Introduction to methods in environmental epidemiology 2

Robbie M Parks, PhD

21st July 2025

Email: robbie.parks@columbia.edu

BlueSky: @robbiemparks

Website: sparklabnyc.github.io

Art by Amy Wolfe

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

Outline from previous lecture

- What is environmental epidemiology?
- Historical milestones
- Major environmental exposures
- Study designs in environmental epidemiology
- Linear regression
- Theories and assumptions
- Key concepts
- Challenges and limitations
- Data structuring

Outline

- Quantile regression
- An example

Linear regression again

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Looks at the average change in one variable (outcome of interest; Y) that corresponds to a given change in the other variable (exposure of interest; X)
- i.e., how does on average Y vary with X?
- How does it estimate the βs?
- “Least squares" -> minimizes the sum of squares $\sum \varepsilon^2$
- What happens when we have values with large residuals ε?

Linear regression again

- Estimates the mean of the outcome conditional on predictors.
- Assumes constant effect across outcome distribution.
- Sensitive to outliers and skewness.
- Often insufficient for full understanding of exposure-outcome relationships. Why might that be?

Linear regression again

- Instead of minimizing the sum of squared differences
- Minimize the sum of absolute differences
- "Least absolute values" $\sum |y_i - (\beta_0 + \beta_1 X_i)|$
  - Note: this is an earlier approach for robust linear regression
  - More recent algorithms: M-estimators (more efficient) and bounded influence methods (trouble with small data)
- Interpretation?
- Now we are estimating how the median of Y varies with X
  - No longer the mean
  - This point will become more clear in the next few slides
- Can these two regressions give (approximately) the same result?

Linear regression again

- The median is the 50th percentile in a distribution
- Can we then look at other percentiles as well?
- When would that be important?
- Distributional effects
- What does that mean?

Introducing quantile regression

- Estimates conditional quantiles (e.g., 10th, 50th, 90th percentile).
- Answers question: "How does exposure affect different parts of the outcome distribution?”.
- More robust to non-normality and outliers.
- Introduced by Koenker & Bassett (1978).

When to use quantile regression

- If there is reason to believe that our exposure has a different impact at the tails of the outcome distribution than the mean/median.
- Or, if we want to evaluate the impact of our exposure across the distribution of the outcome
- In presence of effect modification when we don't know / don’t have information on the modifier to explicitly incorporate in our analyses
  - (Susceptibility)
  - E.g., genetics

When to use quantile regression

- Health outcomes often have skewed distributions (e.g., birth weight, blood pressure, mortality risk scores)
- Exposure may impact the most vulnerable (e.g., upper quantiles of heat-related mortality risk)
- Detects heterogeneous effects missed by OLS

How does it work?

- Remember for median we wanted to minimize

$$\sum |y_i - (\beta_0 + \beta_1 x_i)|$$

- For other quantiles (τ), we just use *assymetric weighting*, with τ as weights

$$\min_{\beta} \left[\tau \sum_{y_i \geq x_i \beta} |y_i - (\beta_0 + \beta_1 x_i)| + (1 - \tau) \sum_{y_i < x_i \beta} |y_i - (\beta_0 + \beta_1 x_i)| \right]$$

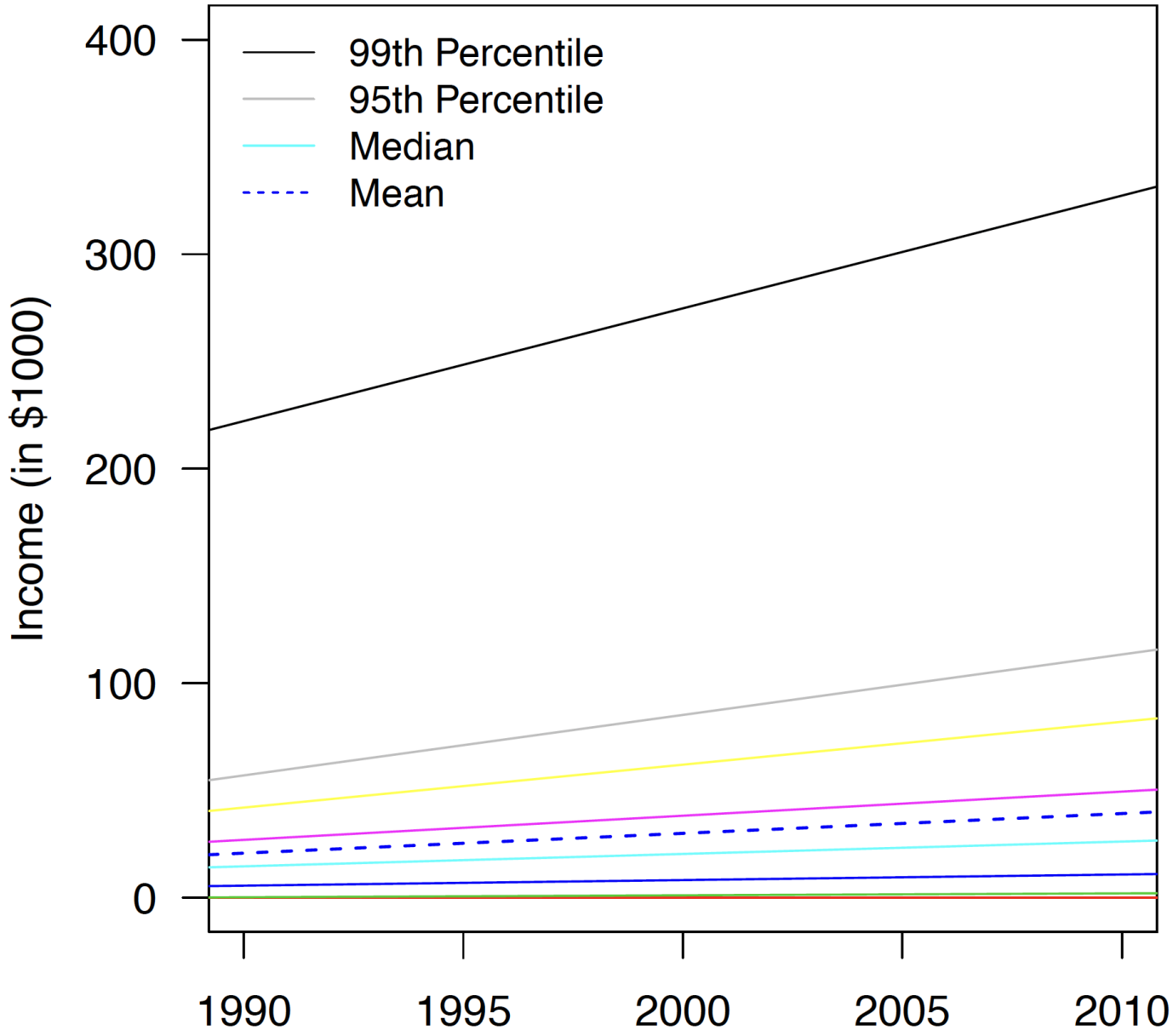
An example

- Occupy Wall Street

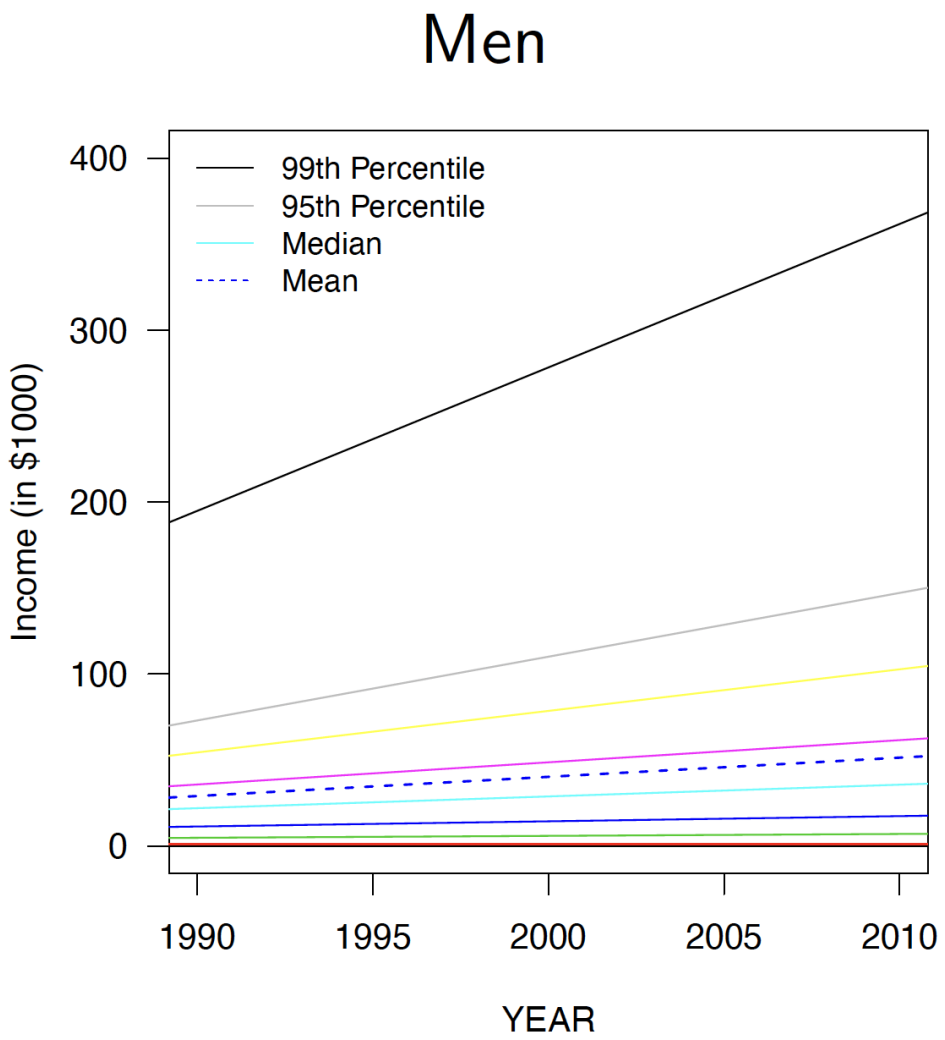
An example

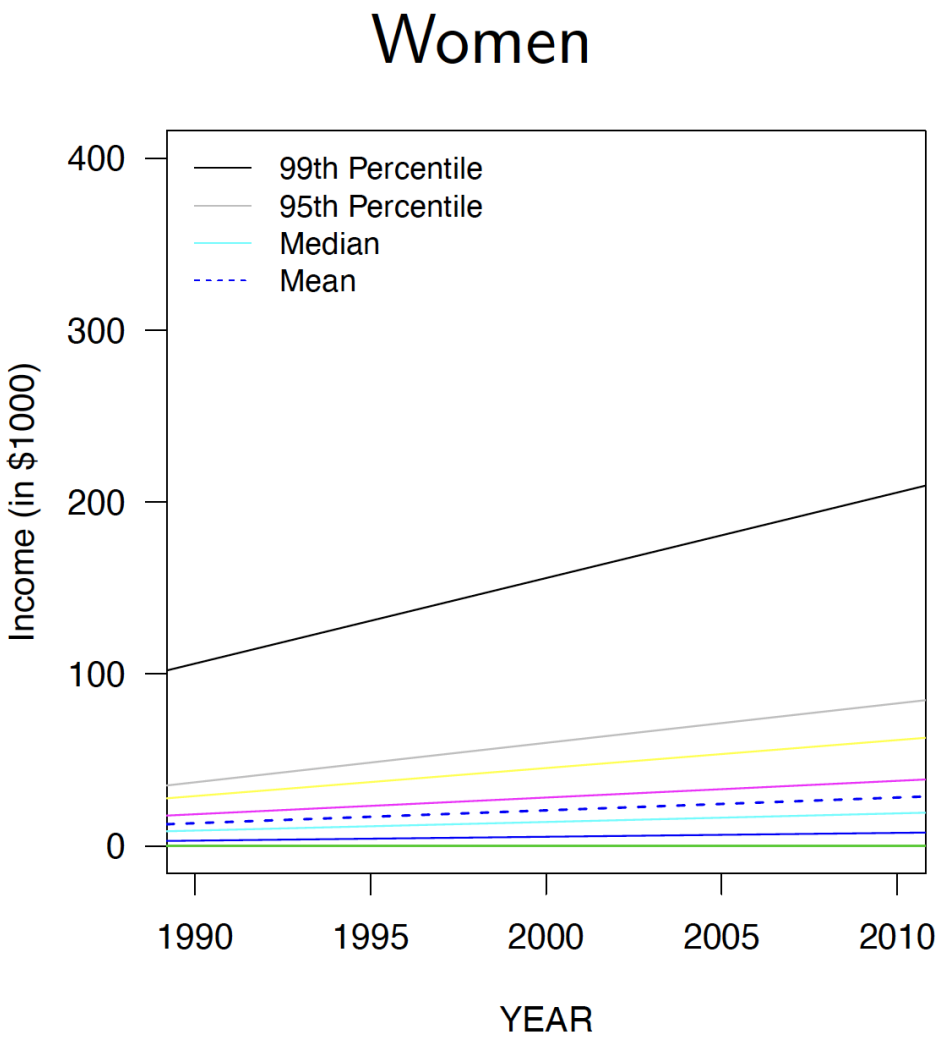
- How has the average income income been increasing over time?
- What has happened to people at the low and high end of the distribution?
- What about the 1%?

An example

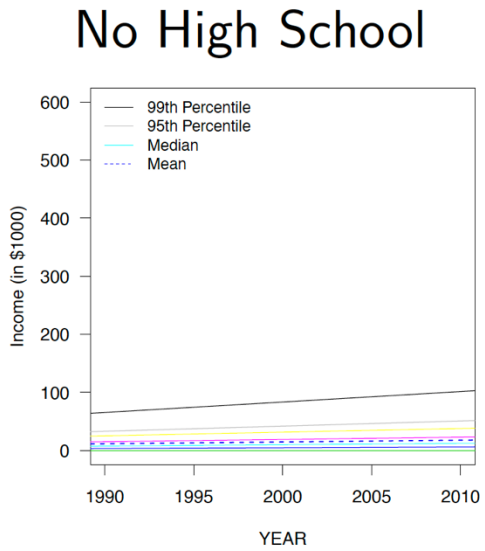

99th Percentile
95th Percentile
Median
Mean
Income (in $1000)
400
300
200
100
0
1990
1995
2000
2005
2010

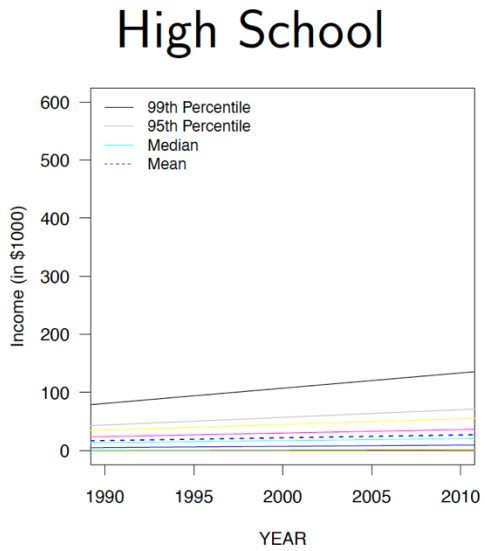
An example

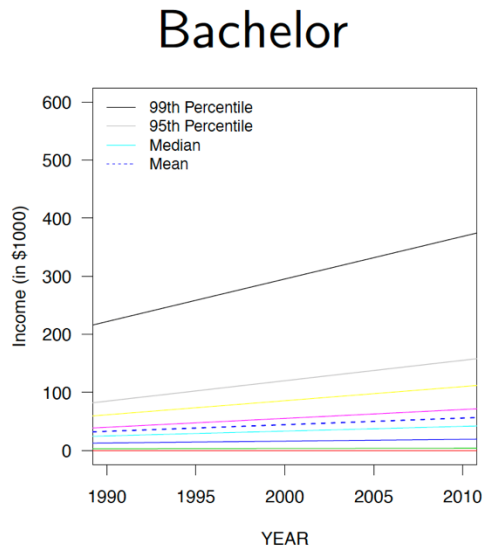

Men
99th Percentile
95th Percentile
Median
Mean
Income (in $1000)
400
300
200
100
0
1990
1995
2000
2005
2010
YEAR

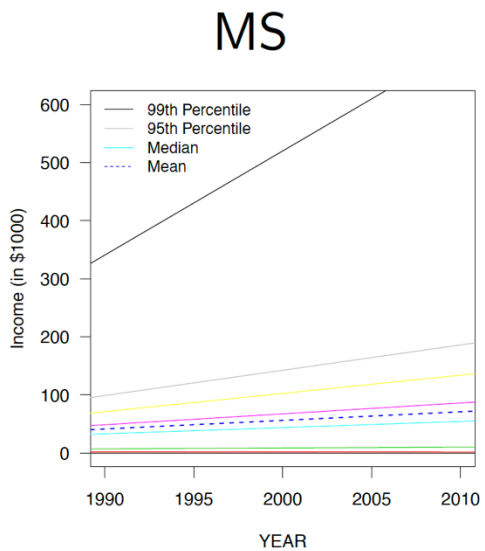

Women
99th Percentile
95th Percentile
Median
Mean
Income (in $1000)
400
300
200
100
0
1990
1995
2000
2005
2010
YEAR

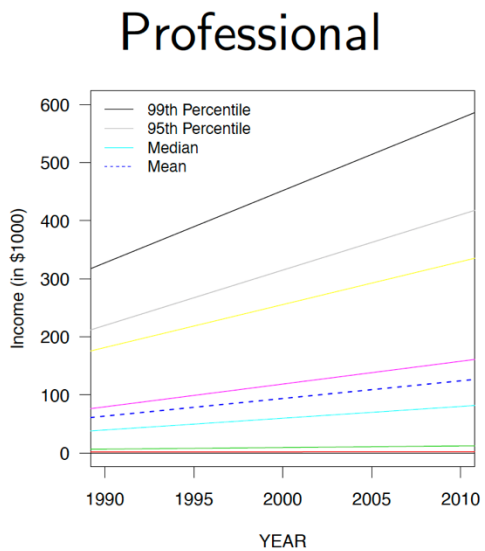
An example

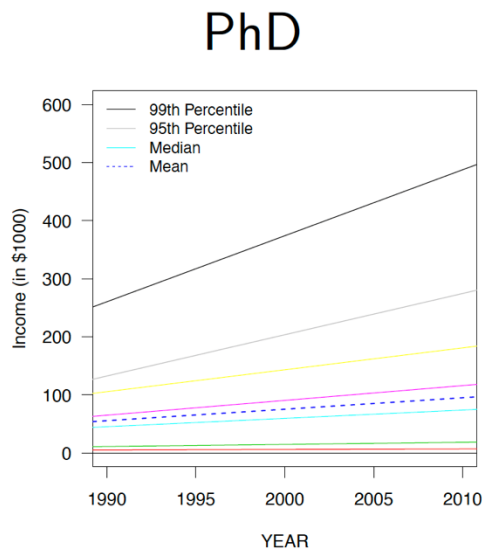

No High School
99th Percentile
95th Percentile
Median
Mean
Income (in $1000)
0
100
200
300
400
500
600
1990
1995
2000
2005
2010
YEAR


High School
99th Percentile
95th Percentile
Median
Mean
Income (in $1000)
0
100
200
300
400
500
600
1990
1995
2000
2005
2010
YEAR


Bachelor
99th Percentile
95th Percentile
Median
Mean
Income (in $1000)
0
100
200
300
400
500
600
1990
1995
2000
2005
2010
YEAR


MS
99th Percentile
95th Percentile
Median
Mean
Income (in $1000)
0
100
200
300
400
500
600
1990
1995
2000
2005
2010
YEAR


Professional
99th Percentile
95th Percentile
Median
Mean
Income (in $1000)
0
100
200
300
400
500
600
1990
1995
2000
2005
2010
YEAR


PhD
99th Percentile
95th Percentile
Median
Mean
Income (in $1000)
0
100
200
300
400
500
600
1990
1995
2000
2005
2010
YEAR

Questions

- Quantile regression
- An example

Getting ready for the lab

- This lab will involve taking some models and concepts from the **Introduction to methods in environmental epidemiology 2** lecture and introduce you to the way quantile regression works with an example from environmental epidemiology.

Application

- How can you imagine applying this learning to your data and your research questions?

Questions

- Questions?

Introduction to methods in environmental epidemiology 2

Robbie M Parks, PhD

21st July 2025

Email: robbie.parks@columbia.edu

BlueSky: @robbiemparks

Website: sparklabnyc.github.io

Art by Amy Wolfe

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH