Descriptive Statistics II

Matti Sarvimäki

Principles of Empirical Analysis
Lecture 3

Course outline and learning objectives

- Data and measurement
  1. introduction, data
  2. today: descriptive statistics
  3. **more descriptive statistics**
- Experimental methods
  1. causality and research designs
  2. statistical significance
  3. statistical power
  4. noncompliance
- Quasi-experimental methods
  1. observational data and quasi-experiments
  2. difference-in-difference (DiD)
  3. regression discontinuity design (RDD)
  4. regression and matching
- Structural methods

- Today's learning objectives. After this lecture you should understand
  1. the meaning of central concepts for conditional descriptive statistics
  2. how to characterize the conditional distributions
  3. how to characterize distributions of more than one variable more generally
  4. key results on recent literature on changes in income distribution

Conditional descriptive statistics

Conditional descriptive statistics

- Conditional descriptives are statistics of a variables *conditional* on another variables
  - The most important: **conditional expectation**

$$\mathbb{E}[Y|X = x]$$

    i.e. expectation of random variable Y when another random variable X takes value x

Conditional descriptive statistics

- Conditional descriptives are statistics of a variables *conditional* on another variables
  - The most important: **conditional expectation**

$$\mathbb{E}[Y|X = x]$$

    i.e. expectation of random variable Y when another random variable X takes value x
  - empirical counterpart: conditional sample average
- All conditional descriptive statistics follow from the **joint distribution** of two or more variables

```
Summary for variables: earn
     by categories of: edul
```

edul	mean	N
Less/unknown	15527	1807
Secodary	22076	2720
Bachelor	32644	1080
Master	42292	346
Lis./PhD	57950	20
Total	23297	5973

Source: FLEED teaching data
`tabstat earn, by(edul) stat(mean N)`
alternatively try: `tabulate edul, sum(earn)`
(see the full code at course website)

Cross tabulation

- A simple, yet efficient way to display (small) data of two variables is **cross tabulation**
  1. the no. rows = no. values that Y can take
  2. the no. columns = no. values that X can take
  3. the cells report no. observations with value (y, x)

edul	woman 0	woman 1	Total
Less/unknown	1,128	894	2,022
Secodary	1,430	1,313	2,743
Bachelor	439	651	1,090
Master	181	185	366
Lis./PhD	17	6	23
Total	3,195	3,049	6,244

Source: FLEED teaching data
`tabulate edul woman`

Joint distribution

- A simple, yet efficient way to display (small) data of two variables is **cross tabulation**
  1. the no. rows = no. values that Y can take
  2. the no. columns = no. values that X can take
  3. the cells report no. observations with value (y, x)
- Alternatively, cross tabulation cells may report the share of observations with value (y, x)

edul	woman 0	woman 1	
Less/unknown	18.07	14.32	
Secodary	22.90	21.03	
Bachelor	7.03	10.43	
Master	2.90	2.96	
Lis./PhD	0.27	0.10	
			100.00

Source: FLEED teaching data
`tabulate edul woman, cell nofreq`

Joint distribution

- A simple, yet efficient way to display (small) data of two variables is **cross tabulation**
  1. the no. rows = no. values that Y can take
  2. the no. columns = no. values that X can take
  3. the cells report no. observations with value (y, x)
- Alternatively, cross tabulation cells may report the share of observations with value (y, x)
- This is the empirical counterpart of the **joint density function**

$$f_{XY}(x, y) = \mathbb{P}(X = x, Y = y)$$

i.e., the probability that random variable X takes the value x *and* that random value Y takes the value y

	woman		
edul	0	1	
Less/unknown	18.07	14.32	
Secodary	22.90	21.03	
Bachelor	7.03	10.43	
Master	2.90	2.96	
Lis./PhD	0.27	0.10	
			100.00

Source: FLEED teaching data
`tabulate edul woman, cell nofreq`

Marginal distribution

- The marginal distribution of Y is defined as

$$f_Y(y) = \sum_{x \in X} f_{XY}(x, y)$$

- This is just probability of Y when not taking the value of X into account

	woman		
edul	0	1	Total
Less/unknown	18.07	14.32	32.38
Secodary	22.90	21.03	43.93
Bachelor	7.03	10.43	17.46
Master	2.90	2.96	5.86
Lis./PhD	0.27	0.10	0.37
			100.00

Source: FLEED teaching data
`tabulate edul woman, cell nofreq`

Marginal distribution

- The marginal distribution of Y is defined as

$$f_Y(y) = \sum_{x \in X} f_{XY}(x, y)$$

- This is just probability of Y when not taking the value of X into account
- Similarly, the marginal distribution of X is

$$f_X(x) = \sum_{y \in Y} f_{XY}(x, y)$$

edul	woman 0	woman 1	Total
Less/unknown	18.07	14.32	32.38
Secodary	22.90	21.03	43.93
Bachelor	7.03	10.43	17.46
Master	2.90	2.96	5.86
Lis./PhD	0.27	0.10	0.37
Total	51.17	48.83	100.00

Source: FLEED teaching data
`tabulate edul woman, cell nofreq`

Conditional distribution

- The conditional distribution of Y is defined as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

i.e., the probability that Y takes value y conditional that X takes value x

edul	woman 0	1	Total
Less/unknown	18.07	14.32	32.38
Secodary	22.90	21.03	43.93
Bachelor	7.03	10.43	17.46
Master	2.90	2.96	5.86
Lis./PhD	0.27	0.10	0.37
Total	51.17	48.83	100.00

Source: FLEED teaching data
`tabulate edul woman, cell nofreq`

Conditional distribution

- The conditional distribution of Y is defined as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

i.e., the probability that Y takes value y conditional that X takes value x

- Example: Probability that a working age woman living in Finland in 2010 had a bachelor degree
  - $\hat{P}(X = w, Y = b) = .1043$
  - $\hat{P}(X = w) = .4883$
  - $\hat{P}(Y = b|X = w) = \frac{.1043}{.4883} \approx .213$
  - where the "hats" indicate that we are using **estimates** of the population probabilities $\mathbb{P}(\cdot)$

	woman		
edul	0	1	Total
Less/unknown	18.07	14.32	32.38
Secodary	22.90	21.03	43.93
Bachelor	7.03	10.43	17.46
Master	2.90	2.96	5.86
Lis./PhD	0.27	0.10	0.37
Total	51.17	48.83	100.00

Source: FLEED teaching data
`tabulate edul woman, cell nofreq`

Conditional expectation

- Let's get back to conditional expectation. When Y is discrete[a], the **conditional expectation function** (CEF) is

$$\mathbb{E}[Y|X=x] = \sum t f_{Y|X}(t|X=x)$$

Summary for variables: earn
by categories of: edul

edul	mean	N
Less/unknown	15527	1807
Secodary	22076	2720
Bachelor	32644	1080
Master	42292	346
Lis./PhD	57950	20
Total	23297	5973

Source: FLEED teaching data
`tabstat earn, by(edul) stat(mean N)`

[a] Continuous version: $\mathbb{E}[Y|X=x] = \int t f_{Y|X}(t|X=x)d(t)$

Conditional expectation

- Let's get back to conditional expectation. When Y is discrete[a], the **conditional expectation function** (CEF) is

$$\mathbb{E}[Y|X = x] = \sum t f_{Y|X}(t|X = x)$$

i.e. **population average of Y holding X fixed**

  - in other words: weighted average of Y, where the weight for of each value of Y is the share of sub-population (for whom $X = x$) with this value of Y

Summary for variables: earn
by categories of: edul

edul	mean	N
Less/unknown	15527	1807
Secodary	22076	2720
Bachelor	32644	1080
Master	42292	346
Lis./PhD	57950	20
Total	23297	5973

Source: FLEED teaching data
`tabstat earn, by(edul) stat(mean N)`

[a] Continuous version: $\mathbb{E}[Y|X = x] = \int t f_{Y|X}(t|X = x)d(t)$

Conditional expectation

- Let's get back to conditional expectation. When Y is discrete[a], the **conditional expectation function** (CEF) is

$$\mathbb{E}[Y|X=x] = \sum t f_{Y|X}(t|X=x)$$

  i.e. **population average of Y holding X fixed**

  - in other words: weighted average of Y, where the weight for of each value of Y is the share of sub-population (for whom $X = x$) with this value of Y

- X can also be a vector, i.e., can include many conditioning variables

```
Summary for variables: earn
     by categories of: edul
```

edul	mean	N
Less/unknown	15527	1807
Secodary	22076	2720
Bachelor	32644	1080
Master	42292	346
Lis./PhD	57950	20
Total	23297	5973

Source: FLEED teaching data
`tabstat earn, by(edul) stat(mean N)`

[a] Continuous version: $\mathbb{E}[Y|X=x] = \int t f_{Y|X}(t|X=x)d(t)$

Conditional expectation

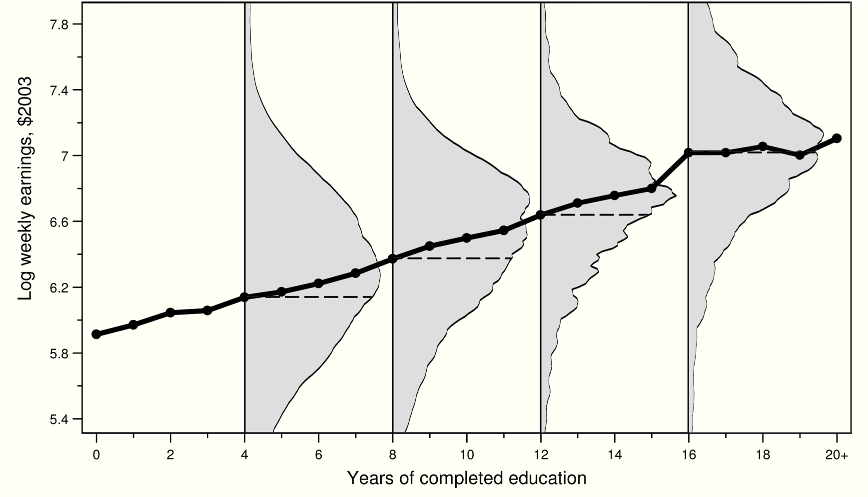


Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

Source: Angrist and Pischke (2009).

Example:
Recent work on the widening U.S. income distribution

Income distribution

- We now have tools to understand the basic results of the income distribution literature
  - group averages
  - changes over the entire distribution
  - extras: top percent shares, social mobility
- Much of this research is based on tax data
  - available over long time periods and many countries, but earlier periods limited to the top (historically, only the rich paid taxes)
  - tax records never capture all income → ongoing work to deal with the missing parts
- Lot's of work also based on surveys, particularly the Labor Force Survey

Source: The Economist, 28 Nov 2019

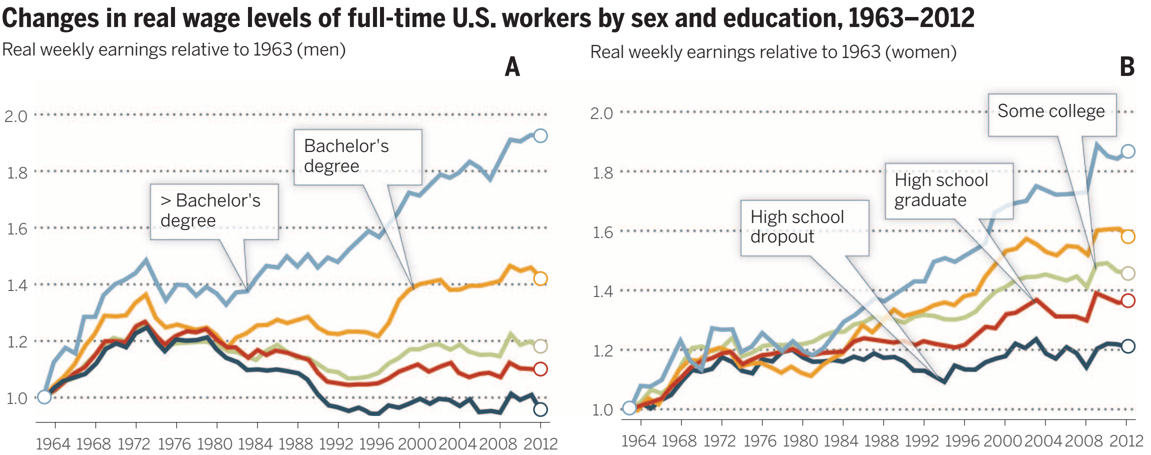
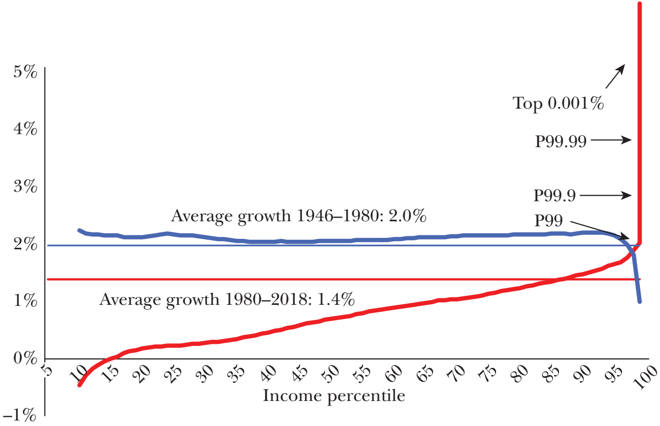


Fig. 6. Change in real wage levels of full-time workers by education, 1963–2012. (**A**) Male workers, (**B**) female workers. Data and sample construction are as in Fig. 3.

Source: Autor (2014), Science.

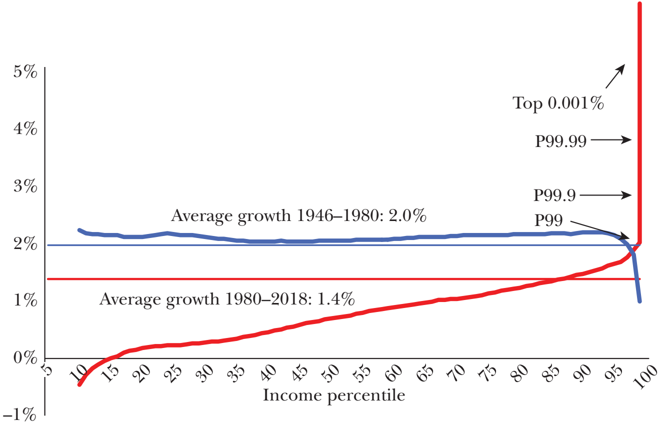
- Estimates over time for $\mathbb{E}[w|E = e, G = G]$, where w is weekly wage, E education level and G is gender. Wages are divided by 1963 group-specific average wages.

Average Annual Income Growth Rates



Source: Saez and Zucman (2019b).
Note: This figure depicts the annual real pre-tax income growth per adult for each percentile in the 1946–1980 period (in blue) and 1980–2018 period (in red). From 1946 to 1980, growth was evenly distributed with all income groups growing at the average 2 percent annual rate (except the top 1 percent which grew slower). From 1980 to 2018, growth has been unevenly distributed with low growth for bottom income groups, mediocre growth for the middle class, and explosive growth at the top.
Source: Saez and Zucman (2020), Journal of Economic Perspectives.

Average Annual Income Growth Rates

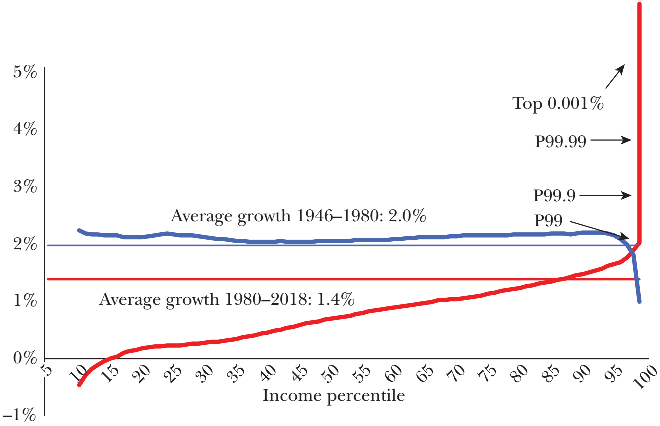


- **1946–1980**: roughly 2% annual income growth across the distribution among "the 99%"

Source: Saez and Zucman (2019b).
Note: This figure depicts the annual real pre-tax income growth per adult for each percentile in the 1946–1980 period (in blue) and 1980–2018 period (in red). From 1946 to 1980, growth was evenly distributed with all income groups growing at the average 2 percent annual rate (except the top 1 percent which grew slower). From 1980 to 2018, growth has been unevenly distributed with low growth for bottom income groups, mediocre growth for the middle class, and explosive growth at the top.

Source: Saez and Zucman (2020), Journal of Economic Perspectives.

Average Annual Income Growth Rates

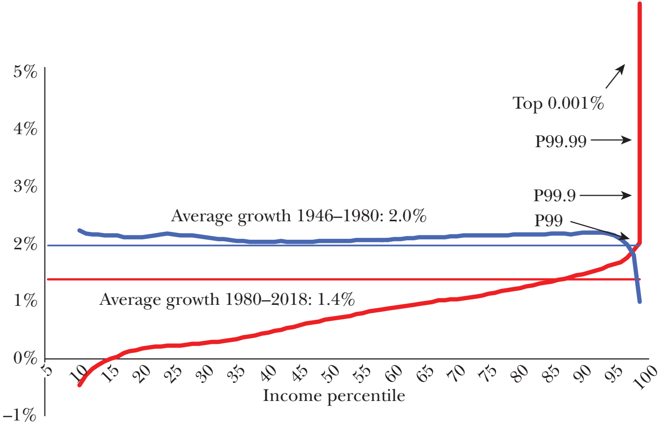


Source: Saez and Zucman (2019b).
Note: This figure depicts the annual real pre-tax income growth per adult for each percentile in the 1946–1980 period (in blue) and 1980–2018 period (in red). From 1946 to 1980, growth was evenly distributed with all income groups growing at the average 2 percent annual rate (except the top 1 percent which grew slower). From 1980 to 2018, growth has been unevenly distributed with low growth for bottom income groups, mediocre growth for the middle class, and explosive growth at the top.

- **1946–1980**: roughly 2% annual income growth across the distribution among "the 99%"
- **1980–2018**: income growth faster among the more wealthy even among "the 99%"; the very top *very* different than the rest

Source: Saez and Zucman (2020), Journal of Economic Perspectives.

Average Annual Income Growth Rates



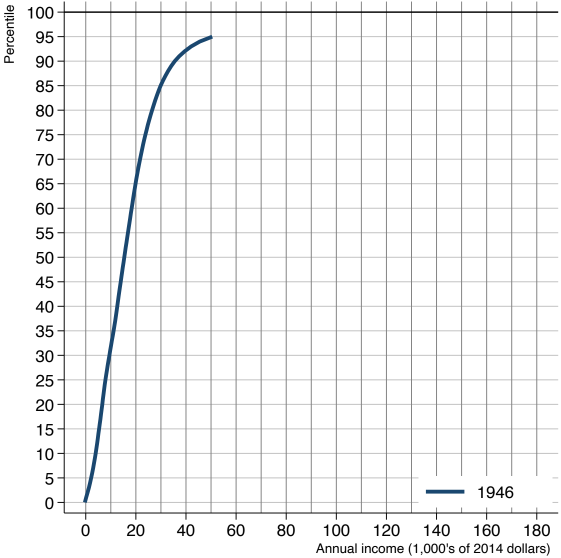
Source: Saez and Zucman (2019b).
Note: This figure depicts the annual real pre-tax income growth per adult for each percentile in the 1946–1980 period (in blue) and 1980–2018 period (in red). From 1946 to 1980, growth was evenly distributed with all income groups growing at the average 2 percent annual rate (except the top 1 percent which grew slower). From 1980 to 2018, growth has been unevenly distributed with low growth for bottom income groups, mediocre growth for the middle class, and explosive growth at the top.

Source: Saez and Zucman (2020), Journal of Economic Perspectives.

- **1946–1980**: roughly 2% annual income growth across the distribution among "the 99%"
- **1980–2018**: income growth faster among the more wealthy even among "the 99%"; the very top *very* different than the rest
- Next: How is this figure constructed?

The U.S. income distribution, 1962–2014, bottom 95 percentiles

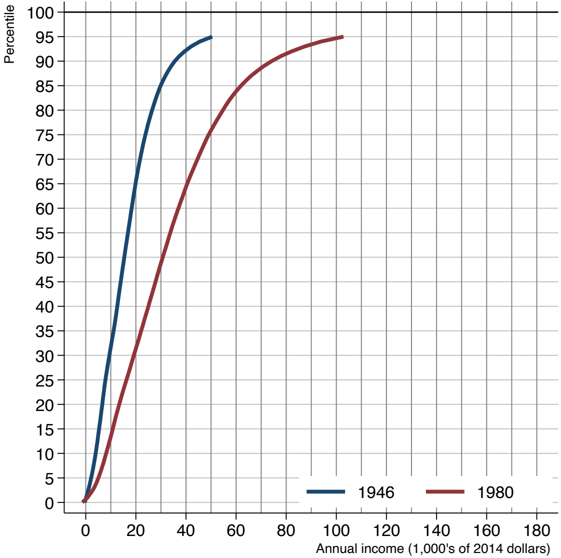
- Let's start with the CDF of income distribution in 1946
  - 90/10 percentile ratio: $\frac{35.5}{3.8} = 9.0$



Source: Piketty, Saez, Zucman (2018) data appendix

The U.S. income distribution, 1962–2014, bottom 95 percentiles

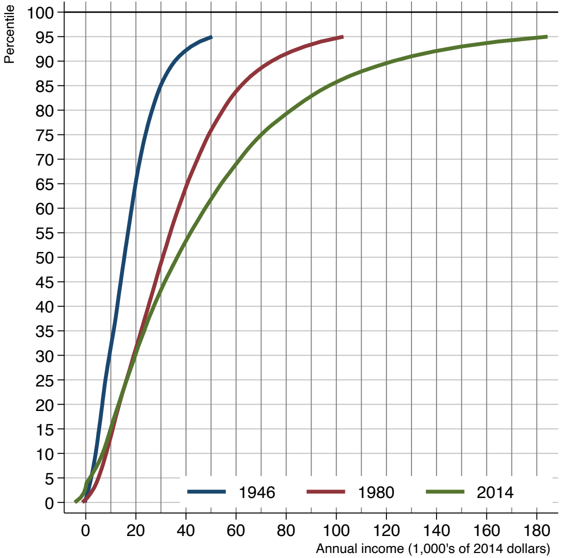
- Let's start with the CDF of income distribution in 1946
  - 90/10 percentile ratio: $\frac{35.5}{3.8} = 9.0$
- Adding the CDF for 1980 income
  - 90/10 percentile ratio: $\frac{74.2}{8.1} = 9.1$



Source: Piketty, Saez, Zucman (2018) data appendix

The U.S. income distribution, 1962–2014, bottom 95 percentiles

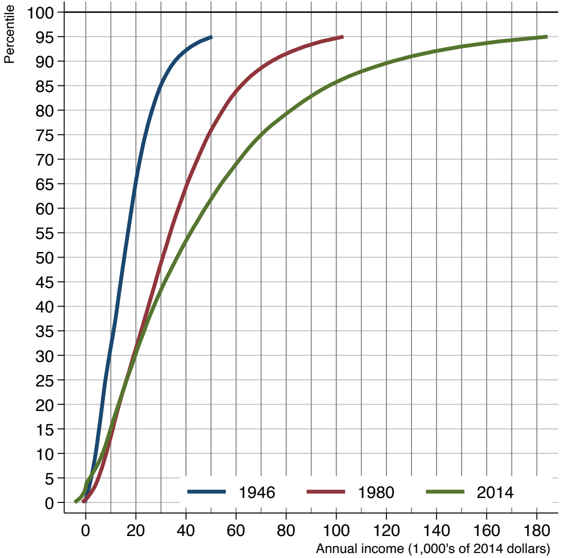
- Let's start with the CDF of income distribution in 1946
  - 90/10 percentile ratio: $\frac{35.5}{3.8} = 9.0$
- Adding the CDF for 1980 income
  - 90/10 percentile ratio: $\frac{74.2}{8.1} = 9.1$
- Adding the CDF for 2014 income
  - 90/10 percentile ratio: $\frac{122.6}{6.7} = 18.2$



Source: Piketty, Saez, Zucman (2018) data appendix

The U.S. income distribution, 1962–2014, bottom 95 percentiles

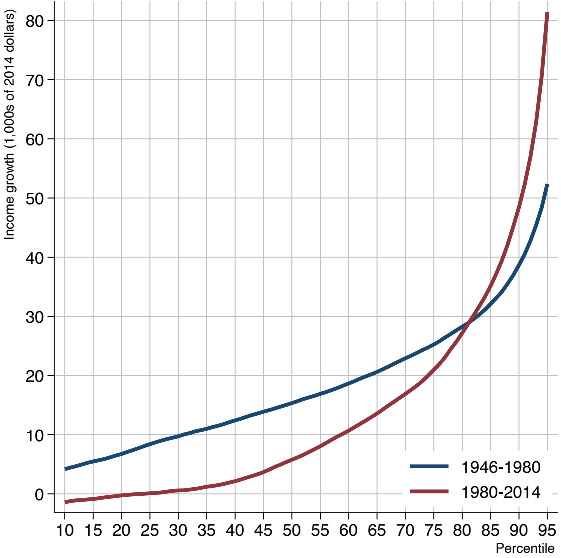
- Let's start with the CDF of income distribution in 1946
  - 90/10 percentile ratio: $\frac{35.5}{3.8} = 9.0$
- Adding the CDF for 1980 income
  - 90/10 percentile ratio: $\frac{74.2}{8.1} = 9.1$
- Adding the CDF for 2014 income
  - 90/10 percentile ratio: $\frac{122.6}{6.7} = 18.2$
- Horizontal distance btw the CDFs = dollar change for each percentile
  - these are not the same *people*; we are comparing percentiles
  - next: from dollar changes to annualized growth rates



Source: Piketty, Saez, Zucman (2018) data appendix

The U.S. income distribution, 1962–2014, bottom 95 percentiles

- Let's first calculate dollar changes
  - i.e., horizontal distance btw CDFs

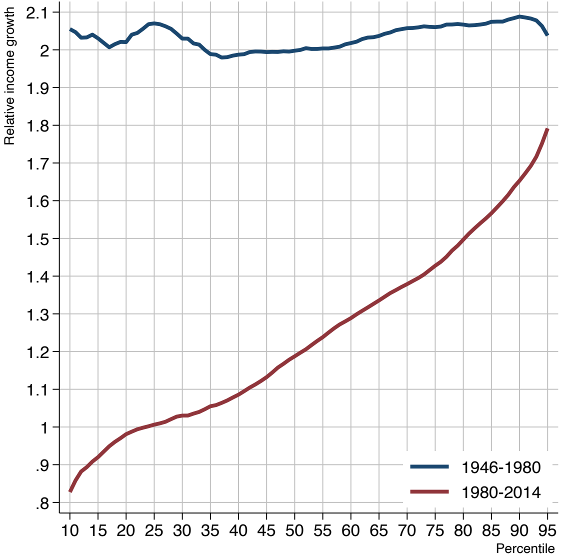


Source: Piketty, Saez, Zucman (2018) data appendix

The U.S. income distribution, 1962–2014, bottom 95 percentiles

- Let's first calculate dollar changes
  - i.e. horizontal distance btw CDFs
- Then: relative change in income between years *a* and *b* for quantile τ

$$G = \frac{Q_b(\tau)}{Q_a(\tau)}$$



Source: Piketty, Saez, Zucman (2018) data appendix

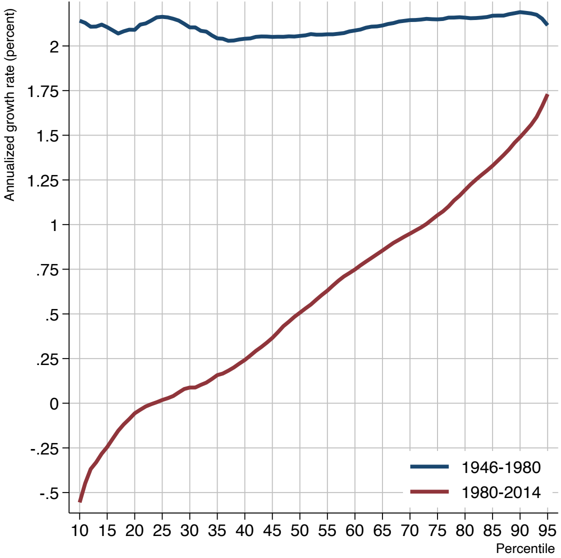
The U.S. income distribution, 1962–2014, bottom 95 percentiles

- Let's first calculate dollar changes
  - i.e. horizontal distance btw CDFs
- Then: relative change in income between years *a* and *b* for quantile τ

$$G = \frac{Q_b(\tau)}{Q_a(\tau)}$$

- Finally: annualization, i.e. annual growth rate g that accumulates to G over 34 years

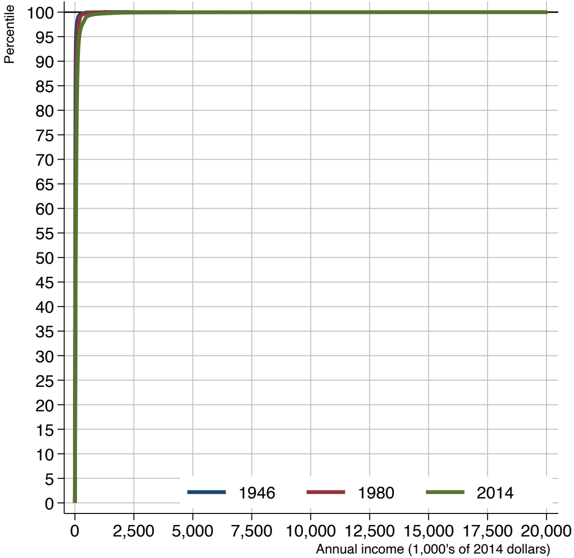
$$(1+g)^{34} = G \Leftrightarrow g = G^{1/34} - 1$$



Source: Piketty, Saez, Zucman (2018) data appendix

The U.S. income distribution, 1962–2014, full distribution

- CDFs for very skewed distributions are uninformative

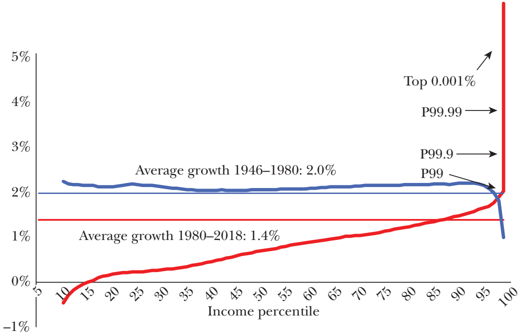


Source: Piketty, Saez, Zucman (2018) data appendix

The U.S. income distribution, 1962–2014, full distribution

- CDFs for very skewed distributions are uninformative ... but changes can nevertheless be made visible

Average Annual Income Growth Rates



Source: Saez and Zucman (2019b).
Note: This figure depicts the annual real pre-tax income growth per adult for each percentile in the 1946–1980 period (in blue) and 1980–2018 period (in red). From 1946 to 1980, growth was evenly distributed with all income groups growing at the average 2 percent annual rate (except the top 1 percent which grew slower). From 1980 to 2018, growth has been unevenly distributed with low growth for bottom income groups, mediocre growth for the middle class, and explosive growth at the top.

Source: Saez and Zucman (2020), Journal of Economic Perspectives.

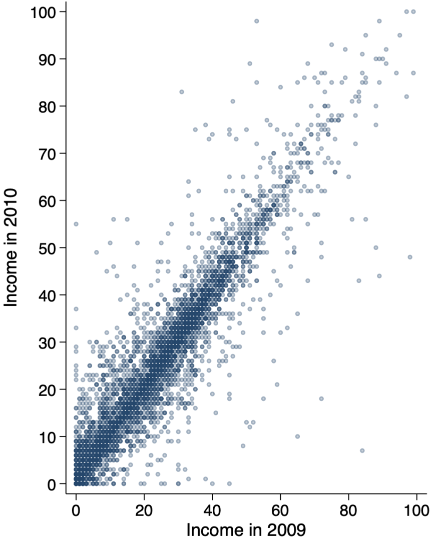
Correlation

Scatter plot

- Conditional expectation is a powerful way to detect how variables are associated with each other

Scatter plot

- Conditional expectation is a powerful way to detect how variables are associated with each other
- An alternative approach is to show all observations and plot two variables against each other
- Example: persistence of income over time
    - **scatter plot**: each dot in this graph shows each individual's income in 2009 and 2010

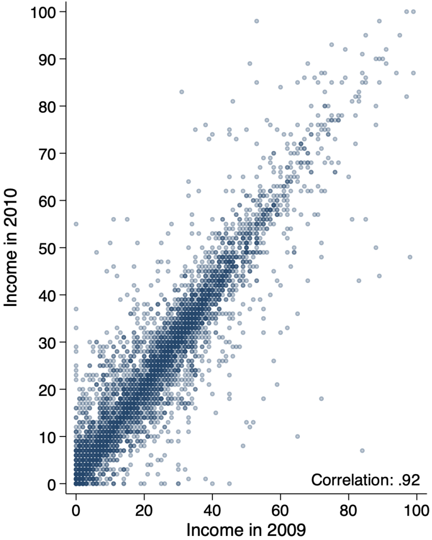


Source: FLEED teaching data

`scatter earn earn_t1, mcolor(navy%25) msize(vsmall)`

Scatter plot

- Conditional expectation is a powerful way to detect how variables are associated with each other
- An alternative approach is to show all observations and plot two variables against each other
- Example: persistence of income over time
  - **scatter plot**: each dot in this graph shows each individual's income in 2009 and 2010
- The best known descriptive statistic to characterize how two variables' values are aligned is **correlation**
  - here, the correlation is 0.92
  - next: what does that mean?



Source: FLEED teaching data

`scatter earn earn_t1, mcolor(navy%25) msize(vsmall)`

Covariance

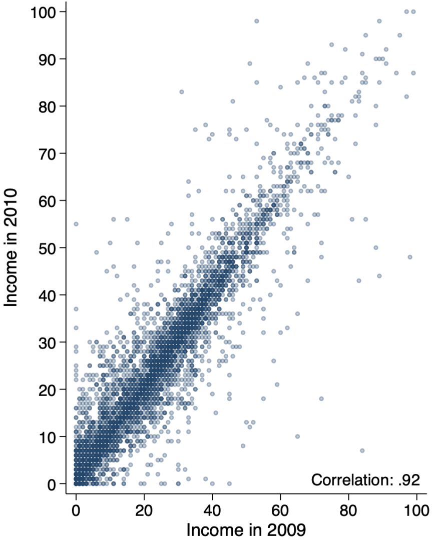
- To get to correlation, we need to first define the **covariance** of Y and X

$$Cov(X, Y) = \mathbb{E}[X - \mathbb{E}(X)]\mathbb{E}[Y - \mathbb{E}(Y)]$$

... and its empirical counterpart

$$\widehat{Cov}(X, Y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Here, the covariance is 256.6
  - a hard number to interpret



Source: FLEED teaching data

```
scatter earn earn_t1, mcolor(navy%25) msize(vsmall)
```

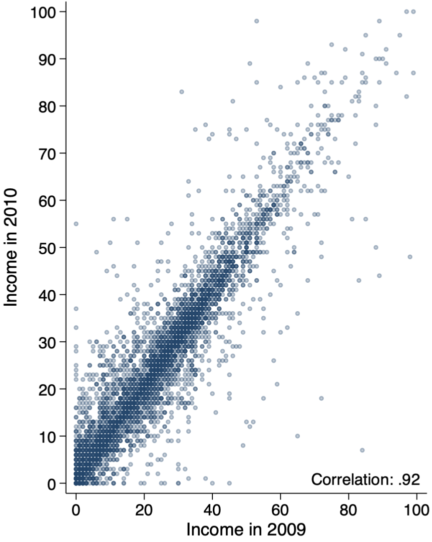
Correlation

- Pearson correlation coefficient is a scaled covariance

$$Cor(X, Y) = \rho_{X,Y} = \frac{Cov(X, Y)}{SD(X)SD(Y)}$$

that varies between $-1 \leq Cor(X, Y) \leq 1$

  - just makes the number easier to interpret



Source: FLEED teaching data

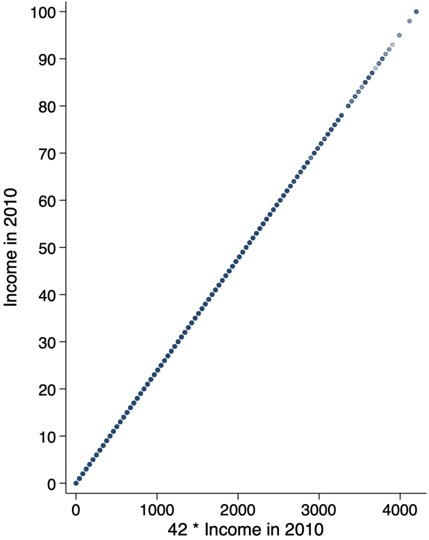
```
scatter earn earn_t1, mcolor(navy%25) msize(vsmall)
```

Correlation

- Pearson correlation coefficient is a scaled covariance

$$Cor(X,Y) = \rho_{X,Y} = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

  that varies between $-1 \leq Cor(X,Y) \leq 1$
  - just makes the number easier to interpret
- More examples
  - correlation 1

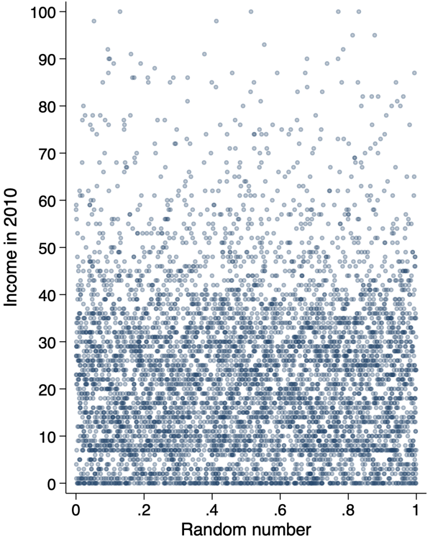


Correlation

- Pearson correlation coefficient is a scaled covariance

$$Cor(X,Y) = \rho_{X,Y} = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

that varies between $-1 \leq Cor(X,Y) \leq 1$
  - just makes the number easier to interpret
- More examples
  - correlation 1
  - correlation 0.009

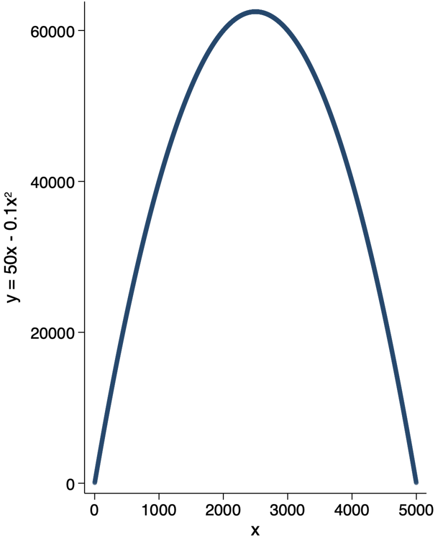


Correlation

- Pearson correlation coefficient is a scaled covariance

$$Cor(X,Y) = \rho_{X,Y} = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

  that varies between $-1 \leq Cor(X,Y) \leq 1$
  - just makes the number easier to interpret
- More examples
  - correlation 1
  - correlation 0.009
  - correlation 0

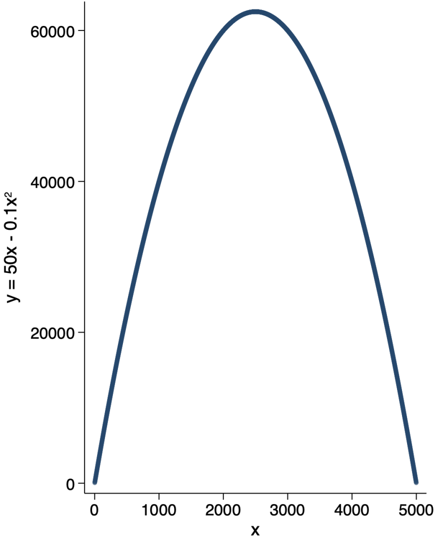


Correlation

- Pearson correlation coefficient is a scaled covariance

$$Cor(X, Y) = \rho_{X,Y} = \frac{Cov(X, Y)}{SD(X)SD(Y)}$$

  that varies between $-1 \leq Cor(X, Y) \leq 1$
  - just makes the number easier to interpret
- More examples
  - correlation 1
  - correlation 0.009
  - correlation 0
- Correlation measures a linear dependence
  - the point: possible to have perfect dependence and zero correlation



Regression

Regression

- A closely related approach for assessing linear dependence: bivariate **regression model**

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Regression

- A closely related approach for assessing linear dependence: bivariate **regression model**

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y is the dependent variable (or outcome)

Regression

- A closely related approach for assessing linear dependence: bivariate **regression model**

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y is the dependent variable (or outcome)
- X is the independent variable (or regressor)
  - **observed** in data

Regression

- A closely related approach for assessing linear dependence: bivariate **regression model**

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y is the dependent variable (or outcome)
- X is the independent variable (or regressor)
  - **observed** in data
- ϵ is the residual (or "error term")
  - represents the relevant **unobserved factors**
  - defined to have $\mathbb{E}[\epsilon] = 0$

Regression

- A closely related approach for assessing linear dependence: bivariate **regression model**

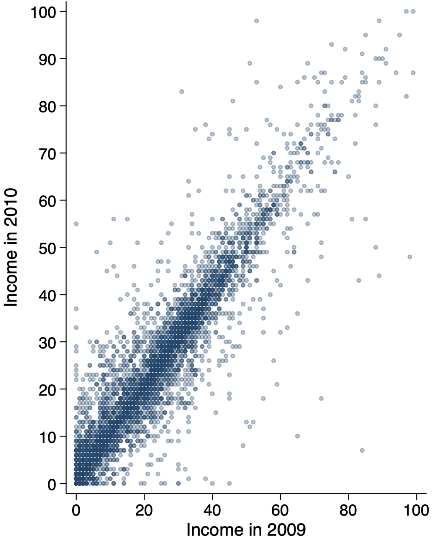
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y is the dependent variable (or outcome)
- X is the independent variable (or regressor)
  - **observed** in data
- ϵ is the residual (or "error term")
  - represents the relevant **unobserved factors**
  - defined to have $\mathbb{E}[\epsilon] = 0$
- parameters: β_0 (constant), β_1 (regression coefficient)

Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- *Question*: How should we set β_0 and β_1 to best describe the data?



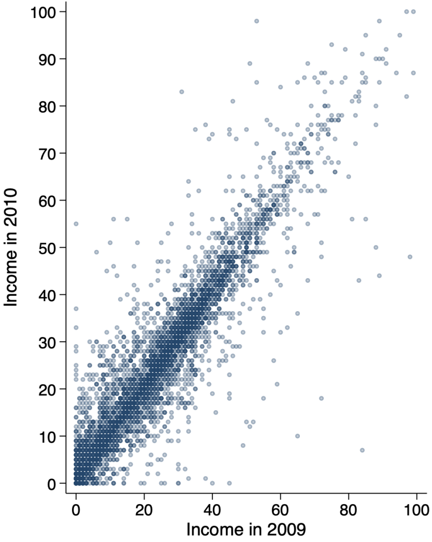
Source: FLEED teaching data
`scatter earn earn_t1`

Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- *Question*: How should we set β_0 and β_1 to best describe the data?
- *One answer*: Ordinary Least Squares (OLS)

$$\arg\min_{\beta_0,\beta 1} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$



Source: FLEED teaching data
`scatter earn earn_t1`

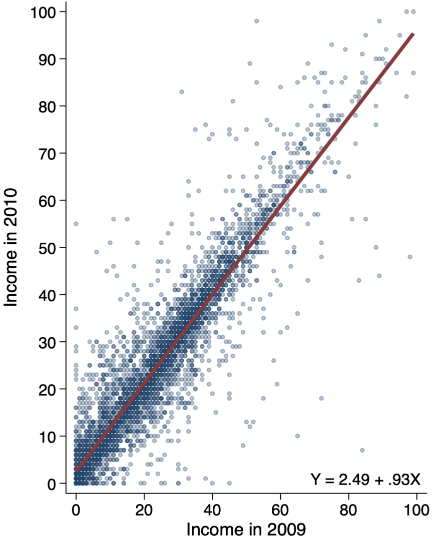
Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- *Question*: How should we set β_0 and β_1 to best describe the data?
- *One answer*: Ordinary Least Squares (OLS)

$$\text{arg min}_{\beta_0,\beta 1} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

- In words: let's find the values of β_0 and β_1 that minimize (the square of) the difference between observed data and regression model's prediction



Source: FLEED teaching data
the code is available at the course's website

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- *Question*: How should we set β_0 and β_1 to best describe the data?
- *One answer*: Ordinary Least Squares (OLS)

$$\text{argmin}_{\beta_0,\beta 1} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

- In words: let's find the values of β_0 and β_1 that minimize (the square of) the difference between observed data and regression model's prediction
  - here, the answer is: $\hat{\beta}_0 = 2.49$, $\hat{\beta}_1 = 0.93$

Source	SS	df	MS			
				Number of obs	=	5,777
				F(1, 5775)	=	33626.24
Model	1390738.85	1	1390738.85	Prob > F	=	0.0000
Residual	238846.737	5,775	41.3587423	R-squared	=	0.8534
				Adj R-squared	=	0.8534
Total	1629585.58	5,776	282.130468	Root MSE	=	6.4311

earn	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
earn_t1	.9383461	.0051171	183.37	0.000	.9283147	.9483776
_cons	2.487598	.1438088	17.30	0.000	2.205679	2.769518

Source: FLEED teaching data
`regress earn earn_t1`

Regression vs correlation

- Turns out that correlation and bivariate regression are closely related, namely:

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)}$$

Regression vs correlation

- Turns out that correlation and bivariate regression are closely related, namely:

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)}$$

- Compare to Pearson correlation coefficient:

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Regression vs correlation

- Turns out that correlation and bivariate regression are closely related, namely:

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)}$$

- Compare to Pearson correlation coefficient:

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

- In our example
  - $\hat{\beta}_0 = 2.49$, $\hat{\beta}_1 = 0.93$
  - $\hat{\rho}_{X,Y} = 0.92$

Regression vs correlation

- Turns out that correlation and bivariate regression are closely related, namely:

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)}$$

- Compare to Pearson correlation coefficient:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

- In our example
  - $\hat{\beta}_0 = 2.49$, $\hat{\beta}_1 = 0.93$
  - $\hat{\rho}_{X,Y} = 0.92$
- Here, $\hat{\rho}_{X,Y} \approx \hat{\beta}_1$ because $Var(X) \approx Var(Y)$

Regression vs correlation

- Turns out that correlation and bivariate regression are closely related, namely:

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)}$$

- Compare to Pearson correlation coefficient:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

- In our example
  - $\hat{\beta}_0 = 2.49$, $\hat{\beta}_1 = 0.93$
  - $\hat{\rho}_{X,Y} = 0.92$
- Here, $\hat{\rho}_{X,Y} \approx \hat{\beta}_1$ because $Var(X) \approx Var(Y)$
- In other applications numerical values may differ ... but this is just a matter of different scaling
  - i.e., both measure essentially the same thing

Regression and Expectation

- If the conditional expectation function (CEF) of Y is linear in X, then:

$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$$

Regression and Expectation

- If the conditional expectation function (CEF) of Y is linear in X, then:

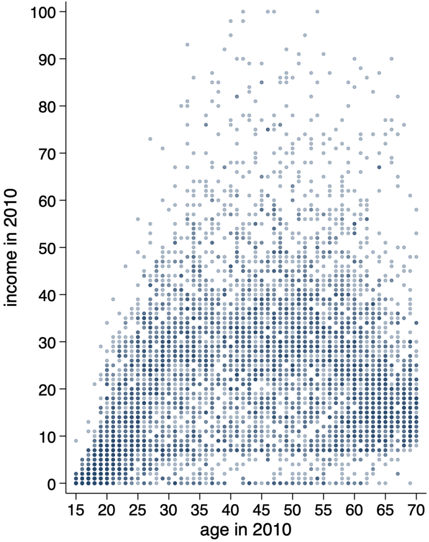
$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$$

- Even if CEF is not linear, regression still provides an approximation
  - specifically, regression is the best minimum mean squared error linear approximation of CEF (more about this in later courses)
  - for many (not all) applications, this is good enough ... particularly when using multivariate regression to make it more flexible (next example)

Example: Age and income

Association between age and income

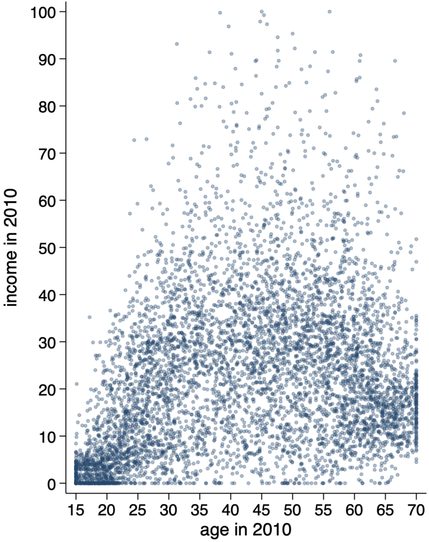
- *Question*: How does income vary with age?
  - scatter plot of the full data



Source: FLEED teaching data
`scatter earn age`

Association between age and income

- *Question*: How does income vary with age?
  - scatter plot of the full data
  - adding a little bit of noise sometimes makes the pattern more visible

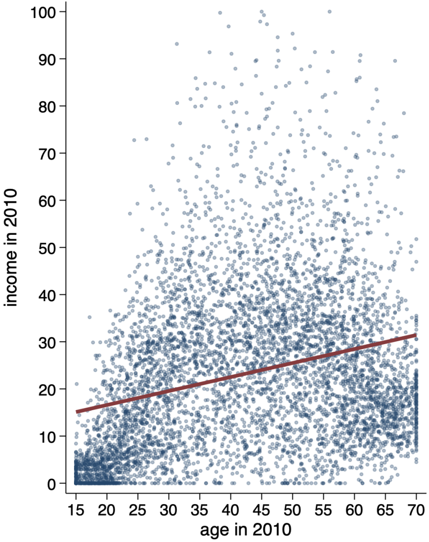


Source: FLEED teaching data

`scatter earn age, jitter(10)`

Association between age and income

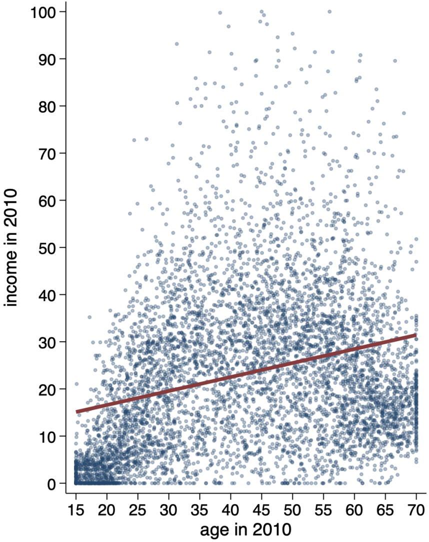
- *Question*: How does income vary with age?
  - scatter plot of the full data
  - adding a little bit of noise sometimes makes the pattern more visible
- Let's use the measures of linear dependence
  - $\hat{\rho}_{X,Y} = 0.28$
  - estimating regression $Y = \beta_0 + \beta_1 X + \epsilon$ yields parameter estimates of $\hat{\beta}_0 = 10,654$, $\hat{\beta}_1 = 297$
    - note that these estimates are in euros, while the figure's y-axis is in thousands of euros



Source: FLEED teaching data
the code is available at the course's website

Association between age and income

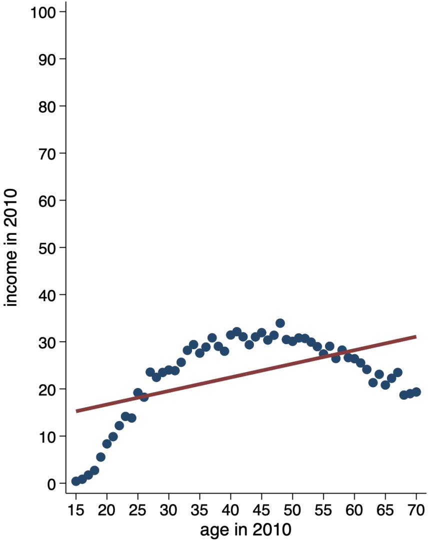
- *Question*: How does income vary with age?
  - scatter plot of the full data
  - adding a little bit of noise sometimes makes the pattern more visible
- Let's use the measures of linear dependence
  - $\hat{\rho}_{X,Y} = 0.28$
  - estimating regression $Y = \beta_0 + \beta_1 X + \epsilon$ yields parameter estimates of $\hat{\beta}_0 = 10,654$, $\hat{\beta}_1 = 297$
- Are these helpful summary statistics?
  - what do they imply for $\mathbb{E}[Y|X = x]$?



Source: FLEED teaching data
the code is available at the course's website

Association between age and income

- *Question*: How does income vary with age?
  - scatter plot of the full data
  - adding a little bit of noise sometimes makes the pattern more visible
- Let's use the measures of linear dependence
  - $\hat{\rho}_{X,Y} = 0.28$
  - estimating regression $Y = \beta_0 + \beta_1 X + \epsilon$ yields parameter estimates of $\hat{\beta}_0 = 10,654$, $\hat{\beta}_1 = 297$
- Are these helpful summary statistics?
  - what do they imply for $\mathbb{E}[Y|X = x]$?
- Compare to sample average by age
  - these are **nonparametric** estimates for $\mathbb{E}[Y|X = x]$
  - any ideas about how to improve the fit?



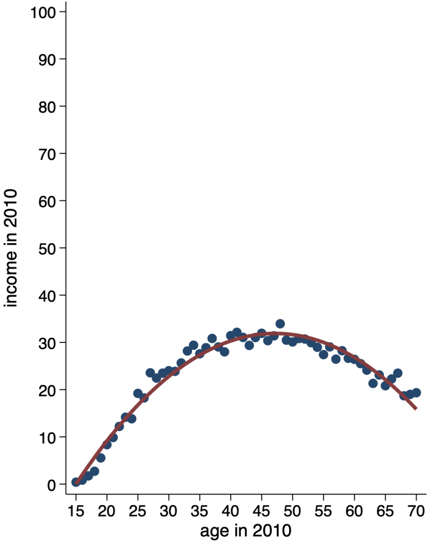
Source: FLEED teaching data
`the code is available at the course's website`

Association between age and income

- Let's use a multivariate regression model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- Now, the estimates that best fit the data best are: $\hat{\beta}_0 = -37,549$, $\hat{\beta}_1 = 2.857$, $\hat{\beta}_2 = -31$



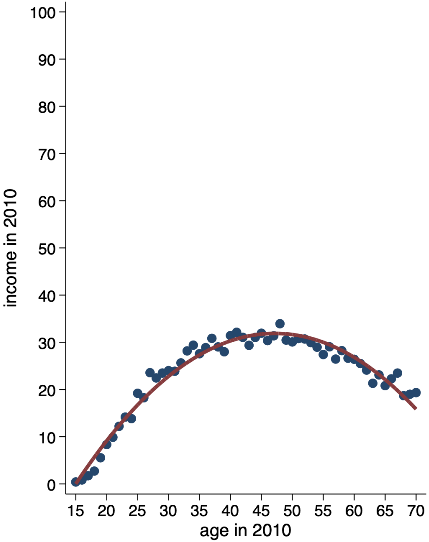
Source: FLEED teaching data
the code is available at the course's website

Association between age and income

- Let's use a multivariate regression model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- Now, the estimates that best fit the data best are:
  $\hat{\beta}_0 = -37,549, \ \hat{\beta}_1 = 2.857, \ \hat{\beta}_2 = -31$
- Are these helpful summary statistics?
  - seems pretty good for approximating $\mathbb{E}[Y|X = x]$ within the 15–70 age range (the figure)
  - less so outside this age range, e.g., suggest that expected income of a new-born would be -37,549€
- General lesson: looking at the data in several ways almost always a good idea



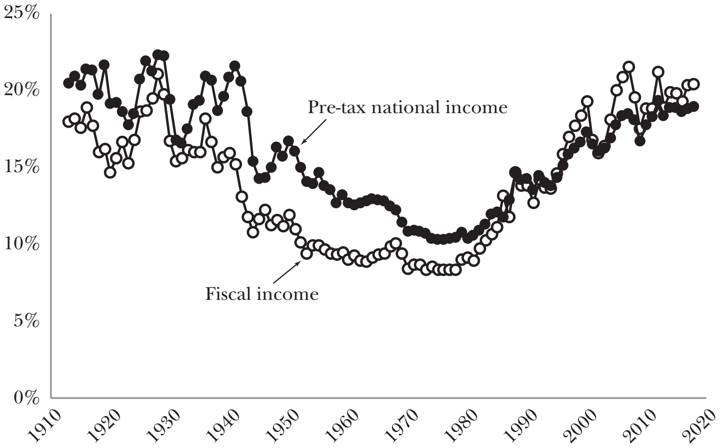
Source: FLEED teaching data
`the code is available at the course's website`

Summary

- Today we learned the basics tools for characterizing joint distributions
- You should now know well the following concepts:
  - joint, marginal and conditional distribution
  - conditional expectation function
  - cross tabulation, scatter plots
  - covariance and correlation
  - regression, ordinary least square (OLS)

Extra 1: Top 1%

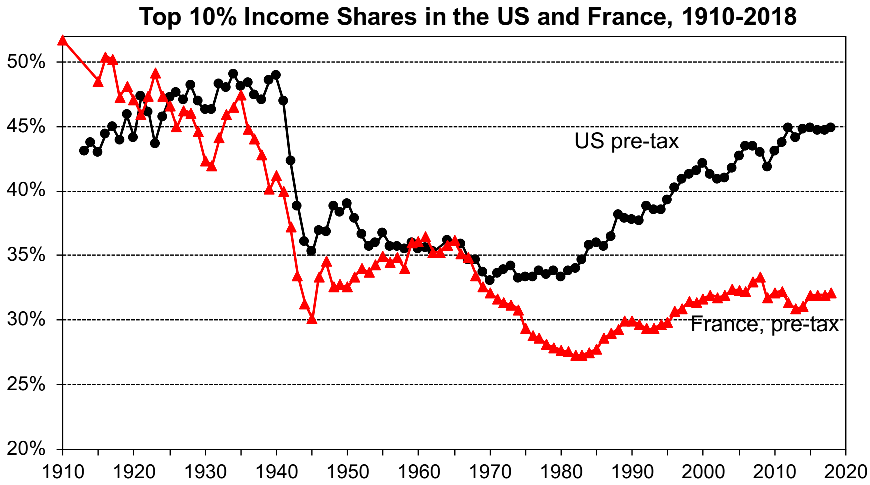
Share of Income Earned by the Top 1 Percent



- US top 1% share based on tax data only and Distributional National Accounts by PSZ

Note: This figure compares the share of fiscal income earned by the top 1 percent tax units (from Piketty and Saez 2003, updated series including capital gains in income to compute shares but not to define ranks, to smooth the lumpiness of realized capital gains) to the share of pre-tax national income earned by the top 1 percent equal-split adults (from Piketty, Saez, and Zucman 2018, updated September 2020, available on WID.world).

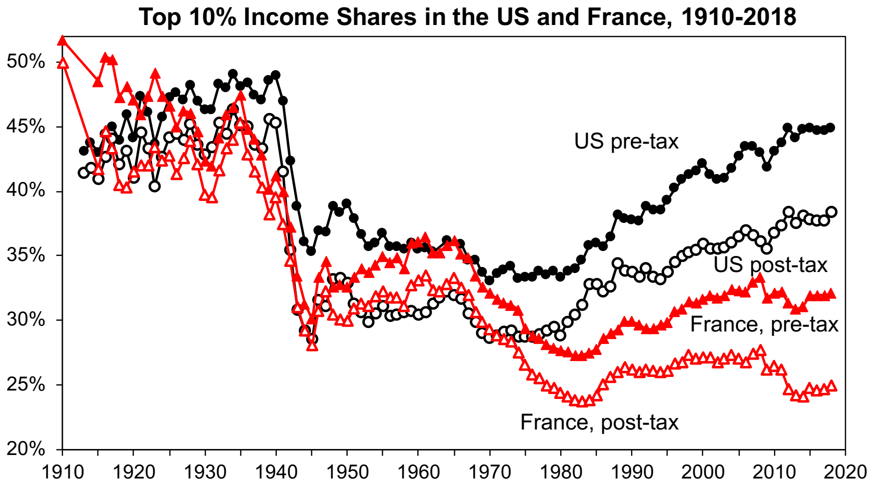
Source: Saez and Zucman (2020), Journal of Economic Perspectives.



Top income shares of pretax national income among adults (income within married couples equally split).
Source is Piketty, Saez, Zucman (2018) for US and Piketty et al. (2020) for France.

- Comparable measures constructed for many countries and made available through the WID database

Source: Saez (2021), AEA Distinguished Lecture.



Top income shares of pretax and posttax national income among adults (income within married couples equally split). Source is Piketty, Saez, Zucman (2018) for US and Piketty et al. (2020) for France.

Source: Saez (2021), AEA Distinguished Lecture.

- Comparable measures constructed for many countries and made available through the WID database
- Taking into account taxes and transfers matters

Extra 2: Intergenerational mobility

Intergenerational mobility

- A complementary way to think about inequality is based on the idea of equality of opportunities
  - the extent to which people compete on a "level playing field" vs. inherit their position

Intergenerational mobility

- A complementary way to think about inequality is based on the idea of equality of opportunities
  - the extent to which people compete on a "level playing field" vs. inherit their position
- An incomplete, but powerful measure

$$\mathbb{E}[p_c | P_p = p_p]$$

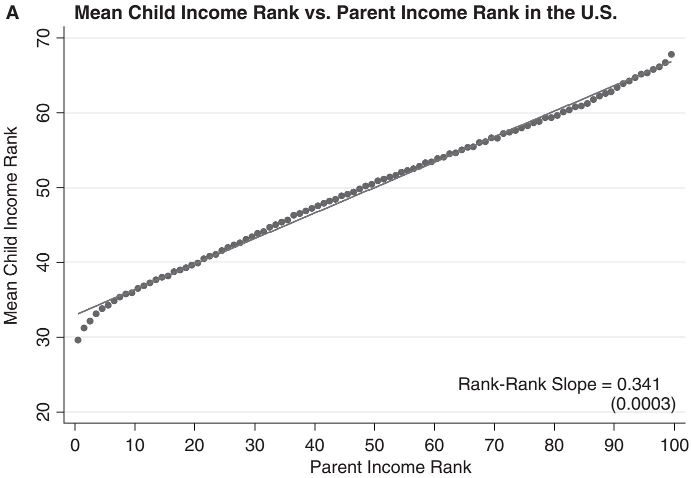
where p_c is the child's position in (lifetime) income distribution and p_p is her parent's position

Intergenerational mobility

- A complementary way to think about inequality is based on the idea of equality of opportunities
  - the extent to which people compete on a "level playing field" vs. inherit their position
- An incomplete, but powerful measure

$$\mathbb{E}[p_c | P_p = p_p]$$

where p_c is the child's position in (lifetime) income distribution and p_p is her parent's position



Children born in 1980–82. Their income is the mean of 2011–2012 family income (when the child is approximately 30 years old). Parent income is mean family income from 1996 to 2000. Children are ranked relative to other children in their birth cohort, and parents are ranked relative to all other parents. *Source:* Chetty, Hendren, Kline and Saez (2014), Quarterly Journal of Economics.

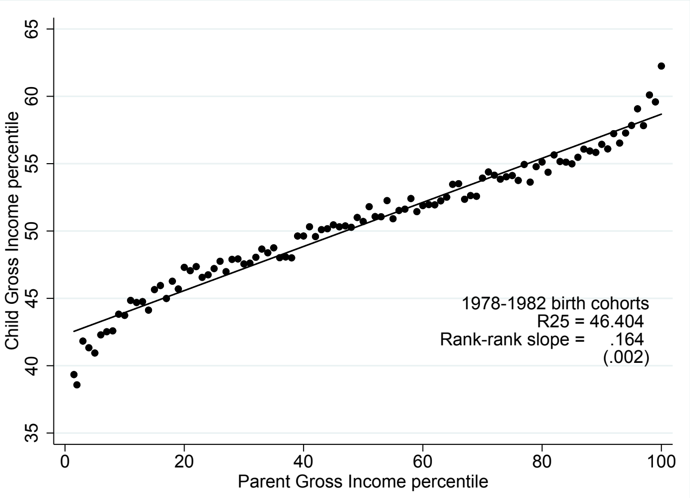
Intergenerational mobility

- A complementary way to think about inequality is based on the idea of equality of opportunities
  - the extent to which people compete on a "level playing field" vs. inherit their position
- An incomplete, but powerful measure

$$\mathbb{E}[p_c | P_p = p_p]$$

where p_c is the child's position in (lifetime) income distribution and p_p is her parent's position

... in Finland



Source: Unpublished, ongoing work.