EC 425/525: Econometrics

Professor Eric Zou

Linear Regression Review

Multiple Linear Regression

- Linear models can be used to estimate the relationship between an outcome Y and a single explanatory variable X. However, often we have multiple explanatory variables $(X_1, X_2, \dots, X_p)$.

- Including multiple explanatory variables in the model can be interesting for multiple reasons:

  1. Prediction: If we are trying to predict Y, using multiple explanatory variables often (but not always!) provides a more accurate prediction than a model with a single explanatory variable.

  2. Inference: We often want to know how a particular explanatory variable X affects Y, but there may be other factors also affecting both X and Y that confound that relationship. A multiple linear regression allows us to "control" for other variables (i.e., holding other variables constant) so that we can isolate the relationship between X and Y alone.

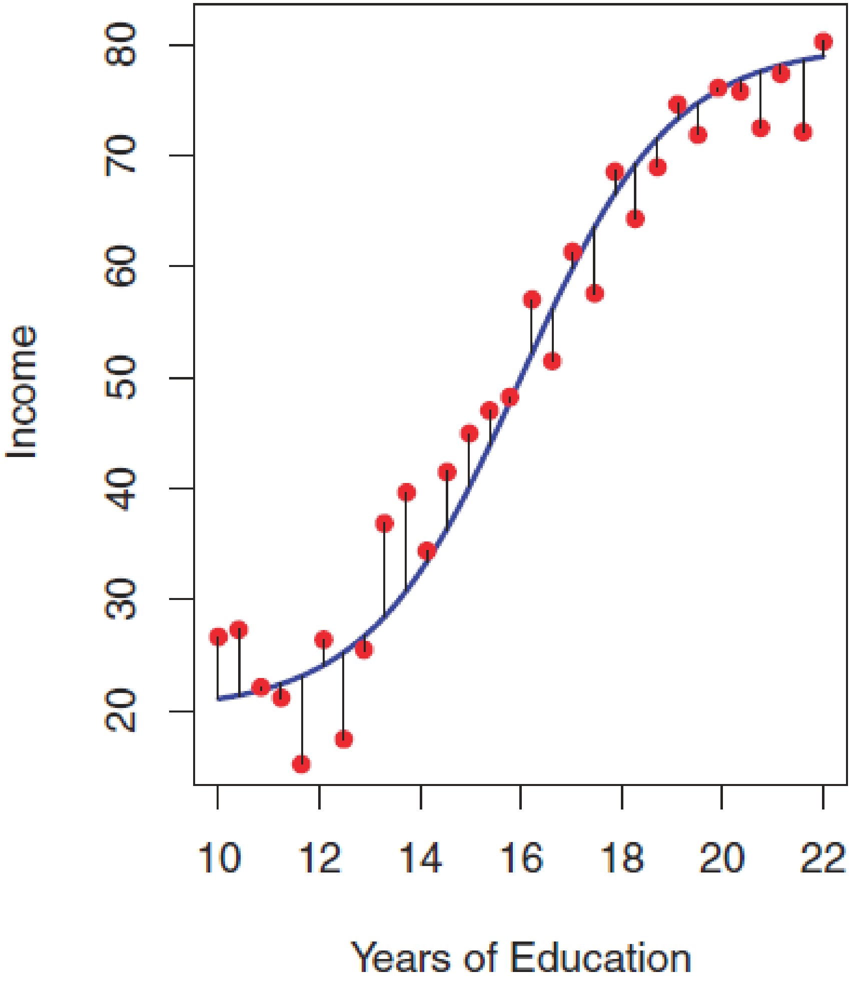
Multiple Linear Regression

- Suppose we are interested in how income is determined by factors such as education, seniority

- Input variables (typically denoted by X)
  - X_1: Education
  - X_2: Seniority
  - May also be called *predictors, independent variables, features*

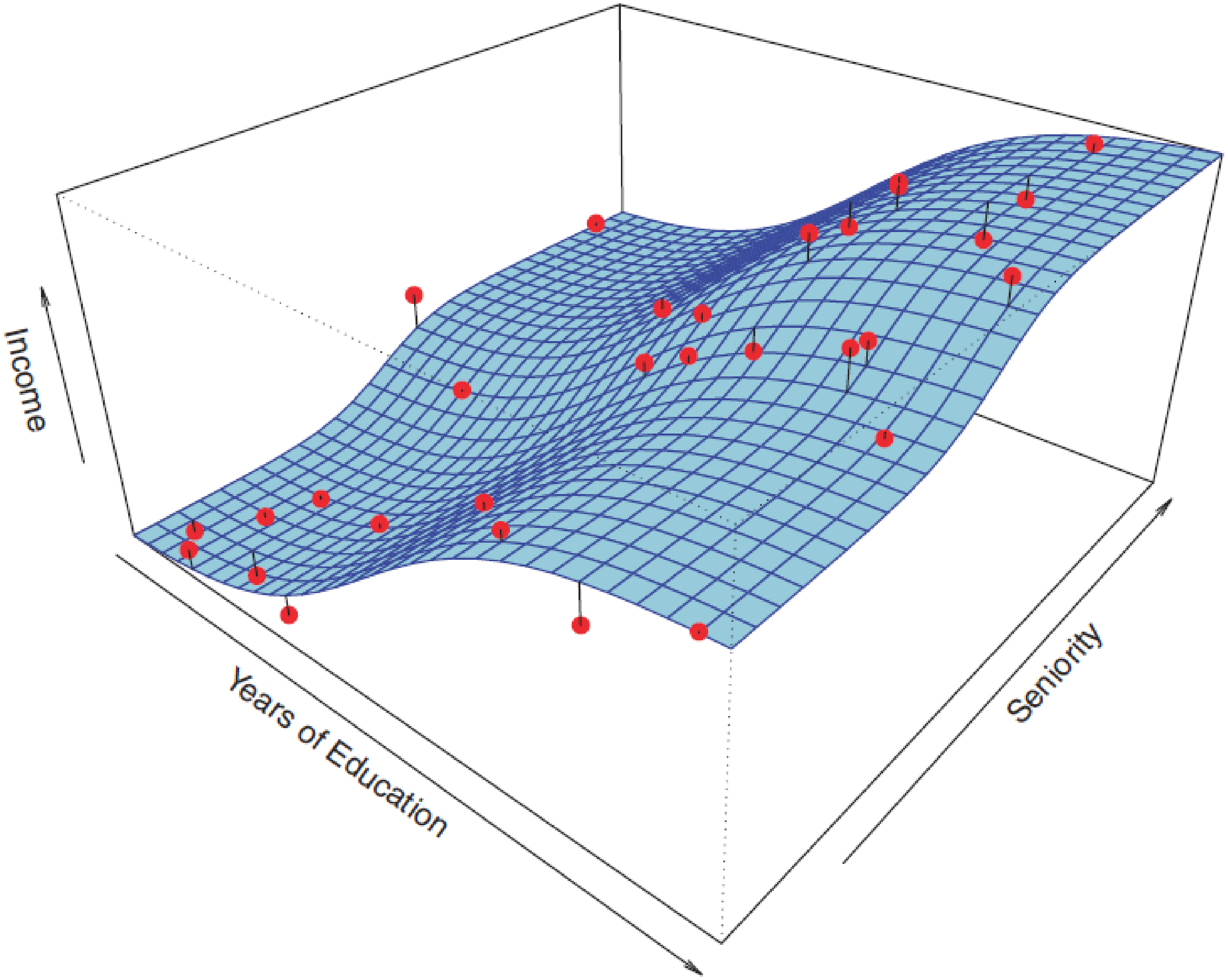
- Output variable (typically denoted by Y)
  - Y: Income
  - Often referred to as the *response* or *dependent variable*

Outcome vs. Single Predictor



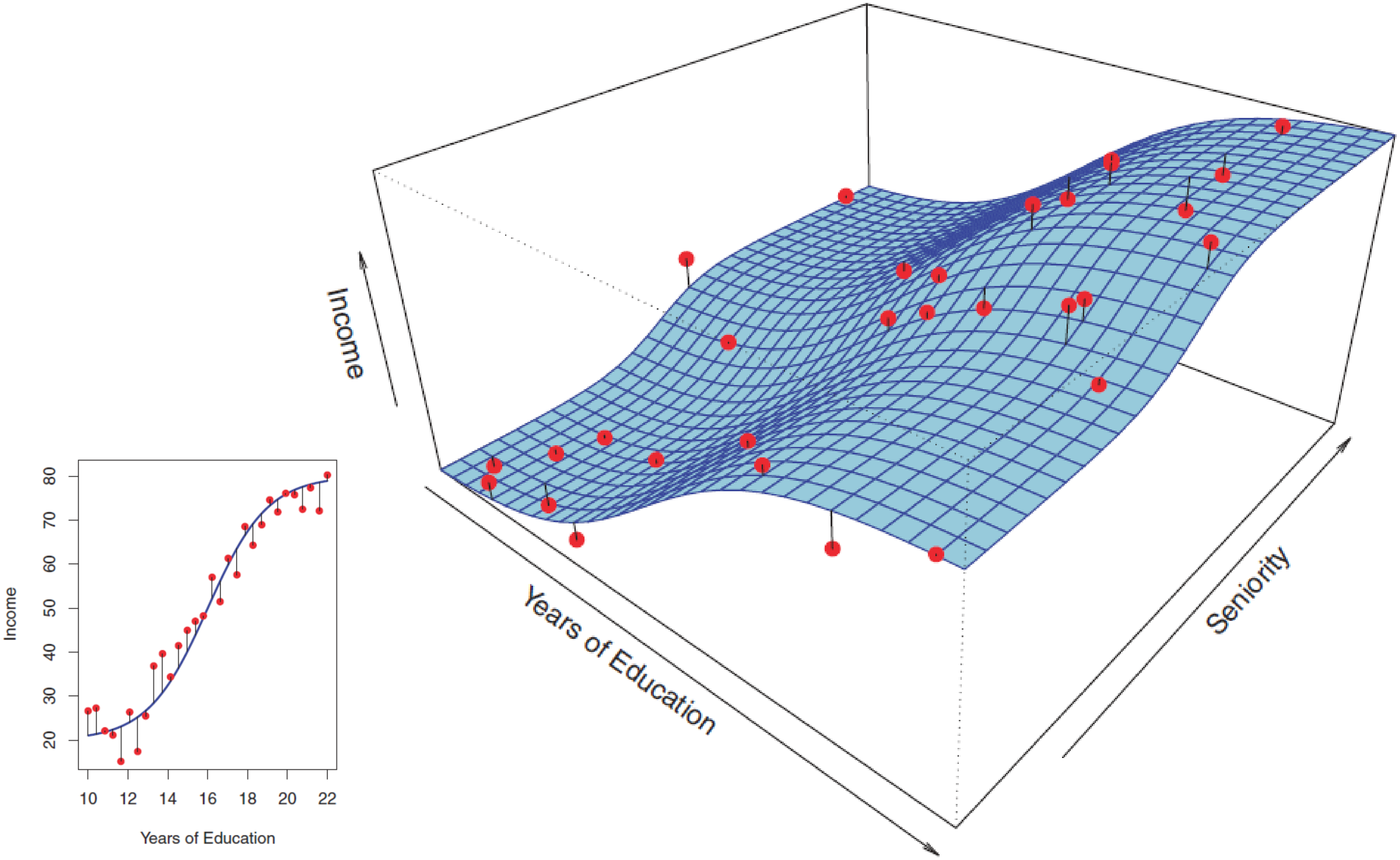
Source: ISL, Figure 2.2

Outcome vs. Multiple Predictors



Source: ISL, Figure 2.3

Outcome vs. Multiple Predictors



Source: ISL, Figure 2.3

Practice Example

- We will use multiple linear regression to analyze the Boston Housing Data data, which consist of housing values in suburbs of Boston taken from the 1970 Census. The data set also contains other information that may affect house prices, such as the crime rate in the area and pollution, etc.
  - Data available on Canvas:
    /Files/data/housing.data
    /Files/data/housing.names

Practice Example

- Can also download from https://archive.ics.uci.edu/ml/machine-learning-databases/housing/
- It's often a good idea to include code you used for raw data downloading in your script
  - Good documentation practice
  - "Tie your own hands" from adding (potentially undesirable) features to the data that induces further errors

```
library(tidyverse)

# Make data directory
dir.create("data_housing")

# Download data
"https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data" %>%
  download.file("data_housing/housing.data")

# Download data dictionary
"https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names" %>%
  download.file("data_housing/housing.names")

# Variable names from housing.names
variables <- c("CRIM", "ZN", "INDUS",  "CHAS", "NOX", "RM", "AGE",
             "DIS", "RAD", "TAX", "PTRATIO", "B", "LSTAT", "MEDV")

# Read in the data
housing_data <- read_table("data_housing/housing.data", col_names = variables)
```

Practice Example

- Estimate a simple linear model where the outcome is the median value of owner-occupied homes (in $1000's) and the explanatory variable is the weighted distances to five Boston employment centers (in miles).

```
require(stargazer)

# Simple linear regression
lm_dis <- lm(MEDV ~ DIS, data = housing_data)

# Show regression output using stargazer function
stargazer(lm_dis, type = "text")
```

```
## ================================
##                      Dependent variable:
##                   -------------------------------
##                            MEDV
## ---------------------------------------------------
## DIS                       1.092***
##                           (0.188)
##
## Constant                 18.390***
##                           (0.817)
##
## ----------------------------------------------------
## Observations               506
## R2                        0.062
## Adjusted R2               0.061
## Residual Std. Error     8.914 (df = 504)
## F Statistic        33.580*** (df = 1; 504)
## =================================
## Note:            *p<0.1; **p<0.05; ***p<0.01
```

Practice Example

- Estimate a simple linear model where the outcome is the median value of owner-occupied homes (in $1000's) and the explanatory variable is the weighted distances to five Boston employment centers (in miles).

```
require(stargazer)

# Simple linear regression
lm_dis <- lm(MEDV ~ DIS, data = housing_data)

# Show regression output using stargazer function
stargazer(lm_dis, type = "text")
```

```
## ================================
##                    Dependent variable:
##                  ---------------------------
##                              MEDV
## -----------------------------------------------
## DIS                        1.092***
##                            (0.188)
##
## Constant                  18.390***
##                            (0.817)
##
## -----------------------------------------------
## Observations                 506
## R2                          0.062
## Adjusted R2                 0.061
## Residual Std. Error    8.914 (df = 504)
## F Statistic       33.580*** (df = 1; 504)
## ================================
## Note:            *p<0.1; **p<0.05; ***p<0.01
```

An increase of 1 unit of DIS (1 mile) increases, on average, the home value by 1.092 thousand dollars.

Practice Example

- Estimate a simple linear model where the outcome is the median value of owner-occupied homes (in $1000's) and the explanatory variable is the weighted distances to five Boston employment centers (in miles).

```
require(stargazer)

# Simple linear regression
lm_dis <- lm(MEDV ~ DIS, data = housing_data)

# Show regression output using stargazer function
stargazer(lm_dis, type = "text")
```

```
## ===============================
##                    Dependent variable:
##                   -------------------------------
##                              MEDV
## ---------------------------------------------------
## DIS                        1.092***
##                            (0.188)
##
## Constant                   18.390***
##                            (0.817)
##
## ---------------------------------------------------
## Observations                 506
## R2                          0.062
## Adjusted R2                 0.061
## Residual Std. Error    8.914 (df = 504)
## F Statistic        33.580*** (df = 1; 504)
## ================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

- We often describe an estimate as "statistically significant" at the 95% confidence level if the estimate is more than 1.96 (i.e., about 2) standard errors away from zero
- 1.092 is (1.092/0.188)=5.8 standard errors away from zero. So, significant at 95% level

Practice Example

- Estimate a simple linear model where the outcome is the median value of owner-occupied homes (in $1000's) and the explanatory variable is the weighted distances to five Boston employment centers (in miles).

```
require(stargazer)

# Simple linear regression
lm_dis <- lm(MEDV ~ DIS, data = housing_data)

# Show regression output using stargazer function
stargazer(lm_dis, type = "text")
```

```
## ===============================
##                   Dependent variable:
##                ------------------------------
##                          MEDV
## -------------------------------------------------
## DIS                      1.092***
##                          (0.188)
##
## Constant                 18.390***
##                          (0.817)
##
## -------------------------------------------------
## Observations               506
## R2                         0.062
## Adjusted R2                0.061
## Residual Std. Error    8.914 (df = 504)
## F Statistic        33.580*** (df = 1; 504)
## ================================
## Note:           *p<0.1; **p<0.05; ***p<0.01
```

- An alternative way to determine if a coefficient is significant at 95% level is by computing the confidence interval (CI) and check if it contains zero

- CI: 1.092 +/- 1.96*0.188 = [0.72, 1.46]. Zero is not part of the CI, therefore DIS is statistically significant at 95%.

Practice Example

- Estimate a simple linear model where the outcome is the median value of owner-occupied homes (in $1000's) and the explanatory variable is the weighted distances to five Boston employment centers (in miles).

```
require(stargazer)

# Simple linear regression
lm_dis <- lm(MEDV ~ DIS, data = housing_data)

# Show regression output using stargazer function
stargazer(lm_dis, type = "text")
```

```
## ================================
##                Dependent variable:
##              -------------------------------
##                          MEDV
## ---------------------------------------------------
## DIS                     1.092***
##                         (0.188)
##
## Constant               18.390***
##                         (0.817)
##
## ---------------------------------------------------
## Observations              506
## R2                       0.062
## Adjusted R2              0.061
## Residual Std. Error   8.914 (df = 504)
## F Statistic       33.580*** (df = 1; 504)
## ================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

- Interpretation: "All else equal, being located further from employment centers is positively correlated with higher median home values."

- Should NOT say distance causes home value to increase

Practice Example

- Estimate multiple linear regressions of MEDV that increasingly “phase in” control variables:
  - Property tax per $10,000

    ```
    lm_tax <- lm(MEDV ~ DIS + TAX, data = housing_data)
    ```

  - Property tax per $10,000, nitric oxides concentration (parts per 10 million)

    ```
    lm_nox <- lm(MEDV ~ DIS + TAX + NOX, data = housing_data)
    ```

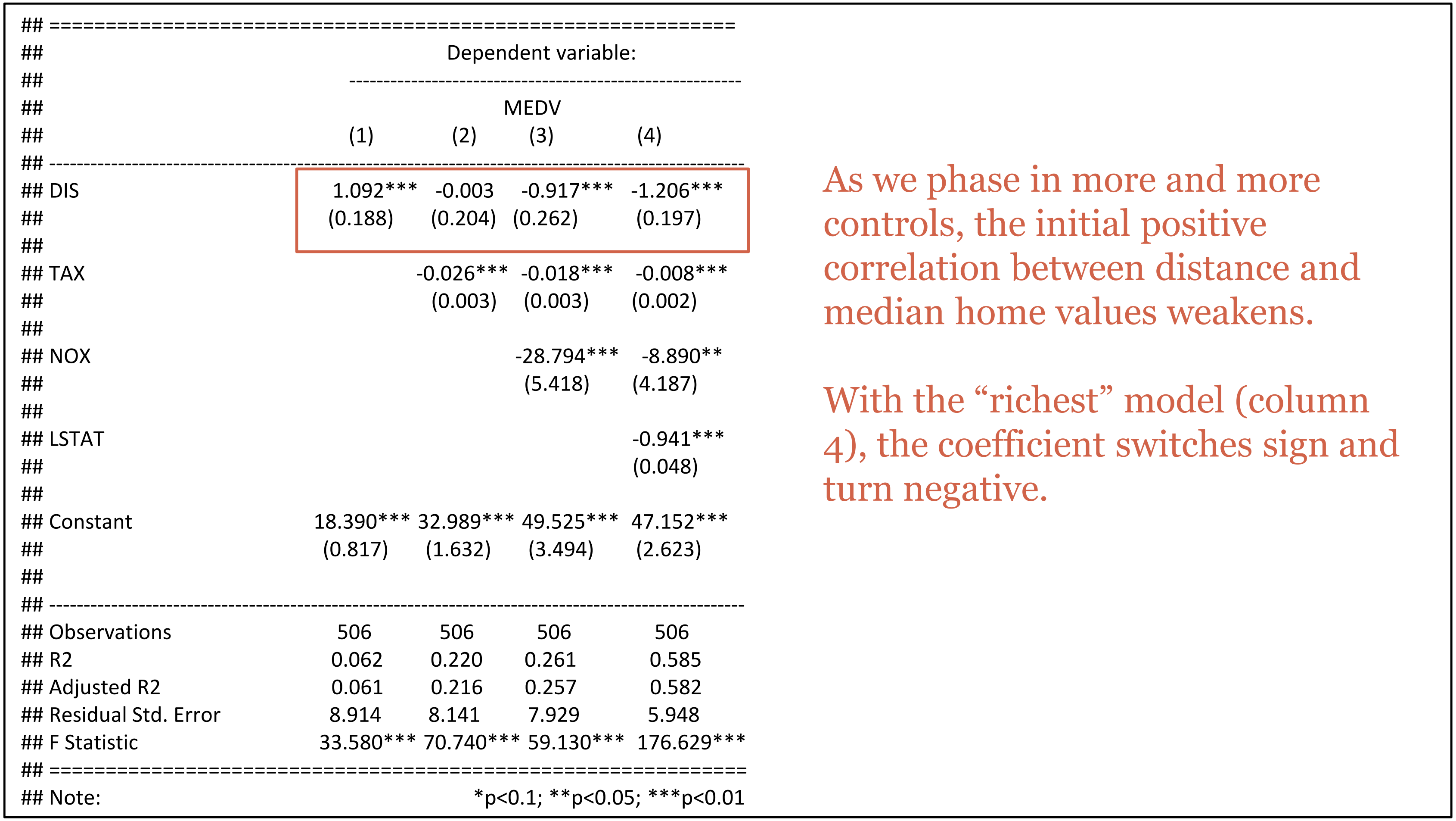
  - Property tax per $10,000, nitric oxides concentration (parts per 10 million), percent of the population with low socioeconomic status

    ```
    lm_ses <- lm(MEDV ~ DIS + TAX + NOX + LSTAT, data = housing_data)
    ```

Practice Example

- Report regressions results from all four models above in the same table

```
stargazer(lm_dis, lm_tax, lm_nox, lm_ses, type = "text", df=FALSE)
```



```
## ==========================================================
##                              Dependent variable:
##                     --------------------------------------
##                                      MEDV
##                      (1)       (2)      (3)        (4)
## ---------------------------------------------------------------
## DIS               1.092***  -0.003   -0.917***  -1.206***
##                   (0.188)   (0.204)  (0.262)    (0.197)
##
## TAX                        -0.026*** -0.018***  -0.008***
##                             (0.003)  (0.003)    (0.002)
##
## NOX                                 -28.794***  -8.890**
##                                       (5.418)   (4.187)
##
## LSTAT                                          -0.941***
##                                                 (0.048)
##
## Constant         18.390*** 32.989*** 49.525***  47.152***
##                  (0.817)   (1.632)   (3.494)    (2.623)
##
## ---------------------------------------------------------------
## Observations       506       506       506        506
## R2                 0.062     0.220     0.261      0.585
## Adjusted R2        0.061     0.216     0.257      0.582
## Residual Std. Error 8.914    8.141     7.929      5.948
## F Statistic      33.580*** 70.740*** 59.130***  176.629***
## ==========================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

As we phase in more and more controls, the initial positive correlation between distance and median home values weakens.

With the “richest” model (column 4), the coefficient switches sign and turn negative.

Practice Example

- Report regressions results from all four models above in the same table

```
stargazer(lm_dis, lm_tax, lm_nox, lm_ses, type = "text", df=FALSE)
```

```
## ===========================================================
##                              Dependent variable:
##                       -----------------------------------
##                                    MEDV
##                          (1)       (2)      (3)       (4)
## -----------------------------------------------------------
## DIS                   1.092***  -0.003   -0.917***  -1.206***
##                       (0.188)   (0.204)  (0.262)    (0.197)
##
## TAX                            -0.026*** -0.018***  -0.008***
##                                 (0.003)  (0.003)    (0.002)
##
## NOX                                      -28.794***  -8.890**
##                                           (5.418)    (4.187)
##
## LSTAT                                               -0.941***
##                                                     (0.048)
##
## Constant             18.390*** 32.989*** 49.525***  47.152***
##                       (0.817)   (1.632)   (3.494)    (2.623)
##
## -----------------------------------------------------------
## Observations            506       506       506        506
## R2                     0.062     0.220     0.261      0.585
## Adjusted R2            0.061     0.216     0.257      0.582
## Residual Std. Error    8.914     8.141     7.929      5.948
## F Statistic          33.580*** 70.740*** 59.130***  176.629***
## ===========================================================
## Note:                             *p<0.1; **p<0.05; ***p<0.01
```

What this tells us is that, in the column 1 model, there are other omitted variables correlated with distance and median value, and their effect is all combined in the only variable -- distance.

In columns 2-4, when we make these other variables explicit and control for them, the negative signal of distance appears.

Practice Example

- Report regressions results from all four models above in the same table

```
stargazer(lm_dis, lm_tax, lm_nox, lm_ses, type = "text", df=FALSE)
```

```
## ==========================================================
##                                Dependent variable:
##                       ----------------------------------------
##                                      MEDV
##                          (1)        (2)       (3)        (4)
## ------------------------------------------------------------------
## DIS                   1.092***   -0.003   -0.917***  -1.206***
##                       (0.188)   (0.204)   (0.262)    (0.197)
##
## TAX                            -0.026***  -0.018***   -0.008***
##                                 (0.003)   (0.003)    (0.002)
##
## NOX                                      -28.794***   -8.890**
##                                           (5.418)    (4.187)
##
## LSTAT                                                -0.941***
##                                                      (0.048)
##
## Constant             18.390*** 32.989*** 49.525***  47.152***
##                       (0.817)   (1.632)   (3.494)    (2.623)
##
## ------------------------------------------------------------------
## Observations            506       506       506        506
## R2                     0.062     0.220     0.261      0.585
## Adjusted R2            0.061     0.216     0.257      0.582
## Residual Std. Error    8.914     8.141     7.929      5.948
## F Statistic          33.580*** 70.740*** 59.130***  176.629***
## ==========================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

We should also not believe even the "richest" model: who knows what else you haven't controlled for

Practice Example

- Report regressions results from all four models above in the same table

```
stargazer(lm_dis, lm_tax, lm_nox, lm_ses, type = "text", df=FALSE)
```

```
## ===========================================================
##                                   Dependent variable:
##                          ---------------------------------------------
##                                          MEDV
##                            (1)       (2)       (3)        (4)
## -----------------------------------------------------------
## DIS                     1.092***   -0.003   -0.917***  -1.206***
##                         (0.188)   (0.204)   (0.262)    (0.197)
##
## TAX                              -0.026***  -0.018***  -0.008***
##                                   (0.003)   (0.003)    (0.002)
##
## NOX                                        -28.794***  -8.890**
##                                              (5.418)    (4.187)
##
## LSTAT                                                 -0.941***
##                                                        (0.048)
##
## Constant               18.390*** 32.989*** 49.525***  47.152***
##                         (0.817)   (1.632)   (3.494)    (2.623)
##
## -----------------------------------------------------------
## Observations              506       506       506        506
## R2                       0.062     0.220     0.261      0.585
## Adjusted R2              0.061     0.216     0.257      0.582
## Residual Std. Error      8.914     8.141     7.929      5.948
## F Statistic           33.580*** 70.740*** 59.130***  176.629***
## ===========================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

This table, however, does NOT tell us that distance has NO causal effect on median home value.

It only tell us that we are not going to be successful trying to estimate the causal effect of distance by simple OLS models.

In other words, we need a better "research design".

Nonlinear Regression

- We will mostly work with linear regression models for this course
  - Mathematically tractable; nice statistical (especially large-sample) properties
  - Most state-of-the-science causal inference tools still work in the linear world

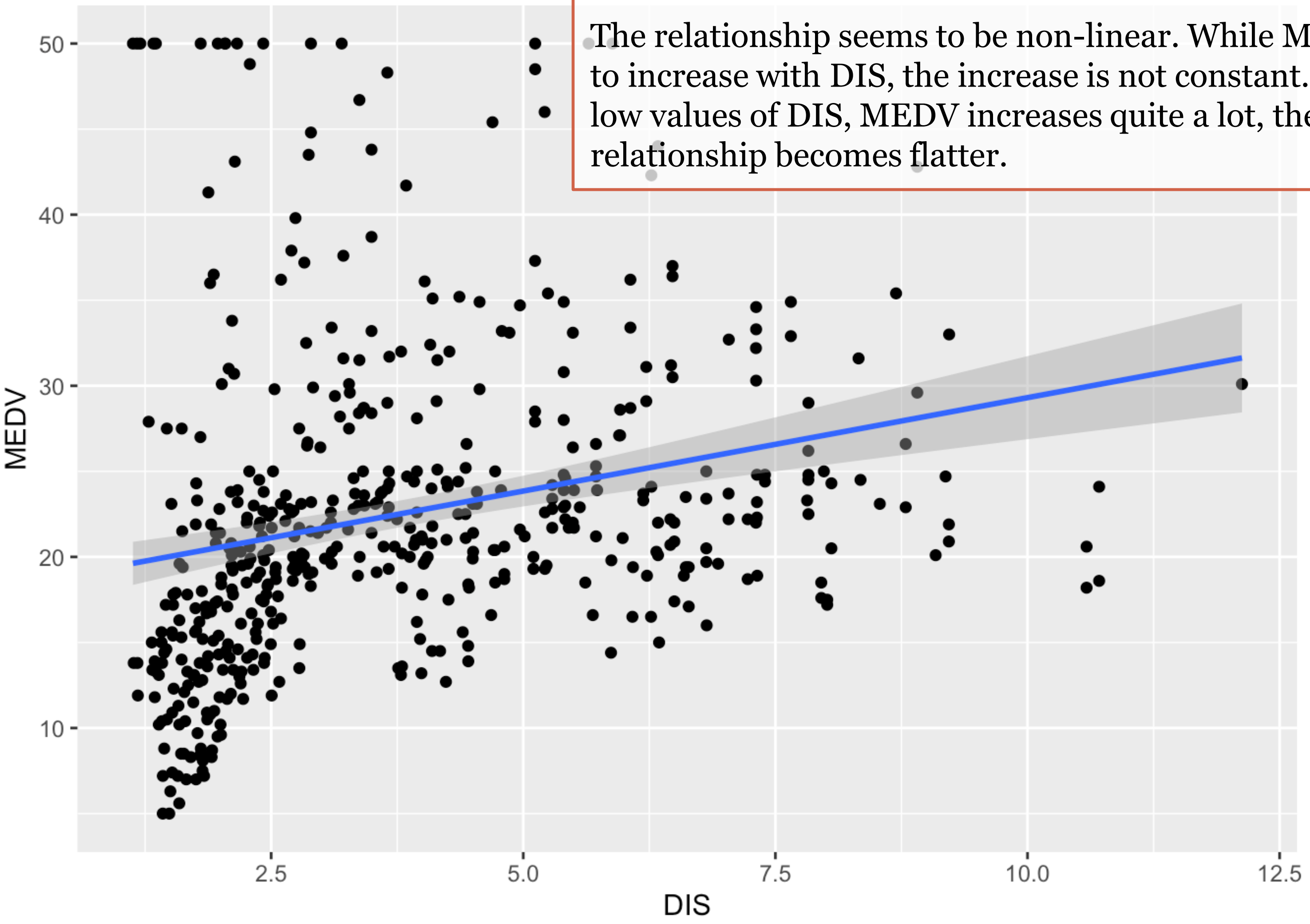
- But, we will work quite often with nonlinear models when it comes to visualization

Practice Example

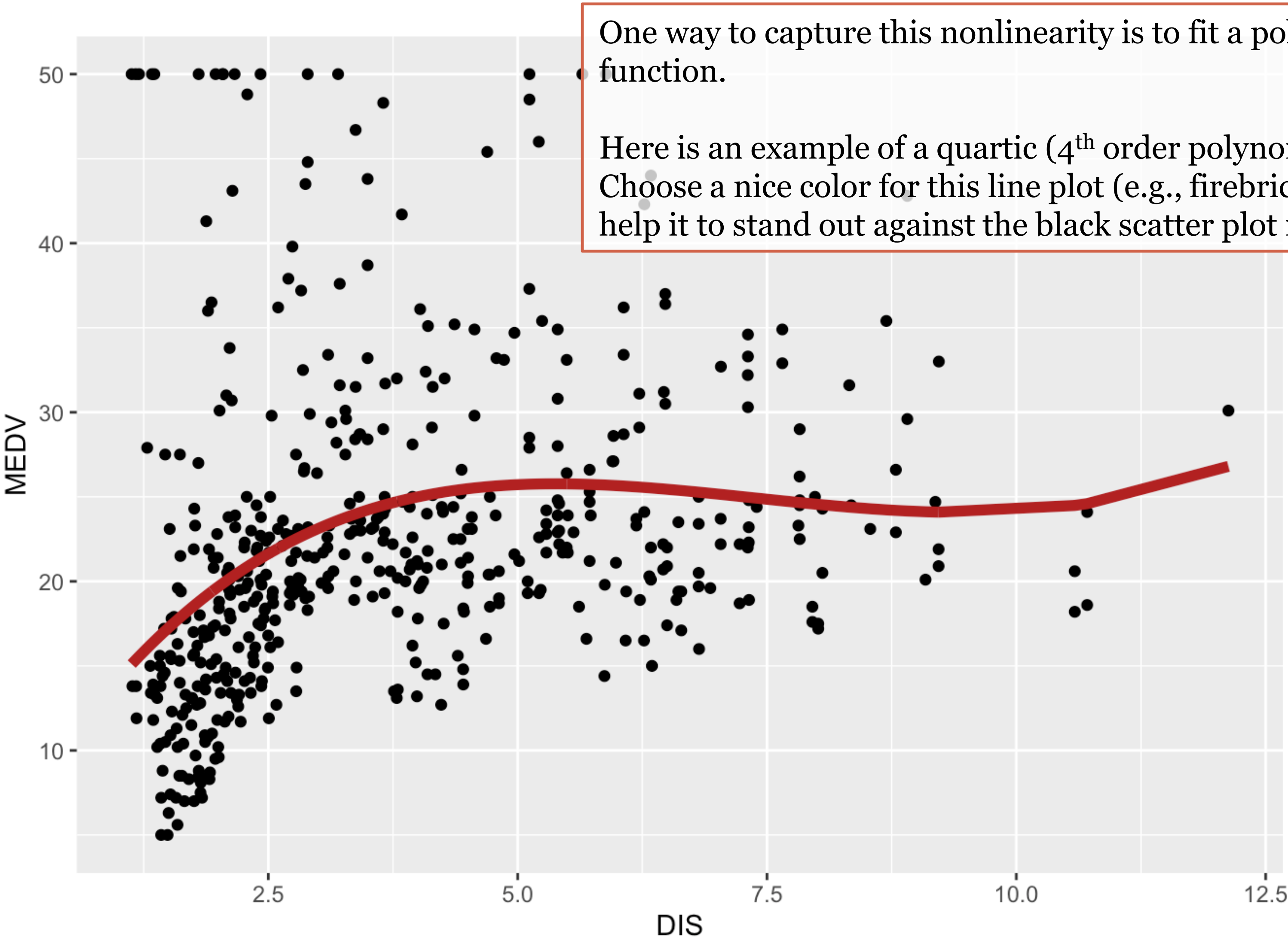
- Suppose we are interested in visualizing the (raw) correlation between median value of owner-occupied homes (in $1000's) and the distances to employment centers

The linear line seems a bit “out of place” in this case

The relationship seems to be non-linear. While MEDV seems to increase with DIS, the increase is not constant. For very low values of DIS, MEDV increases quite a lot, then the relationship becomes flatter.



```
# Scatter plot
housing_data %>%
  ggplot(aes(x=DIS, y=MEDV)) +
  geom_point() +
  geom_smooth(method = "lm")
```



One way to capture this nonlinearity is to fit a polynomial function.

Here is an example of a quartic (4th order polynomial) fit. Choose a nice color for this line plot (e.g., firebrick red) to help it to stand out against the black scatter plot markers.

```
# Add DIS_2-DIS_4 to the data frame
housing_data <- housing_data %>%
  mutate(DIS_2 = DIS^2, DIS_3 = DIS^3, DIS_4 = DIS^4)

# Estimate the model
lm_dis_man <- lm(MEDV ~ DIS + DIS_2 + DIS_3 + DIS_4, data = housing_data)

# Scatter plot with model predicted values
housing_data %>%
  ggplot(aes(x = DIS, y = MEDV)) +
  geom_point() +
  geom_line(aes(y = lm_dis_man$fitted.values), color = "firebrick", size = 2)
```

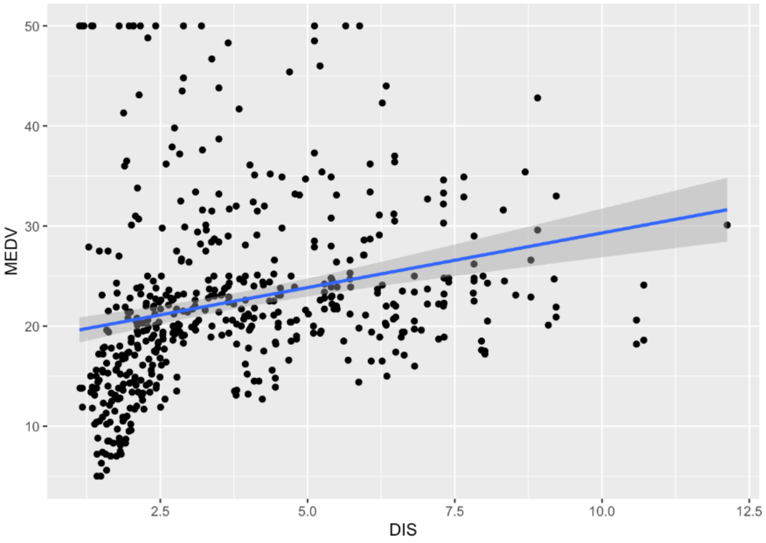
Practice Example

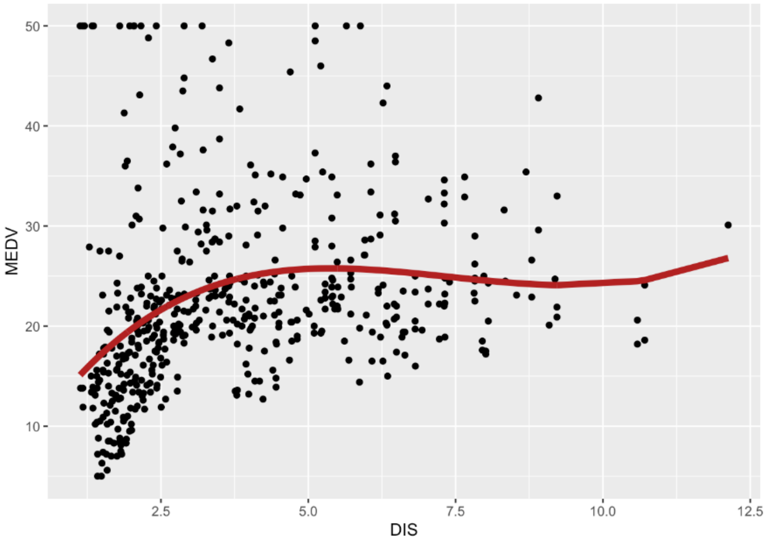
- Exercise:
  1. Produce a binscatter plot version of the relationship between median house value and distance.
  2. Explain why the linear and polynomial fits represent a “parametric” approach, while the binscatter fit represents a “non-parametric” approach

Visualization

- There are many different ways to present the same data/results, and people often react differently to different presentations.

- This course focuses on causal inference, so we won’t talk a lot about choosing the right functional form. But, we will spend some time on thinking carefully about good/bad/controversial visualization tools.





Remarks on Code

- Through these review slides, I have shown you my expectation on how you should document your codes in empirical projects. For example:

```
# Add DIS_2-DIS_4 to the data frame
housing_data <- housing_data %>%
  mutate(DIS_2 = DIS^2, DIS_3 = DIS^3, DIS_4 = DIS^4)

# Estimate the model
lm_dis_man <- lm(MEDV ~ DIS + DIS_2 + DIS_3 + DIS_4, data = housing_data)

# Scatter plot with model predicted values
housing_data %>%
  ggplot(aes(x = DIS, y = MEDV)) +
  geom_point() +
  geom_line(aes(y = lm_dis_man$fitted.values), color = "firebrick", size = 2)
```

- Most econ exercises can be done with simple codes. Don't *try* to be fancy. Simpler is just better.

- Clearly document what you did for each paragraph of code.

- Give variables good names, so that I can read what you are doing directly off of your codes
  - Most people can guess what's "DIS_4", especially if you are working on a polynomial exercise
  - Nobody should spend time figuring out what's "_X1" in your code

Remarks on Code

- These review codes are presented in R. But again, you are welcome to use Stata programming instead.

Application of Binscatter Plot in Causal Inference

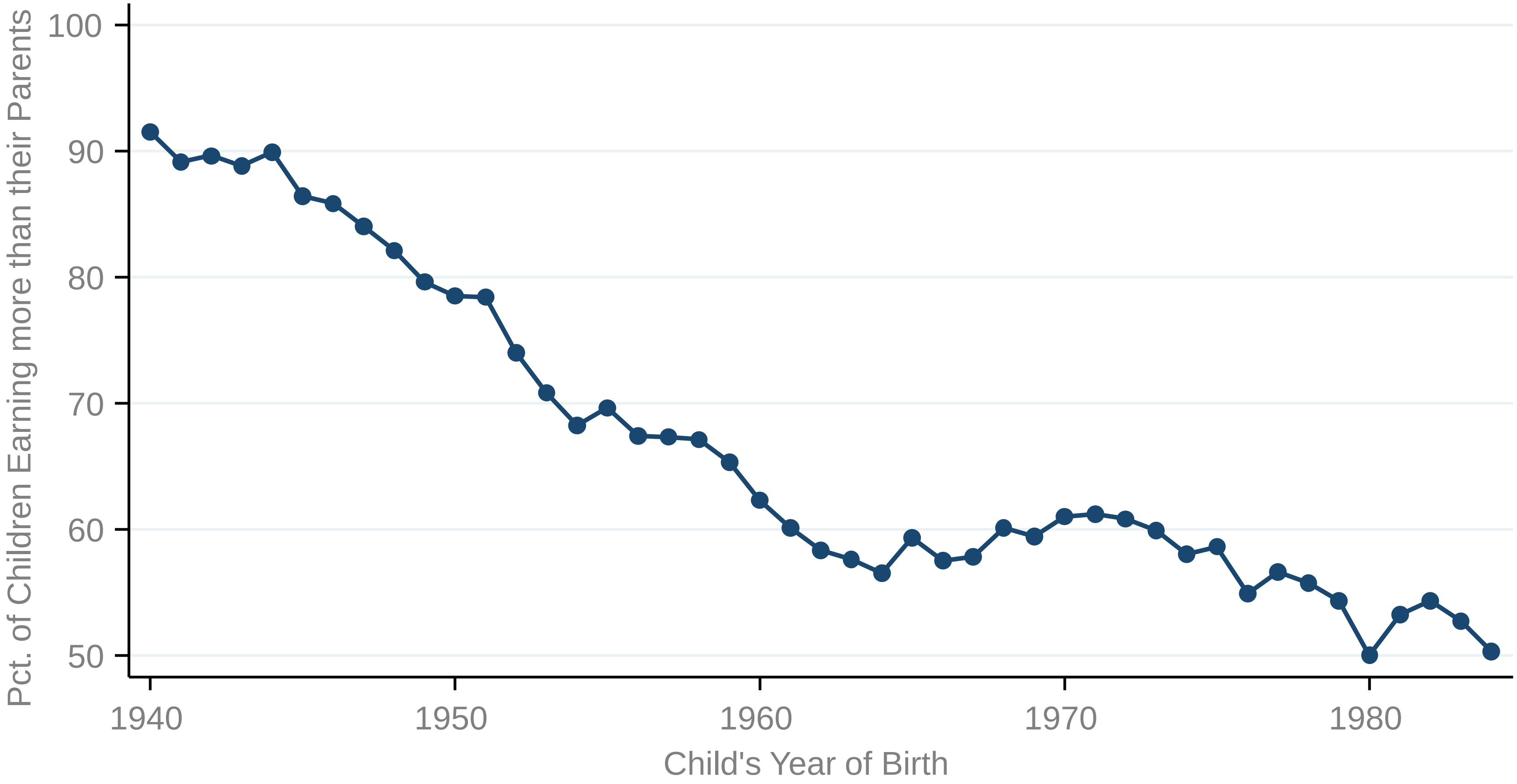
- Binscatter plots are promoted in a series of papers by Professor Raj Chetty and coauthors that use tax return data to study the "causal effect of cities" on intergenerational mobility in the U.S.
  - Chetty, Raj, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang. "The fading American dream: Trends in absolute income mobility since 1940." Science 356, no. 6336 (2017): 398-406.
  - Chetty, Raj, John N. Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. *The opportunity atlas: Mapping the childhood roots of social mobility*. No. w25147. National Bureau of Economic Research, 2018.
  - Chetty, Raj, and Nathaniel Hendren. "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects." *The Quarterly Journal of Economics* 133, no. 3 (2018): 1107-1162.

- [Mandatory "watching"] Stanford Micro Lecture: Reviving the American Dream with Raj Chetty (14 minutes) https://www.youtube.com/watch?v=g0MeZ4EAY3w

- Longer version: 2018 Childx: Keynote by Raj Chetty https://www.youtube.com/watch?v=0GYLngfhonw

Upward Mobility in the United States

Percent of Children Earning More than Their Parents, by Year of Birth



Source: Chetty, Grusky, Hell, Hendren, Manduca, Narang (Science 2017)

Why is Upward Mobility Declining?

- Central policy question: why are children's chances of climbing the income ladder falling in America?
  - And what can we do to reverse this trend…?
- Difficult to answer this question based solely on historical data on macroeconomic trends
  - Numerous changes over time make it hard to test between alternative explanations
  - Problem: only a handful of data points

Differences in Opportunity Across Local Areas

- How do children's chances of moving up vary across areas in America?
  - Are there some areas where kids do better than others? If so, what lessons can we learn from them?

- Recent studies have used big data to measure how upward mobility varies based on where children *grow up*

Data Sources and Sample Definitions

- Data sources: Anonymized Census data (2000, 2010, ACS) covering U.S. population linked to federal income tax returns from 1989-2015

- Link children to parents based on dependent claiming on tax returns

- Target sample: Children in 1978-83 birth cohorts who were born in the U.S. or are authorized immigrants who came to the U.S. in childhood

- Analysis sample: 20.5 million children, 96% coverage rate of target sample

Measuring Parents' and Children's Incomes in Tax Data

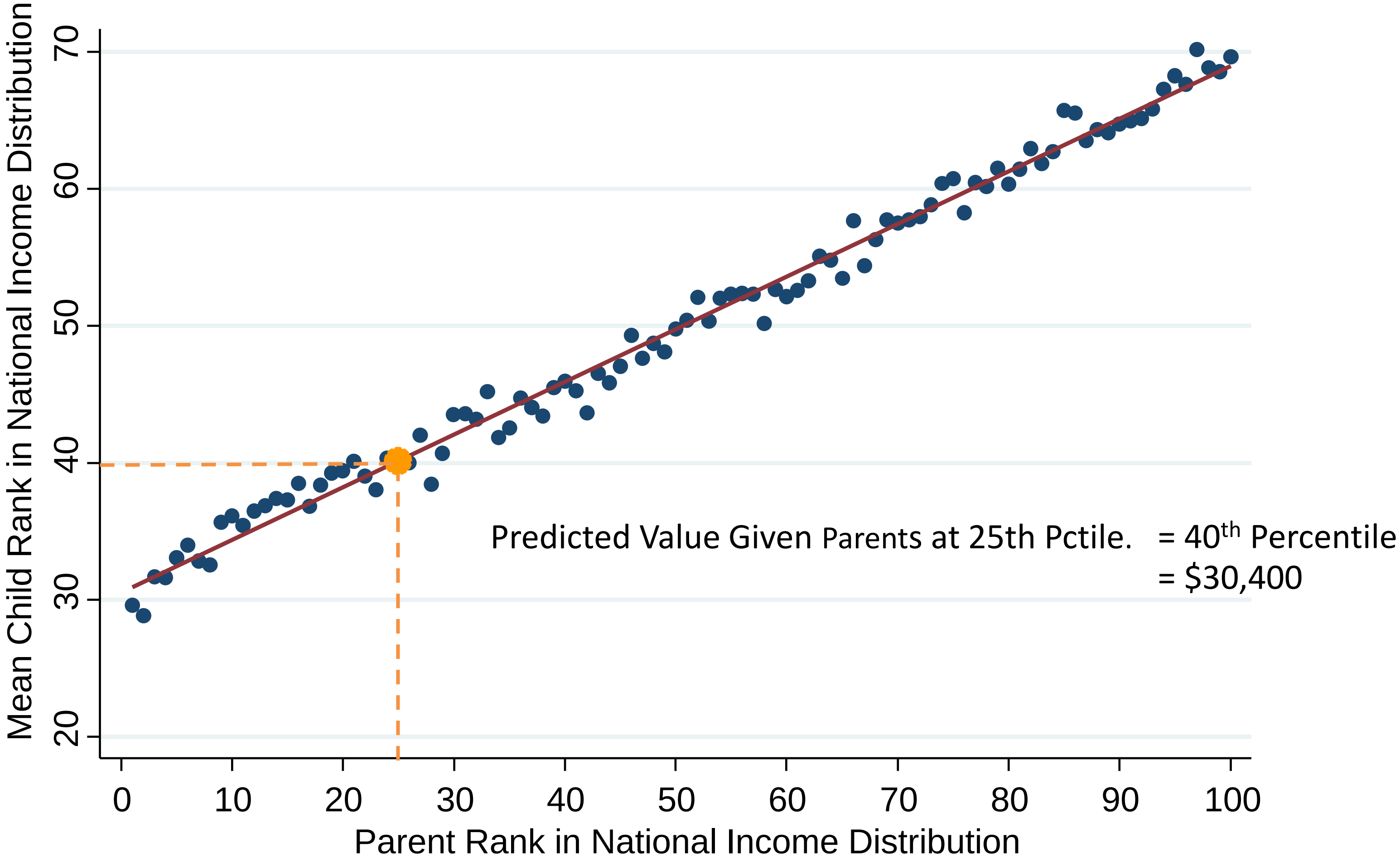
- Parents' household incomes: average income reported on Form 1040 tax return from 1994-2000

- Children's incomes measured from tax returns in 2014-15 (ages 31-37)

- Focus on percentile ranks in national distribution:
  - Rank children relative to others born in the same year and parents relative to other parents

Intergenerational Income Mobility for Children Raised in Chicago

Average Child Household Income Rank vs. Parent Household Income Rank



Predicted Value Given Parents at 25th Pctile. = 40th Percentile
= $30,400

Source: Chetty, Hendren, Kline, Saez 2014

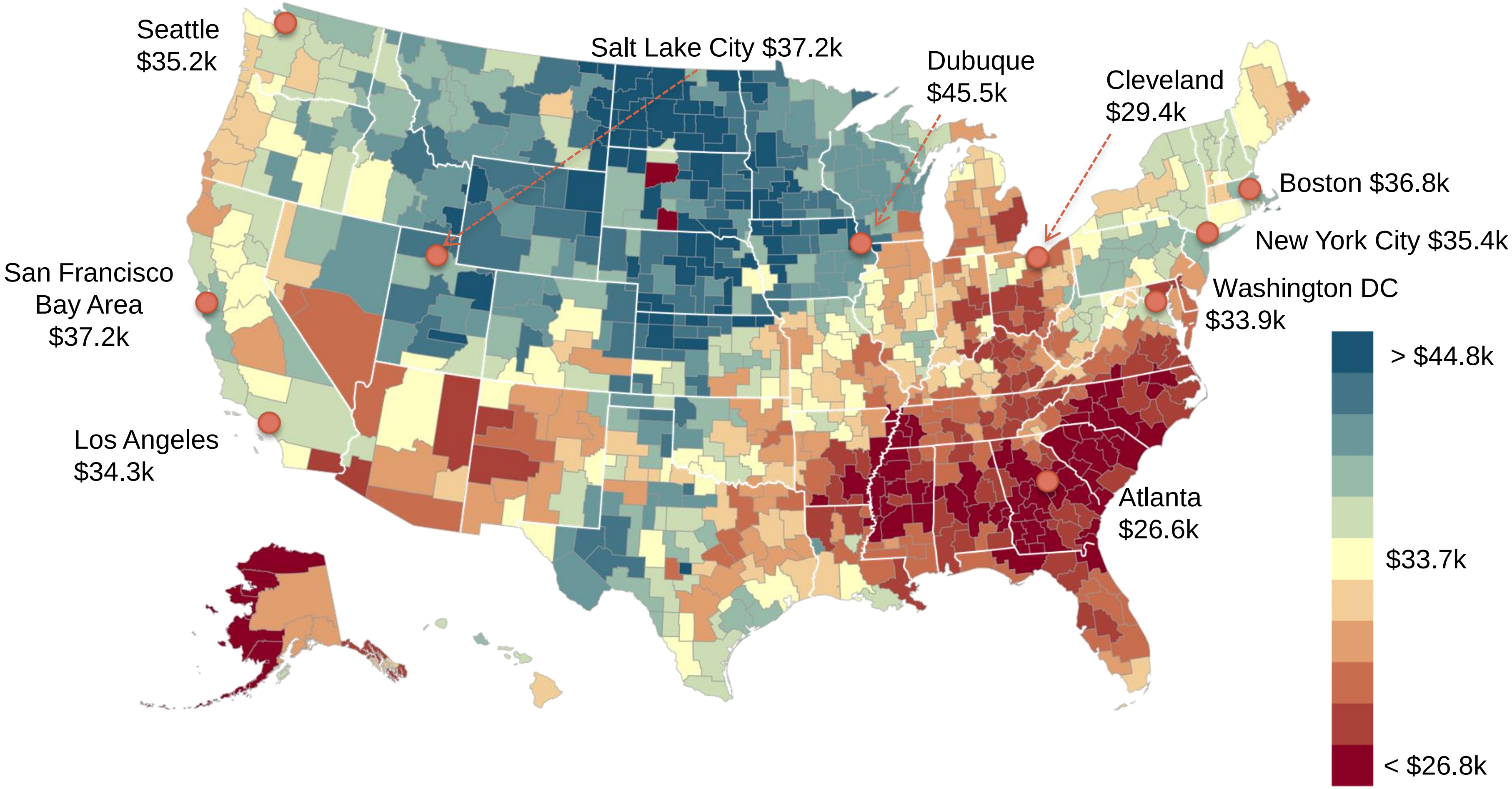
Estimating Children's Average Outcomes by Counties

- A key finding is that there is substantial cross-city variation in upward mobility even conditional on parental income

- That is, among families with a household income of $27,000 (25th percentile of the national income distribution), kids grew up in Salt Lake City, UT may have a much better chance to earn more than their parents do than kids grew up in Atlanta, GA.

The Geography of Upward Mobility in the United States

Average Household Income for Children with Parents Earning $27,000 (25th percentile)



Note: Blue = More Upward Mobility, Red = Less Upward Mobility

Source: The Opportunity Atlas. Chetty, Friedman, Hendren, Jones, Porter 2018

Causal Effects of Neighborhoods vs. Sorting

- Two very different explanations for variation in children's outcomes across areas:
  - Sorting: different people live in different places
  - Causal effects: places have a causal effect on upward mobility for a given person

Identifying Causal Effects of Neighborhoods

- Ideal experiment: randomly assign children to neighborhoods and compare outcomes in adulthood

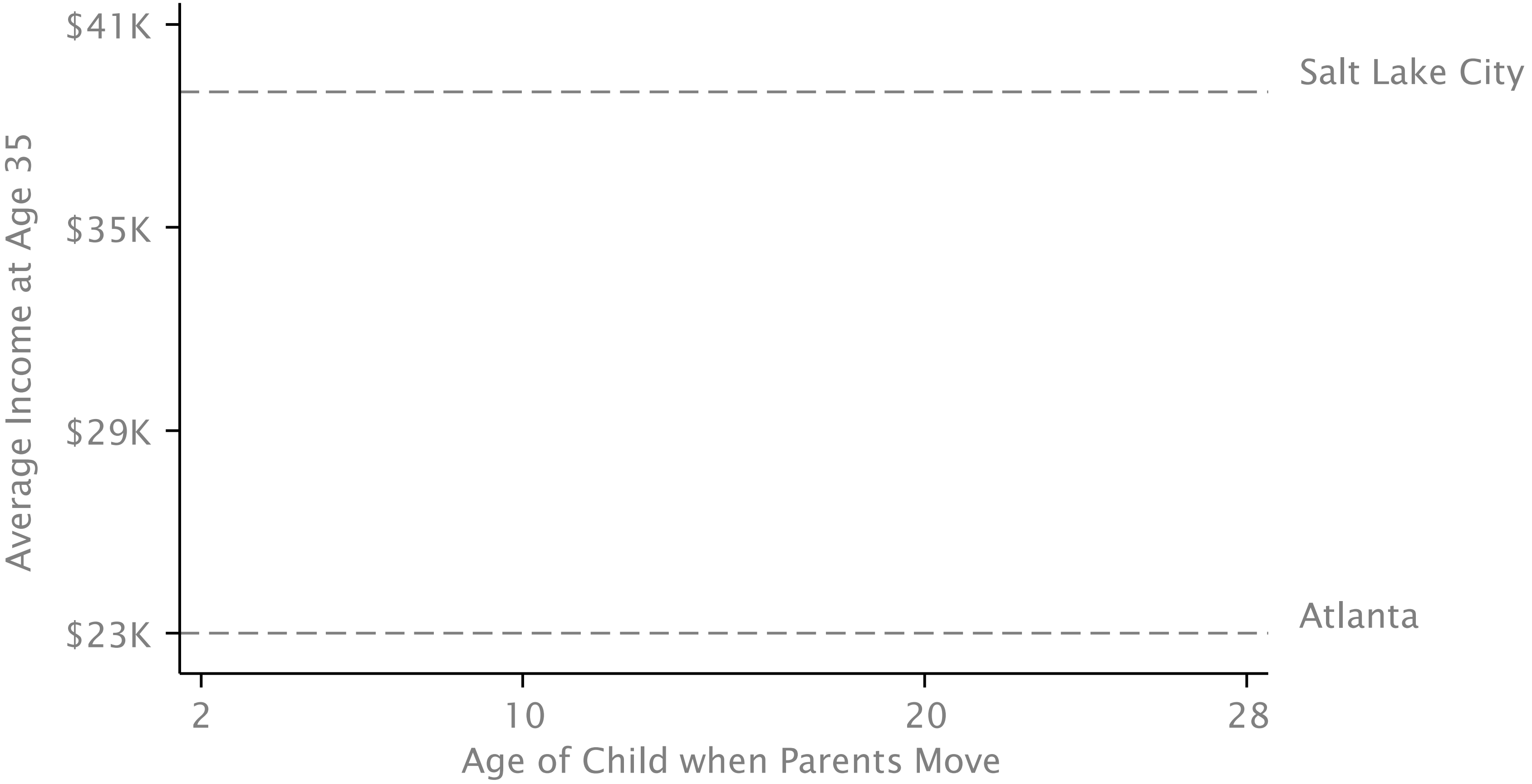
- Approximate this experiment using a quasi-experimental design

  - Study 3 million families who move across areas in observational data

  - Key idea: exploit variation in age of child when family moves to identify causal effects of environment

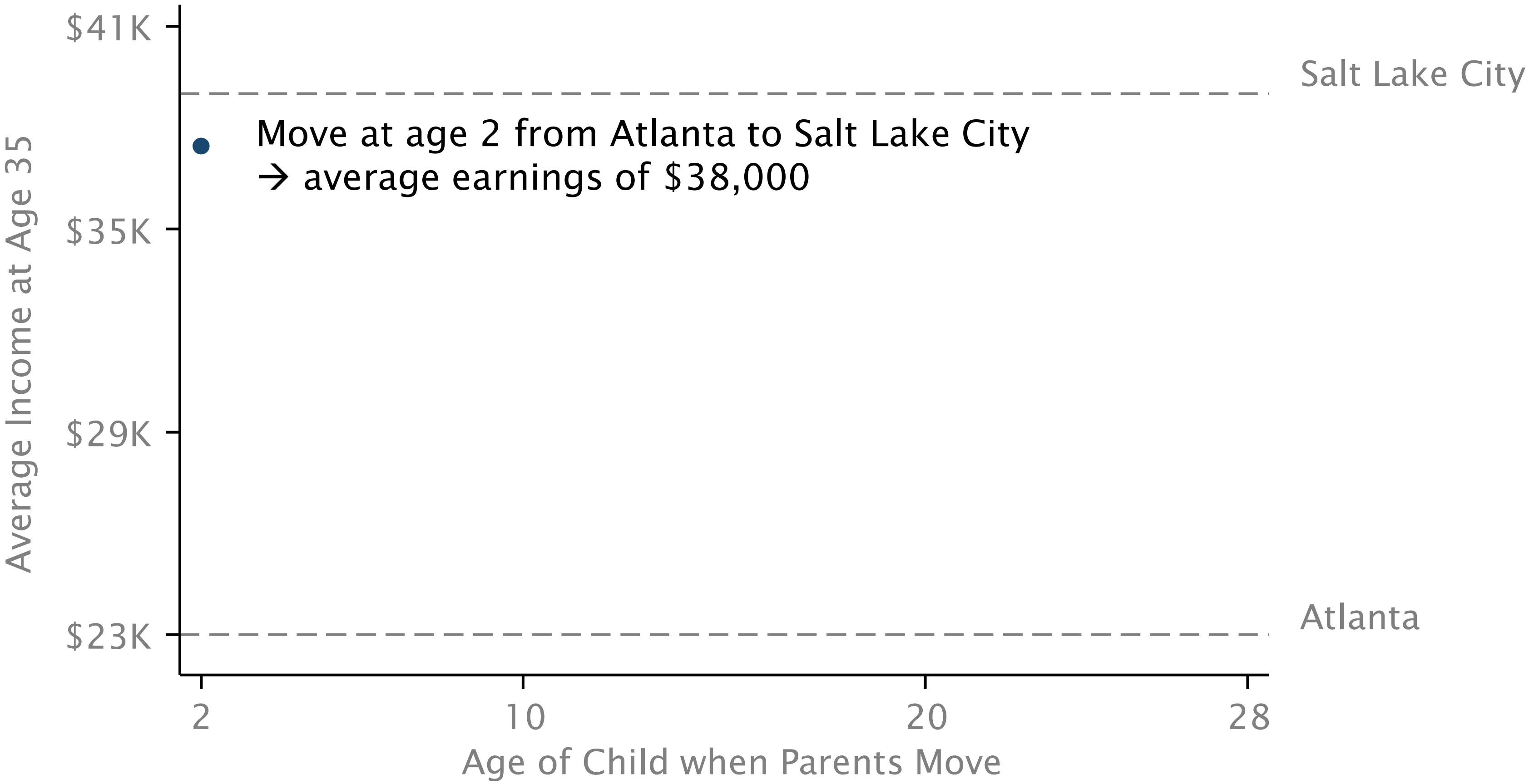
Income Gain from Moving to a Better Neighborhood

By Child's Age at Move



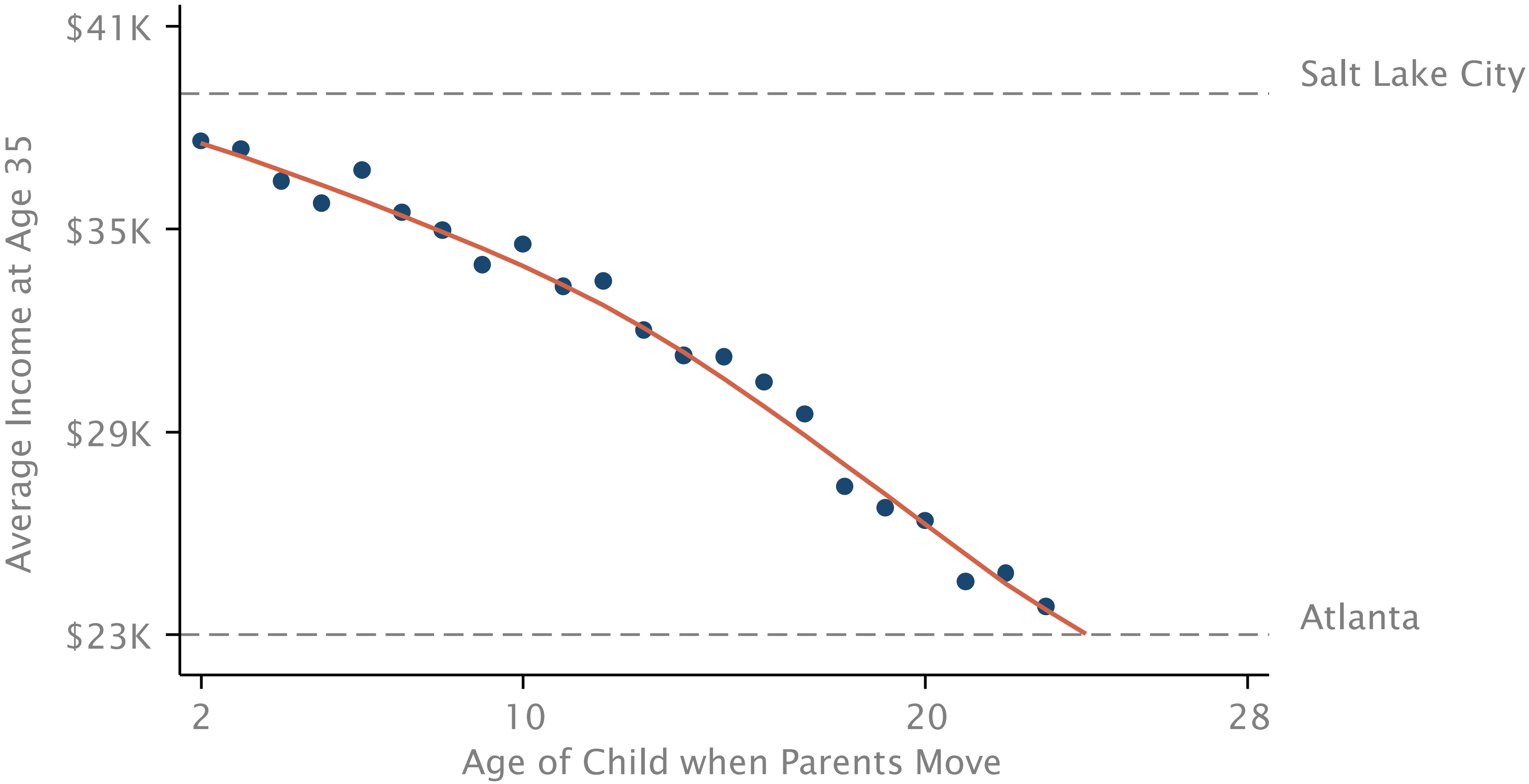
Income Gain from Moving to a Better Neighborhood

By Child's Age at Move



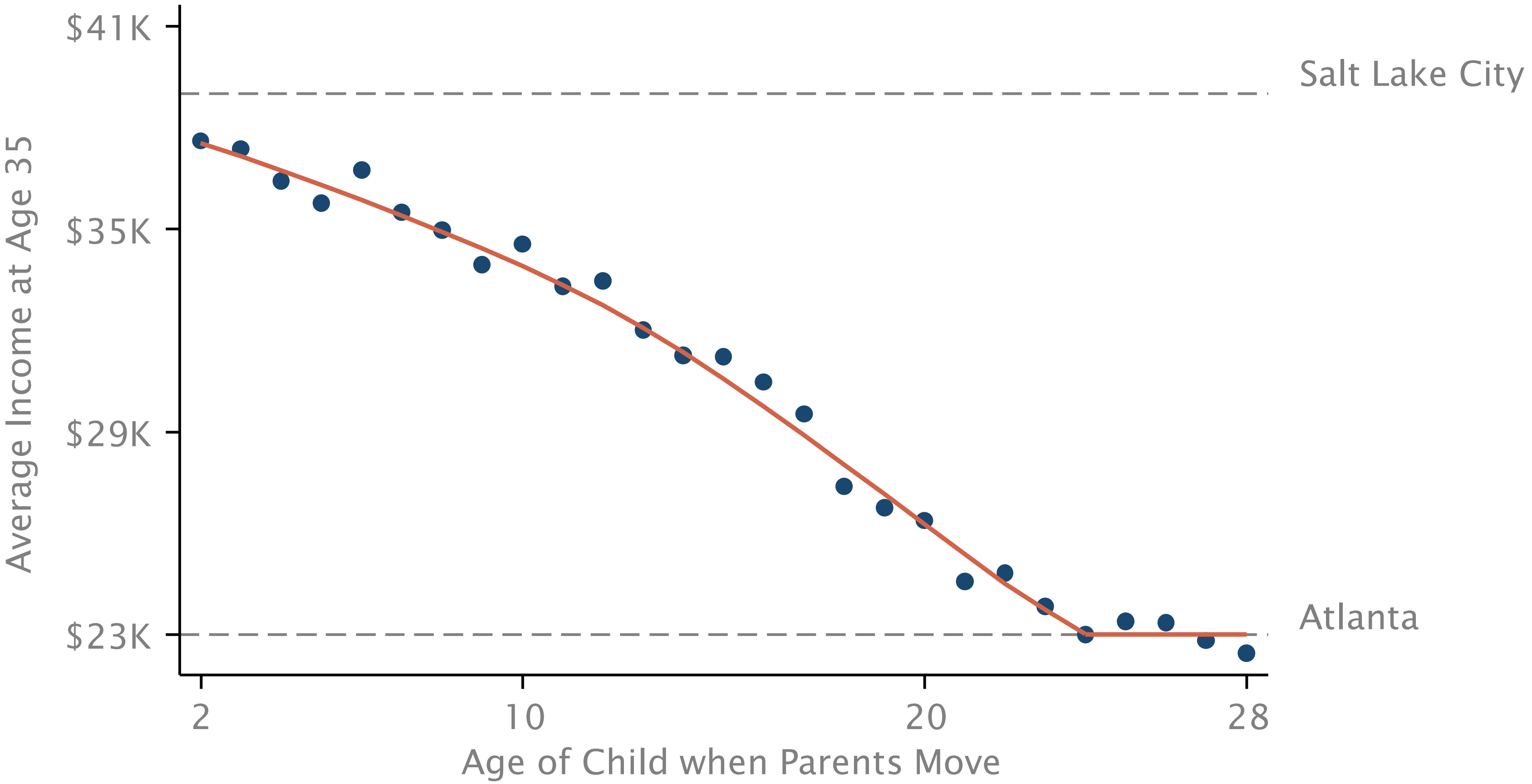
Income Gain from Moving to a Better Neighborhood

By Child's Age at Move



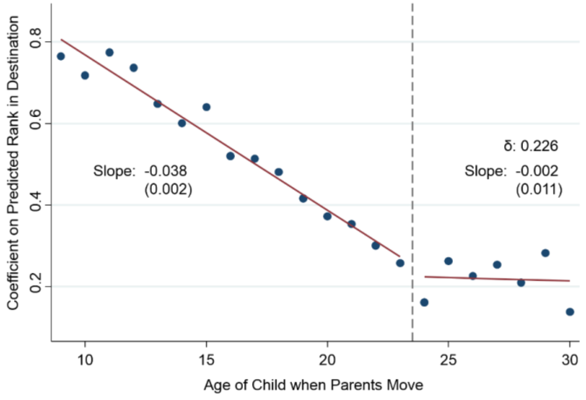
Income Gain from Moving to a Better Neighborhood

By Child's Age at Move



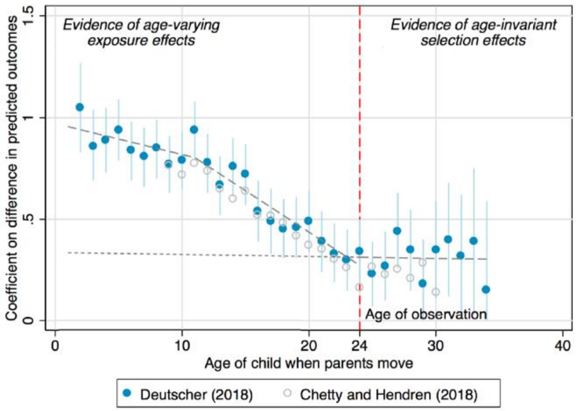
Childhood Exposure Effects Around the World

United States



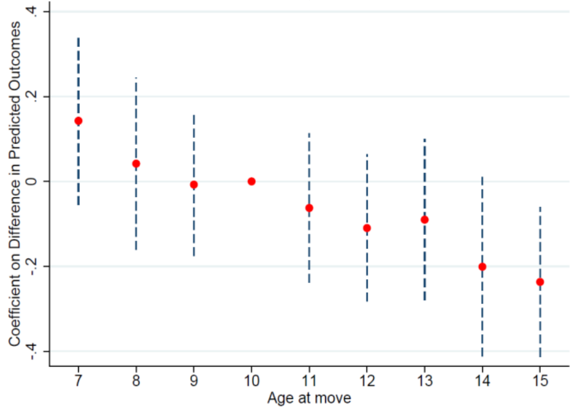
Source: Chetty and Hendren (QJE 2018)

Australia



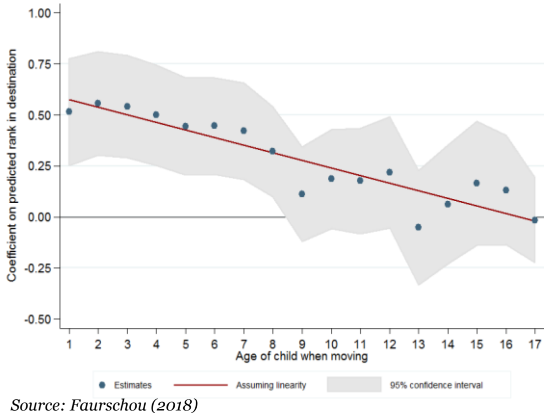
Source: Deutscher (2018)

Montreal, Canada



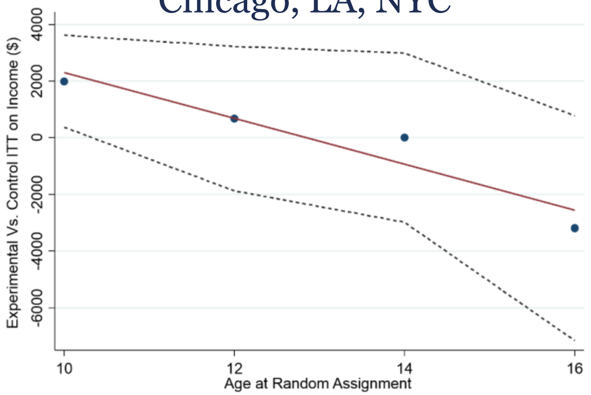
Source: Laliberté (2018)

Denmark



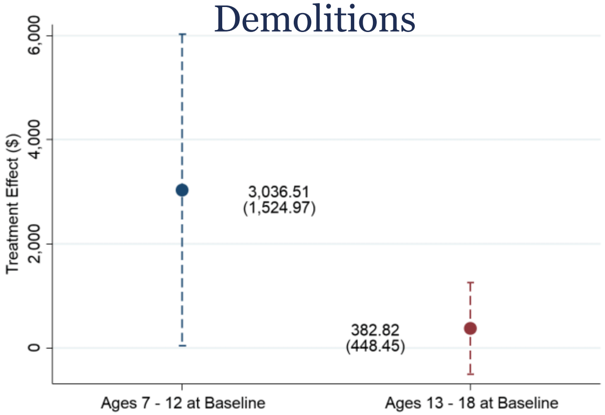
Source: Faurschou (2018)

MTO: Baltimore, Boston, Chicago, LA, NYC



Source: Chetty, Hendren, Katz (AER 2016)

Chicago Public Housing Demolitions



Source: Chyn (AER 2018)

Identifying Causal Effects of Neighborhoods

- Two approaches to evaluating validity of this assumption:

  1. Compare siblings' outcomes to control for family effects

  2. Use differences in neighborhood effects across subgroups to implement "placebo" tests

     – Ex: some places (e.g., low-crime areas) have better outcomes for boys than girls

     – Move to a place where boys have high earnings → son improves in proportion to exposure but daughter does not

- Conclude that about two-thirds of the variation in upward mobility across areas is due to causal effects