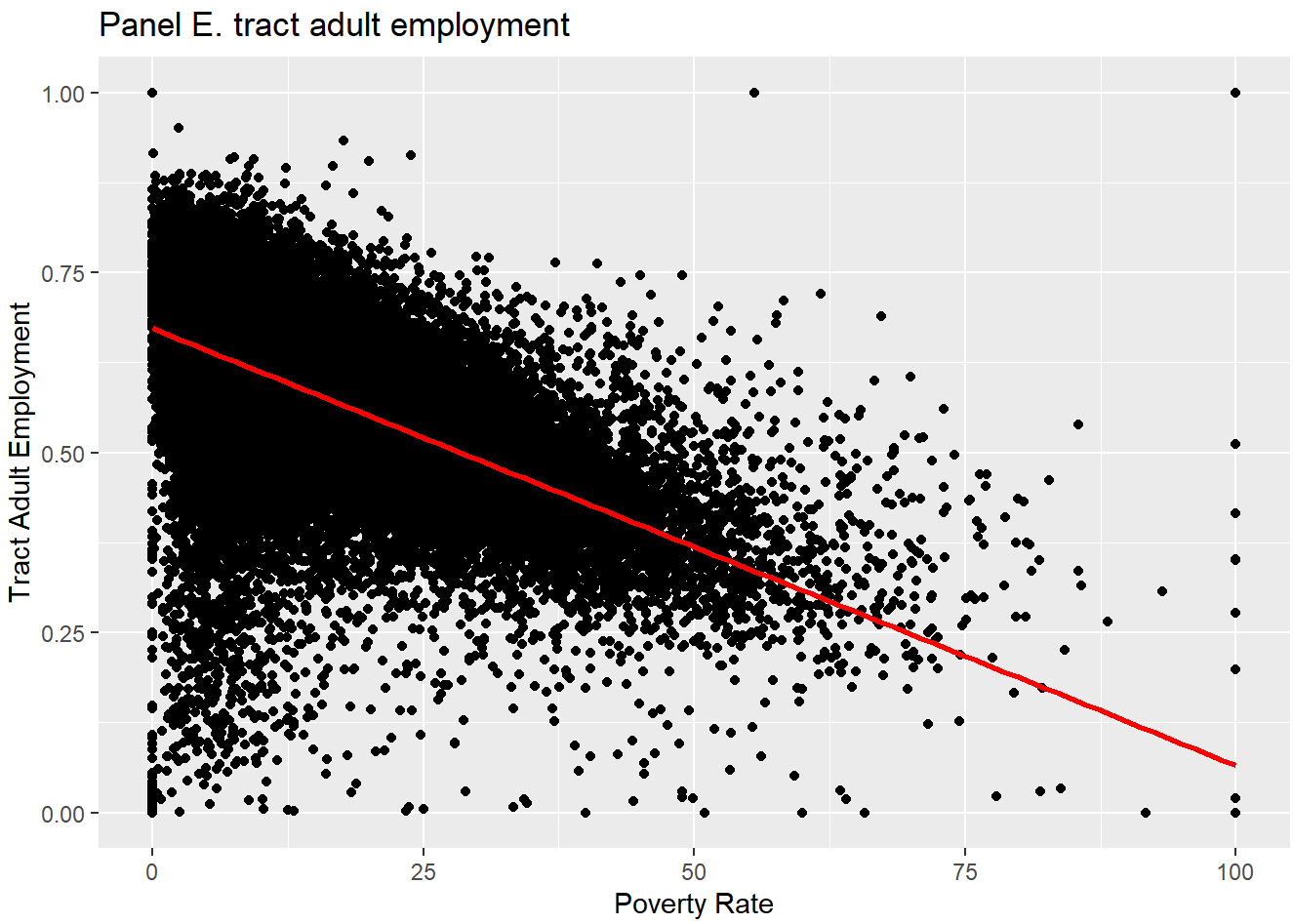
```
Call:
lm(formula = emp2000 ~ pov_rate, data = tract_adult_emp_pov)

Coefficients:
(Intercept)     pov_rate
   0.672416    -0.006069
```

```
# Normal plot with ggplot2
ggplot(tract_adult_emp_pov, aes(x = pov_rate, y = emp2000)) +
  geom_point() +
```

```
  stat_smooth(formula = y ~ x, method = "lm", se = FALSE, colour = "red", linetype = 1) +
  labs(x = "Poverty Rate", y = "Tract Adult Employment", title = "Panel E. tract adult em
```

```
Warning: Removed 1707 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

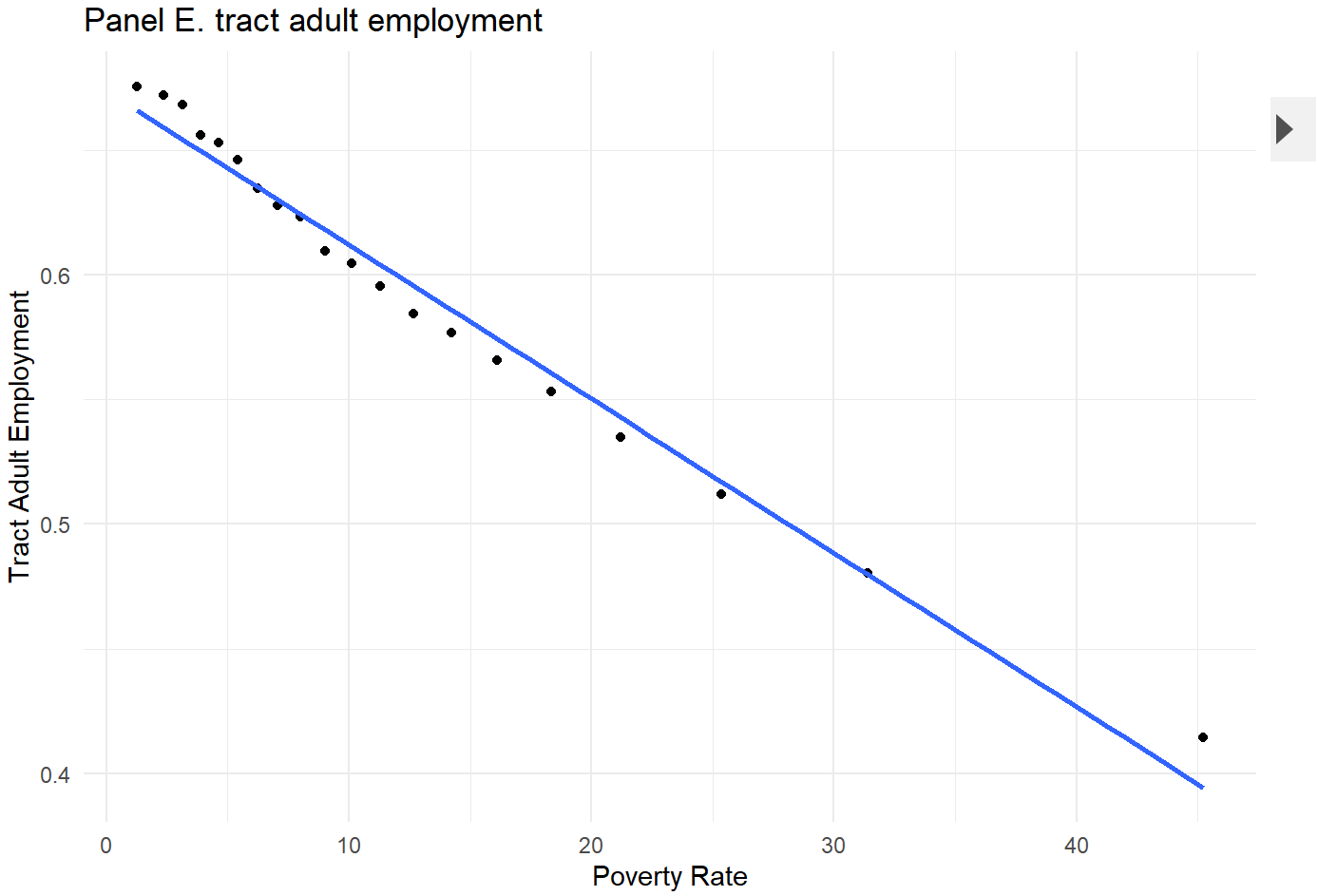
```
Warning: Removed 1707 rows containing missing values or values outside the scale range
(`geom_point()`).
```




```
# Create quantiles based on poverty rate (pov_rate) and assign observations to bins
tract_adult_emp_pov = tract_adult_emp_pov %>% mutate(bin = ntile(pov_rate, n=20))

# Calculate mean values of emp2000 and pov_rate for each bin
new_tract_adult_emp_pov = tract_adult_emp_pov %>% group_by(bin) %>% summarise(emp2000_me

# Create the binned scatter plot with a linear regression line
ggplot(new_tract_adult_emp_pov, aes(x = pov_rate_mean, y = emp2000_mean)) +
  geom_point()+
  geom_smooth(method = "lm", se = FALSE) +
```

```
            labs(x = "Poverty Rate", y = "Tract Adult Employment", title = "Panel E. tract adult e
            theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



5.3.0.2 **Tract Upward Mobility**

```
        #sixth dataset
        #Panel F. tract upward mobility

        # Loading and inspecting the data
        head(tract_upward_mobility_pov)
```

```
# A tibble: 6 × 6
  tract county state kfr_pooled_p25 fips  pov_rate
```

```
  <dbl>  <dbl> <dbl>          <dbl> <chr>    <dbl>
1 20100      1     1         27621. 01001    12.7
2 20200      1     1         22303. 01001    22.7
3 20300      1     1         28215. 01001     7.66
4 20400      1     1         33331. 01001     4.55
5 20500      1     1         34633. 01001     3.68
6 20600      1     1         23583. 01001    15.2
```

```r
# Fitting a linear regression model
model<-lm(kfr_pooled_p25 ~ pov_rate , data = tract_upward_mobility_pov)
show(model)
```

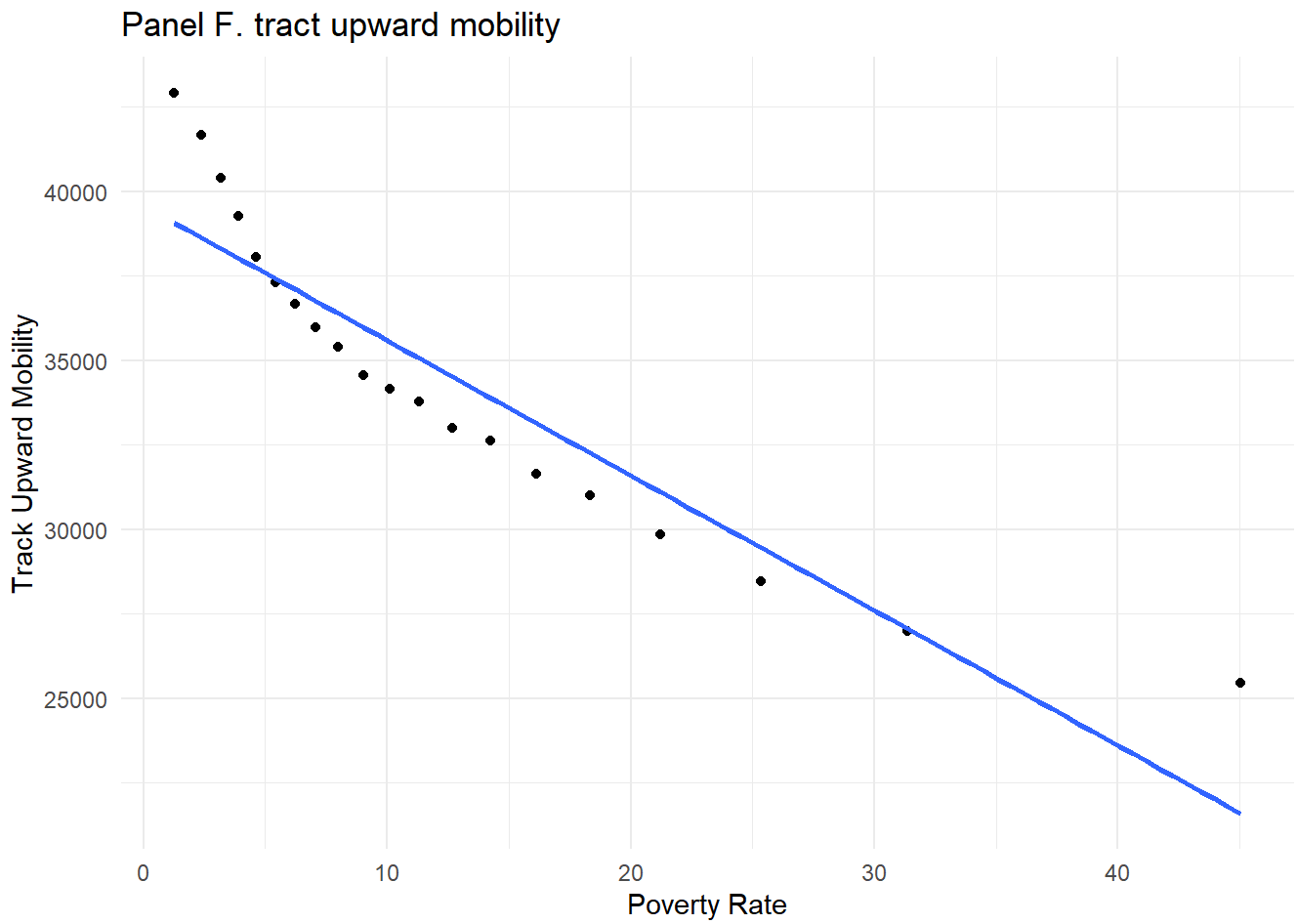
```
Call:
lm(formula = kfr_pooled_p25 ~ pov_rate, data = tract_upward_mobility_pov)

Coefficients:
(Intercept)     pov_rate
    39465.0       -391.5
```

```r
# Normal plot with ggplot2
ggplot(tract_upward_mobility_pov, aes(x = pov_rate, y = kfr_pooled_p25)) +
  geom_point() +
  stat_smooth(formula = y ~ x, method = "lm", se = FALSE, colour = "red", linetype = 1) +
  labs(x = "Poverty Rate", y = "Track Upward Mobility", title = "Panel F. tract upward m
```

```
Warning: Removed 1355 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 1355 rows containing missing values or values outside the scale range
(`geom_point()`).
```


```
# Binned scatter plot with binsreg
# Create quantiles based on poverty rate (pov_rate) and assign observations to bins
tract_upward_mobility_pov = tract_upward_mobility_pov %>% mutate(bin = ntile(pov_rate, n

# Calculate mean values of kfr_pooled_p25 and pov_rate for each bin
new_tract_upward_mobility_pov = tract_upward_mobility_pov %>% group_by(bin) %>% summaris

# Create the binned scatter plot with a linear regression line
ggplot(new_tract_upward_mobility_pov, aes(x = pov_rate_mean, y = kfr_pooled_p25_mean)) +
  geom_point()+
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Poverty Rate", y = "Track Upward Mobility", title = "Panel F. tract upward m
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



5.3.0.3 **District Test Score**

```
#seventh datasets
#Panel G. district test scores

# Loading and inspecting the data
head(district_test_scores_pov)
```

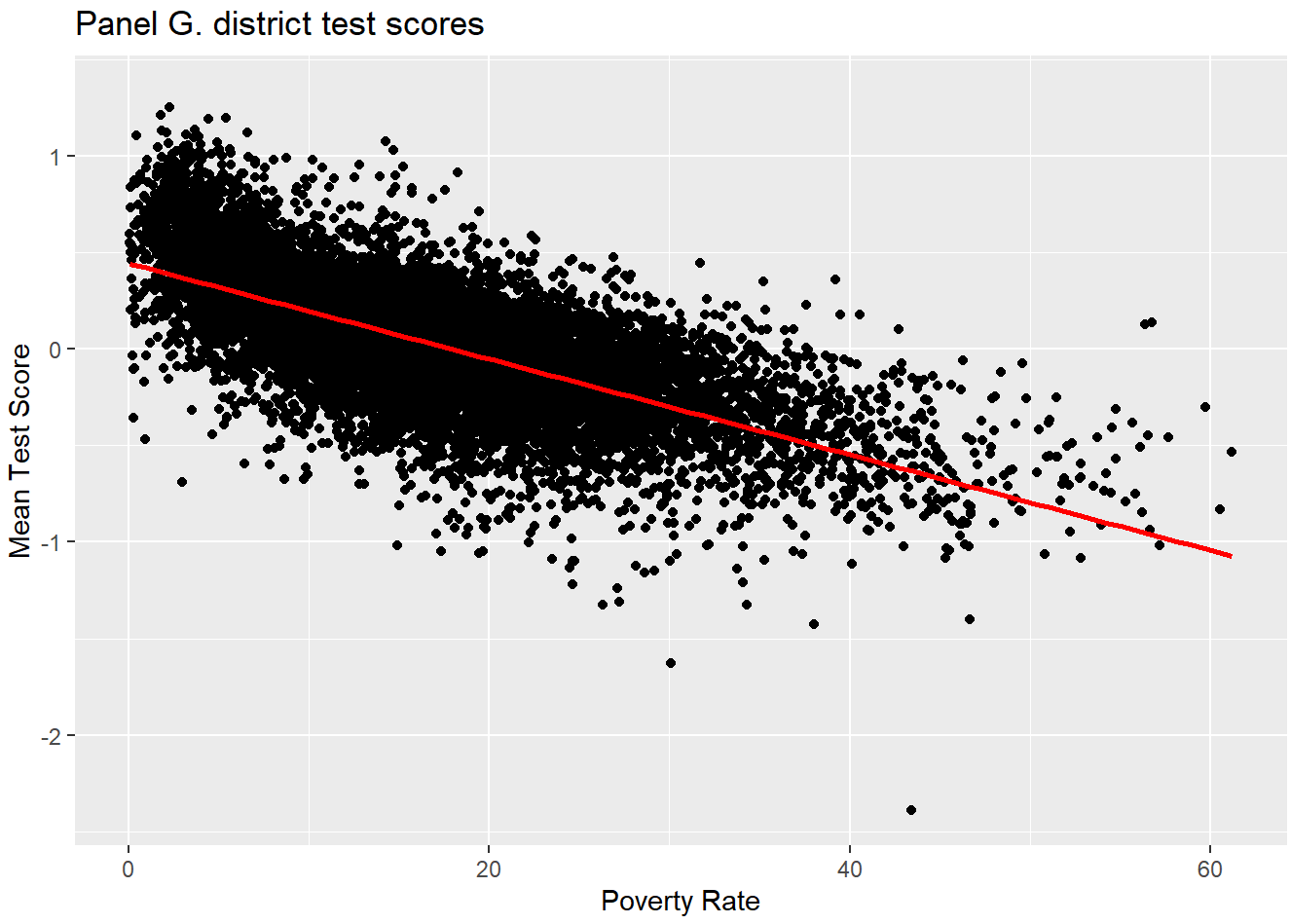
```
# A tibble: 6 × 4
  leaidC  mn_avg_ol pov_rate   bin
  <chr>       <dbl>    <dbl> <int>
1 0100002    NA        NA       NA
2 0100005    -0.291    39.4     20
3 0100006    -0.187    26.0     18
4 0100007     0.229     8.60     4
5 0100008     0.475     7.11     3
6 0100009    NA        NA       NA
```

```
# Fitting a linear regression model
model<-lm(mn_avg_ol ~ pov_rate , data = district_test_scores_pov)
show(model)
```

```
Call:
lm(formula = mn_avg_ol ~ pov_rate, data = district_test_scores_pov)

Coefficients:
(Intercept)     pov_rate
     0.4451      -0.0248
```

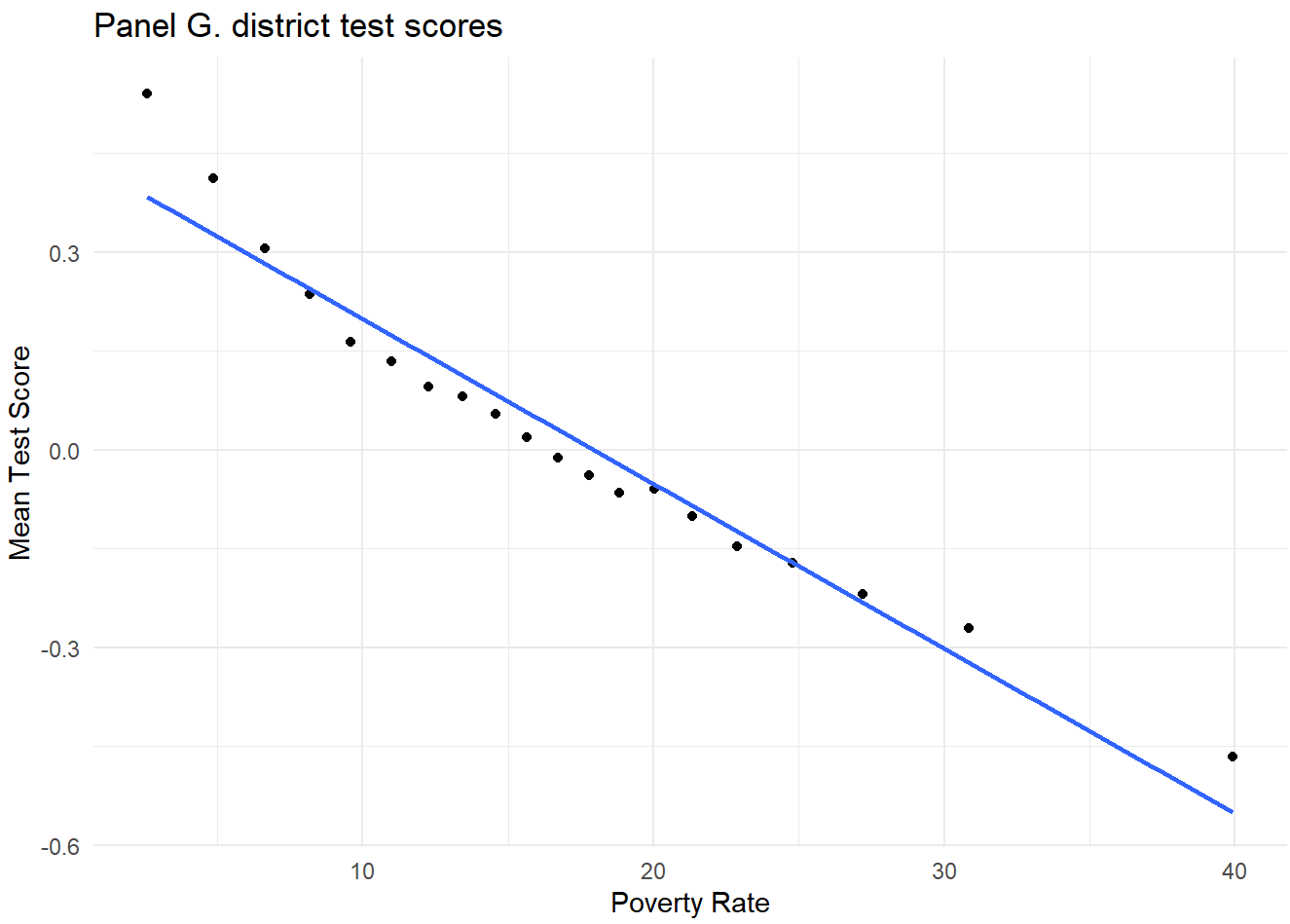
```
# Normal plot with ggplot2
ggplot(district_test_scores_pov, aes(x = pov_rate, y = mn_avg_ol)) +
  geom_point() +
  stat_smooth(formula = y ~ x, method = "lm", se = FALSE, colour = "red", linetype = 1) +
  labs(x = "Poverty Rate", y = "Mean Test Score", title = "Panel G. district test scores
```

```
Warning: Removed 559 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 559 rows containing missing values or values outside the scale range
(`geom_point()`).
```



```r
# Binned scatter plot with binsreg
# Create quantiles based on poverty rate (pov_rate) and assign observations to bins
district_test_scores_pov = district_test_scores_pov %>% mutate(bin = ntile(pov_rate, n=2

# Calculate mean values of mn_avg_ol and pov_rate for each bin
new_district_test_scores_pov = district_test_scores_pov %>% group_by(bin) %>% summarise(

# Create the binned scatter plot with a linear regression line
ggplot(new_district_test_scores_pov, aes(x = pov_rate_mean, y = mn_avg_ol_mean)) +
  geom_point()+
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Poverty Rate", y = "Mean Test Score", title = "Panel G. district test scores
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



5.3.0.4 Critique of Bin Scatter Plot Usage

While bin scatter plots can simplify visualization, they may also obscure important details. Here, I'll outline why displaying all 700 observations, despite being crowded, is preferable to using bin scatter plots in this context.

5.3.0.5 The Issue with Bin Scatter Plots

Bin scatter plots average data within each bin, often removing outliers. This can make the relationship appear cleaner and more convincing than it truly is. By averaging, bin scatter plots might suggest a perfect explanation of the data, which is misleading. The process of binning cuts out valid information. Outliers and data variability, which can be critical for understanding the true nature of the data, are lost. The averages might align closely with the regression line, giving a false impression of a higher R-squared value.

While the raw data plot is crowded, it provides a complete picture, including variability and outliers. Simplifying visualization with bins is not necessary when the data can be displayed in full. This preserves the integrity of the information presented.

5.3.0.6 Importance of R-squared

In our analysis, the R-squared value is around 0.50. This means only 50% of the variation is explained by the poverty rate, with the remaining 50% unexplained. An R-squared of 1 indicates that all data points lie exactly on the regression line, while an R-squared of 0 suggests no explanatory power. The proximity of data points to the regression line influences the R-squared value. The closer the points, the higher the R-squared. In bin scatter plots, the apparent closeness of data points to the line might inflate the perceived explanatory power, which is not the case with raw data plots.

5.3.0.7 Recommendations

If necessary, averages can be calculated within each poverty rate category manually or using functions in R. This should be done transparently to highlight the presence of outliers and variability. Presenting raw data, even if crowded, is crucial for maintaining transparency and integrity. Visual clutter can be managed with thoughtful use of plotting techniques, ensuring that important data points and variations are not lost.

Although the paper briefly mentions the use of binned scatter plots, it is also beneficial to utilize the Stata solution provided with the replication package as a reference. Employing Stata can provide a clearer understanding of the methodology they used, which in turn can streamline the replication process in R. This approach can be particularly helpful for reducing the time and effort required to achieve similar results in R.

5.4 Table1

```r
# Table 1, Column 1: Employment Rate (CZ)
model1 <- lm(emp2000 ~ pov_rate, data = cz_adult_emp_pov)

summary(model1)
```

```
Call:
```

```
lm(formula = emp2000 ~ pov_rate, data = cz_adult_emp_pov)

Residuals:
      Min        1Q    Median        3Q       Max
-0.159547 -0.023766  0.000398  0.025363  0.166849

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6969800  0.0043892  158.79   <2e-16 ***
pov_rate    -0.0082076  0.0002825  -29.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04371 on 739 degrees of freedom
Multiple R-squared:  0.5331,    Adjusted R-squared:  0.5325
F-statistic: 843.9 on 1 and 739 DF,  p-value: < 2.2e-16
```

```r
mean_emp2000 <- mean(cz_adult_emp_pov$emp2000, na.rm = TRUE)


# Table 1, Column 2: Life Expectancy (CZ)
model2 <- lm(le_agg ~ pov_rate, data = cz_life_expect_pov)

summary(model2)
```

```
Call:
lm(formula = le_agg ~ pov_rate, data = cz_life_expect_pov)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3316 -0.7548 -0.0786  0.6652  3.8920

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.111297   0.119402  704.44   <2e-16 ***
pov_rate    -0.109057   0.007946  -13.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.015 on 593 degrees of freedom
Multiple R-squared:  0.2411,    Adjusted R-squared:  0.2398
F-statistic: 188.4 on 1 and 593 DF,  p-value: < 2.2e-16
```

```r
mean_le_agg <- mean(cz_life_expect_pov$le_agg, na.rm = TRUE)


# Table 1, Column 3: Upward Mobility (CZ)
model3 <- lm(kfr_pooled_p25_cz ~ pov_rate, data = cz_upward_mobility_pov)
```

```
summary(model3)
```

```
Call:
lm(formula = kfr_pooled_p25_cz ~ pov_rate, data = cz_upward_mobility_pov)

Residuals:
   Min     1Q Median     3Q    Max
-18366  -4241  -1114   3059  33404

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40839.97     650.66  62.767   <2e-16 ***
pov_rate     -371.49      41.88  -8.869   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6480 on 739 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.09621,   Adjusted R-squared:  0.09499
F-statistic: 78.67 on 1 and 739 DF,  p-value: < 2.2e-16
```

```
mean_kfr_pooled_p25_cz <- mean(cz_upward_mobility_pov$kfr_pooled_p25_cz, na.rm = TRUE)


# Table 1, Column 4: Test Scores (School Districts)
model4 <- lm(mn_avg_ol ~ pov_rate, data = district_test_scores_pov)

summary(model4)
```

```
Call:
lm(formula = mn_avg_ol ~ pov_rate, data = district_test_scores_pov)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7568 -0.1467  0.0067  0.1553  1.1025

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4450659  0.0046465   95.78   <2e-16 ***
pov_rate    -0.0247970  0.0002413 -102.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2469 on 12599 degrees of freedom
  (559 observations deleted due to missingness)
Multiple R-squared:  0.4559,    Adjusted R-squared:  0.4559
F-statistic: 1.056e+04 on 1 and 12599 DF,  p-value: < 2.2e-16
```

```r
mean_mn_avg_ol <- mean(district_test_scores_pov$mn_avg_ol, na.rm = TRUE)

# Table 1, Column 5: Employment Rate (Tract)
model5 <- lm(emp2000 ~ pov_rate, data = tract_adult_emp_pov)

summary(model5)
```

```
Call:
lm(formula = emp2000 ~ pov_rate, data = tract_adult_emp_pov)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67242 -0.04165  0.00640  0.05222  0.93447

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.724e-01  5.130e-04  1310.7   <2e-16 ***
pov_rate    -6.069e-03  3.031e-05  -200.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08983 on 72414 degrees of freedom
  (1707 observations deleted due to missingness)
Multiple R-squared:  0.3563,    Adjusted R-squared:  0.3563
F-statistic: 4.009e+04 on 1 and 72414 DF,  p-value: < 2.2e-16
```

```r
mean_tract_emp2000 <- mean(tract_adult_emp_pov$emp2000, na.rm = TRUE)


# Table 1, Column 6: Upward Mobility (Tract)
model6 <- lm(kfr_pooled_p25 ~ pov_rate, data = tract_upward_mobility_pov)

summary(model6)
```

```
Call:
lm(formula = kfr_pooled_p25 ~ pov_rate, data = tract_upward_mobility_pov)

Residuals:
   Min     1Q Median     3Q    Max
-36504  -4746   -828   3729  68438

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39464.981     40.186   982.1   <2e-16 ***
pov_rate     -391.460      2.396  -163.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6964 on 71921 degrees of freedom
  (1355 observations deleted due to missingness)
Multiple R-squared:  0.2706,    Adjusted R-squared:  0.2706
F-statistic: 2.668e+04 on 1 and 71921 DF,  p-value: < 2.2e-16
```

```
mean_tract_kfr_pooled_p25 <- mean(tract_upward_mobility_pov$kfr_pooled_p25, na.rm = TRUE
```

The code snippet you've provided fits linear models with county fixed effects using the `felm` function from the `lfe` package and obtains clustered standard errors for the models. By using `felm` and `vcovCL` from the `lfe` package, you're able to fit linear models with fixed effects and calculate clustered standard errors, which are robust against heteroscedasticity and serial correlation within clusters (counties, in this case). These steps ensure that your model estimations account for potential clustering effects within counties, providing more reliable standard errors for inference.

```
if (!require(lfe)) install.packages("lfe")
```

```
Loading required package: lfe
```

```
Loading required package: Matrix
```

```
library(lfe)

# Fit the models with county fixed effects using lfe
model5_cluster <- felm(emp2000 ~ pov_rate | county | 0 | county, data = tract_adult_emp_
model6_cluster <- felm(kfr_pooled_p25 ~ pov_rate | county | 0 | county, data = tract_upw

if (!require(sandwich)) install.packages("sandwich")
```

```
Loading required package: sandwich
```

```
library(sandwich)
# Obtain clustered standard errors
cluster_se_model5 <- sqrt(diag(vcovCL(model5_cluster, cluster = ~ county)))
cluster_se_model6 <- sqrt(diag(vcovCL(model6_cluster, cluster = ~ county)))
```

The `stargazer` package in R is commonly used to create well-formatted regression tables from various model objects.

Standard Errors (`se`): If you have clustered standard errors (`clustered_se_5` and `clustered_se_6`), you can uncomment and use the `se` argument to include them in the table.

- `stargazer()` function is used to generate the regression table.
- Arguments:

- `model1, model2, ..., model6_cluster`: These are the regression model objects (`lm` or `felm`) that you want to include in the table.
- `type = "text"`: Specifies that the output format should be plain text.
- `title`: Title of the table.
- `dep.var.labels`: Labels for the dependent variables.
- `covariate.labels`: Label for the covariate.
- `add.lines`: Additional lines to add to the table. Here, it includes the level of analysis, controls, and mean values for each outcome.
- `omit.stat`: Specifies which statistics to omit from the table (in this case, F-statistic, standard error of the regression, and adjusted R-squared).
- `no.space = TRUE`: Removes extra spaces in the output for a cleaner appearance.

```
if (!require(stargazer)) install.packages("stargazer")
```

```
Loading required package: stargazer


Please cite as:

 Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

 R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(stargazer)

# Create Table 1
stargazer(
  model1, model2, model3, model4, model5_cluster, model6_cluster,
  type = "text",
  title = "Table 1: Associations between Adult and Child Outcomes and Neighborhood Pover
  dep.var.labels = c("Adult employment Rate(2000)", "Life expectancy", "Upward Mobility,
  covariate.labels = c("Poverty Rate"),
  add.lines = list(
    c("Level of Analysis", "CZ", "CZ", "CZ", "School District", "Tract", "Tract"),
    c("Controls", "None", "None", "None", "None", "County FE ", "County FE "),
    c("Mean", round(mean_emp2000, 3), round(mean_le_agg, 3), round(mean_kfr_pooled_p25_c
  ),
  #se = list(NULL, NULL, NULL, NULL, clustered_se_5, clustered_se_6),
  omit.stat = c("f", "ser", "adj.rsq"), #This indicates that the F-statistic, standard e
  no.space = TRUE
)
```

Table 1: Associations between Adult and Child Outcomes and Neighborhood Poverty

	Dependent variable:			
	Adult employment Rate(2000)	Life expectancy	Upward Mobility, p25 parents	Test-based achievement
	OLS	OLS	OLS	OLS
	felm	felm		
	(1)	(2)	(3)	(4)
	(5)	(6)		
Poverty Rate	-0.008***	-0.109***	-371.486***	-0.025***
	-0.006***	-393.294***		
	(0.0003)	(0.008)	(41.884)	(0.0002)
	(0.0001)	(12.073)		
Constant	0.697***	84.111***	40,839.970***	0.445***
	(0.004)	(0.119)	(650.664)	(0.005)
Level of Analysis	CZ	CZ	CZ	School District
	Tract	Tract		
Controls	None	None	None	None
	County FE	County FE		
Mean	0.578	82.575	35469.045	0.018
	0.594	34443.482		
Observations	741	595	741	12,601
	72,416	71,923		
R2	0.533	0.241	0.096	0.456
	0.388	0.325		

Note:
*p<0.1; **p<0.05; ***p<0.01

This table presents regression estimates where the dependent variable is a measure of adult or child outcomes within a specific geographic area, such as commuting zones (CZ), school districts, or census tracts. The independent variable of interest is the area's poverty rate, defined as the fraction of residents living below the poverty line. Here is a summary of the data sources and the outcomes measured in each column:

- **Columns 1 and 5:** Dependent variables are measures from the 2000 Decennial Census.
- **Column 2:** Life expectancy measure from Chetty et al. (2016a,b) based on Social Security Administration mortality data.

- **Columns 3 and 6:** “Upward Mobility” measure from the Opportunity Atlas (Chetty et al. 2020a,b), representing the mean household income rank at ages 31-37 for children whose parents were at the 25th percentile of the national income distribution.

- **Column 4:** Test-based achievement measure from the Stanford Education Data Archive (SEDA), which estimates mean test scores on a cohort-standardized scale using data from the National Assessment of Educational Progress (NAEP) as detailed in Fahle et al. (2019).

The poverty rates are primarily sourced from the 2000 Decennial Census, except for Column 4, which uses poverty rates averaged over 2007–2016 from the American Community Survey. Standard errors in Columns 5 and 6 are clustered at the county level.

The key findings highlight the relationship between higher poverty rates and poorer outcomes, such as lower adult employment rates, reduced life expectancy, decreased upward mobility, and lower academic achievement. Notably, the correlations persist even when controlling for county fixed effects, indicating that the observed associations are not merely due to broad differences across metropolitan areas but are evident at more granular levels like census tracts.

6 GitHub

https://mahsaabdollahim.github.io/Data_Science_Presentation

Uploading Files to GitHub Repository To ensure the reproducibility and accessibility of our project’s data and code, we have uploaded all relevant files to a publicly accessible GitHub repository. The following steps outline the procedure used to upload these files:

Repository Creation: 1. A new repository was created on GitHub to host the project’s files. This was accomplished by navigating to the GitHub website (https://github.com) and logging into the associated account. A new repository was then created and was named as MahsaAbdollahiM.github.io and initialized with a README file.

2. Local Repository Cloning: The newly created repository was cloned to a local machine to facilitate file management and version control. This was done using the Git command-line interface with the following commands:

git clone https://github.com/MahsaAbdollahiM/MahsaAbdollahiM.github.io.git cd MahsaAbdollahiM.github.io

3. Adding Files to the Repository: The project files were then copied into the local repository directory. These files include all data sets, analysis scripts, and supplementary materials necessary for replication of the study. The files were staged for commit using the following command:

git add .

4. Committing and Pushing Changes: A commit was created to record the addition of these files to the repository, accompanied by a descriptive message. The commit was then pushed to the remote repository on GitHub. This was executed with the following commands:

git commit -m “Initial commit with project files” git push origin main Verification: The presence of the uploaded files was verified by accessing the GitHub repository through a web browser at the following URL: https://github.com/MahsaAbdollahiM/MahsaAbdollahiM.github.io/tree/main. This step ensured that all files were correctly uploaded and accessible.

7 References

Cattaneo, M.D., Crump, R.K., Farrell, M.H. and Feng, Y., 2024. Online Appendix On Binscatter.

Lost Stats. (n.d.). *Binned scatter plots and binning*. Retrieved from https://lost-stats.github.io/Presentation/Figures/binscatter.html .

8 Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Addi tionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and author ship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published. I have read the Handbook of Academic Writing by Hildebrandt and Nelke [2019] and have endeavored to comply with the guidelines and standards set forth therein. I acknowledge that the university may use plagiarism detection software to check my thesis. I agree to cooperate with any investigation of suspected plagiarism and to provide any additional information or evidence requested by the university. The report includes: □ About 4000 words (+/- 500). □ A title page with personal details (name, email, matriculation number). □ An abstract. □ A bibliography, created using BibTeX with APA citation style. □ The complete R code required to reproduce the results. □ Detailed instructions on data acquisition and importation into R. □ An introduction to guide the reader and a conclusion summarizing the work and discussing potential future extensions. □ All significant resources used in the report and R code development. □ The filled out Affidavit. □ A concise description of the successful use of Git and GitHub, as detailed here: make_a_pull_request. □ A concise description of the presentation published on GitHub. The project submission includes: □ The .qmd file(s) of the report. □ The _quarto.yml file of the report. □ The .pdf file of the report. □ The standalone .html file of the report. □ All necessary files (not available online) to reproduce the report and the R code. □ The standalone .html file of the presentation.

Mahsa Abdollahi Mirzanagh, Cologne, 22.07.2024