

PSC 101 Multiple Regression examples

Heart attack data:

We have three measures of interest here. Our dependent variable “attack” measures the level of indicators for heart attacks. The higher the values, the greater the risk of heart attack. Half of our sample gets an experimental drug treatment, while half does not. If we look at the data, we see that the people receiving the treatment in fact do worse on the measure of interest (remember, higher values are worse).

```
. tab treat, summ(attack)
```

Summary of Heart attack indicators			
treatment	Mean	Std. Dev.	Freq.
0	.44812433	.1599955	250
X	.50499594	.14360768	250
Total	.47656013	.1545146	500

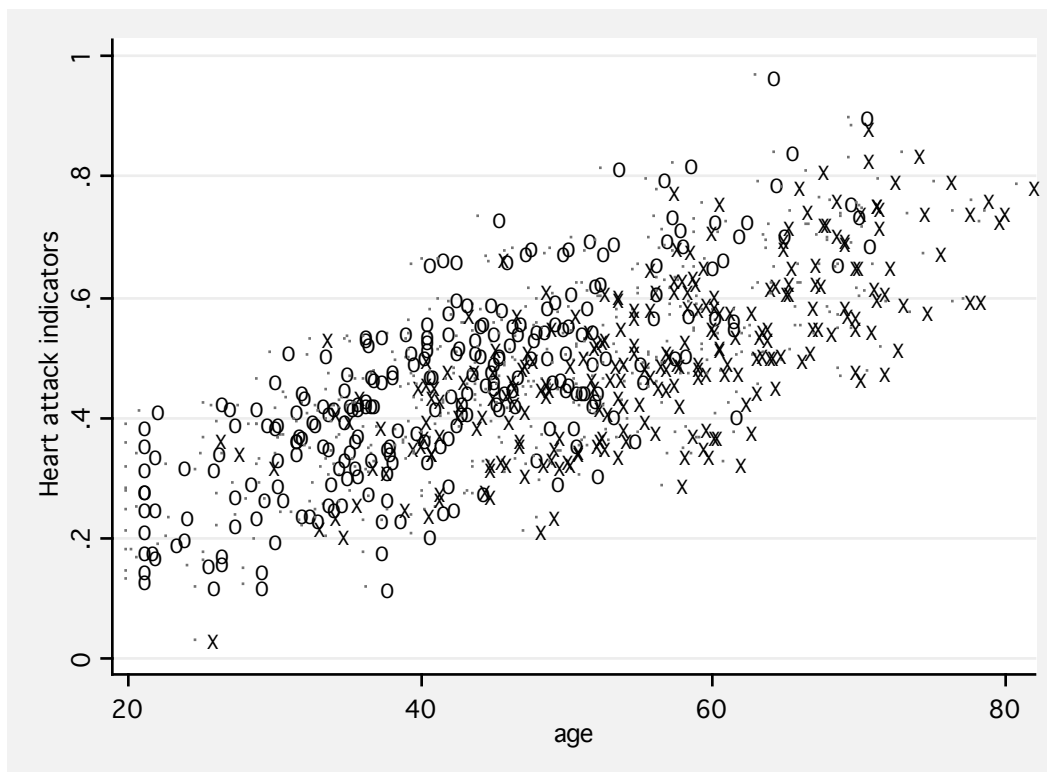
Maybe the drug makes people worse off. There is not random assignment, however, so we can't assume that the treatment and control groups are equal in every way prior to receiving the treatment. One thing in particular to examine is the age of the two groups. When we do that we see:

```
. tab treat, summ(age)
```

Summary of age			
treatment	Mean	Std. Dev.	Freq.
0	40.375356	11.333267	250
X	54.832344	11.394994	250
Total	47.60385	13.462594	500

Aha! The group getting the treatment is much older on average than the group getting no treatment. Now we have two candidate independent variables—whether a subject received the treatment, and the subject's age. What we would like to do is to compare how *subjects of similar ages* fare under treatment and control. One way to see that is with a scatter plot. Here, we mark the points by treatment (X) and no treatment (0).

```
twoway scatter attack age, msym(point) mlabel(treat)
```



What should you be looking for if you are trying to see if there is a pattern in the data? If the treatment has an effect (positive or negative), we should see that for most ranges of the age along the x-axis, the Os should tend to have higher values than the Xs (if the drug is beneficial) OR the Os should tend to have lower values than the Xs (if the drug is detrimental). The former seems to be the case in the scatter plot above. We can see if the insight generated by the scatter plot holds up under multiple regression.

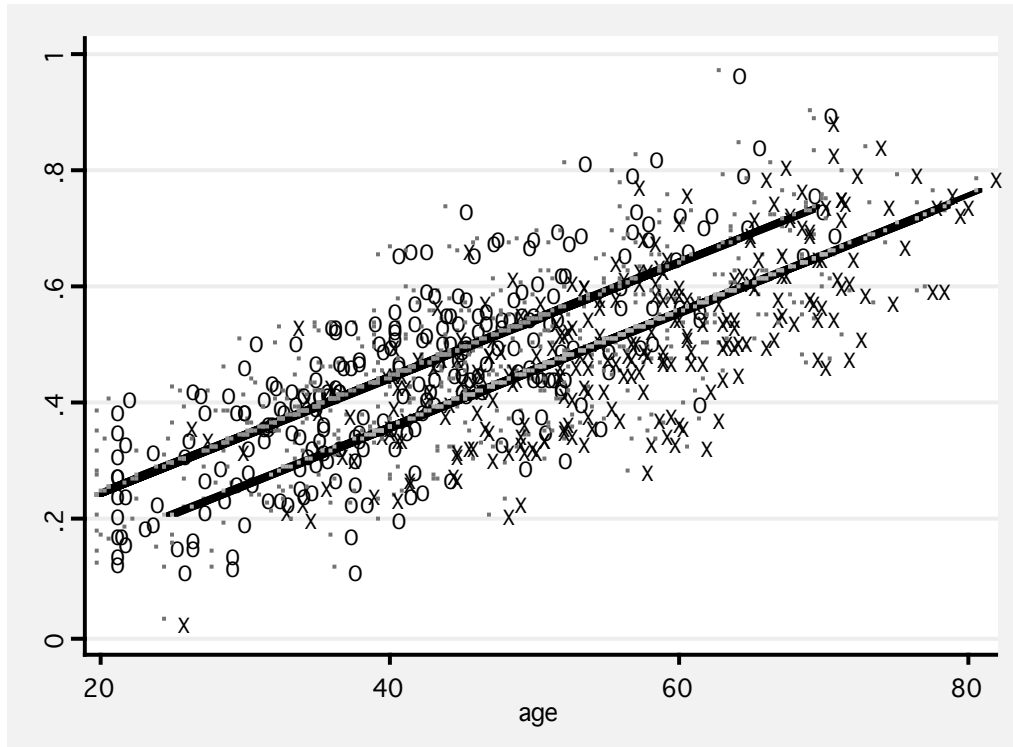
```
. reg attack age treat
```

Source	SS	df	MS	Number of obs =	500
Model	6.64454921	2	3.3222746	F(2, 497) =	313.38
Residual	5.26895757	497	.010601524	Prob > F =	0.0000
Total	11.9135068	499	.023874763	R-squared =	0.5577
				Adj R-squared =	0.5560
				Root MSE =	.10296

attack	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0098503	.000406	24.26	0.000	.0090526 .010648
treatment	-.0855338	.0109208	-7.83	0.000	-.1069904 -.0640771
_cons	.0504158	.0176387	2.86	0.004	.0157602 .0850714

The insight does appear to hold. Controlling for the effects of the treatment, the heart attack indicators increase with age (you can see this by looking at the sign of the slope coefficient on age, which is positive and equal to a little less than .01.) The effect of the treatment, controlling

for age, is negative, which we can interpret here as meaning that the drug reduced the risk of heart attack. It decreased it on average by .085 units, which is a little more than half a standard deviation effect. We can show all the effects at the same time by adding our predicted values (our y-hats) to the graph:



What we end up with is a scatter plot with two regression lines. The upper line corresponds to the control group (the 0s) and the lower line corresponds to the treatment group (the Xs). The lines are parallel, as they have the same slope of $-.0855338$. The equations for the two lines are not identical, however. They are parallel lines, so have different y-intercepts. The two lines have equations as follows.

$$\hat{y}_i = \alpha + \hat{B}_1 * age + \hat{B}_2 * treatment$$

$$treatment = 1 : \hat{y}_i = \hat{\alpha} + \hat{B}_1 * age + \hat{B}_2 * 1$$

$$treatment = 1 : \hat{y}_i = .050 + .0098_1 * age - .0855 * 1$$

$$treatment = 0 : \hat{y}_i = \hat{\alpha} + \hat{B}_1 * age + \hat{B}_2 * 0$$

$$treatment = 0 : \hat{y}_i = \hat{\alpha} + \hat{B}_1 * age$$

$$treatment = 0 : \hat{y}_i = .050 + .0098_1 * age - .0855 * 0$$

$$treatment = 0 : \hat{y}_i = .050 + .0098_1 * age$$

By writing out the equations this way, you can see what's going on with the treatment variable (which is called a dummy, dichotomous, or indicator variable). The difference in the two

regression lines is captured by the coefficient on the treatment variable (\hat{B}_2 , estimated to be -.0855), which drops out when the treatment variable equals zero.

Education, income, and Gender data:

Now, a much briefer treatment of a similar problem. Say we have data on average salaries for men and women as below, with the female variable equaling 1 for females and 0 for males. This first summary and the regression show that women are making about \$1,500 less than males on average, and the difference in mean salary is statistically significant at conventional levels.

```
. tab female, summ(inc)
```

female	Mean	Std. Dev.	Freq.
0	32287.939	4394.2125	240
1	30759.868	3954.9275	260
Total	31493.342	4236.8316	500

```
. reg inc fem
```

Source	SS	df	MS	Number of obs =	500
Model	291408242	1	291408242	F(1, 498) =	16.75
Residual	8.6660e+09	498	17401630.2	Prob > F =	0.0000
Total	8.9574e+09	499	17950741.6	R-squared =	0.0325
				Adj R-squared =	0.0306
				Root MSE =	4171.5

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
female	-1528.071	373.4115	-4.09	0.000	-2261.727 -794.4152
_cons	32287.94	269.2709	119.91	0.000	31758.89 32816.99

But in our sample, females average slightly more years of schooling:

```
. tab female, summ(educ)
```

female	Mean	Std. Dev.	Freq.
0	12.403096	1.582904	240
1	12.604497	1.5044325	260
Total	12.507824	1.5443323	500

This means that if education level affects income level, in our original regression, we are underestimating the effect of gender on salaries, since the gender variable will have to “do the work” of both the education variable and the gender variable. These result suggest that we underestimate the salary gap by \$175. Where does that figure come from? I took the difference of the two coefficients on the female variable (-1703 and -1528).

```
. reg inc edu fem
```

Source	SS	df	MS			
Model	1.1952e+09	2	597618757	Number of obs	=	500
Residual	7.7622e+09	497	15618073.5	F(2, 497)	=	38.26
Total	8.9574e+09	499	17950741.6	Prob > F	=	0.0000
				R-squared	=	0.1334
				Adj R-squared	=	0.1299
				Root MSE	=	3952

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educa	873.3289	114.8017	7.61	0.000	647.7724	1098.885
female	-1703.96	354.5129	-4.81	0.000	-2400.489	-1007.431
_cons	21455.96	1446.567	14.83	0.000	18613.82	24298.1