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] "rows: 9066"
```

```
##  [1] "ID"               "AGE_HEAD"         "AGE_SPOUSE"       "SEX_HEAD"
##  [5] "SEX_SPOUSE"       "EDU_HEAD"         "EDU_SPOUSE"       "OCC_HEAD"
##  [9] "OCC_SPOUSE"       "STATE"            "MARITAL"          "HOUSEHOLD_SIZE"
## [13] "CHILDREN"         "SPOUSE_PRESENT"   "WEIGHT"           "INCOME"
## [17] "HEAD_LABOR"       "SPOUSE_LABOR"     "WEEKS_HEAD"       "WEEKS_SPOUSE"
## [21] "WEEKS_OUT_HEAD"   "WEEKS_OUT_SPOUSE" "EXP"              "GAS"
## [25] "FOOD_HOME"        "FOOD_DELIV"       "FOOD_OUT"         "RENT"
## [29] "EDUC_EXP"         "CHILDCARE"        "TRIPS"            "WEALTH"
## [33] "HOUSE_VALUE"      "MORTGAGE"         "MORT_PAY"         "HOME_EQUITY"
## [37] "BUSINESS_VAL"     "PROPERTY_TAX"
```

Question 2

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: Income

The average household income is $78,265.69. This is substantially lower the $142k reported by the SCF in 2022 (Section 1.2, slide 9). One possible reason for this difference is that the SCF surveys a small sample, focusing on high-income individuals, resulting in a less representative mean. The PSID has a larger sample, with less emphasis on the top of the income distribution.

```
## [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. The households with negative income are likely in debt.

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

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

Question 5: spouse age

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

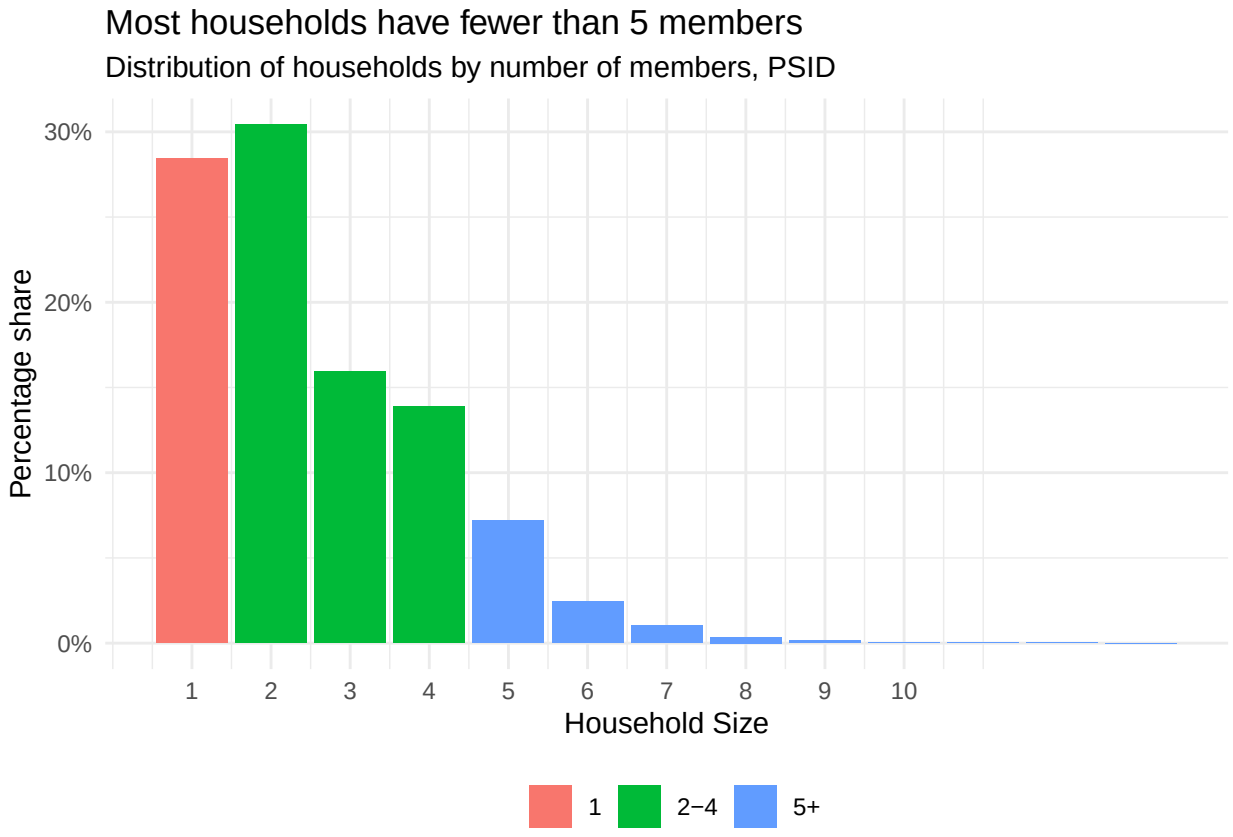
```
## [1] "Mean spouse age: 45.64"
```

Question 6: household members

Households with only 1 member make up 28.45% of the data. Households with 5 or more members make up 11.29%. Compared to the histogram shown in class, the PSID has fewer 2-member households and more 3-5 member households. In both datasets, about 28% of households have one member. The PSID dataset has a smaller share of two-member households: only 30.43% compared to the nearly 35% shown on slide 2 [what dataset is this?]. 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.

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



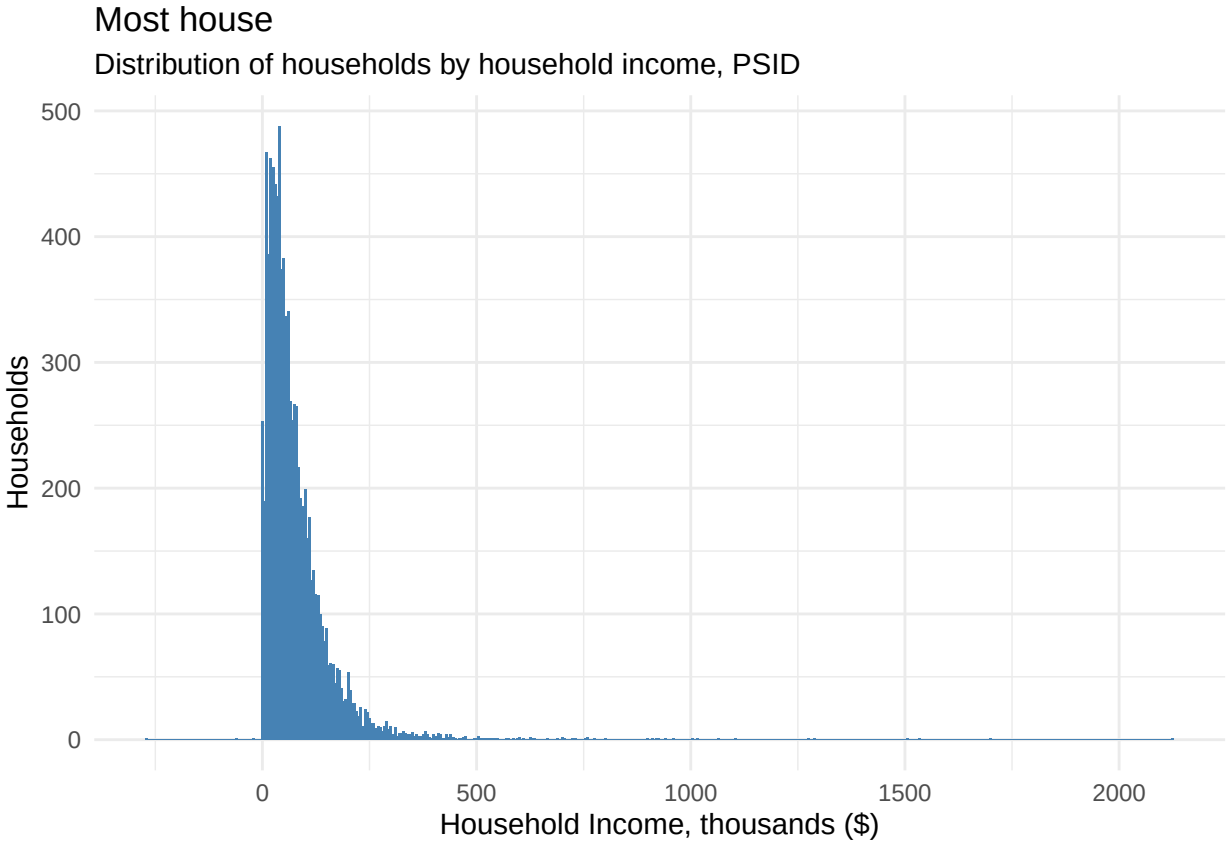
Part 2: Income Distribution

Question 1

The following histogram shows the distribution of household incomes in the PSID, with values above the 99th percentile removed. 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. A small minority of households surveyed have higher household incomes, up to $2.13 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

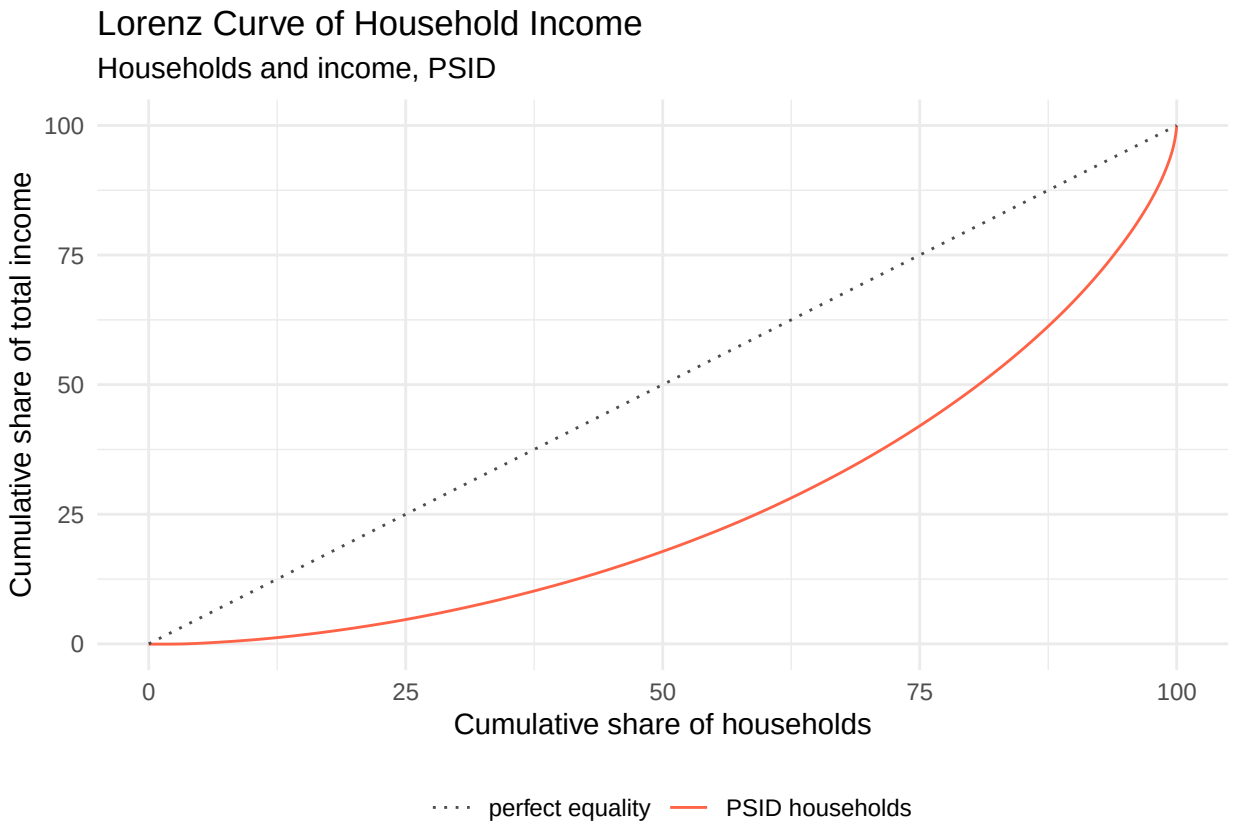
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-267900	28000	55090	78266	100681	2125100



Question 2

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

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 account for just above 8%. In this dataset, we can tell that higher-earning households account for more than their proportional share of household income.



? ## Question 3 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 _____ compared to the _____ derived from the SCF.

```
## [1] "Coefficient of variation of household income: 1.15"
```

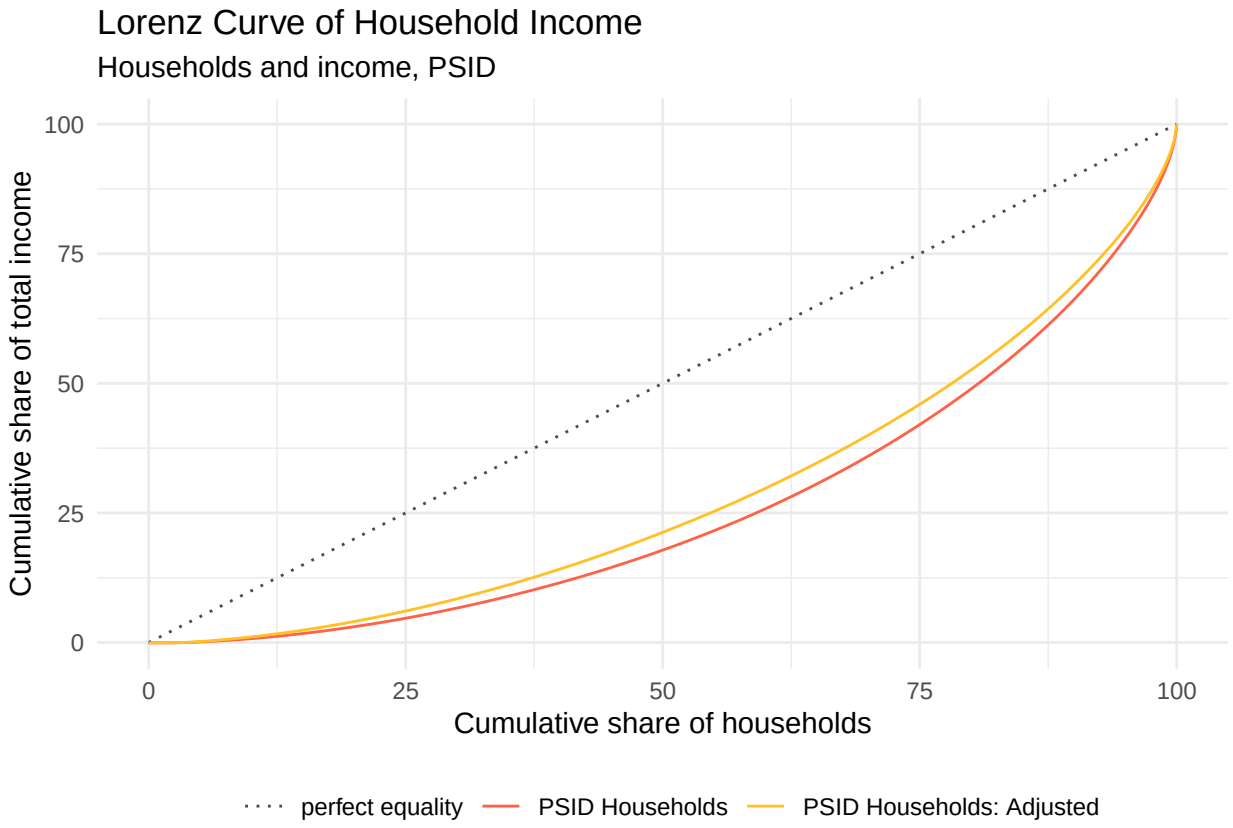
#? ## Question 4 When we adjust this curve for the number of household heads present, the distance from the line of perfect equality decreases. On the yellow curve representing the relationship between adjusted income share and population share, we see that the bottom 50% of households account for 21% 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.

```
psid_lorenz_adj <- psid %>%
  mutate(spouse_adjust = ifelse(SPOUSE_PRESENT == 1, 2, 1),
    income_per_head = INCOME/spouse_adjust) %>%
    arrange(income_per_head) %>%
  mutate(
    INC_SHARE = income_per_head / sum(income_per_head),
    HHS_SHARE = 1 / nrow(.),
    CUM_HH_SHARE = 100*cumsum(HHS_SHARE),
    CUM_INC_SHARE = 100*cumsum(INC_SHARE))

ggplot() +
  geom_line(data = psid_lorenz, aes(x = CUM_HH_SHARE, y = CUM_INC_SHARE, color = "PSID Households")) +
  geom_line(data = perfect_equality, aes(x = x, y = y, color = "perfect equality"), linetype = "dotted"
  geom_line(data = psid_lorenz_adj, aes(x = CUM_HH_SHARE, y = CUM_INC_SHARE, color = "PSID Households:
  theme_minimal() +
```

```
  scale_color_manual(values = c("gray30", "tomato", "goldenrod1"), name = "") +
  labs(
    title = "Lorenz Curve of Household Income",
    subtitle = "Households and income, PSID",
    x = "Cumulative share of households",
    y = "Cumulative share of total income"
  ) +
  theme(legend.position = "bottom")
```



Question 5

The table below displays the 30th, 50th, 90th, and 99th percentiles of household income in the dataset. The wide gap between the 50th and 90th and 50th and 99th percentiles give us a sense of the income inequality among the households, but percentiles themselves are insufficient.

```
data.frame(percentile = c("30th", "50th", "90th", "99th"),
           value = c(paste("$",round(quantile(psid$INCOME, 0.3), 2), sep=""),
                     paste("$",round(quantile(psid$INCOME, 0.5), 2), sep=""),
                     paste("$",round(quantile(psid$INCOME, 0.9), 2), sep=""),
                     paste("$",round(quantile(psid$INCOME, 0.99), 2), sep=""))) %>%
  as_tibble() %>%
  pander(caption = "PSID Household Income Percentiles")
```

Table 3: PSID Household Income Percentiles The percentile ratios in the table reveal a markedly unequal income distribution. Households at the top earn several times more than those in the middle or bottom. Specifically, households at the 90th percentile earn nearly three times the median income, while those at the 99th percentile earn 7.2 times the median, indicating extreme concentration of income at the very top. The 2.67 ratio between the 30th and 10th percentiles shows that inequality at the lower end exists but is considerably smaller than that among top earners.

percentile	value
30th	$33107.5
50th	$55090
90th	$161188
99th	$396420

```
data.frame(percentile_ratio = c("90-30", "90-50", "30-10", "99-50"),
           value = c(round(quantile(psid$INCOME, 0.9)/quantile(psid$INCOME, 0.3), 2),
                     round(quantile(psid$INCOME, 0.9)/quantile(psid$INCOME, 0.5), 2),
                     round(quantile(psid$INCOME, 0.3)/quantile(psid$INCOME, 0.1), 2),
                     round(quantile(psid$INCOME, 0.99)/quantile(psid$INCOME, 0.5), 2))) %>%
  as_tibble() %>%
  pander(caption = "PSID Household Income Percentile Ratios")
```

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

- Interpretation of ratios
- Comparison with SCF

Question 6

- Share of income by quintile
- Average income by quintile
- Interpretation of results
- Comparison with SCF (lecture)

Question 7

- Average income and income share for top 1%
- Comparison with SCF

Part 3: Labor Income

Question 1

- Create variable of total labor income of household
- Earnings and share of earnings for each quintile

Question 2

- Compare distribution of household earnings vs income
- Compare share of quintiles
- Coefficient of variation
- Lorenz curve of labor earnings and total household income
- Compare relative inequality in labor earnings vs total household income

Question 3

- Create variable of share of labor earnings in total household income
- Average share of labor earnings

Question 4

- Average share of labor earnings by quintile of income distribution
- Share of labor earnings for top 1%
- Compare both to SCF
- Comment on differences

Question 5

- Filter by households where household head has positive work weeks and positive earnings
- Weekly wage of household head
- Variance of log of earnings, wages, weeks worked
- Relative contribution of hours and wages to inequality in labor earnings?
- Caveats to this calculation?

Question 6

- Regression of log wages of household head on age, ageˆ2, education, occupation
- Residuals
- Variance of residuals
- Share of inequality explained by age, education and occupation?