```
***** Preliminaries
capture log close _all  // Closes any log if open //

cd "C:/Users/AN.4271/Dropbox/HHS 651/Assignments/Assignment 1/" /* Sets the Stata Working Directory (note the forward slashes.
 Stata will try to work with either forward or backwards slashes, but Windows-style back
 slashes sometimes interfere with functionality, so forward slashes are preferred. */

 log using "assignment1log", text replace   /* Starts a text-type log file called
                        "assignment1log" */

************             HHS 651: Assignment 1          ********************
***************    Stata Solutions - Andrew Proctor      ********************



********** Data Manipulation

**** Import Dataset CSV File
import delimited using "prgswep1.csv", clear

**** Question 1:  Describe Dataset
    describe, short

    /*  Discussion: There are 4,469 observations (individuals) and 1,328
    variables in the dataset.  */


**** Question 2:  Explanatory Variables

    **** 2a. Gender (gender_r)
        *** Explore Gender Variable
        codebook gender_r  //  View storage format of variable 'gender_r' //

        *** Create a "Female" Indicator Variable
        gen female = (gender_r == 2) if !missing(gender_r)
            /*  For individuals whose gender is listed in gender_r, assigns a
            value of 1 for female if gender is equal to 2, 1 if not.  Missing
            values in gender_r would also appear as missing in the female
            variable.*/

        tabulate female  // Displays the freq/percent of each value of "female."

        /*
        Discussion: The variable "gender_r" represents the gender listed
        for each variable.  When the CSV file was read into Stata, the variable
        was interpreted as a 'numeric' type variable.   50.41% of observations
        are male, 49.59% female, and there are no missing observations.
        */

        *** Note: Another way to create the female indicator variable would be:
        //  gen female_alt = 0 if !missing(gender_r)
```

```
        //  replace female_alt = 1 if ( gender_r == 2 & !missing(gender_r))
        //  tabulate female_alt

    **** 2b.  Years of Schooling (yrsqual)
        *** Explore 'Years of Schooling' Variable
        codebook yrsqual // View storage format of variable 'j_q04a' //
        tabulate yrsqual
            /* Since 'yrsqual' is a string-variable, only the first
            9 values are shown using the codebook command.  Using tabulate, we
            see some of the observations have a missing value "D" -  which means
            "Don't Know" according to the downloaded codebook. */

        ***** Format- Years of Schooling Variable
        replace yrsqual =  ".d" if yrsqual == "D"
            /* Since we need to format the variable as a numeric (quantitive)
            variable, we need to Stata to interpret the missing values
            correctly.  Missing values in Stata are denoted my ".", where
            letters can follow the "." to indicate what type of missing data we
            have.  So we change "D" to ".d". */

        destring(yrsqual), gen(yearsch)
            /* Now, we need to Stata to convert the variable to numeric,
            by parsing the text (string) values as numbers. */

        tabulate yearsch // Check to make sure no more missing values.

        tabulate yearsch, missing /* Note: You can see missing values again in
                              tabulate by using option, ", missing" */

        summarize yearsch  // Produces basic descriptive statistics for 'age'

        /*
        Discussion:  The variable "yrsqual" is a derived measure of years
        of schooling.  The variable was stored in Stata as a "string" type of
        variable (Why?  Because some observations take on the non-numeric "D"
        value).  After converting the variable to numeric, we see the mean is
        12.33, with std. dev. of 2.57, min of 6 and max of 20. There are 2
        missing observations.
        */

    **** 2c. Age (age_r)
        *** Explore Gender Variable
        codebook age_r  //  View storage format of variable 'gender_r' //

        rename age_r age  /* Rename 'age_r' to 'age' (not necessary,
                    but makes regression more understandable later */

        *** Generate 'Potential Experience' Variable
        gen potent_exper = max(0,age - 19) /* Generates a 'Potential Experience"
                                variable, equal to age - 19 for
                                individuals who are at least 19,
```

```
                              0 otherwise. */

        summarize potent_exper, detail

        /*
        Discussion:  The variable "age_r" is a derived measure of age (in years)
        of the individual.  The variable is stored in Stata as numeric and there
        are no missing observations. Using "summarize, detail" we see that the
        mean is 22.03 years and median (50th percentile) is 23 years.
        */

    **** 2d.  Cognitive Ability (using pvpsl1)
        *** Explore 'Problem-solving scale score' Variable
        codebook pvpsl1 // View storage format of variable 'pvpsl1' //

        *** Generate Quantile of Cognitive Ability
        egen cogn_rank = rank(pvpsl1) if !missing(pvpsl1) /* Rank of individuals'
                                    pvpsl1 if known. */

        egen count_cogn = count(pvpsl1) if !missing(pvpsl1) /* Total number of
                                     nomissing observations
                                     for pvpsl1. */

        *** Percentile Rank
        gen cogn_samp_pctile  = ((cogn_rank -1) / (count_cogn - 1)) * 100

        /*
        Discussion:  The variable "pvpsl1" is a derived measure of an
        individuals' problem solving ability.  The variable is stored as a
        numeric variable in Stata and there are 506 missing observations.
        */

**** Question 3:  Dependent Variable (Monthly Earnings Quintile)
    codebook monthlyincpr // View storage format of variable 'earnhrbonus' //

    *** Explore 'Employment Status'
    codebook monthlyincpr

    recode monthlyincpr (1 = 5) (2 = 17.5) (3 = 37.5) (4 = 62.5) (5 = 82.5) ///
        (6 = 95), gen(income_pctile)

    * Alternate recode
//  gen income_pctile = .
//  replace income_pctile = 5 if (monthlyincpr == 1 & !missing(monthlyincpr))
//  replace income_pctile = 17.5 if (monthlyincpr == 2 & !missing(monthlyincpr))
//  replace income_pctile = 37.5 if (monthlyincpr == 3 & !missing(monthlyincpr))
//  replace income_pctile = 62.5 if (monthlyincpr == 4 & !missing(monthlyincpr))
//  replace income_pctile = 82.5 if (monthlyincpr == 5 & !missing(monthlyincpr))
//  replace income_pctile = 95 if (monthlyincpr == 6 & !missing(monthlyincpr))

    replace income_pctile = 0 if c_d05 ==2 // Assign value of 0 for unemployed.
```

```

    drop if c_d05 ==3 | c_d05 == 4 // Drop if not in labor market or unknown.

    codebook income_pctile // Check number of missing values of new var.

    /*
    Discussion:  The number missing observations for "monthlyincpr" is  1,236.
    The number of missing observations for the revised measure is 122.
    */

*** Question 4:  Regression Analysis
    *** 4a:  Regress Income Rank on Cognitive Ability, Potential Experience, and Female Gender
    reg income_pctile cogn_samp_pctile potent_exper i.female if ///
        ((age >= 30) & (age <= 65))
        /*
           Note:  A more concise way to write the condition for age in this
           interval is to use the command inrange as follows (I will use
           inrange in the remainder of the solution).

           Additionally, an alternative to use any 'if' condition in the
           regression whatsoever would be the command:
           "keep if inrange(age, 30,65)" but deleting observations outside this
           range is both unnecessary and would make things more difficult if you
           want to do further analysis on the full sample. */


    reg income_pctile cogn_samp_pctile potent_exper i.female if ///
        inrange(age, 30, 65)

        /*
        Discussion:

        The coefficient on cogn_samp_pctile implies that a one percentile
        increase in cognitive ability is estimated to shift an individual's percentile
        of earnings up by .3391429 (that is, .3391429 percentage points if
        percentile is expressed on a 0-1 scale).

        The coefficient on  potent_exper implies that a one year increase in
        potential experience is estimated to increase ones' percentile
        of earnings by .4132204 percentage points.

        The coefficient on female suggests that being female is estimated to
        increase the percentile of income by 12.38118 percentage points,
        compared to being a male.

        The constant estimate suggests that that the predicted percentile
        of income for a male (female = 0) with 0 years of potential experience
        and in the 0th percentile of cognitive ability is the 37th percentile.
        */


```

```
*** 4b: Add Exper^2 and Age
reg income_pctile cogn_samp_pctile c.potent_exper##c.potent_exper ///
    i.female age if inrange(age, 30, 65)

/*
    Discussion:

    Age:  The age variable is omitted.  If you look at the top of the
    regression output, it notes that age is omitted because of
    collinearity (Stata automatically detects perfect collinearity and drops
    one of the collinear variables.  Age here is a linear function of potential
    experience and the constant, since age = potentexper + 19. This is a violation
    of the MLR Assumption 3, which is simply "no perfect collinearity."

    Square of Potential Experience:  The quadratic of experience is
    negative and significant.  This indicates that the benefit of an
    additional year of experience is diminishing as the years of
    experience one already has increases.  Omission of a relevant quadratic
    term like this is a common example of the mispecification of functional
    form that is a violation of MLR Assumption 4 (zero conditional mean) for
    estimating the true model.

    R^2:  The R^2 in the second model is higher than the first (0.1668 as
    opposed to 0.1551), indicating adding the square of experience increases
    the total amount of explained variation in income percentile.  R^2 will
    never decrease with the addition of subsequent variables.  To see this,
    note that R^2= 1 - (Sum of Squared Residuals / Total Sum of Squares).
    Everything except the Sum of Squared Residuals are the same across
    the two models, and since the second model contains all predictors from
    the firt model, the sum of squared residuals will be no greater than in
    the first model.

*/

*** 4c: Compare School Years vs Cognitive Ability
reg income_pctile cogn_samp_pctile potent_exper i.female if inrange(age, 30, 65)
scalar R2model4a = e(r2_a)  // Save R^2 as a scalar.  (Also in reg output)

reg income_pctile yearsch potent_exper i.female if inrange(age, 30, 65)
scalar R2model4c = e(r2_a) // Save R^2 as a scalar.  (Also in reg output)

display R2model4a - R2model4c /* Displays difference in R^2 output. Note: For
                 the assignmnent, you could just compare them
                 from the regression output of each model. */

/*
Discussion:

The two models perform nearly identically, with the
regression model from 4(a) explaining .066432% more of the variation in
income quintiles.
```

```

    (Not graded) Potential Problems with Either Model:
    The two models preview common challenges in applied econometrics we will
    discuss in subsequent lectures.  As you can see from the covariance matrix
    below, Cov(cogn_samp_pctile, yearsch) is not equal to zero, and both appear
    likely to affect incomes, implying omitted variable bias (i.e. a violation
    of MLR Assumption 4).  One response would be to control for both cognitive
    ability and schooling.  But this brings up an issue from Ch.3: endogeneity.
    The basic idea is that OLS is biased if you include explanatory variables
    that are caused by other variables in the model.  If cognitive ability
    increases years of schooling, then years of schooling is endogenous when you
    both are in the model. Equally, one might imagine that, as individual gains
    more years of schooling, their cognitive ability increases.  If this is
    true, cognitive ability is also endogenous to schooling (when two variables
    causally influence each other, this is a particular type of endogenity called
    simultaneity).

    */

    correlate cogn_samp_pctile yearsch, covariance


**** Extra Question for three person groups

    **** Question 5(a) Explore Structure of the variable "g_q03h" - which is
    ** 'Skill use work - Numeracy - How often - Use advanced math or statistics'
    codebook g_q03h

        /*
            From looking at 'math use at work' with the codebook command, we
            see that this variable takes on only 9 unique values, meaning that
            all values are displayed by Codebook.  From this, we can see right
            away that we have the following 'Missing value' indicators that need
            to be relabelled: 'D', 'N', 'R', and 'V'.
        */

    **** Question 5(b) Suitably reformat g_q03h and provide the mean and
    **** standard deviation using the original vaue scheme.

        *** Recode Missing Values for g_q03h
        replace g_q03h = ".d" if g_q03h=="D"
        replace g_q03h = ".n" if g_q03h=="N"
        replace g_q03h = ".r" if g_q03h=="R"
        replace g_q03h = ".v" if g_q03h=="V"

        *** Convert g_q03h to a numeric variable by destringing
        destring g_q03h, replace

        *** Produce summary statistics for g_q03h using original coding of
        *** use frequencies
        summarize g_q03h
```

```

        /*
            The mean (pre-transformation) of this variable is 1.287818 and the
            standard deviation is 0.7269503.
        */

    **** Question 5(c) - Recode g_q03h so that the values represents number of
    *** times each month an individual uses advanced math or statistics at work
        recode g_q03h (1 = 0) (1 = 0.5) (3 = 2.5) (4 = 12) (5 = 20) ///
                , gen(mathuseatwork)

        /*
            This question highlights a common problem in applied work, which is
            that survey data often uses an ordinal or interval approach to
            asking retrospectative information.  You as the researcher must then
            decide how to make that interpretable numerically and justify it.

            In assigning values here myself, I assume that individuals work
            4 5-day work weeks per month, for a total of 20 work days. So if an
            individual reports they use math at work "everyday," (5 in the old
            schema) that equates to 20 days per month.

            "Never" (1 in original coding) is straightforwardly represented as
            0 times per month.

            For less than once a month (1), I code this as
            as the midpoint between 0 and 1, i.e. 0.5 days per month.

            For less than once a week but at least once a month (3), this
            should be less than four (i.e. at most 3) according to my
            assumptions about a 4 week work month, but greater than 1.  I again
            use the midpoint of (1,3), that is is 2.5 days per month.

            For at least once a week but not every day (4), this again should be
            less than 20 but less than 4.  So once again taking the midpoint of
            (4,20), I code this as 12 days per month.
        */

        **** Summarize recoded math use at work variable
        summarize mathuseatwork

        /*
            The mean of the variable after transforming it to be more directly
            interpretable is .7665993  and the standard deviation is 2.663051.
        */

    **** Question 5(d) - Regressions relating to a math use at work -> cognitive
    **** ability -> income pathwawy.

        *** Question 5(d)(i) Regression of Cognitive Ability on math use at work
        reg cogn_samp_pctile mathuseatwork if (inrange(age, 30, 65) & (c_d05==1))
```

```
        *** Question 5(d)(ii)Regression of Earnings Pctile on Cognitive Ability
        reg income_pctile cogn_samp_pctile if (inrange(age, 30, 65) & (c_d05==1))

        *** Question 5(d)(iii) Regression of Earnings Pctile on math use at work
        reg income_pctile mathuseatwork if (inrange(age, 30, 65) & (c_d05==1))

        /*
            Discussion:

            Regression 5(d)(i) suggests that for each additional day per month
            that an individual uses advanced math at work, their percentile of
            cognitive ability increases by  1.723182, which is statistically
            significant (p-value < 0.01).  It's not immediately required for
            this question, but you may note that these estimates seem almost
            implausibly high - as we will discuss further in 5(f).

            Regression 5(d)(ii), like analysis in question 4, suggests that
            cognitive ability has a positive impact on earning, with a
            1 percentile increase in positive ability estimated to increase
            earnings percentile by 0.2753122, which is statistically
            significant (p-value < 0.01). If both this relationship and the
            relationship from 5(d)(i) are indeed correct, then math use at
            work should have a direct effect on earnings percentile via this
            pathway.

            Regression 5(d)(iii) estimates that cognitive ability does indeed
            have an effect earnings percentile - in fact even larger than the
            estimated effect through the cognitive ability - earnings pathway.
            An increase in math use of work by once a month is estimated to
            increase earnings percentile by  1.811131, which is statistically
            significant (p-value < 0.01).  Again, these results are implausibly
            high - raising the spector of reverse caualty / endogeneity and
            foreshadowing 5(f).
        */


    **** Question 5(e) - Regressions relating to an erroneous math use at work
    **** -> years of schooling -> income pathway.

        *** Question 5(e)(i) Regression of years of schooling on math use at work
        reg yearsch mathuseatwork  if (inrange(age, 30, 65) & (c_d05==1))

        *** Question 5(e)(i) Regression of income percentile on years of schooling
        reg income_pctile yearsch  if (inrange(age, 30, 65) & (c_d05==1))

    /*
            Discussion:

            Regression 5(e)(i) estimates that math use at work
            has a positive, statistically significant effect on years of
```

```
                schooling.  Regression 5(e)(ii) then suggests that years
                of schooling has a positive, statistically significant effect on
                earnings percentile.

                This would point to a second causal pathway
                for math use at work to effect earnings, but thinking about
                regression 5(e)(i) - it doesn't make any sense under our assumptions.
                If schooling strictly predates math use at work, then math use at
                work cannot effect schooling.  Instead, what we very likely have is
                reverse causality - an individual's schooling instead affects their
                math use at work.  To see that a coefficient will be different from
                zero when the true relationship runs in reverse of what is estimated,
                consider the expression for Beta in terms of the sample correlation
                and standard deviations:

                    - For regression of y on x, the coefficient on x is:
                          beta_x = Corr(x,y) * (StdDev_x / StdDev_y)
                    - And for the regression of x on y, the coefficient on y is:
                          beta_y = Corr(x,y) * (StdDev_y / StdDev_x)

                Since the fraction (StdDev_y / StdDev_x) and it's inverse are always
                strictly positive, then for nonzero Corr(x,y), running regression
                in the 'wrong' direction (from y to x) will always yield a nonzero
                coefficient with the same sign as the effect in the right direction
                (from x to y).

                To demonstrate this argument, we run a regression interchanging
                our dependent and independent variables in 5(e)(i).

            */

        *** Demonstrating that regression can't tell us the direction of causality
            reg mathuseatwork yearsch

        **** Question 5(f) - Inference from 5(d) in light of 5(e)

        /*
                Discussion:
                In 5(e), we see a rather stark case where causality cannot run in
                the direction estimated by OLS, where math use at work is estimated
                to increase years of schooling that predates work.

                This same concern is likely to extend to the relationship in 5(d).
                Individuals with higher cognitive ability are probably more likely
                to work in jobs with greater use of advanced math.  In general,
                there is likely to be the same issue of simultaneity in the
                relationship between math use at work and congitive ability.

                Generally, this question highlights the difficulty in finding good
                variables where there is no concern about OVB or reverse caulality.

```

```
            Specifically, extending the logic from 5(d), it seems reasonable to
            believe that higher paying jobs may often require greater use of
            mathematics - irrespective of someone's aptitude or qualifications.
            Hence, rather than higher math use 'causing' higher earnings, higher
            earnings in these situations would be 'causing' more math use.  But
            since more math use might actually have the effect we originally
            hypothesized - increasing congitive ability and thereby leading to
            greater earnings - it's hard to disentangle these two effects.

            The potentially problematic nature of the relationship between math
            use at work and cognitive abiltiy highlights another possible
            challenge to the regression we have specified in 4(a):  while
            cognitive ability is likely to influence earnings, earnings may also
            be affecting the measurement of cognitive ability through higher
            math use at better paid jobs.

            Note:  Questions 5 is meant to get at the questions of
            reverse causality and simultaneity more in-depth.  The timing of
            effects problem in 5(e) is meant especially to highlight that
            causality can't run in the direction specified.  But it is also
            possible to make a critique centered entirely around more typical
            ommited variable bias (OVB).  Students who don't address reverse
            causality but instead make a clear and well-reasoned analysis to
            this question using OVB will still earn full credit.
        */
*****
log close _all
```