Section #2

Outline

1. Econometrics
   (a) Simple (Univariate) regressions
   (b) Multivariate regressions
   (c) Time trends
   (d) Dummy/indicator variables
   (e) Fixed effects

1 Reduced Form Regressions

1.1 Simple (univariate) regressions

One way to think about regressions is as a way to summarize data. Under certain conditions, we can also interpret regressors as having a causal impact on the dependent variable. In the following section, I use data from Bailey's (2006) paper "More Power to the Pill" which we will read later in this course. The data she uses come from supplements to the Current Population Survey (CPS) administered in various years from 1979 to 1995. I will make her data that I use here and my Stata files available on my website so that you can play with the data yourself.

Let's start by considering a simple scatter plot that shows the average number of children ever born to a woman, by age. Each point in the scatter plot below is an age $\times$ survey-year observation.

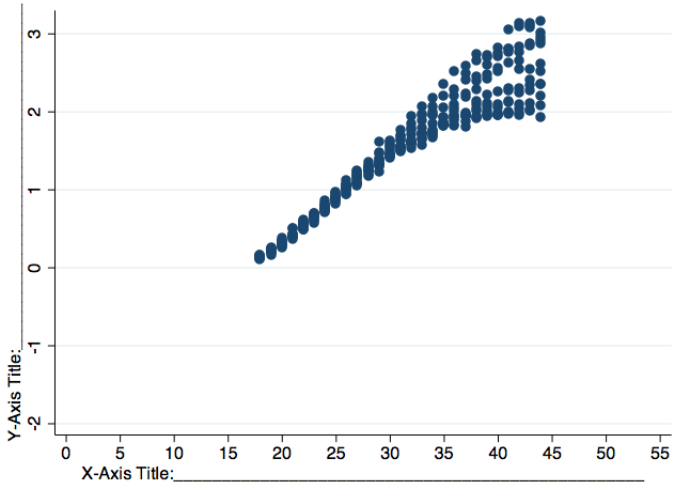

Figure 1: _____

1. First, it's important to understand what is being plotted. Give the figure a title and add titles to the axes in the indicated space.

2. Based on simply eyeing the data, draw a linear line that best fits the data. Approximately, what is the intercept of this line? Approximately, what is the slope of this line? Recall the equation for a linear line is

$$y = \alpha + \beta \cdot x$$

where α is the intercept, β is the slope, and y and x are the variables represented on the y- and x- axes, respectively.

3. Similarly, draw a quadratic line that best fits the data. In this case, the equation for a quadratic line will be given by

$$y = \alpha + \beta_1 \cdot x + \beta_2 \cdot x^2$$

4a. Compare two women: one who is 25 and one who is 40. They are alike in all other regards. Using the linear and quadratic functions you drew on the plot above, what is the **expected**, also known as **predicted value**, for the number of children born. You don't need to do any math here, simply follow the line or curve and look at the values on the x- and y- axes.

4b. Next, let's use our models to make an "out of sample" prediction. Assume there is a third woman aged 55. Again, compare predictions from your linear and quadratic models. Which model—the linear model or the quadratic model—do you think is a better fit for the data, and why?

5. How does what you did above compare to a simple linear regression that regresses the average number of children ever born on the age of the mother? A regression is simply a more formal way to find the best fitting line (or curve) that minimizes deviations from the actual data, known as **errors** or **residuals**. Consider the regression output below. How does the constant term and the slope in the linear model given in column 1 compare with the equation for the linear line you drew above?

Table 1: Simple regression

	(1) No. Children	(2) No. Children
Age	0.0982***	0.185***
	(0.00157)	(0.0132)
Age^2		-0.00140***
		(0.000212)
Constant	-1.565***	-2.829***
	(0.0501)	(0.196)
Observations	324	324
R^2	0.924	0.933

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6. The R^2 (**R-squared**) of a model tells you the percentage of the variation in the dependent variable that is explained by your model; thus the R^2 is always between 0 and 100%. Technically, adding any additional regressors will improve the model fit. The R^2 that is reported is technically an "Adjusted R^2", which penalizes regressions for adding in too many variables that do not add anything significant to the model fit. Which R^2 is higher? How does this compare do your conclusion in 4.b. about which model better fit the data?

1.2 Multivariate regressions

7. Now, suppose we think there might be a linear time trend. That is, we might posit that women are having fewer babies over time. Now we have two independent variables: age and time. We can draw this in three dimensions as shown below:

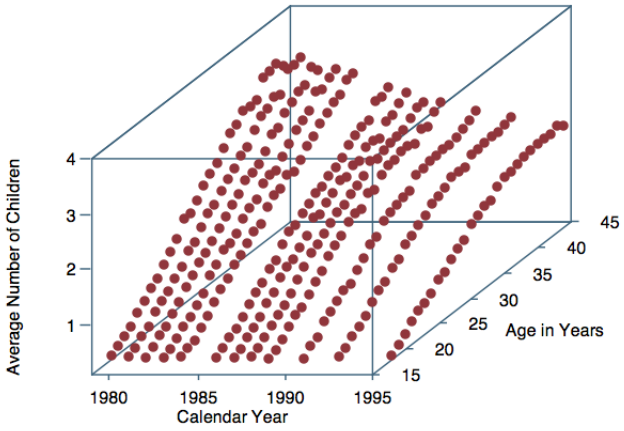

Figure 2: 3D Scatter

Can you see in the 3D scatter what is happening to the average number of children born over time at all ages?

1.2.1 Time trends

We can easily add a linear time trend, ***t***, to our regression model that takes on the value 0 in the first year (1979), and increases by 1 with every year thereafter.

1.2.2 Dummy/indicator Variables

What if we also think that marriage is an important variable affecting childbirth that we also want to consider in our regression. Instead of collapsing the data to the age $\times$ year level as I did above, let's work with the full microdata now.

Let's code a variable ***Married*** that takes on the value of 1 if a woman is married, and 0 otherwise. Note that this variable take on two discrete levels. We call variables that take on {0,1} values **indicator** or **dummy variables**. In addition to our marriage variable, we can also code a variable ***Single*** that takes on a value of 1 if a woman is single and 0 otherwise.

Adding more variables to the right-hand side of the regression makes it impossible to draw since we can't draw in more than 3 dimensions, but the interpretation of the regression output remains the same as it did above. Regressions controlling for time with a linear trend and marital status are shown below.

Interpret these new coefficients.

Use Column 3 of Table 2 to compare two individuals: one who is unmarried in 1980 and one who is married in 1990. They are alike in all other regards. How many fewer children do women in 1990 have, on average?

Now do the same thing using Column 4. How do the results differ?

What would happen if I put both the ***married*** and ***single*** dummy variables in the same regression? (Hint: this is known as the "Dummy Variable Trap")

Table 2: Simple regression

	(1) No. Children	(2) No. Children	(3) No. Children	(4) No. Children
Age	0.196***	0.197***	0.221***	0.219***
	(0.00237)	(0.00237)	(0.00232)	(0.00232)
Age^2	-0.00155***	-0.00156***	-0.00235***	-0.00231***
	(0.0000383)	(0.0000383)	(0.0000370)	(0.0000370)
t	-0.0280***	-0.0534***	-0.0959***	-0.0138***
	(0.000434)	(0.00145)	(0.00414)	(0.00402)
t^2		0.00168***	0.00545***	-0.000804***
		(0.0000915)	(0.000224)	(0.000209)
Married			0.521***	
			(0.0129)	
single				-0.728***
				(0.0154)
Constant	-2.820***	-2.778***	-3.011***	-2.673***
	(0.0351)	(0.0351)	(0.0373)	(0.0376)
Observations	382957	382957	382923	382923
R^2	0.268	0.268	0.196	0.197

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1.2.3 Fixed effects

When we turn a continuous variable (or more accurately, discrete variables that take on many values) into separate dummy variables, we call these "fixed effects." You can interpret these as separate slopes for each value the variable takes on. Remember when we tried to impose a linear or quadratic model above? Well, fixed effects allow for a more flexible specification that doesn't require us to make a call about the underlying structural relationship. Consider the following regression with age and year fixed effects (what happens if we also included AgeXYear Fixed Effects?):

Table 3: Fixed Effects regressions

	(1) No. Children	(2) No. Children
Age	0.197***	
Age^2	-0.00156***	
year=1980	-0.0414***	-0.0413***
year=1981	-0.0755***	-0.0758***
year=1982	-0.135***	-0.135***
year=1983	-0.163***	-0.163***
year=1985	-0.274***	-0.274***
year=1986	-0.291***	-0.292***
year=1987	-0.309***	-0.309***
year=1988	-0.341***	-0.341***
year=1990	-0.360***	-0.361***
year=1992	-0.388***	-0.389***
year=1995	-0.423***	-0.424***
19		0.0783***
20		0.178***
21		0.286***
22		0.403***
23		0.522***
24		0.650***
25		0.777***
26		0.913***
27		1.031***
28		1.156***
29		1.277***
30		1.403***
31		1.488***
32		1.592***
33		1.682***
34		1.775***
35		1.853***
36		1.953***
37		2.017***
38		2.116***
39		2.163***
40		2.209***
41		2.305***
42		2.367***
43		2.447***
44		2.472***
Constant	-2.788***	0.338***
Observations	382957	382957
R^2	0.268	0.269

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Let's plot the age and year fixed effects and see what they look like:

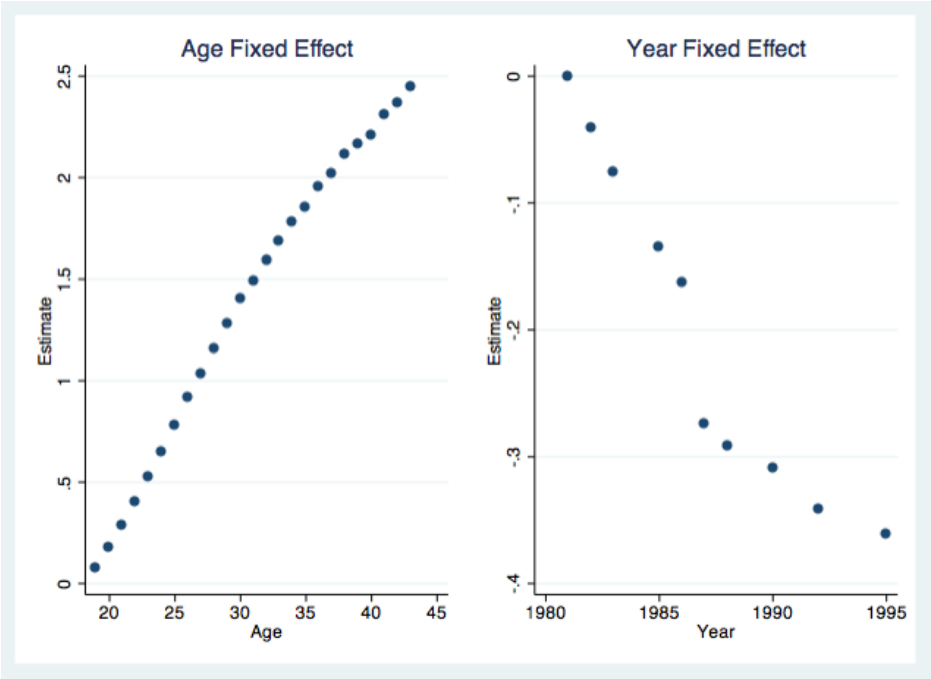

Figure 3: Estimated Age and Year Fixed Effects

Interpret these fixed effects.

Note that, often times, the fixed effects are not explicitly reported in a regression table because they are not the variable of interest, because there are so many of them and/or because they may not be meaningful in some cases. Instead, authors often note in the regression table which fixed effects are being used.

1.3 Group exercise with a more grown-up regression!

Now that we've gone over the basics, let's get together in groups of 3-4 to see if we can help each other interpret some more advanced regression output.

Consider the following regression, which replicates Table III in Bailey (2006). [1]

Check out the number of observations. The unit of observation is at the individual level. The dependent variable takes on a value of 1 if a woman has had a child before the indicated age, and is 0 otherwise. ELA stands for "Early Legal Access", and indicates whether the state that the women lives in had ELA to the pill or abortion, as indicated. Bailey uses this table to try and answer whether ELA had a causal effect on the probability of childbirth.

[1] Note: Results differ from the published paper. I use a linear probability model (LPM) while Bailey uses a Probit model. I maintain the original sample and weights to make the results more comparable to the published paper. While her original data and code are unavailable, Bailey has noted likely errors with her initial sample construction and issued an addendum with updated results that can be found on her website.

Table 4: Replication of Table III in Bailey (2006)

	(1)	(2)	(3)	(4)	(5)
	Age 22	Age 22	Age 22	Age 19	Age 36
ELA to pill	-0.0829[b]	-0.0855[b]	-0.0683[c]	-0.0141	-0.00219
	(0.0329)	(0.0339)	(0.0363)	(0.0407)	(0.00810)
ELA to abortion			0.0247	-0.00366	0.00658
			(0.0466)	(0.0341)	(0.0210)
ELA to pill X ELA to abortion			-0.0862	-0.0945	-0.0201[c]
			(0.0526)	(0.0651)	(0.0117)
Constant	0.620[a]	0.737[a]	0.738[a]	0.253[a]	0.994[a]
	(0.0370)	(0.0301)	(0.0301)	(0.0386)	(0.00640)
Observations	106527	106527	106527	106527	106527
R^2	0.029	0.038	0.038	0.034	0.018
State FE	X	X	X	X	X
Year of Birth FE	X	X	X	X	X
State X Year of Birth FE		X	X	X	X

Standard errors in parentheses

Standard errors clustered at the state level.

ELA = Early Legal Access.

Dependent variable is 1 if the first birth is before the indicated age, and zero otherwise.

[c] $p < 0.1$, [b] $p < 0.05$, [a] $p < 0.01$

1. Go through the table and use asterisks to indicate which variables are statistically significant at the 90 % confidence level, 95% confidence level, and at the 99% confidence level.

2. Does the constant have any economic meaning here? What else would we need to know in order to construct predicted values?

3. Interpret the coefficient in column 1 (think hard about the units). How does it differ from the interpretation of the coefficient in column 2?

4. California and New York had ELA to the pill and abortion by 1972. Putting aside differences across states and state time trends captured by the fixed effects, what is the difference in the effect on childbirth by age 22 attributed to these ELA laws, compared with a state like Missouri, which was a laggard in ELA to the pill and abortion?

5. Using the regression results, what can we say about the effect of early access of the pill and abortion on women who are 19 years old?

6. The effect on women's birth rate by age 36 is not statistically significant. We also like to consider **practical significance** (also called **economic significance**): Given our estimated coefficient on this sample of woman, ask yourself how important or useful the explanatory variable is in determining the dependent variable. In other words, do you think the result on "ELA to pill" would be very economically significant, even if it were statistically significant? (Hint: construct a confidence interval around the coefficient). If you are not sure, pretend that the estimate and standard error were each 100 times smaller.

7. It looks like the pill had a large and statistically-significant effect on women aged 22. Based on the size of the results in column 5, the regression suggests that there was no real difference on the probability of having children by the time women reached middle age, after controlling for state fixed effects, year of birth fixed effects, and trends within states over time. Can you come up with a story that would be consistent with this evidence?

8. If you were to test differences in birth rates of women before and after access to the pill and abortion, what two additional variables would you want to include? What effect would you expect each of those variables to have?