14.03/14.003 Problem Set 4
Fall 2014

Professor David Autor

Due Thursday, 11/6/2014 at 8pm EST

1 Trade and production (30 points)

Walter White lives in the desert in relative isolation. He only produces (for personal amusement) red phosphorus (R) and lye (Y), and every day, he has 20 hours to spend making them. He produces (in grams)

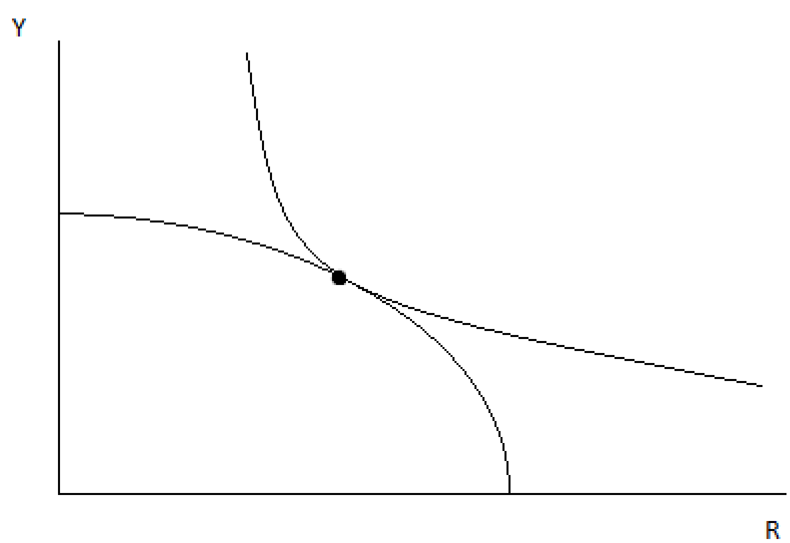
$$R = \sqrt{L_R}$$
$$Y = \frac{\sqrt{L_Y}}{2}$$

and has utility $U = RY^2$. His total time spent cannot exceed 20: $L_R + L_Y = 20$.

1. Write Walter's constraint on amount of red phosphorus and lye he can consume. Set up his maximization problem with this constraint and solve. What is his utility at this optimum? On a graph, draw Walter's production possibility frontier (PPF) and his optimal level of production. **(7 points)**
   Walter's production constraint $L_R + L_Y = 20$ can be rewritten by substituting $R^2 = L_R$and $4Y^2 = L_Y$, so we get $R^2 + 4Y^2 = 20$. His maximization problem is then:

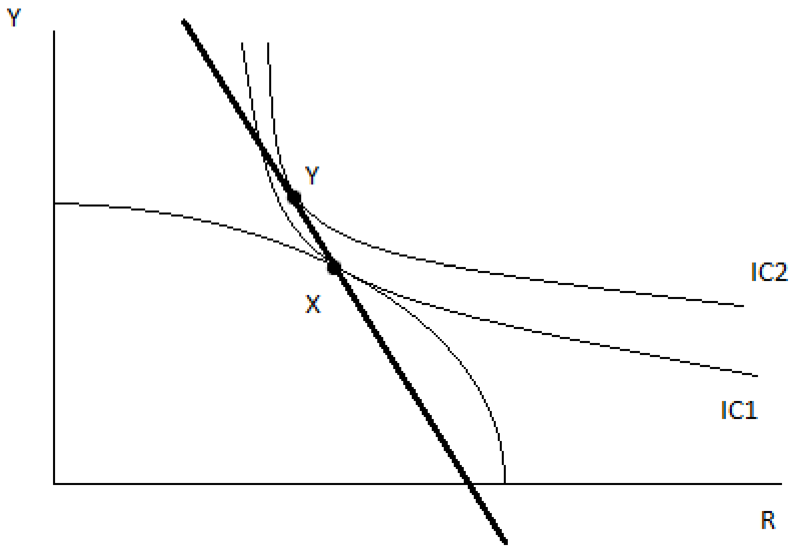
$$\max RY^2$$
$$s.t. \quad R^2 + 4Y^2 = 20$$

and we can solve by substituting: $\max R(20 - R^2)/4$**. The FOC is** $5 - 3R^2/4 = 0$**, or** $R^2 = 20/3$**,** $R = \sqrt{20/3}$**. And using the budget constraint, we get the optimal Y is** $Y = \sqrt{10/3}$**. His utility is** $U = RY^2 = \sqrt{20/3}(10/3) =$ **8.606.**



2. Walter's former student Jesse invents the International (Recreational Chemicals) Trade Machine, which can take in any amount of phosphorus to produce $1/p$ times that amount of lye, or take any number of lye and produce p times that number of red phosphorus. On your graph, for a single p draw Walter's new budget constraint if he produces the previous optimal amount of R and Y but can also use International Trade and state the approximate value of p for your line. How does a change in p affect your budget constraint? **(4 points)**
The slope of the budget line in the graph is equal to $1/p$. Any change in p changes the slope of the line passing through our optimal point, which changes Walter's budget constraint. Any change in p will change Walter's optimal consumption bundle, even holding his production fixed.

3. Without doing any algebra, for what values of p will the International Trade Machine increase Walter's payoff at this level of production? On your graph indicate where a new optimal point might be for the p that you drew(without solving). **(4 points)**
Any value of p will weakly increase our welfare. If the price ratio p is equal to $(2\sqrt{2})$ then the price ratio line will be exactly tangent to the optimal point we found in a) and there will no be gains from trade. In the graph, for $p = 0.5$, the optimal point moves from X to Y.



4. Walter's old optimal production is no longer optimal if there is trade. Walter can increase his utility if he were to produce in a way that maximizes the value of his goods in the International Trade Machine. If the machine will trade 1 *tweaker* for each gram of red phosphorus and p *tweakers* for a gram of lye, setup Walter's maximization where he maximizes his number of tweakers using his time constraint. Solve the problem and draw this point and the budget constraint on your graph. What is his income (in tweakers)? **(5 points)**
The optimal production will be at a point where the production possibility frontier is tagent to the price ratio. The PPF has slope$\frac{\frac{\partial P}{\partial Y}}{\frac{\partial P}{\partial R}}$. We find a point where $\frac{\frac{\partial P}{\partial Y}}{\frac{\partial P}{\partial R}} = p$. $\frac{8Y}{2R} = p$ or $Y = pR/4$. Substituting this condition into the PPF constraint, we get that on the production possibilities frontier, $R + p^2R^2/4 = 20$ or $R = \left(\frac{80}{p^2+4}\right)^{1/2}$ worth 1 tweaker each, and and $Y = p\left(\frac{80}{p^2+4}\right)^{1/2}/4$ worth p tweakers each. His total amount of tweakers is $I = \left(\frac{80}{p^2+4}\right)^{1/2} + p^2\left(\frac{80}{p^2+4}\right)^{1/2}/4$.

5. With the tweakers that Walter now has, he can then trade them back to the International Trade Machine in exchange for red phosphorus and lye at the same ratio (1 and p respectively). What is his new budget constraint (Should be in terms of tweakers rather than hours)? Solve his new utility maximization problem. **(5 points)**
We found that Walter's optimal production bundle was worth I tweakers total. He now has I tweakers to spend to maximize his utility $u = RY^2$. The maximization problem is

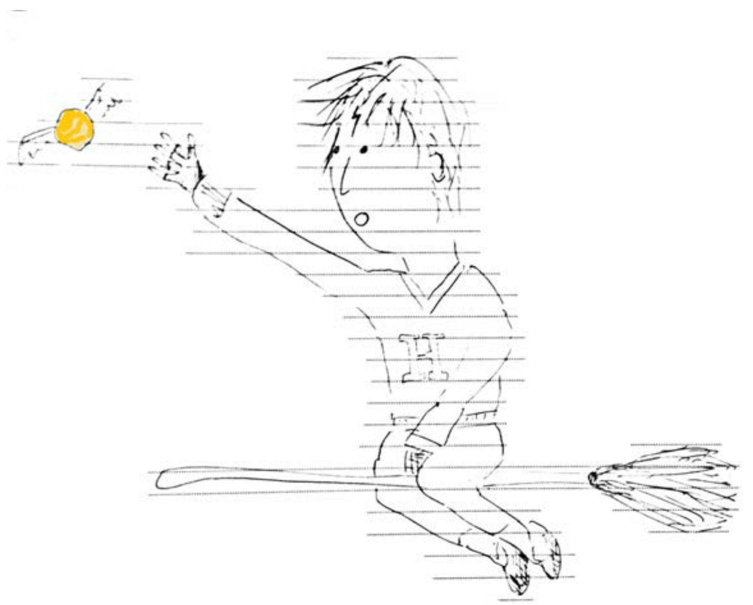
$$\max RY^2$$
$$s.t. \quad R + pY = I$$

Solving the maximization problem, we get that Walter spends $I/3$ tweakers on R to consume $I/3$ grams of R. He spends $2I/3$ tweakers on Y, to consume $2I/(3p)$ grams of Y.

6. Assume now that $p = 2$. How much time does international trade save Walter? Calculate the number of hours he would have worked (without trade) in order to have utility equal to the optimal utility under trade. **(5 points)**

At p=2, Walter's income is equal to 6.32. He consumes 2.106 R and 2.106 Y, for total utility of 9.34. We want to find the amount of time Walter has to work in order to match this level of utility. If he works K hours, his maximization problem would be $\max RY^2$s.t. $R^2 + 4Y^2 = K$, and he would consume $R = \sqrt{K/3}$ and $Y = \sqrt{K/6}$. This gives him utility $K\sqrt{K/3}/6$. The level of K that makes his utility equal to 9.34 is 21.12.

2 Long Question: Quidditch and Congestion [30 points]



Oliver Wood organizes quidditch practices each week, but ever since Harry Potter became famous, attendance has been too high. On Tuesday evenings, each student at Hogwarts decides whether to go to quidditch practice, getting utility $u(N) = h + \quad N - \theta N^2$ where N is the number of people at the practice and $\quad, \theta > 0$. Students who don't attend quidditch practice ride unicorns in the Forbidden Forest instead, getting fixed utility $\bar{F}$ (which does not depend on how many people choose that option). There are 100 students in total. Assume throughout that $h + \frac{\quad^2}{4\theta} > \bar{F} > h$. [Note: you should ignore discreteness throughout this question.]

1. What exactly is the externality in this context? Is it positive, negative, or ambiguous? [Hint: sketch out the $u(N)$ function.] **(5 points)**

   *[**Solution:** the $u(N)$ function is strictly concave – initially increasing in N and then decreasing. On the increasing section, each individual has a positive externality on others when they decide to play quidditch – by going to the practice, they also raise the utility of others that go. But on the decreasing section, each individual has a negative externality on others. Intuitively we can think of these two effects as being because quidditch players want (a) enough players to form two teams, (b) not so many players that the pitch is over-crowded, or that people have to be left out.]*

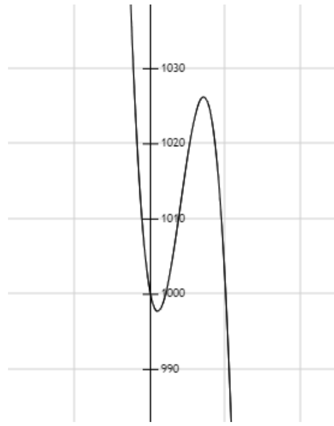
2. If there is no cost of attending quidditch practice, how many will attend quidditch practice? [Hint: there are three equilibria – explain each, and use a graph if it helps. Assume that individuals are 'small' in the sense that if they move from hide-and-seek to quidditch, they don't take into account their effect on N and thus on $u(N)$ (this is similar to assuming that firms are price-takers).] **(7 points)**

   *[**Solution:** there are three equilibria. In each equilibrium no person wants to switch to the other activity.*

   Equilibrium 1: all 100 students play hide-and-seek in the Forbidden Forest, getting fixed utility $\bar{F}$. This is an equilibrium since none want to switch to quidditch as $u(N=0) = h < \bar{F}$.

   Equilibria 2 and 3a: these are characterized by all students being indifferent between both activities, which is the case when

$$\bar{F} = h + \quad N_e - \theta N_e^2 \Rightarrow \theta N_e^2 - \quad N_e + \left(\bar{F} - h\right) = 0$$



This equation has two solutions for N_e, one on the upward-sloping part of $u(N)$ (low N_e) and one on the downward-sloping part (high N_e). We can use the quadratic formula to solve for these two solutions, i.e. we have that

$$N_e = \frac{\quad \pm \sqrt{\ ^2 - 4\theta\left(\bar{F} - h\right)}}{2\theta}$$

Note that we could have $N_e > 100$, in which case we just have an corner solution equilibrium with all students playing quidditch. But with the parameter values from part 4. this is never the case – we get $N_e = 10$ or 2.

Equilibrium 3b: technically if $u(100) > \bar{F}$ equilibrium 3a doesn't exist since the $u(N)$ function only crosses $\bar{F}$ once. Instead our third equilibrium has all 100 students playing quidditch, each getting utility $u(100)$, and no-one wants to switch. I ignore this corner solution equilibrium from now on.]

3. What N maximizes the combined welfare of Hogwarts students? Call this N^* **(7 points)**.

   *[**Solution:** we choose N to maximize $W(N) = (100 - N)\bar{F} + Nu(N)$ which is*

   $$(100 - N)\bar{F} + N\left(h + \quad N - \theta N^2\right)$$

   which is not strictly concave, but its maximum given $N > 0$ has $W'(N) = 0$ (we can see this by graphing the function – see below). So we set the first derivative equal to zero:

   $$W'(N) = h - \bar{F} + 2 \quad N - 3\theta N^2 = 0$$

   $$\Rightarrow 3\theta N^2 - 2 \quad N + \left(\bar{F} - h\right) = 0$$

   so again using the quadratic formula we get two possible solutions:

   $$N^* = \frac{2 \quad \pm \sqrt{4 \ ^2 - 12\theta\left(\bar{F} - h\right)}}{6\theta}$$

   The graphical example below uses $\bar{F} = 10$, $h = 5$, $\quad = 3$ and $\theta = \frac{1}{4}$ (we can see that these parameters satisfy the assumptions in the question that $h + \frac{^2}{4\theta} > \bar{F} > h$).

   We can see from the graph that there are two turning points with $W'(N)$, but it is the larger N which maximises welfare, so we choose

   $$N^* = \frac{2 \quad + \sqrt{4 \ ^2 - 12\theta\left(\bar{F} - h\right)}}{6\theta}$$

   And technically if $N^* > 100$, the welfare-maximising $N = 100$ since this is the upper bound on the number of students. However, with the parameter values below this will not be the case.]

4. Suppose we start off in an equilibrium with $N > N^*$ (famous Harry Potter is leading to Quidditch congestion). What price can Oliver charge to dissuade enough people that only N^* attend in equilibrium? Assume utility is linear in the price, i.e. $u(N) = h + \quad N - \theta N^2 - p$, and use that $\bar{F} = 10$, $h = 5$, $\quad = 3$ and $\theta = \frac{1}{4}$. Can any other price do better (i.e.

be a further Pareto improvement)? **(6 points)**

*[**Solution:** needs to be the case that* $\bar{F} = h + \quad N^* - \theta (N^*)^2 - p$ *so follows that*

$$\begin{aligned} p &= h - \bar{F} + \quad N^* - \theta (N^*)^2 \\ &= h - \bar{F} + \quad \left(\frac{2 \quad + \sqrt{4 \quad ^2 - 12\theta \left(\bar{F} - h\right)}}{6\theta} \right) - \theta \left(\frac{2 \quad + \sqrt{4 \quad ^2 - 12\theta \left(\bar{F} - h\right)}}{6\theta} \right)^2 \end{aligned}$$

given the parameter values we have $N^* \approx 7.1$*. It follows that*

$$p \approx -5 + 3\,(7.1) - \frac{1}{4}\,(7.1)^2 = 3.6975$$

The price is increasing in Harry Potter's fame h*, and falling in utility from hide-and-seek* $\bar{F}$*. Note that there are still three equilibria – but the far right equilibria has* N^**. If instead we chose a price to maximise the utility of those playing quidditch we would be down to only two equilibria – this is because doing so would set the* $u(N)$ *curve tangent to* $\bar{F}$ *at its max, whereas otherwise it crosses* $\bar{F}$ *twice.*

No other price can do better – this is not only a Pareto improvement, it is welfare-maximising (since it ensures N^* *play quidditch, which we know maximises welfare).]*

5. Suppose the revenue from charging a price is distributed back to students lump sum (i.e. each student receives $\frac{pN^*}{100}$ – again assume that this enters linearly into utility). Does Wood's pricing policy lead to a Pareto improvement (recall that we started off in the equilibrium with $N > N^*$)? **(5 points)**

   *[**Solution:** yes – in the initial equilibrium all students had utility* $\bar{F}$*. In the new one,* $100 - N^*$ *students get* $\bar{F} + \frac{pN^*}{100}$ *and the* N^* *students going to quidditch practice get the same. Since no one has been made worse off, and at least one person has been made better off, this is a Pareto improvement.]*

3 Causal Inference using Instrumental Variables

3.1 Theory

A key objective in empirical analysis is to credibly estimate the causal effect of a variable of interest X on an outcome Y. In this problem set, we will apply an important technique discussed in lecture for estimating causal relationships in (typically) non-experimental settings. That tool is instrumental variables (IV, also known as Two Stage Least Squares, or 2SLS). We will use instrumental variables here to study the classic economic question of whether and by how much schooling raises earnings. Concretely, we will measure the causal effect (in percentage terms) of an additional year of education on workers' earnings.

Let Y_i represent log(earnings) of individual i. To simplify matters, assume that schooling is binary (people either go to school, or not go to school). Let S_i represent whether individual i went to school or not. That is, $S_i \in \{0, 1\}$ with $S_i = 1$ indicating that i went to school and $S_i = 0$ indicating that i did not go to school.

1. Using the notation developed throughout the semester, write down the average treatment effect on the treated. Call this T^*

   This is simply

   $$T^* = E\left[(Y_{1i} - Y_{0i})\,|S_i = 1\right]$$

   where Y_{1i} *is the potential log(earnings) of individual* i *if they go to school and* Y_{0i} *is the potential log(earnings) of the same individual if they do not go to school.*

2. Suppose you estimated

   $$E\left[Y_i|S_i = 1\right] - E\left[Y_i|S_i = 0\right]$$

   Explain, along with mathematical notation, that this estimate is unlikely to be causal.

The usual algebra (which should be quite familiar to all of you by now)

$$\begin{aligned}
E[Y_i|S_i=1] - &E[Y_i|S_i=0] \\
&= E[Y_{1i}|S_i=1] - E[Y_{0i}|S_i=0] \\
&= E[Y_{1i}|S_i=1] - E[Y_{0i}|S_i=1] \\
&\quad + E[Y_{0i}|S_i=1] - E[Y_{0i}|S_i=0] \\
&= T^* + \underbrace{E[Y_{0i}|S_i=1] - E[Y_{0i}|S_i=0]}_{\textit{Selection bias}}
\end{aligned}$$

The selection bias in this comparison is unlikely to be zero because individuals who choose to go to school are different from individuals who do not go to school. For example, rich people might like schooling and go to school and they would have earned more even without schooling (by, for example, working in their ancestral firm).

Suppose you have an "instrument" Z_i that is randomly assigned but affects schooling S_i.

3. You estimate
$$FS = E[S_i|Z_i=1] - E[S_i|Z_i=0]$$
Is this estimate a causal estimate of the instrument on schooling S_i? Explain your answer using formal notation.

   Yes, this estimate is the causal estimate of the instrument on schooling. Doing the same decomposition, you will find that
$$\begin{aligned}
FS =& E[(S_{1i} - S_{0i})|Z_i=1] \\
&+ E[S_{0i}|Z_i=1] - E[S_{0i}|Z_i=0]
\end{aligned}$$
   Given that the instrument Z_i was random, $E[S_{0i}|Z_i=1] = E[S_{0i}|Z_i=0]$ and the selection bias disappears.

4. Then you estimate
$$RF = E[Y_i|Z_i=1] - E[Y_i|Z_i=0]$$
Is this estimate a causal estimate of the instrument on earnings? Explain your answer using formal notation.

   Yes, this estimate is the causal estimate of the instrument on earnings. Doing the same decomposition, you will find that
$$\begin{aligned}
RF =& E[(Y_{1i} - Y_{0i})|Z_i=1] \\
&+ E[Y_{0i}|Z_i=1] - E[Y_{0i}|Z_i=0]
\end{aligned}$$
   Given that the instrument Z_i was random, $E[Y_{0i}|Z_i=1] = E[Y_{0i}|Z_i=0]$ and the selection bias disappears.

5. Then you use the estimates in parts 3 and 4 to compute the following
$$\frac{RF}{FS} = \frac{E[Y_i|Z_i=1] - E[Y_i|Z_i=0]}{E[S_i|Z_i=1] - E[S_i|Z_i=0]}$$
   (a) What is this estimate called?
   This is the Wald Estimate.
   (b) Can you interpret this as an unbiased estimate of T^*? What *additional* assumption(s), if any, do you have to make in order to do so?
   Yes, we can interpret this as an unbiased estimate of T^* under the following additional assumptions
      i. *(First stage). The instrument does affect schooling. That is:* $E[S_i|Z_i=1] \neq E[S_i|Z_i=0]$
      ii. *(Exclusion restriction). The instrument does not affect schooling through any channel other than schooling.*

3.2 Empirical

1. Load **census70.dta** into Stata. You will notice that all the relevant variables have been labeled. Regress log(weekly wages) on completed education.

   (a) Is the coefficient statistically different from zero? How do you know?
   (b) Interpret the coefficient on education.
   The regression result is as follows:

```
. regress LWKLYWGE EDUC

      Source |       SS       df       MS              Number of obs =  247199
-------------+------------------------------           F(  1,247197) =50948.11
       Model |  17917.6603     1  17917.6603           Prob > F      =  0.0000
    Residual |  86935.3595247197  .351684525           R-squared     =  0.1709
-------------+------------------------------           Adj R-squared =  0.1709
       Total |   104853.02247198  .424166133           Root MSE      =  .59303

------------------------------------------------------------------------------
    LWKLYWGE |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        EDUC |   .0801112   .0003549   225.72   0.000     .0794156    .0808068
       _cons |    4.23443     .00425   996.33   0.000       4.2261     4.24276
------------------------------------------------------------------------------
```

The coefficient is 0.08 and is statistically different from zero. This is because the t-statistic (225.72) is greater than 1.96 (the critical value) and the p-value is 0. We can reject the null hypothesis of no effect at 95% significance.

An increase of education by 1 year is associated with an increase in weekly wages of 8 percent. To see this, note that

$$\log(y) = S + \epsilon$$

$$= \frac{\partial \log(y)}{\partial S} = \frac{1}{y}\frac{dy}{ds}$$

Discretizing the derivative

$$= \frac{\left(\frac{\Delta y}{y}\right)}{\Delta s} = \frac{\Delta y}{y}$$

when $\Delta s = 1$

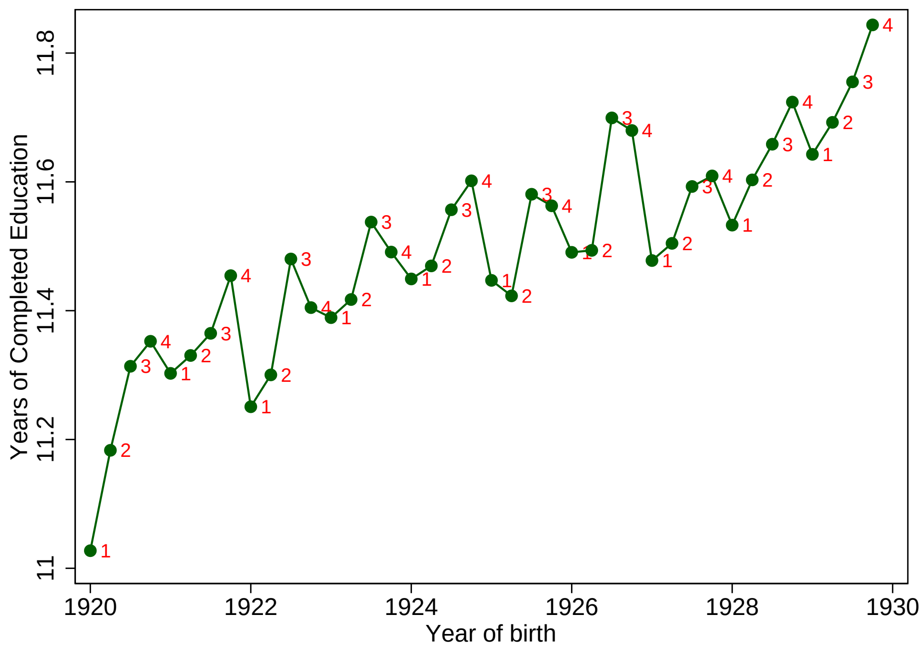
Joshua Angrist and Alan Krueger noticed that in the US children born in different months of the year start schooling at different ages. This is because school districts require a student to have turned age 6 by January 1 of the year in which he or she enters the school. Compulsory schooling laws require children to remain in school until their 16th (or 17th, depending on the state of residence) birthday. The interaction between the school entry requirement and compulsory schooling laws compel children born in certain months to attend school longer than students born in other months.

For example, consider two children: child A born on January 2, 1920 and child B born on December 2, 1920. Both children have to start school in September 1926 because on January 1, 1927, both A and B will be 6 years old. Note that A is 11 months older than B while starting school. Child A can quit school in January 2, 1936 when he turns 16 and would be in the middle of his 10th year at school. But child B cannot quit school until December 2, 1936 when he turns 16 and would be in the middle of his 11th year at school.

2. Let's check whether this pattern holds true in the data. Load **census70.dta** into Stata. Compute average education by each year -quarter of birth. You can use the command `collapse` to do so. Make sure that you have both `YOB` and `QOB` inside the `by()` option. Now generate a single variable which will have both year and quarter of birth information. Plot average education against this variable. Comment on the shape of the plot. For which quarters are completed education the lowest?

   [Hint: the new variable could be `YOB + (QOB-1)/4`]

   You will notice that individuals born in quarter 1 tend to have the lowest educational attainment. The trend by year of birth is increasing, suggesting that recent-borns have higher level of education in general. But within each year of birth cohort, those born in the first quarter tend to have the lowest level of education and those born in 3rd or 4th quarter tend to have the highest level of education.



3. Load the **census70.dta** again (as your dataset has changed after the `collapse` command). Find the average education and average log weekly wages for each quarter of birth.

   This tabulation just confirms the result from the graph above. This shows that individuals born in the first quarter have the lowest level of education and as well as wages whereas individuals born on the third and fourth quarter have higher level of education as well as wages

```
. table QOB, c(mean EDUC mean LWKLYWGE)
```

Quarter of birth	mean(EDUC)	mean(LWKLYWGE)
1	11.3996	5.148471
2	11.4434	5.152764
3	11.556	5.161561
4	11.5754	5.157816

4. Generate a variable `first` that is 1 if the individual is born in the first quarter and 0 otherwise.

   (a) What is the average education attained by those born in the first quarter,

   (b) by those born in other quarters?

   The average education of those born in the first quarter is 11.53 years. The average for those born in other quarters is 11.40 years.

```
. table first, c(mean EDUC)
```

first	mean(EDUC)
0	11.5252
1	11.3996

(c) What is the difference?

(d) What is the t-statistic of the difference? Is this difference statistically significant?
The difference is -0.126 years. You can do all of this through a regression. The results are shown below.

```
. reg EDUC first

      Source |       SS       df       MS              Number of obs =  247199
-------------+------------------------------           F(  1,247197) =   65.29
       Model |  737.149181     1  737.149181           Prob > F      =  0.0000
    Residual |  2791131.65247197  11.2911227           R-squared     =  0.0003
-------------+------------------------------           Adj R-squared =  0.0003
       Total |   2791868.8247198   11.294059           Root MSE      =  3.3602

------------------------------------------------------------------------------
        EDUC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       first |  -.1255553   .0155391    -8.08   0.000    -.1560115   -.0950991
       _cons |   11.52515   .0078214  1473.53   0.000     11.50982    11.54048
------------------------------------------------------------------------------
```

As seen in the table, the difference has a standard error of 0.016 and a t-statistic of 8.08. The associated p-value is zero. The difference is statistically significant at the 95% (over even higher) significance test.

(e) Assuming that QOB is random, which estimate from section 3.1 does this correspond to?
Assuming that QOB is random, this is the FS estimate from section 3.1 as it examines the effect of a randomly assigned instrument on schooling S.

5. Repeat the same exercise as in part 4 with log(weekly wages).

(a) What is the average log(wages) for those born in first quarter?

(b) for those born in other quarters?
As the table below shows, the mean log(weekly wages) for those born in the first quarter is 5.16 and for those born in other quarters is 5.14

```
. table first, c(mean LWKLYWGE)

----------------------
    first | mean(LWKLYWGE)
----------+-----------
        0 |    5.15745
        1 |   5.148471
----------------------
```

(c) What is the difference?

(d) Is this difference statistically significant?
The difference is 0.00898 (or 0.90 percent) and is statistically significant at 95% significance level. As before, you can find this out from a regression. The t-statistic is 2.98 and the associated p-value is 0.003.

```
. reg LWKLYWGE first

      Source |       SS       df       MS              Number of obs =  247199
-------------+------------------------------           F(  1,247197) =    8.89
       Model |  3.76989396     1  3.76989396           Prob > F      =  0.0029
    Residual |   104849.25247197  .424152599           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |   104853.02247198  .424166133           Root MSE      =  .65127

------------------------------------------------------------------------------
    LWKLYWGE |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       first |  -.0089789   .0030117    -2.98   0.003    -.0148818   -.0030759
       _cons |    5.15745   .0015159  3402.17   0.000     5.154479    5.160421
------------------------------------------------------------------------------
```

(e) Assuming that QOB is random, which estimate from section 3.1 does this correspond to?

Assuming that QOB is random, this is the RF estimate from section 3.1 as it examines the effect of a randomly assigned instrument on the outcome Y.

6. Now, compute the Wald estimator. Is the Wald-estimator statistically significant? [Hint: You will need to use Stata command `ivregress 2sls` to compute the standard errors and t-statistics]

   The Wald estimator, as seen in part 5, is the ratio of the RF to the first stage FS. You can see directly that it is 0.0715. To test whether this estimator is significant, we need to do two-staged-least squares (2SLS). The result of this estimator is shown below.

```
. * Part 6
. ivregress 2sls LWKLYWGE (EDUC=first)

Instrumental variables (2SLS) regression             Number of obs =  247199
                                                     Wald chi2(1)  =   10.69
                                                     Prob > chi2   =  0.0011
                                                     R-squared     =  0.1689
                                                     Root MSE      =  .59373
```

LWKLYWGE	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
EDUC	.0715133	.0218682	3.27	0.001	.0286525	.1143741
_cons	4.333248	.251341	17.24	0.000	3.840629	4.825867

You will notice that the estimate is exactly the ratio of the RF to the FS. The standard error is 0.0219 and the t-statistic is 3.27. The associated p-value is 0. This estimate is statistically different from zero at 95% (or even higher) significance level.

7. What are the assumptions needed to interpret the Wald estimator as the causal effect of schooling on wages. Are all the assumptions satisfied? Explain.

   The assumptions are:

   - *There is a first stage. This assumption is satisfied as we checked in part 4*
   - *Exclusion condition. This means that the quarter of birth of an individual does not affect schooling by any channels other than schooling. This assumption will be satisfied if parents do not systematically time the quarter of birth of their children. This assumption will fail if rich parents like to have summer/autumn babies (It probably makes sense if you are a professor and want to minimize the loss of working days?). Then these children would have had more years of schooling simply because their parents are richer and not just because of the compulsory education regulation.*

8. Suppose that your assumptions in 7 are satisfied. Interpret the coefficient given by the Wald-estimator.

   If all the assumption in part 7 are satisfied, then an increase in one year of schooling ***causes*** *weekly wages to go up by 7.15 percent.*

9. Propose an economic hypothesis on why your estimate from 1 is different from that in 6

   One potential reason could be that rich parents want their children to go to school, which causes them to have more schooling. On the other hand, rich parents have better connections and can find better jobs for their children when they grow up. This will lead them to have higher wages. This kind of selection bias will overestimate the effect. Mathematically,

$$E\left[Y_{0i}|S_i = 1\right] > E\left[Y_{0i}|S_i = 0\right]$$

   which makes the selection bias positive leading the comparison in part 1 to overestimate the causal relationship estimated in part 6.

Example do-file

```
/*
        Do file for 14.03 Problem Set 4
        Written by: Maheshwor Shrestha
        Modified: 10/23/2014
*/

*Boilerplates
clear
set scheme s1color

cd "C:\Users\mahesh\Dropbox (MIT)\14.03-Fall-2014\p-sets\ps4"

use census70, clear


**Part 1, simple regression
regress LWKLYWGE EDUC

**Part 2
** Compute the mean education by yob-qob
collapse (mean) EDUC, by(YOB QOB)
* generate year is quarters
gen yq = YOB + (QOB-1)/4
label var EDUC "Years of Completed Education"
label var yq "Year of birth"
* produce the scatterplot
sc EDUC yq, ml(QOB) c(l) mlabsize(small) mlabcolor(red)
graph export figpart2.pdf, replace

** reload the data as it has changed after the collapse command.
use census70, clear

*Part 3
table QOB, c(mean EDUC mean LWKLYWGE)

** generate an indicator variable for those born in first quarter
gen first = QOB==1
* Part 4 a/b
table first, c(mean EDUC)
* Part 4 c/d - do the test through a regression.
** could have done a t-test as well.
reg EDUC first

* Part 5
table first, c(mean LWKLYWGE)
* do the test through a regression. Could do a t-test as well.
reg LWKLYWGE first

* Part 6
* Wald estimate. Do 2SLS to compute standard errors correctly.
ivregress 2sls LWKLYWGE (EDUC=first)
```