ECON4225 Homework 1

Anna Duan

2025-10-06

Part 1: Overview of the Dataset

Question 1: household and variable count

This dataset has 9066 households and 38 variables, including the household ID.

```
## [1] "PSID row count: 9066"
```

```
## [1] "PSID variable count: 38"
```

Question 2: age of household head

The average age of the household head is 46.25. The highest age is 102 and the lowest age is 18.

```
## [1] "Mean age: 46.25"
```

```
## [1] "Highest age: 102"
```

```
## [1] "Lowest age: 18"
```

Question 3: average household income

The average household income is $78,265.69. This is lower than the $142,000 reported by the SCF (based on class slides for Section 1.2, slide 9). One possible reason for this difference is that the SCF intentionally oversamples wealthy households to collect information about the income and wealth patterns at the top of the income distribution. In addition to SCF's intentional focus on a high-income subset of the population, other differences that may create differences in the mean household income collected include SCF's smaller sample size (6k, compared to PSID's 10k) and lower frequency.

SCF also surveys a smaller (6k) group compared to PSID (10k).

```
## [1] "Mean household income: 78265.69"
```

Question 4: income range

The household incomes in the dataset have a wide range: the lowest is -$267,900 and the highest is $2,125,100. According to the codebook, the households with negative incomes likely incurred business or farm losses. This may include losses from business investments, farming, or other financial decisions.

```
## [1] "Lowest household income: -267900"
```

```
## [1] "Highest household income: 2125100"
```

Question 5: age of household head's spouse

Of the 4,523 households with a spouse present, the average age of the spouse is 45.64 years.

```
## [1] "Mean spouse age: 45.64"
```

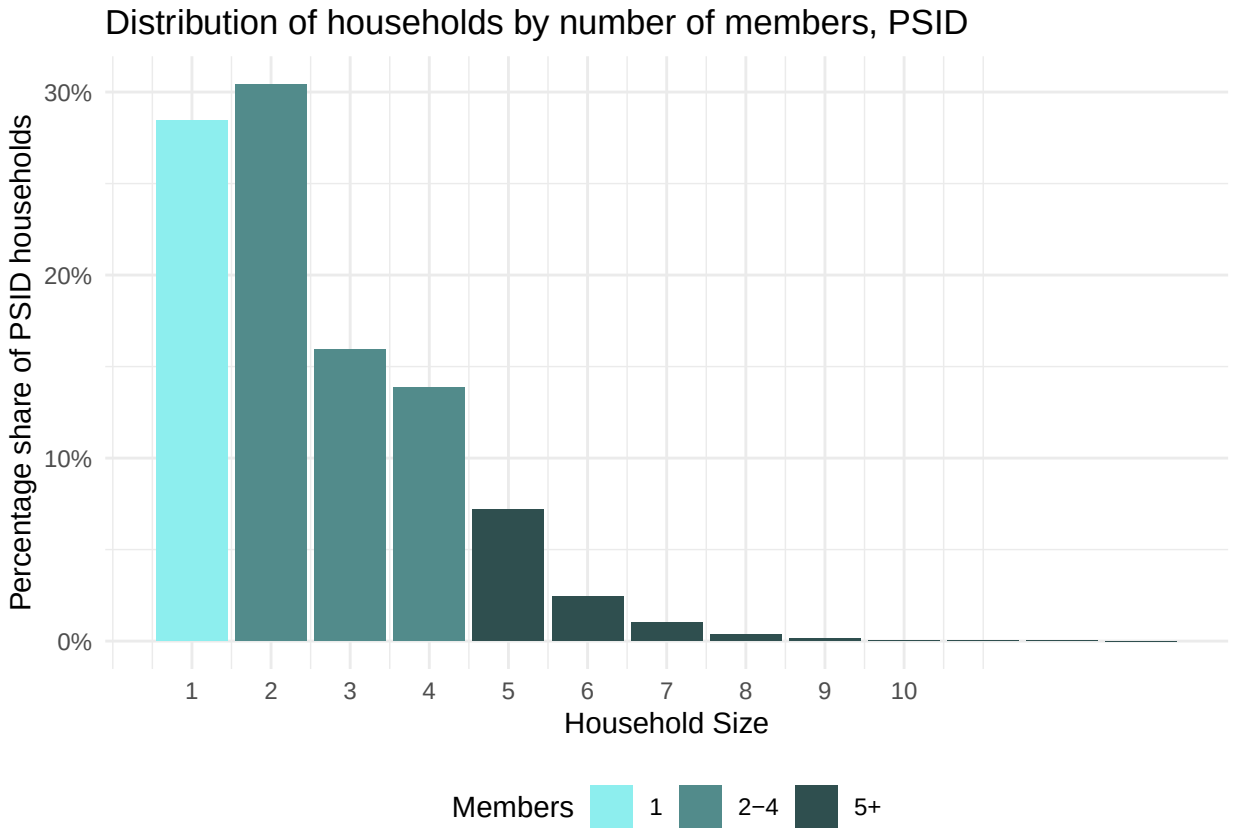
Question 6: household size

The average household in the dataset has 2.6 members. Households with only 1 member make up 28.45% of the data. Households with 5 or more members make up 11.29%.

Table 1: Percentage share of PSID households by size

size_group	count	pct
1	2579	28.45
2–4	5463	60.26
5+	1024	11.29

The histogram below shows the distribution of households by member count in the PSID dataset. Compared to the SCF histogram shown in class, the PSID histogram has fewer 2-member households and more 3-5 member households. The PSID dataset has 30.43% two-member households compared to the nearly 35% shown on Slide 2 of Section 1.1. PSID has a slightly larger share of households with 3 members: 15.95%, compared to the slide's 15%; a 13.88% share of 4-member households, compared to the slide's 12.5%; and a 7.2% share of 5-member households, compared to just past 5% in the slide.



Part 2: Income Distribution

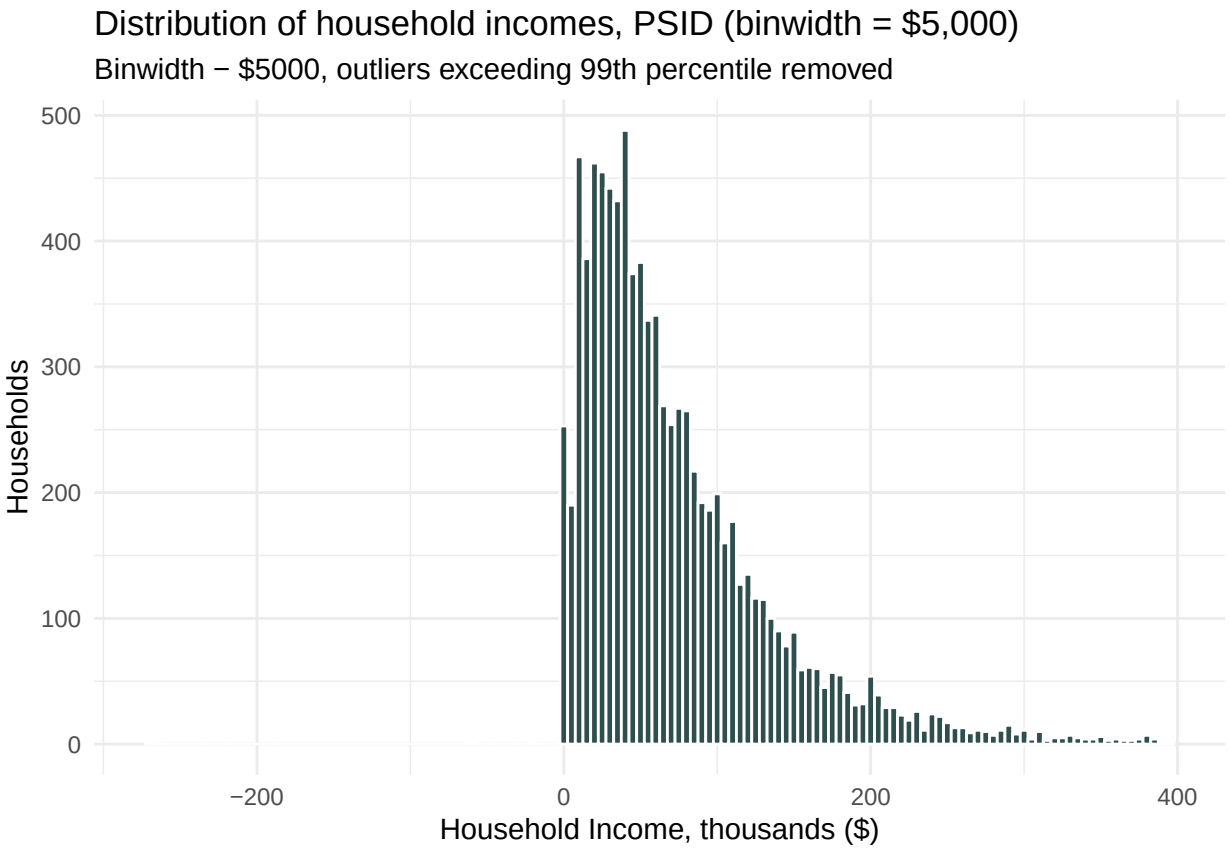
Question 1: distribution of household income

The following histogram shows the distribution of household incomes in the PSID, with outliers exceeding the 99th percentile ($396,420). The distribution is right-skewed, as the median of $55,090 is lower than the mean of $78,266. The majority of households earn less than $100,682, which is the 75th percentile. 25% of

households surveyed have higher household incomes, up to $2.125 million. 137 households have no income, and three households have negative income, with the lowest income being -$267,900.

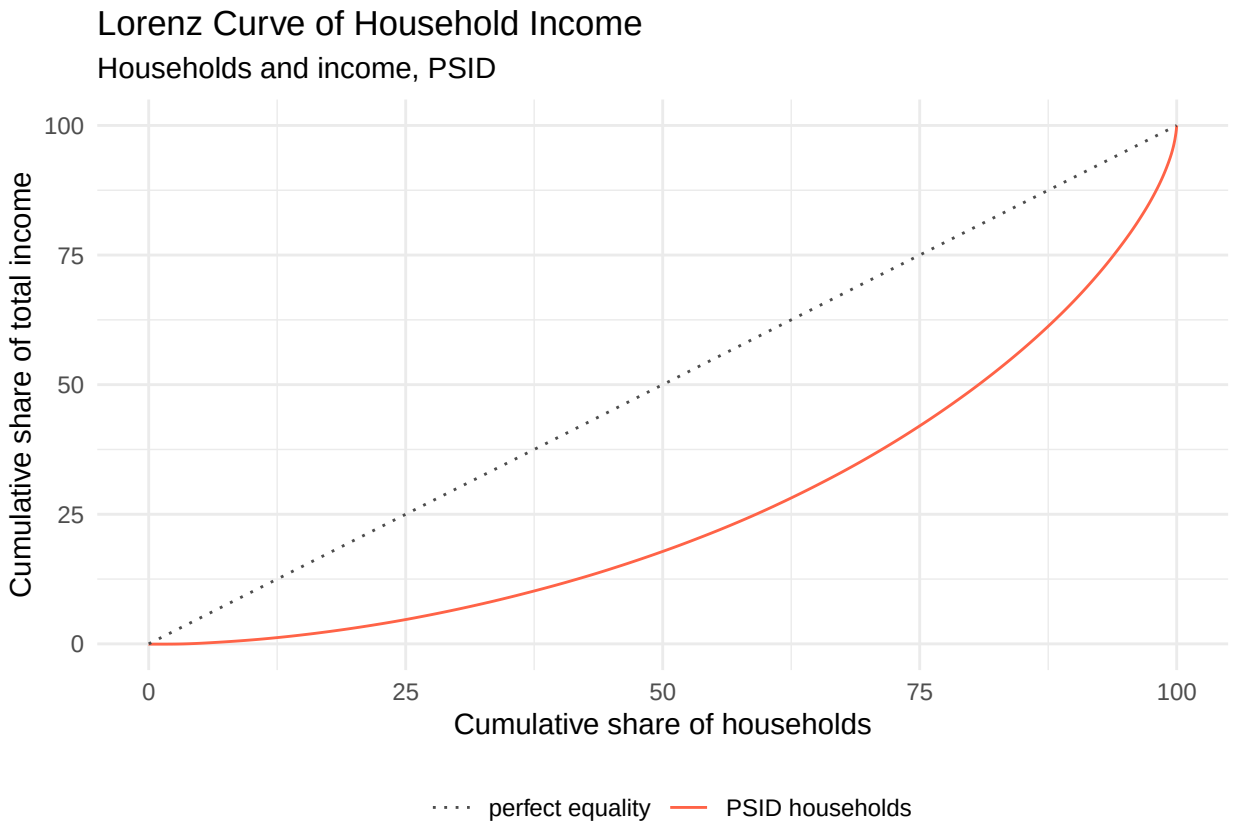
Table 2: Distribution of Household Incomes

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-267900	28000	55090	78266	100681	2125100



Question 2: household income Lorenz curve

The following Lorenz curve visualizes the cumulative share of households (x) against the cumulative share of income (y). The dotted line is what the curve would look like at perfect equality, where the bottom 50% of households possess 50% of total household income. The solid red line represents the relationship between cumulative household share and cumulative income share in reality.



We see on this curve that the lowest earning 50% of households in the PSID account for only 18% of total household income, and that the bottom 75% of households account for 42% of income. By contrast, the top 10% of highest earning households account for about a third of all household income, and the top 1% of households make around 8%. In this dataset, we can tell that high-earning households account for more than their proportional share of household income, indicating income inequality within the sample.

Question 3: total household income coefficient of variation

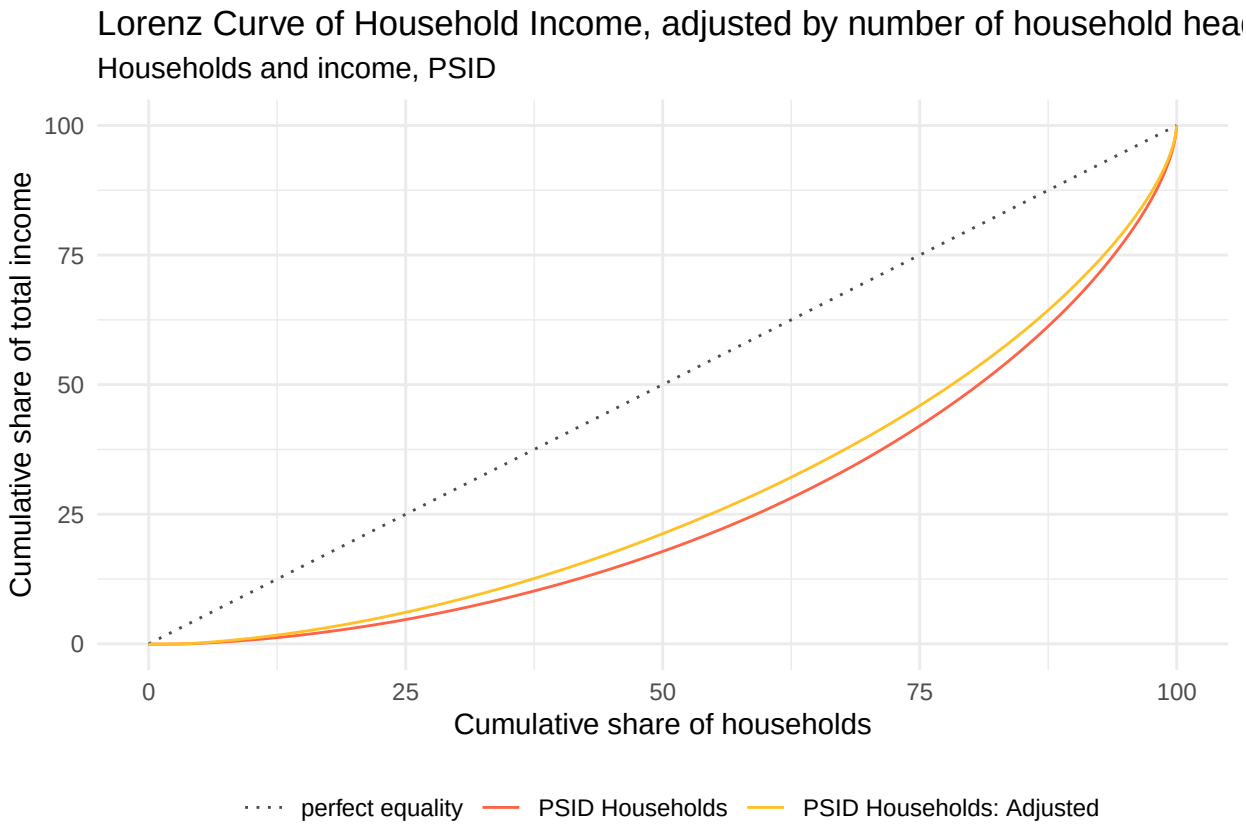
To further quantify the level of income inequality in the dataset, we can calculate the coefficient of variation by dividing the standard deviation in household income by the mean household income. This produces a coefficient of variation of 1.15, which is lower compared to the SCF. Based on slide 14 of Section 1.2 in the lecture slides, the coefficient of variation derived from the SCF was 4.61 in 1989, 4.83 in 2019, and 5.31 in 2022. This tells us that there is less income inequality in the PSID sample than in the SCF samples.

```
## [1] "Coefficient of variation of household income: 1.152"
```

Question 4: adjusted total household income Lorenz curve

When we adjust household income based on the number of household heads present, our Lorenz curve moves closer to the line of perfect equality, meaning that adjusting household income by the number of household heads decreases income inequality. This is intuitive: we are now adjusting for the number of potential earners in a household, allowing for a more fair comparison between households with one and two heads.

On the yellow curve representing the relationship between adjusted income share and household share, we see that the bottom 50% of households account for 21% of income, compared to the 18% on the non-adjusted curve. This indicates that some of the inequality among households may be attributed to differences in household size and the number of earners.



Question 5: household income percentile ratios

The table below displays the 30th, 50th, 90th, and 99th percentiles of household income in the dataset. Based on these percentile, we know that 30% of the sampled households earn $33,107.50 or less per year; 50% earn $55,090 or less; 90% earn $161,188 or less; and 99% earn $396,420 or less. The wide gap between the 50th and the 90th percentile is a clear marker of income inequality.

Table 3: PSID Household Income Percentiles

percentile	value
30th	$33,107.50
50th	$55,090.00
90th	$161,188
99th	$396,420

The percentile ratios in the table further illustrate a highly unequal income distribution, where households at the top earn several times more than those in the middle or bottom. The 90-30 ratio is 4.87, meaning households at the top earn 4.87 times what households at the bottom earn. The 90-50 ratio is 2.93, meaning households at the top earn 2.93 times what households in the middle earn. The 30-10 ratio is 2.67, meaning there is income inequality within the bottom as well, with those at the 30th percentile earning 2.67 times what those at the 10th percentile earn. The 99-50 ratio is 7.2, meaning that households at the very top of the distribution earn 7.2 times what households in the middle earn, indicating extreme concentration of income at the very top.

The numbers in the SCF are even more drastic: in 2022, the 90-50 ratio was 3.54 and the 99-50 ratio was 17.05. This may be due to SCF's sampling again: by over-sampling the ultra-wealthy, the SCF likely captures much higher incomes in the upper end of its distribution than the PSID does.

Table 4: PSID Household Income Percentile Ratios

percentile_ratio	value
90-30	4.87
90-50	2.93
30-10	2.67
99-50	7.2

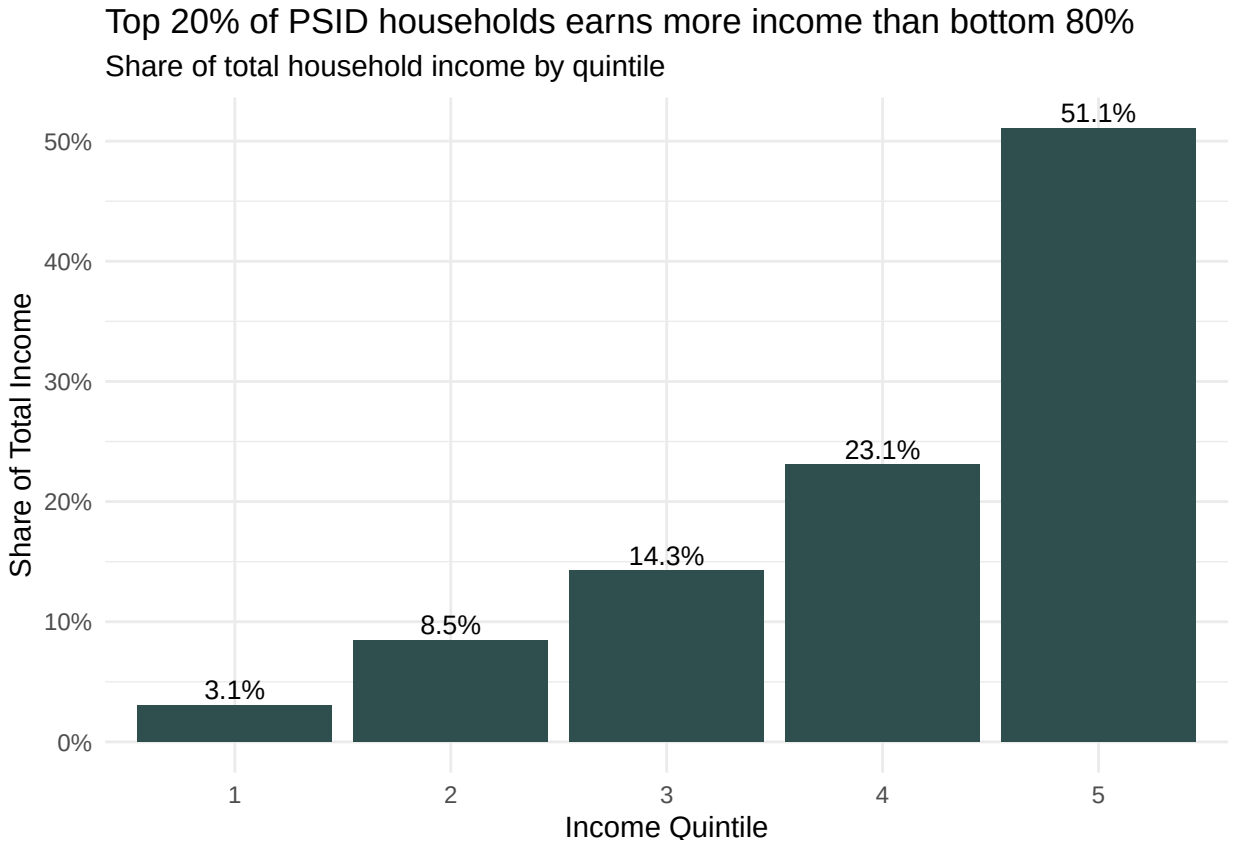
Question 6: mean household income and share by quintile

The following table further illustrates the income inequality among PSID households: the highest disparities between consecutive quintiles are between quintiles 1 and 2 and quintiles 4 and 5. Quintile 2 has a mean income of $33,188, 2.76 times the first quintile's mean income of $12,003. The second quintile's income share of 8.5% is 2.74 times that of the first quintile (3.1%). The fifth quintile's income of $199,894 is 2.21 times that of the fourth quintile ($90,408). The fifth quintile's income share of 51.1% is 2.21 times that of the fourth quintile. It's also notable that the income share of the top quintile is greater than that of the 4 lower quintiles combined. Taken together, this shows the high overall income inequality among PSID households, as well as the heightened inequality at the two ends of the distribution.

Compared to SCF (Section 1.2, slide 9), the average income is lower for all quintiles in the PSID. The first four quintiles have a higher income share in the PSID, while the fifth quintile has a higher income share in the SCF. This pattern once again speaks to the difference in sampling across the two surveys.

Table 5: Share of total household income and mean income by quintile

quintile	income_share	mean_income
1	3.1%	$12,003
2	8.5%	$33,188
3	14.3%	$55,883
4	23.1%	$90,408
5	51.1%	$199,894



Question 7: mean total household income and share for the top 1%

As discussed in class, looking at mean incomes and income shares by quintiles can obscure the inequality within the top of the distribution. Looking at the top 1% of households, we see that their mean income ($639,974) is 8.81 times that of the bottom 99% ($72,570), and 3.2 times that of the top quintile ($199,894). In terms of income share, the top 1% accounts for 8% of all household income, making up nearly one-sixth of the top quintile's share. Even within the top quintile, there is drastic inequality between the highest and lowest earners.

Compared to the SCF (based on Section 1.2 Slide 9), the top 1% has a lower income share and mean income in the PSID (SCF reported 22.4% and $3.18 million in 2022).

Table 6: Share of total household income and mean income held by the top 1%

group	income_share	mean_income
bottom 99%	92%	$72,570
top 1%	8%	$639,974

Part 3: Labor Income

Question 1: household earnings share and mean by quintile

We can create a variable for the total labor income by adding the LABOR_HEAD and LABOR_SPOUSE. Looking at total labor earnings of households, the inequality within the bottom of the distribution is much higher. The table below lists the mean household earnings and share of household earnings by quintile. The

second quintile of households ($14,976) earns 516 times the labor income of the first quintile ($29). The earnings inequality at the top of the distribution is also higher than household income inequality, but by a smaller margin.

Table 7: Share of total household labor income and mean labor income by quintile

quintile	labor_income_share	mean_labor_income
1	0.0%	$24
2	5.0%	$14,908
3	12.9%	$38,925
4	24.6%	$71,716
5	57.5%	$170,274

Question 2: total household income vs household earnings by quintile

Side by side, we can see that the distribution of earnings share is more unequal. The bottom quintile's share of earnings rounds to zero, while the bottom quintile's share of total household income is 3.1%. Quintiles 2 and 3 also hold smaller shares of earnings than they do total income. Quintile 4 has a slightly higher share of earnings than total income, and quintile 5's share of earnings is higher than total income by 6.3 percentage points. Compared to the distribution of total income, earnings is more unevenly distributed, with higher percentage shares in the top 2 quintiles.

Table 8: Total income vs earnings quintile shares

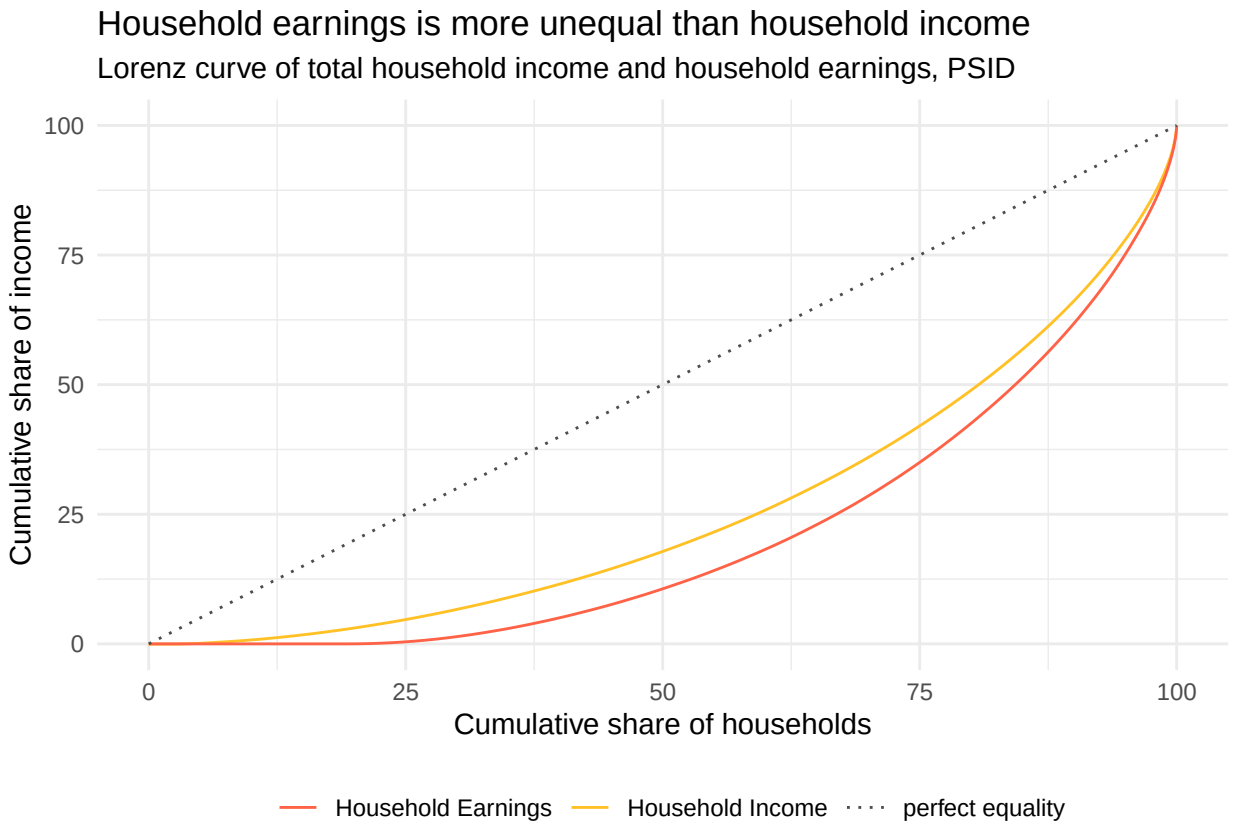
quintile	total_income_share	earnings_share
1	3.1%	0.0%
2	8.5%	5.0%
3	14.3%	12.9%
4	23.1%	24.6%
5	51.1%	57.5%

Calculating the coefficient of variation for both metrics confirms this: earnings have a coefficient of variation of 1.36, compared to 1.15 for household income. This indicates that there is more inequality in earnings than in income.

Table 9: Coefficient of Variation: Household Income and Earnings

metric	cv
Household Income	1.15
Household Earnings	1.36

This trend is seen when plotted on a Lorenz curve, as well. The curve representing cumulative households against cumulative earnings is more far away from the line of perfect equality, indicating that its distribution is more unequal than that of household income. As discussed in class, redistributive policies like social insurance and means-tested transfers may be contributing to the household income of households at the bottom of the income distribution who do not have labor income. Consequently, the bottom quintile can have a higher share of total household income than household earnings.



Question 3: average share of total household income from earnings

The average share of earnings in household income is 68.51%, indicating that households typically receive just over two-thirds of their income from labor income. This figure is excluding the 137 households with incomes of 0, as their share would be undefined.

```
## [1] "Mean share of earnings in total household income: 68.51"
```

Question 4: mean and share of labor earnings by quintile

In the dataset, higher income households derive a larger share of their income from labor earnings. The most stark difference is between quintile 1 and 2: the mean share of labor is 46.6% in the former, and increases by 18.9 percentage points in quintile 2. The difference in the earnings share between the remaining quintiles is modest: between quintile 2 and 5, the total increase is less than that between quintile 1 and 2.

? SCF.

Compared to SCF, _____–

Table 10: Share of total household income from labor earnings by quintile, PSID

quintile	mean_labor_share
1	46.6%
2	65.5%
3	71.9%
4	76.8%

quintile	mean_labor_share
5	80.2%

When we pull out the top 1% of households, however, we see that they get 77.1% of their income from labor earnings, higher than the overall mean. This is unexpected, as we discussed in class that the top earning households earn a higher share of capital income, and rely less on labor. Compared to SCF, this figure _____.

Table 11: Share of total household income from labor earnings, top 1%, PSID

group	mean_labor_share
bottom 99%	68.4%
top 1%	77.1%

Question 5: weekly wage, contribution of hours and wages to labor earnings inequality

In the dataset, there are 6,971 households where the household head has positive labor earnings and weeks of work. The mean weekly wage for these household heads is $1,117.74.

```
## [1] 6971
```

```
## [1] 1117.736
```

A variance decomposition shows that most earnings inequality arises from wage dispersion: the variance of log wages (0.87) is much larger than the variance of log weeks worked (0.17). The positive covariance (0.15) indicates that individuals with higher wages also tend to work more weeks, further amplifying overall earnings inequality. These results suggest that differences in wages contribute much more to earnings inequality than differences in weeks worked.

However, a caveat to this decomposition is that it is restricted to household heads with positive weeks worked and positive labor earnings. By excluding those who did not work or earn, the approach may understate the role of weeks worked in earnings inequality, as those at the bottom of the earnings distribution could be working very few weeks or not at all due to limited employment opportunities.

```
## Error: object 'cov_log' not found
```

Error: ! object ‘cov_log’ not found Backtrace: 1. … %>% pander(“Weekly wages log decomposition results”) 3. tibble::tibble(. . .) 4. tibble:::tibble_quos(xs, .rows, .name_repair) 5. rlang::eval_tidy(xs[[j]], mask) Warning message: In grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x, xy, : for ‘2–4’ in ‘mbcsToSbcs’: - substituted for – (U+2013)

Question 6

A linear regression of the log-transformed weekly wage of the household head on the head’s age, age-squared, education, and occupation provides further information about the drivers of wage inequality. The observables in this regression explain 44% of the variation in log-transformed weekly wages, while the residuals (residual standard error = 0.7243) explain 56%.

```
## [1] "Share of inequality explained by observables: 0.44"
```

```
## [1] "Share of inequality explained by residuals: 0.56"
```