Problem Set 1: OLS Review

EC 421: Introduction to Econometrics

Solutions

DUE Upload your answer on Canvas *before* midnight on Sunday, 19 April 2020.

IMPORTANT You must submit **two files**:

1. your typed responses/answers to the question (in a Word file or something similar)
2. the R script you used to generate your answers. Each student must turn in her/his own answers.

README! The data in this problem set come from the 2018 American Community Survey (ACS), which I downloaded from IPUMS. The last page has a table that describes each variable in the dataset(s).

OBJECTIVE This problem set has three purposes: (1) reinforce the metrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality within econometrics/regression.

INTEGRITY If you are suspected of cheating, then you will receive a zero. We may report you to the dean.

Setup

Q01. Load your packages. You'll probably going to need/want `tidyverse` and `here` (among others).

Answer:

```
# Load packages using 'pacman'
library(pacman)
p_load(tidyverse, here)
```

Q02. Now load the data. I saved the same dataset as two different formats:

- an `.rds` file: use a function that reads `.rds` files—for example, `readRDS()` or `read_rds()` (from the `readr` package in the `tidyverse`.
- a `.csv` file: use a function that reads `.csv` files—for example, `read.csv()` or `read_csv()` (from the `readr` package in the `tidyverse`.

Answer:

```
# Load data
ps_df = here("001-data.csv") %>% read_csv()
```

Q03. Check your dataset. How many observations and variables do you have? *Hint:* Try `dim()`, `ncol()`, `nrow()`.

Answer:

```
# Check dimensions
dim(ps_df)
```

```
#> [1] 25000    12
```

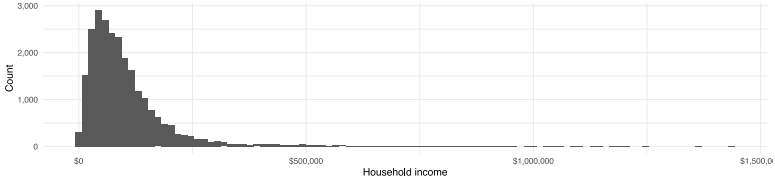
We have 25,000 observations (rows) on 12 variables (columns).

Getting to know your data

Q04. Plot a histogram of households' income (variable: `hh_income`). *Note:* Household income is in tens of thousands of dollars (so a value of `8` implies an income of $80,000.)

Answer:

```
# Create the histogram of HH income using ggplot2
ggplot(data = ps_df, aes(x = hh_income * 10000)) +
geom_histogram(bins = 100) +
scale_x_continuous("Household income", labels = scales::dollar) +
scale_y_continuous("Count", labels = scales::comma) +
theme_minimal(base_size = 10)
```



Q05. What are the mean and median levels of household income? Based upon this answer and the previous histogram, is household income (fairly) evenly distributed or is it more skewed? Explain your answer.

Answer:

```
# Check summary of hh income
summary(ps_df$hh_income)
```

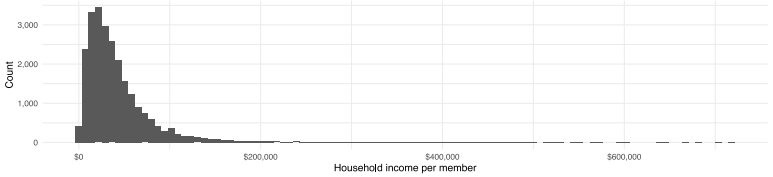
```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   0.004   4.600   8.000  10.616  13.000 143.600
```

Households' mean income in the data is approximately $106,160 and median household income is approximately $80,000. Based upon this information—and especially based upon the histogram in the previous problem—we can see that household income in this sample is extremely skewed. The right tail of the distribution extends quite far.

Q06. Create a histogram of household income per capita—meaning the household's income divided by the number of individuals in the household. Does dividing by the number of individuals in the household change your understanding of the income distribution? Explain your answer.

Answer:

```
# Create the histogram of HH income using ggplot2
ggplot(data = ps_df, aes(x = hh_income / hh_size * 10000)) +
geom_histogram(bins = 100) +
scale_x_continuous("Household income per member", labels = scales::dollar) +
scale_y_continuous("Count", labels = scales::comma) +
theme_minimal(base_size = 10)
```



Dividing by the number of members in a household does not really change our understanding of income—it is still extremely skewed.

Q07. Run a regression that helps summarize the relationship between household income and household size. Interpret the results of the regression—the meaning of the coefficient(s). Comment on the coefficient's statistical significance.

Answer: You have a lot of options here. I'm going to regress the log of income on the level of household size.

```
# Regression
est07 = lm(log(hh_income) ~ hh_size, data = ps_df)
# Results
est07 %>% broom::tidy()
```

```
#> # A tibble: 2 x 5
#>   term        estimate std.error statistic   p.value
#>   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)    1.67    0.0114      146.  0.
#> 2 hh_size        0.125   0.00356      35.2 8.97e-265
```

The estimated coefficient in this log-linear model suggests that a one-person increase in a household's size is associated with a 12.5% increase in household income.

Q08. Explain why you chose the specification you chose in the previous question.

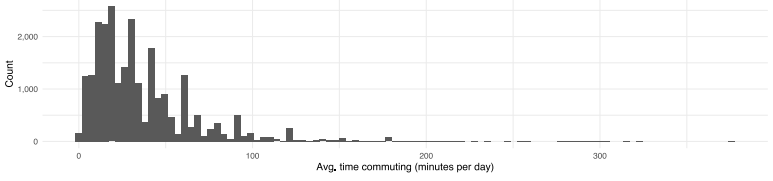
- Was it linear, log-linear, log-log?
- What was the outcome variable?
- What was the explanatory variable?
- Why did you make these choices?

Answer: I chose I log-linear specification to allow household size to be associated with *percent* changes in income (rather than level changes)—and because logging a variable can compress its distribution (and we know income is very skewed).

Q09. Plot a histogram of the time households spend commuting each day (the variable `time_commuting` is the average commuting time for a household). Is the distribution of commute time more or less equitable than income? Explain.

Answer:

```
# Create the histogram of HH income using ggplot2
ggplot(data = ps_df, aes(x = time_commuting)) +
geom_histogram(bins = 100) +
scale_x_continuous("Avg. time commuting (minutes per day)", labels = scales::comma) +
scale_y_continuous("Count", labels = scales::comma) +
theme_minimal(base_size = 10)
```



This distribution still appears to be pretty inequitable: it has a lot of variance, and there is still a lot of skew in the distribution. Perhaps the skew is a bit lower than with income.

Regression refresher: Varying the specification

Q10. Linear specification Regress average commute time (`time_commuting`) on household income (`hh_income`). Interpret the coefficient and comment on its statistical significance.

Answer:

```
# Regress commute time on income
est10 = lm(time_commuting ~ hh_income, data = ps_df)
# Results
est10 %>% broom::tidy()
```

```
#> # A tibble: 2 x 5
#>   term         estimate std.error statistic  p.value
#>   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)   34.9      0.301     116.   0.
#> 2 hh_income      0.175    0.0203      8.64 6.20e-18
```

Our estimated coefficient suggests that a one-unit increase in household income (an increase of $10,000) is associated with an increase in commute time of approximately 0.2 minutes. This coefficient is statistically significant at the 5% level (though not very economically meaningful—the magnitude of the coefficient is quite small).

Q11. Log-linear specification Regress the log of average commute time on household income. Interpret the coefficient and comment on its statistical significance.

Answer:

```
# Log-linear regression
est11 = lm(log(time_commuting) ~ hh_income, data = ps_df)
# Results
est11 %>% broom::tidy()
```

```
#> # A tibble: 2 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)  3.18     0.00806      395.  0.
#> 2 hh_income    0.00689  0.000545      12.6 1.59e-36
```

With this log-linear specification, our coefficient estimate suggests that a one-unit increase in household income (an increase of $10,000 dollars) is associated with an increase in commute time of approximately 0.7%. This coefficient is still statistically significant at the 5% level (and still small in absolute magnitude).

Q12. Log-log specification Regress the log of average commute time on the log of household income. Interpret the coefficient and comment on its statistical significance.

Answer:

```
# Log-linear regression
est12 = lm(log(time_commuting) ~ log(hh_income), data = ps_df)
# Results
est12 %>% broom::tidy()
```

```
#> # A tibble: 2 x 5
#>   term           estimate std.error statistic   p.value
#>   <chr>             <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)       2.98    0.0142      210.  0.
#> 2 log(hh_income)    0.138   0.00645      21.3 5.97e-100
```

With this log-log specification, our coefficient estimate suggests that a one-percent increase in household income is associated with an increase in commute time of approximately 0.138 percent. This coefficient is still statistically significant at the 5% level (and still small in absolute magnitude).

Multiple linear regression and indicator variables

Q13. Regress average commute time on household income **and** the share of the individuals in the household who are non-white ethnicities (`hh_share_nonwhite`). Interpret the intercept and coefficient and comment on their statistical significance. Also compare your results to Q10. Has anything changed?

Answer:

```
# Log-linear regression
est13 = lm(time_commuting ~ hh_income + hh_share_nonwhite, data = ps_df)
# Results
est13 %>% broom::tidy()
```

```
#> # A tibble: 3 x 5
#>   term              estimate std.error statistic  p.value
#>   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)         34.5      0.330     105.   0.
#> 2 hh_income            0.179    0.0204      8.80 1.48e-18
#> 3 hh_share_nonwhite    1.31     0.522       2.51 1.21e- 2
```

The intercept (34.5 minutes) tells us the expected (or estimated) commute time for a household with zero income and that is 0% nonwhite.

Our coefficient on household income (`hh_income`) tells us that a one-unit increase in household income (an increase of $10,000 dollars) is associated with an increase in commute time of approximately 0.18 minutes **holding everything else constant**. This coefficient is still statistically significant at the 5% level (and still small in absolute magnitude). This coefficient is still statistically significant. It hasn't changed much relative to **Q10**.

Our coefficient on the share of the household that represents a non-white ethnicity (`hh_share_nonwhite`) tells us the expected difference in commute time between entirely non-white households (`hh_share_nonwhite = 1`) and entirely white households (`hh_share_nonwhite = 0`) **holding all other variables constant**. Specifically, we find that non-white households, on average, have a 1.31-minute longer commute time (holding all other variables constant). This coefficient is statistically significant at the 5% level.

Q14. Regress average commute time on the indicator variable for whether a household moved in the last year (`i_moved`). Interpret the intercept and coefficient and comment on their statistical significance.

Answer:

```
# Log-linear regression
est14 = lm(time_commuting ~ i_moved, data = ps_df)
# Results
est14 %>% broom::tidy()
```

```
#> # A tibble: 2 x 5
#>   term        estimate std.error statistic p.value
#>   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
#> 1 (Intercept)  36.7        0.233  158.       0
#> 2 i_moved      -0.0329     0.536   -0.0613   0.951
```

The intercept (36.7) tells us the average commute for households that did not move in the last year (`i_moved = 0`).

The coefficient on `i_moved` (-0.03) tells us the difference in commute time between households that moved and households that did not move. This coefficient is not statistically significant—meaning we do not find a significant difference in commute time between households that moved and households that did not move.

Q15. Add the share of the household that represents a non-white ethnicity (`hh_share_nonwhite`) to the regression in Q14. *Note:* Your outcome variable is still average household commute time, but you should now have two explanatory variables. Interpret the intercept and coefficient and comment on their statistical significance.

Answer:

```
# Log-linear regression
est15 = lm(time_commuting ~ i_moved + hh_share_nonwhite, data = ps_df)
# Results
est15 %>% broom::tidy()
```

```
#> # A tibble: 3 x 5
#>   term              estimate std.error statistic p.value
#>   <chr>                <dbl>     <dbl>     <dbl>   <dbl>
#> 1 (Intercept)        36.5        0.261   140.     0
#> 2 i_moved            -0.0706     0.536    -0.132  0.895
#> 3 hh_share_nonwhite   0.973      0.522     1.86   0.0622
```

The intercept (36.5) tells us the average commute for households that did not move in the last year (`i_moved = 0`) and who are 0% nonwhite.

The coefficient on `i_moved` (-0.07) now tells us the difference in commute time between households that moved and households that did not move **holding the share non-white constant**. This coefficient is still not statistically significant—meaning we do not find a significant difference in commute time between households that moved and households that did not move.

Our coefficient on the share of the household that represents a non-white ethnicity (`hh_share_nonwhite`) tells us the expected difference in commute time between entirely non-white households (`hh_share_nonwhite = 1`) and entirely white households (`hh_share_nonwhite = 0`) **holding all other variables constant**. Specifically, we find that non-white households, on average, have a 0.97-minute longer commute time **holding moving history constant**. This coefficient is marginally statistically significant—significant at the 10% level but not at the 5% level.

Q16. Did adding this second explanatory variable change the coefficient of the first variable at all? What does that tell you? Explain your answer.

Answer: The coefficient on `i_moved` changed a little, but not much. This coefficient still is not statistically signficiant.

Because the coefficient did not change much when we added a new variable, we know that our new variable (the share of the household that is not white) is pretty uncorrelated with the original variable (the indicator for whether the household moved in the last year).

Q17. Now add the interaction between your two explanatory variables in Q16 and re-run the regreation. (You should have an intercept and three coefficients—the two variables plus their interaction.) Interpret the coefficient on the interaction and comment on its statistical significance.

Answer:

```
# Log-linear regression
est17 = lm(time_commuting ~ i_moved + hh_share_nonwhite + i_moved:hh_share_nonwhite, data = ps_
# Results
est17 %>% broom::tidy()
```

```
#> # A tibble: 4 x 5
#>   term                       estimate std.error statistic p.value
#>   <chr>                         <dbl>     <dbl>     <dbl>   <dbl>
#> 1 (Intercept)                  36.4       0.267   136.      0
#> 2 i_moved                       0.579     0.632     0.916   0.360
#> 3 hh_share_nonwhite             1.48      0.583     2.53    0.0113
#> 4 i_moved:hh_share_nonwhite    -2.53      1.31     -1.94    0.0524
```

There are a couple of ways to think about the coefficient on the interaction. First, we can interpret this coefficient as testing whether there's a different effect of moving for white households and non-white households. Another way to interpret this coefficient is whether there's a different effect of ethnicity on households I have recently moved or have not recently moved. However you were thinking about it, this interaction asks whether the effects of moving and ethnicity depend on each other.

The coefficient on this interaction is statistically significant at the 10% level, and it's nearly significant at the 5% level. It seems like the effect of moving differs for non-white and white households.

Q18. Did including the interaction change your understanding of the relationship between the variables? Explain.

Answer: Yes. The coefficient on ethnicity became statistically significant and larger. We also found significant evidence of an interaction between moving and ethnicity. Both of these changes affect the way that we interpret the relationship between our outcome variable (commute time) and the explanatory variables.

Q19. Regress the indicator for whether the household has a smartphone (`i_smartphone`) on the household's income (`hh_income`) and the share of the household's individuals who represent non-white ethnicities (`hh_share_nonwhite`). Interpret the intercept and coefficients. Comment on their statistical significance.

Answer:

```
# Regression
est19 = lm(i_smartphone ~ hh_income + hh_share_nonwhite, data = ps_df)
# Results
est19 %>% broom::tidy()
```

```
#> # A tibble: 3 x 5
#>   term              estimate std.error statistic  p.value
#>   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)        0.910    0.00242     376.   0.
#> 2 hh_income          0.00261  0.000149     17.5  3.07e-68
#> 3 hh_share_nonwhite -0.00564  0.00382      -1.47 1.40e- 1
```

We interpret the intercept 0.91 as the percentage of households that on a cell phone when the household income is zero and the share non-white is also zero.

The coefficient on household income tells us that for every unit of increase ($10,000 in household income) we expect the share of households who own a smart phone to increase by 0.26% percent **holding everything else constant**.

The coefficient on the share of the household who represent non-white ethnicities tells us that the probability a household owns a cellphone is 0.56% percent lower than white households **holding everything else (*e.g.*, income) constant**.

The bigger picture

Q20. In the last regression (Q19), should we be concerned about omitted-variable bias? Explain your answer and provide an example of a potential omitted variable if you are concerned about omitted-variable bias.

Answer: Yes. There's a great potential for omitted variable problems in this setting. One example: geography. In different areas, there is different access to sell funds. Across different geographic regions, there are also differences in income and ethnicity. We are not controlling for geography or other things that are caused by/related to geography.

Q21. Is R-squared a good measure of model performance? Explain your answer.

Answer: Maybe, but probably not. It tells us the share of the variation in our outcome variable that we are able to explain. But it also mechanically increases as we more add more explanatory variables. So it can help us in knowing how much of our outcome variable we are explaining, but it does not help us to choose explanatory variables.

Q22. Define the term *standard error*.

Answer: Standard error is the standard deviation of an estimator's distribution.

Q23. What does our assumption of *exogeneity* require?

Answer: Our assumption of exogeneity requires that the explanatory variables are uncorrelated with our disturbances. With math: $E[u|x] = 0$.

Q24. What does it mean for an estimator to be *unbiased*?

Answer: An estimator is unbiased if the meaning of its distribution is equal to the perimeter that it is estimating. In math (for estimator $\hat{\theta}$ and parameter θ): $E[\hat{\theta}] = \theta$.

Q25. What does it mean for an estimator to be *more efficient than another estimator*?

Answer: Estimator 1 is more efficient than estimator 2 if the distribution of estimator 1 has smaller variance than the distribution of estimator 2.

Description of variables and names

Variable	Description
`fips`	County FIPS code
`hh_size`	Household size (number of people)
`hh_income`	Household total income in $10,000
`cost_housing`	Household's total reported cost of housing
`n_vehicles`	Household's number of vehicles
`hh_share_nonwhite`	Share of household members identifying as non-white ethnicities
`i_renter`	Binary indicator for whether any household members are renters
`i_moved`	Binary indicator for whether a household member moved in prior 1 year
`i_foodstamp`	Binary indicator for whether any household member participates in foodstamps
`i_smartphone`	Binary indicator for whether a household member owns a smartphone
`i_internet`	Binary indicator for whether the household has access to the internet
`time_commuting`	Average time spent commuting per day by each household member (minutes)

In general, I've tried to stick with a naming convention. Variables that begin with `i_` denote binary indicatory variables (taking on the value of `0` or `1`). Variables that begin with `n_` are numeric variables.