# Problem Set 1: OLS Review

## EC 421: Introduction to Econometrics

*Solutions*

**DUE** Upload your answer on Canvas *before* midnight on Saturday, 23 January 2021.

**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.

If you are using RMarkdown, you can turn a single file, but it must be a `html` or `pdf` file with **both** your R code **and** your 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. **Cheating includes copying from your classmates, from the internet, and from previous assignments.**

# Setup

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

**Answer:**

```
# Load packages using 'pacman'
library(pacman)
p_load(tidyverse, patchwork, 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: As .rds
ps_df = here("001-data.rds") %>% read_rds()
# Load data: As 'csv'
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] 5000   19
```

We have 5,000 observations (rows) on 19 variables (columns).

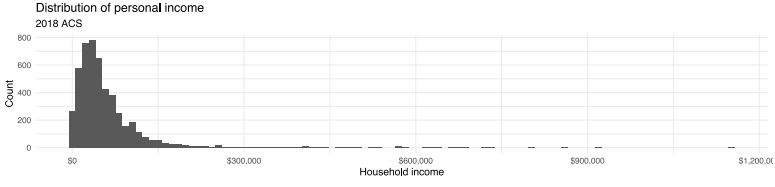
# Getting to know your data

**Q04.** Plot a histogram of individuals' personal income (variable: `personal_income`). *Note:* Household income is in tens of thousands of dollars (so a value of `3` implies an income of $30,000.)

Don't forget to label your plot's axes. A title wouldn't be be, either.

**Answer:**

```
# Create the histogram of HH income using ggplot2
ggplot(data = ps_df, aes(x = personal_income * 10000)) +
geom_histogram(bins = 100) +
scale_x_continuous("Household income", labels = scales::dollar) +
scale_y_continuous("Count", labels = scales::comma) +
ggtitle("Distribution of personal income", "2018 ACS") +
theme_minimal(base_size = 10)
```




**Q05.** Compare the distributions of personal income for (1) women *vs.* men and (2) black individuals *vs.* white individuals. Are the differences at the extremes of the distribution or at the center (*e.g.*, mean and median)?

*Note:* Your answer should include four histograms (women, men, black, and white).

*Hints*

- There is an indicator for female in the data called `i_female`. There are also indicators for *black* and *white* names `i_black` and `i_white`.
- You can take a subset of a variable using the `filter()` variable from the `tidyverse`. *E.g.*, to take find all married individuals in the `ex_df` dataset, you could use `filter(ex_df, i_married == 1)`.

**Answer:**

```
# Summary of women
ps_df %>% filter(i_female == 1) %>% select(personal_income) %>% summary()
```

```
#>  personal_income
#>  Min.   : 0.016
#>  1st Qu.: 2.000
#>  Median : 3.700
#>  Mean   : 4.819
#>  3rd Qu.: 6.000
#>  Max.   :72.800
```

```
# Summary of men
ps_df %>% filter(i_female == 0) %>% select(personal_income) %>% summary()
```

```
#>  personal_income
#>  Min.   :  0.0004
#>  1st Qu.:  2.7775
#>  Median :  5.0000
#>  Mean   :  7.3765
#>  3rd Qu.:  8.5000
#>  Max.   :115.1000
```

```
# Summary of black
ps_df %>% filter(i_black == 1) %>% select(personal_income) %>% summary()
```

```
#>  personal_income
#>  Min.   : 0.016
#>  1st Qu.: 1.907
#>  Median : 3.500
#>  Mean   : 4.849
#>  3rd Qu.: 6.000
#>  Max.   :72.800
```

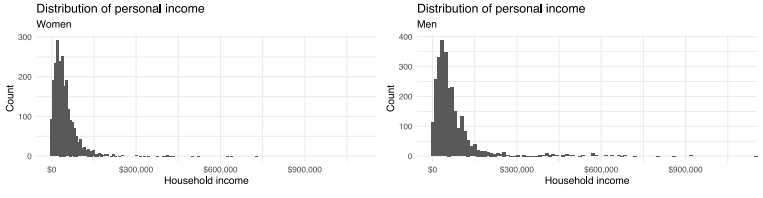
```
# Summary of white
ps_df %>% filter(i_white == 1) %>% select(personal_income) %>% summary()
```

```
#>  personal_income
#>  Min.   : 0.0004
#>  1st Qu.: 2.5000
#>  Median : 4.5000
#>  Mean   : 6.4268
#>  3rd Qu.: 7.5000
#>  Max.   :91.9000
```

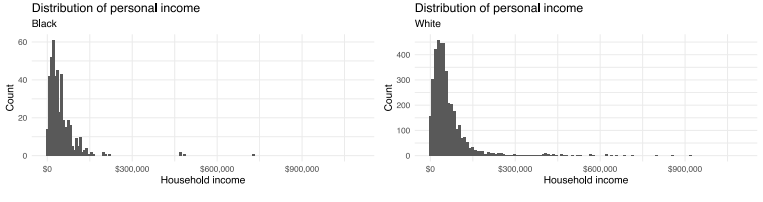
The personal income distributions (in this sample) for women and men differ throughout distribution. For example, at their means, we see a difference between women and men of $48,000 and $74,000, respectively. The distribution of income for men appears to be shifted right (to higher values).

The distribution of income for black individuals in the sample is lower than the distribution of white individuals—across the distribution. The means of the two groups differ by approximately $16,000—*i.e.*, $48,000 *vs.* $64,000.

```
# Histogram: Women
hist_female = ggplot(data = filter(ps_df, i_female == 1), aes(x = personal_income * 10000)) +
geom_histogram(bins = 100) +
scale_x_continuous("Household income", labels = scales::dollar) +
scale_y_continuous("Count", labels = scales::comma) +
ggtitle("Distribution of personal income", "Women") +
theme_minimal(base_size = 10)
# Histogram: Men
hist_male = ggplot(data = filter(ps_df, i_male == 1), aes(x = personal_income * 10000)) +
geom_histogram(bins = 100) +
scale_x_continuous("Household income", labels = scales::dollar) +
scale_y_continuous("Count", labels = scales::comma) +
ggtitle("Distribution of personal income", "Men") +
theme_minimal(base_size = 10)
# Histogram: Black
hist_black = ggplot(data = filter(ps_df, i_black == 1), aes(x = personal_income * 10000)) +
geom_histogram(bins = 100) +
scale_x_continuous("Household income", labels = scales::dollar) +
scale_y_continuous("Count", labels = scales::comma) +
ggtitle("Distribution of personal income", "Black") +
theme_minimal(base_size = 10)
# Histogram: White
hist_white = ggplot(data = filter(ps_df, i_white == 1), aes(x = personal_income * 10000)) +
geom_histogram(bins = 100) +
scale_x_continuous("Household income", labels = scales::dollar) +
scale_y_continuous("Count", labels = scales::comma) +
ggtitle("Distribution of personal income", "White") +
theme_minimal(base_size = 10)
# Print the figures
hist_female + hist_male & coord_cartesian(xlim = c(0, 1.1e6))
```



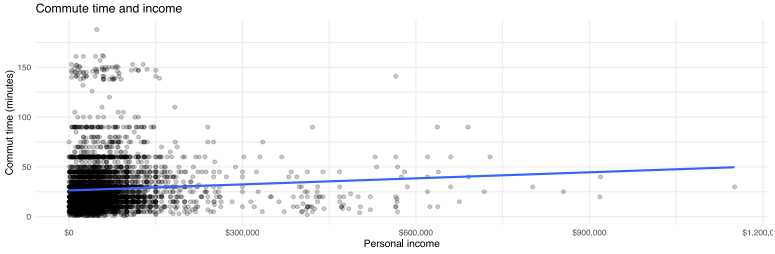

```
hist_black + hist_white & coord_cartesian(xlim = c(0, 1.1e6))
```



**Q06.** Create a scatterplot (AKA: dot plot) with commute time (`time_commuting`, which the length of the individual's morning commute, in minutes) on the `y` axis and personal income on the `x` axis.

**Answer:**

```
# Create the histogram of HH income using ggplot2
ggplot(data = ps_df, aes(x = personal_income * 10000, y = time_commuting)) +
geom_point(alpha = 0.2) +
geom_smooth(method = lm, se = F) +
scale_x_continuous("Personal income", labels = scales::dollar) +
scale_y_continuous("Commut time (minutes)", labels = scales::comma) +
ggtitle("Commute time and income") +
theme_minimal(base_size = 10)
```




**Q07.** Based upon your plot in **Q06**: If we regressed commute time on income, do you think the coefficient on income would be *positive* or *negative*? **Explain** your answer.

**Answer:** It's a bit difficult to say, but it looks like there are a lot of obsevations near the origin—*i.e.*, the regression line will start near the origin and then will slope slightly upward toward the extreme observations on the high end of the income distribution. That said, some individuals with lower incomes are making very long commutes that we basically do not observe at higher income levels.

**Q08.** Run a regression that helps summarize the relationship between commute length and personal income. 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 commute time on the level of personal income.

```
# Regression
est08 = lm(log(time_commuting) ~ personal_income, data = ps_df)
# Results
est08 %>% broom::tidy()
```

```
#> # A tibble: 2 x 5
#>   term            estimate std.error statistic      p.value
#>   <chr>              <dbl>     <dbl>     <dbl>        <dbl>
#> 1 (Intercept)      2.97      0.0146     204.   0
#> 2 personal_income  0.00816   0.00146      5.57 0.0000000264
```

The estimated coefficient in this log-linear model suggests that a $10,000 increase in personal income is associated with a 0.816% increase in commute time.

**Q09.** 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 income to be associated with *percent* changes in commute length (rather than level changes)—and because logging a variable can compress its distribution (commute lengths appear to be skewed). Percent changes also help us put things "in perspective"—helping us understand whether a 5 minute increase is "big."

# Regression refresher: Varying the specification

*Note:* In this section, when I ask you to "comment on the statistical significance," I want you to tell me whether the coefficient is significantly different from zero at the 5% level. You do not need write out the full hypothesis test.

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

**Answer:**

```
# Regress commute time on income
est11 = lm(time_commuting ~ personal_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)       26.4      0.419      62.9  0
#> 2 personal_income    0.202    0.0421      4.81 0.00000156
```

Our estimated coefficient suggests that a one-unit increase in personal 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: about 12 seconds).

**Q11.** Did the sign of the coefficient on personal income surprise you based upon your figure in **06**? Explain.

Answer Perhaps this surprised you a bit, but notice that there are **a lot** of observations down near the origin in **06**.

**Q12. Log-linear specification** Regress the log of commute time on personal income. Interpret the slope coefficient and comment on its statistical significance.

**Answer:**

```
# Log-linear regression
est12 = lm(log(time_commuting) ~ personal_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)      2.97      0.0146     204.   0
#> 2 personal_income  0.00816   0.00146      5.57 0.0000000264
```

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 0.8%. This coefficient is still statistically significant at the 5% level (and still small in absolute magnitude).

**Q13. 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
est13 = lm(log(time_commuting) ~ log(personal_income), data = ps_df)
# Results
est13 %>% broom::tidy()
```

```
#> # A tibble: 2 x 5
#>   term                 estimate std.error statistic  p.value
#>   <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)            2.89      0.0183    158.   0.
#> 2 log(personal_income)   0.0952    0.0106      8.95 5.06e-19
```

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 0.095 percent. This coefficient is still statistically significant at the 5% level (and still small in absolute magnitude).

# Multiple linear regression and indicator variables

*Note:* We're now moving to thinking about the time at which an individual leaves her home to go to work (`time_depart`). This variable is measured in minutes from midnight (so smaller values are earlier in the day).

**Q14.** Regress departure time (`time_depart`) on the indicator for female (`i_female`) **and** the indicator for whether the individual was married at the time of the sample (`i_married`). Interpret the intercept and **both** coefficients (commenting on their statistical significances).

**Answer:**

```
# Log-linear regression
est14 = lm(time_depart ~ i_female + i_married, data = ps_df)
# Results
est14 %>% broom::tidy()
```

```
#> # A tibble: 3 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    509.       5.15     98.9  0.
#> 2 i_female        21.3      5.58      3.82 1.36e- 4
#> 3 i_married      -48.2      5.63     -8.56 1.48e-17
```

The intercept (approximately 508.9 minutes past midnight, which is roughly 8.5 hours past midnight) tells us the expected time of departure when the other explanatory variables are 0. Thus, the intercept tells us the expected time of departure for unmarried men (when `i_married = 0` and `i_female = 0`).

Our coefficient for female (`i_female`) tells us the difference in average departure time for women and men is 21.32 minutes (meaning in this sample women, on average, leave for work later than men) **holding everything else constant**. This coefficient is still statistically significant at the 5% level.

Our coefficient on whether the individual is married (`i_married`) the average difference in departure time between married and unmarried individuals in the sample **holding all other variables constant**. Specifically, we find that married individuals, on average, leave for work 48.21 minutes **earlier** than their unmarried counterparts (holding all other variables constant). This coefficient is statistically significant at the 5% level.

**Q15.** What would need to be true for `age` to cause omitted-variable bias. Explain the requirements and whether you think they are likely to cause bias in this setting.

**Answer:** For age to cause bias as an omitted variable, it must (1) have an effect on time of departure and (2) correlate with an included variable. The first requirement seems possible, as sleep and work tendencies change with age. The second requirement also seems at least possible, as marriage status could be correlated with age.

**Q16.** Add `age` to the reression you ran in **Q14**. Do the results of this new regression suggest that `age` was causing omitted-variable bias? Explain your answer.

**Answer:**

```
# Log-linear regression
est16 = lm(time_depart ~ i_female + i_married + age, data = ps_df)
# Results
est16 %>% broom::tidy()
```

```
#> # A tibble: 4 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)   557.       9.17      60.7  0.
#> 2 i_female       22.2      5.57       4.00 6.50e- 5
#> 3 i_married     -36.5      5.91      -6.18 6.97e-10
#> 4 age            -1.26     0.201     -6.26 4.11e-10
```

It does seem like age might have been causing some omitted-variable bias. When we include `age` in the regression, the coefficient on marriage changes considerably, and the coefficient on age is statistically significant (and economically large). The fact that there is a large and significant relationship between departure time and age is at least consistent with age affecting departure time (the first requirement for omitted-variable bias). The fact the the coefficient on `i_married` changes suggests that marriage and age are correlated (the second requirement for omitted-variable bias).

**Q17.** Now regress departure time on `i_female`, `i_married`, **and their interaction**. (You should have an intercept and three coefficients: the two variables and their interaction.) Interpret the coefficient on the interaction and comment on its statistical significance.

*Hint:* In R you can get an interaction using `:`, for example, `lm(y ~ x1 + x2 + x1:x2, data = fake_df)`.

**Answer:**

```
# Log-linear regression
est17 = lm(time_depart ~ i_female + i_married + i_female:i_married, data = ps_df)
# Results
est17 %>% broom::tidy()
```

```
#> # A tibble: 4 x 5
#>   term                estimate std.error statistic  p.value
#>   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)          512.        6.14    83.3   0.
#> 2 i_female              16.4       8.49     1.93  5.40e- 2
#> 3 i_married            -52.4       7.80    -6.72  2.07e-11
#> 4 i_female:i_married     8.73     11.3      0.774 4.39e- 1
```

There are a couple of ways to think about the coefficient on the interaction. Likely the clearest: We can interpret this coefficient as asking whether marriage (`i_married`) has different effects for women and men. For example, if marriage causes men to go to work earlier and women to go to work later, then this coefficient would be positive. Interpreted this way, this coefficient says that, on average, marriage causes women to go to work slightly less early (48.21 minutes) relative to the marriage effect on men, holding all else constant.

Notice that the main effect of marriage on time of departure (the non-interacted effect) is large, negative, and significant. This interaction is much smaller, positive, and not statistically significant.

**Q18.** For this last regression, we are going to do something totally different. Our outcome variable is going to be an indicator for whether the individual has internet access (`i_internet`). Regress this internet-access variable on a two explanatory variables: (1) an indicator for whether the household's location is considered urban `i_urban` (vs. rural) and (2) an indicator for whether the individual is black (`i_black`).

Interpret the intercept and coefficients. Comment on their statistical significance.

**Answer:**

```
# Regression
est18 = lm(i_internet ~ i_urban + i_black, data = ps_df)
# Results
est18 %>% broom::tidy()
```

```
#> # A tibble: 3 x 5
#>   term         estimate std.error statistic  p.value
#>   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)   0.945    0.00475    199.   0.
#> 2 i_urban       0.0277   0.00600      4.62 3.92e- 6
#> 3 i_black      -0.0726   0.0102      -7.09 1.53e-12
```

continued...

When the outcome variable is an indicator variable, we interpret the coefficients as percentages (sometimes referred to as *shares*).

The intercept tells us the percentage of individuals who have internet access when the other variables are zero—meaning for non-urban, non-black individuals. Thus, approximately 94.5% of rural, non-black individuals have internet access in the sample.

The coefficient on `i_urban` tells us the urban *vs.* rural gap in internet access (in this sample). Thus, urban individuals are 2.77% (percentage points) more likely to have internet access than their rural counterparts, **holding everything else constant**.

The coefficient on `i_black` tells us the difference in internet access between black and non-black individuals in the sample. Specifically, we find the black individuals in the sample are 7.26% less likely to have internet access relative to non-black individuals **holding everything else constant**.

Both of the coefficients are statistically significant at the 5% level.

# The bigger picture

Write short answers to each of these questions. No math needed: Just explain your reasoning.

**Q19.** Define the term *standard error*.

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

**Q20.** For exogeneity, we write $E[u|x] = 0$. What does this mathematical expression mean for the relationship between $u$ and $x$?

**Answer:** This expression means that our disturbance $u$ cannot have *any* relationship with the variable $x$.

## Description of variables and names

| Variable | Description |
|---|---|
| `state` | State abbreviation |
| `i_urban` | Binary indicator for whether home county is 'urban' |
| `age` | The individual's age (in years) |
| `i_asian` | Binary indicator for whether the individual identified as Asian |
| `i_black` | Binary indicator for whether the individual identified as Black |
| `i_hispanic` | Binary indicator for whether the individual identified as Hispanic |
| `i_indigenous` | Binary indicator for whether the individual identified with a group indigenous to North Am. |
| `i_white` | Binary indicator for whether the individual identified as White |
| `i_female` | Binary indicator for whether the individual identified as Female |
| `i_male` | Binary indicator for whether the individual identified as Male |
| `i_grad_college` | Binary indicator for whether the individual graduated college (estimated) |
| `i_married` | Binary indicator for whether the individual was married at the time of the sample |
| `personal_income` | Total (annual) personal income (tens of thousands of dollars) |
| `i_foodstamps` | Binary indicator for whether the individual uses 'foodstamps' (SNAP) |
| `i_health_insurance` | Binary indicator for whether the individual has health insurance |
| `i_internet` | Binary indicator for whether the individual has access to the internet |
| `time_depart` | The time that the individual typically leaves for work (in minutes since midnight) |
| `time_arrive` | The time that the individual typically arrives at work (in minutes since midnight) |
| `time_commuting` | The length of time that the individual typically travels to work (in minutes) |

I've tried to stick with a naming convention. Variables that begin with `i_` denote binary/indicator variables (taking on the value of `0` or `1`).