

# Problem Set 1

## Econometrics Review

**EC 421:** Introduction to Econometrics

Due *before* midnight (11:59pm) on Sunday, 24 April 2021

**DUE** Your solutions to this problem set are due *before* 11:59pm on Sunday, 24 April 2021 on [Canvas](#). **Your answers must include two files (1)** your responses/answers to the question (e.g., a Word document) and **(2)** the R script you used to generate any answers in R. 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.**

## Problem 2: Getting started

### 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("ps-001-data.rds") %>% read_rds()
# Load data: As 'csv'
ps_df = here("ps-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] 8000 24
```

We have 8,000 observations (rows) on 24 variables (columns).

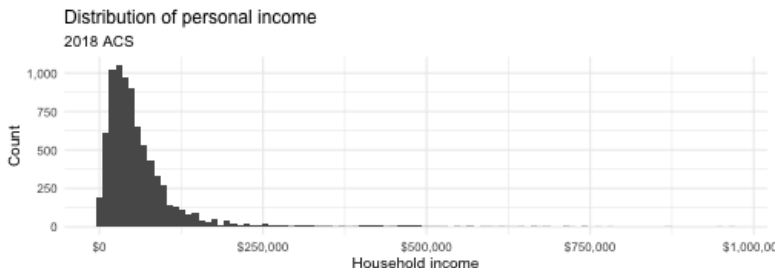
## Getting to know your data

**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 bad, 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.   : 0.0004
#> 1st Qu.: 2.2000
#> Median : 3.8350
#> Mean   : 5.0210
#> 3rd Qu.: 6.2000
#> Max.   :96.7000
```

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

#> personal_income
#> Min.   : 0.0004
#> 1st Qu.: 3.0088
#> Median : 5.2000
#> Mean   : 7.6376
#> 3rd Qu.: 8.6000
#> Max.   :95.0000
```

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

#> personal_income
#> Min.   : 0.060
#> 1st Qu.: 2.137
#> Median : 3.500
#> Mean   : 4.504
#> 3rd Qu.: 5.400
#> Max.   :71.400
```

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

#> personal_income
#> Min.   : 0.0004
#> 1st Qu.: 2.8000
#> Median : 4.8000
#> Mean   : 6.6993
#> 3rd Qu.: 7.8000
#> Max.   :96.7000
```

```
# Summary of all citizens
ps_df %>% filter(i_citizen == 1) %>% select(personal_income) %>% summary()

#> personal_income
#> Min.   : 0.0004
#> 1st Qu.: 2.7000
#> Median : 4.6000
#> Mean   : 6.5329
#> 3rd Qu.: 7.5175
#> Max.   :96.7000
```

```
# Summary of all noncitizens
ps_df %>% filter(i_citizen == 0) %>% select(personal_income) %>% summary()

#> personal_income
#> Min.   : 0.032
#> 1st Qu.: 1.742
#> Median : 2.900
#> Mean   : 4.606
#> 3rd Qu.: 5.500
#> Max.   :48.300
```

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.

The personal income distributions (in our sample) for citizens and non-citizens. Interestingly, non-citizens have a larger minimum in our sample than our citizens do. At their means, the two groups differ by 19,269. Citizens earn 65,329 on avg., while non-citizens earn 46,060 on avg.

```

# 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)

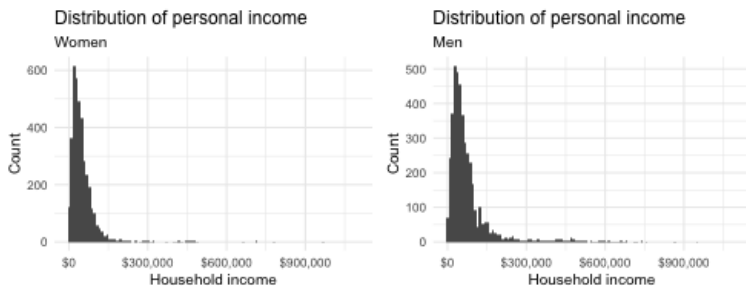
# Histogram: Men
hist_male = ggplot(data = filter(ps_df, i_male == 1), aes(x = personal_income * 10000)) +
  geom_histogram(bins = 100) +
  scale_x_continuous("Household income", labels = scales::dollar) +
  scale_y_continuous("Count", labels = scales::comma) +
  ggtitle("Distribution of personal income", "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)

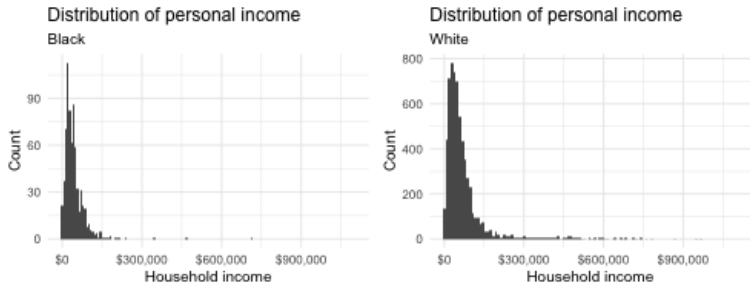
# Histogram: White
hist_white = ggplot(data = filter(ps_df, i_white == 1), aes(x = personal_income * 10000)) +
  geom_histogram(bins = 100) +
  scale_x_continuous("Household income", labels = scales::dollar) +
  scale_y_continuous("Count", labels = scales::comma) +
  ggtitle("Distribution of personal income", "White") +
  theme_minimal(base_size = 10)

# Print the figures
hist_female + hist_male & coord_cartesian(xlim = c(0, 1.1e6))

```



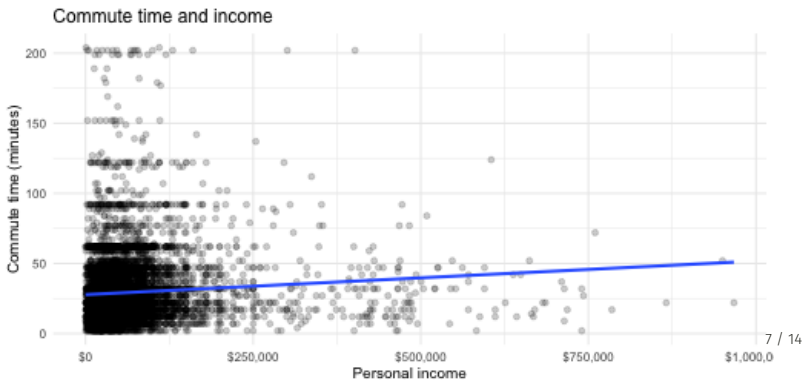
```
hist_black + hist_white & coord_cartesian(xlim = c(0, 1.1e6))
```



**Q06.** Create a scatterplot (AKA: dot plot) with commute time (`time_commuting`, which is 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("Commute 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 observations 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 log of personal income.

```
# Regression
est08 = lm(log(time_commuting) ~ log(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)         2.96      0.0160    185.     0.
#> 2 log(personal_income) 0.0948    0.00924   10.3 1.43e-24
```

The estimated coefficient in this log-linear model suggests that a 1 percent increase in personal income is associated with a 9.482% 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-log 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." This particular specification represents the elasticity of commute time with respect to income, which tells us how willing workers are in our sample to commute for higher wages.

## 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)         27.8      0.349     79.5    0.
#> 2 personal_income      0.240     0.0353     6.79 1.16e-11
```

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 14 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)         3.04      0.0115    264.    0.
#> 2 personal_income    0.00815  0.00116    7.01 2.63e-12
```

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)         2.96      0.0160    185.    0.
#> 2 log(personal_income) 0.0948  0.00924   10.3 1.43e-24
```

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_male`) **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_male + 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)  520.      4.19     124.    0
#> 2 i_male      -24.6     4.45     -5.52  0.0000000346
#> 3 i_married   -26.6     4.56     -5.83  0.0000000560
```

The intercept (approximately 520.4 minutes past midnight, which is roughly 8.7 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 -24.58 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 26.63 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`, as well as expected earnings. The second requirement also seems at least possible, as marriage status could be correlated with `age`.

**Q16.** Add `age` to the regression 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)  522.      8.43     61.9    0
#> 2 i_female      25.1     4.45      5.64  0.0000000180
#> 3 i_married   -23.6     4.64     -5.09  0.000000358
#> 4 age         -0.614    0.171     -3.59  0.000339
```

**Q17.** Now regress departure time on `i_male`, `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_male + i_married + i_male: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)         516.      5.01    103.      0
#> 2 i_male              -16.3      7.14    -2.29  0.0222
#> 3 i_married           -19.7      6.54    -3.01  0.00258
#> 4 i_male:i_married    -13.5      9.13    -1.48  0.140
```

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 negative. Interpreted this way, this coefficient says that, on average, marriage causes men to go to work slightly earlier (26.63 minutes) relative to the marriage effect on women, 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 similarly sized, but 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 a citizen (`i_citizen`).

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

**Answer:**

```
# Regression
est18 = lm(i_internet ~ i_urban + i_citizen, 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)         0.872      0.0106     82.5      0
#> 2 i_urban              0.0302     0.00541    5.58 0.0000000243
#> 3 i_citizen            0.0557     0.0101    5.54 0.0000000306
```

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 87.2% 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 3.02% (percentage points) more likely to have internet access than their rural counterparts, **holding everything else constant**.

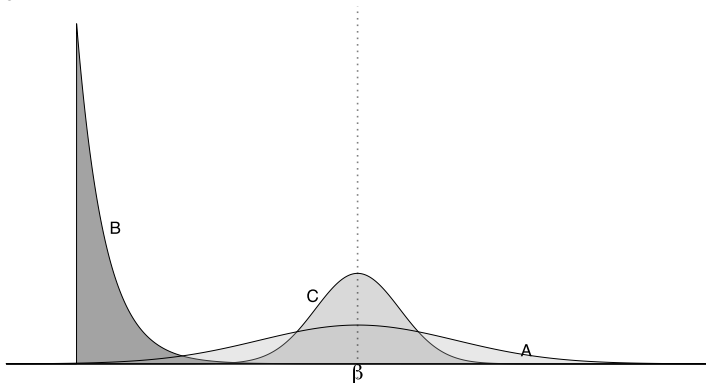
The coefficient on  $i\_citizen$  tells us the difference in internet access between citizens and non-citizens in the sample. Specifically, we find the citizens in the sample are 5.57% more likely to have internet access relative to non-citizens **holding everything else constant**.

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

## The bigger picture

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

Figure 1



**Note** This figure shows the distributions of three estimators (A, B, and C) that each estimate the unknown parameter  $\beta$ .  $E[A] = \beta - 3$ ,  $E[B] = \beta$ ,  $E[C] = \beta$

**Q19a.** Are the estimators in Figure 1 (above) unbiased? Which ones? *Hint:* There may be more than one.

**Answer:** A, C

**Q19b.** Which of the estimators in Figure 1 (above) has the minimum variance?

**Answer:** B.

**Q19c.** Which of the estimators in Figure 1 (above) is the best (minimum variance) unbiased estimator?

Variable	Description
state	State abbreviation
age	The individual's age (in years)
i_urban	Binary indicator for whether home county is 'urban'
i_citizen	Binary indicator for whether the individual is a citizen (naturalized or born.)
i_noenglish	Binary indicator for whether the individual speaks English
i_only_english	Binary indicator for whether the individual speaks ONLY English
i_drive_to_work	Binary indicator for whether the individual drives to work or takes a personal car
i_asian	Binary indicator for whether the individual identified as Asian
i_black	Binary indicator for whether the individual identified as Black
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
i_grad_highschool	Binary indicator for whether the individual graduated high school
i_married	Binary indicator for whether the individual was married at the time of the sample
i_married_mult	Binary indicator for whether the individual has been married multiple times at the time of the sample
personal_income	Total (annual) personal income (tens of thousands of dollars)
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).