SO5032 Lecture 5

Brendan Halpin

March 6, 2023

Outline

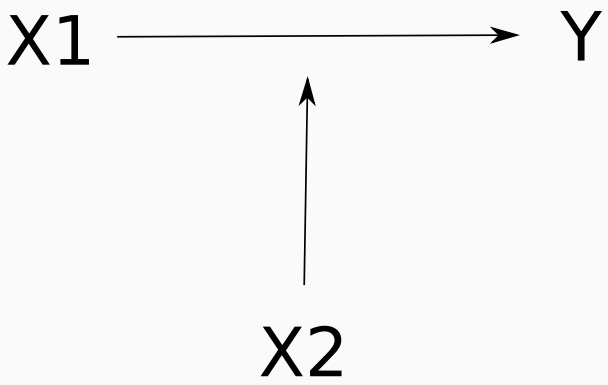
SO5032 Lecture 5

SO5032 Lecture 5

Interaction effects in regression

Interactions

- An interaction effect is where the effect of one variable on Y changes depending on the value of another



Income, hours and gender

```
. reg income hours i.sex
```

Source	SS	df	MS	Number of obs	=	7,945
				F(2, 7942)	=	794.96
Model	1.8935e+09	2	946761687	Prob > F	=	0.0000
Residual	9.4586e+09	7,942	1190962.07	R-squared	=	0.1668
				Adj R-squared	=	0.1666
Total	1.1352e+10	7,944	1429021.17	Root MSE	=	1091.3

income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hours	33.96065	1.123629	30.22	0.000	31.75804	36.16326
sex						
female	-337.0889	26.44232	-12.75	0.000	-388.9228	-285.255
_cons	787.1759	45.73595	17.21	0.000	697.5214	876.8304

For men

```
. reg income hours if sex==1
```

Source	SS	df	MS		
				Number of obs	= 3,818
				F(1, 3816)	= 204.70
Model	344180174	1	344180174	Prob > F	= 0.0000
Residual	6.4162e+09	3,816	1681398.47	R-squared	= 0.0509
				Adj R-squared	= 0.0507
Total	6.7604e+09	3,817	1771128.3	Root MSE	= 1296.7

income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hours	28.71923	2.007313	14.31	0.000	24.78372	32.65474
_cons	983.9722	78.23438	12.58	0.000	830.587	1137.357

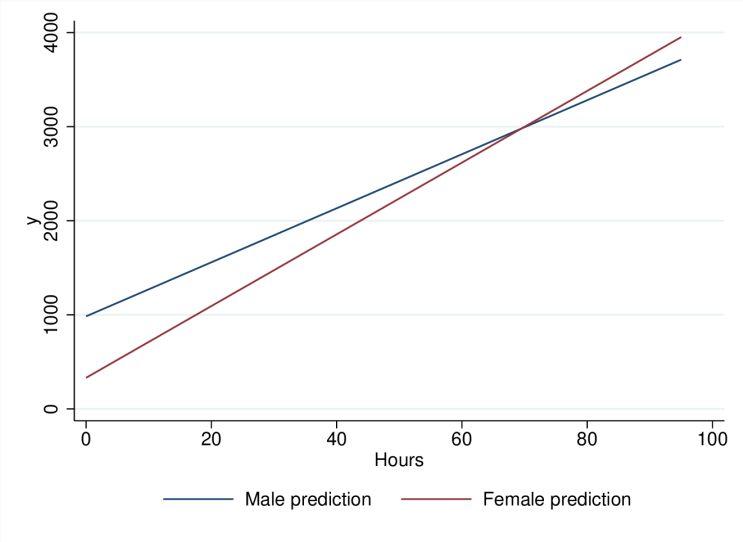
For women

```
. reg income hours if sex==2
```

Source	SS	df	MS		
Model	764315243	1	764315243	Number of obs	= 4,127
Residual	3.0218e+09	4,125	732568.614	F(1, 4125)	= 1043.34
				Prob > F	= 0.0000
Total	3.7862e+09	4,126	917634.7	R-squared	= 0.2019
				Adj R-squared	= 0.2017
				Root MSE	= 855.9

income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hours	38.11874	1.180121	32.30	0.000	35.80507	40.43241
_cons	330.7275	36.40158	9.09	0.000	259.3607	402.0942

Different effects



Interaction in regression

- We can capture interaction effects with a regression model of this form:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

- That is, a 1-unit increase in X_1 leads to a $\beta_1 + \beta_3 X2$ increase in $\hat{Y}$
- Equivalently, a 1-unit increase in X_2 leads to a $\beta_1 + \beta_3 X1$ increase in $\hat{Y}$

Interaction between hours and sex

- Simplest example: one variable is binary

$$\hat{Y}_m = \beta_0 + \beta_1 X_1 + \beta_2 \times 0 + \beta_3 X_1 \times 0$$

$$\hat{Y}_f = \beta_0 + \beta_1 X_1 + \beta_2 \times 1 + \beta_3 X_1 \times 1$$

One-unit increase

If X_1 increases by 1 unit, $\hat{Y}$ changes:

$$\Delta \hat{Y}_m = \beta_1$$

$$\Delta \hat{Y}_f = \beta_1 + \beta_3$$

Stata: by hand

- First create an interaction variable:

```
genfemale = sex == 2
gen intvar = hours*female
```

- Then fit the regression:

```
reg incom hours female intvar
```

Results

```
. gen female = sex==2

. gen intvar = female*hours

. reg income hours sex intvar
```

Source	SS	df	MS	Number of obs	=	7,945
				F(3, 7941)	=	536.82
Model	1.9141e+09	3	638027348	Prob > F	=	0.0000
Residual	9.4381e+09	7,941	1188523.12	R-squared	=	0.1686
				Adj R-squared	=	0.1683
Total	1.1352e+10	7,944	1429021.17	Root MSE	=	1090.2

income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hours	28.71923	1.687655	17.02	0.000	25.41098	32.02747
sex	-653.2448	80.47524	-8.12	0.000	-810.9974	-495.4921
intvar	9.399515	2.260017	4.16	0.000	4.969287	13.82974
_cons	1637.217	139.4834	11.74	0.000	1363.793	1910.641

Stata's formula syntax

- But more convenient to use Stata's formula syntax

```
reg income c.hours##i.sex
```

- `i.sex` means treat `sex` as categorical
- `c.hours#i.sex` creates the interaction between hours (continuous, `c.`) and sex
- `c.hours##i.sex` puts both the interaction and the first order terms in the model

Same results using Stata's formula syntax

```
. reg income c.hours##i.sex
```

Source	SS	df	MS	Number of obs	=	7,945
				F(3, 7941)	=	536.82
Model	1.9141e+09	3	638027348	Prob > F	=	0.0000
Residual	9.4381e+09	7,941	1188523.12	R-squared	=	0.1686
				Adj R-squared	=	0.1683
Total	1.1352e+10	7,944	1429021.17	Root MSE	=	1090.2

income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hours	28.71923	1.687655	17.02	0.000	25.41098	32.02747
sex						
female	-653.2448	80.47524	-8.12	0.000	-810.9974	-495.4921
sex#c.hours						
female	9.399515	2.260017	4.16	0.000	4.969287	13.82974
_cons	983.9722	65.7758	14.96	0.000	855.0344	1112.91

Predictions

Sex	Hrs	β_0	β_1	β_2	β_3	$\hat{y}$
M	0	1637.217	+ 0*28.71923	+ 0*-653.2448	+ 0*0*9.399515	= 1637.217
M	80	1637.217	+ 80*28.71923	+ 0*-653.2448	+ 80*0*9.399515	= 3934.7554
F	0	1637.217	+ 0*28.71923	+ 1*-653.2448	+ 0*1*9.399515	= 983.9722
F	80	1637.217	+ 80*28.71923	+ 1*-653.2448	+ 80*1*9.399515	= 4033.4718

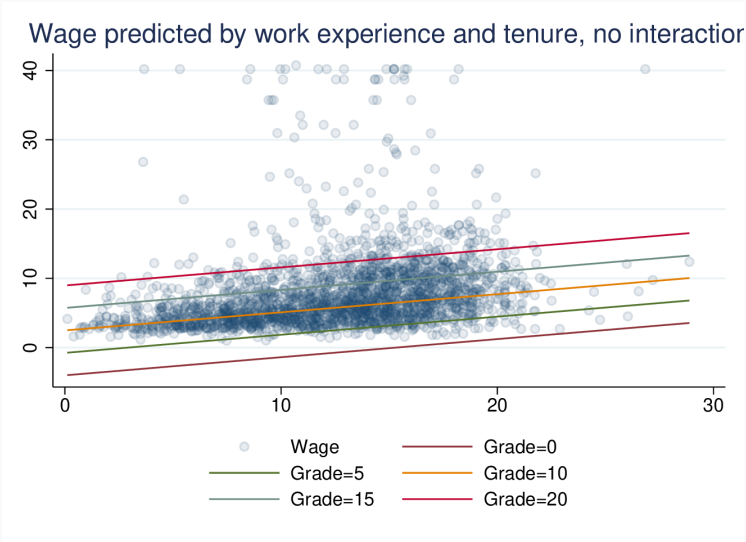
Interactions between two continuous variable

```
. reg wage c.ttl_exp##c.grade
```

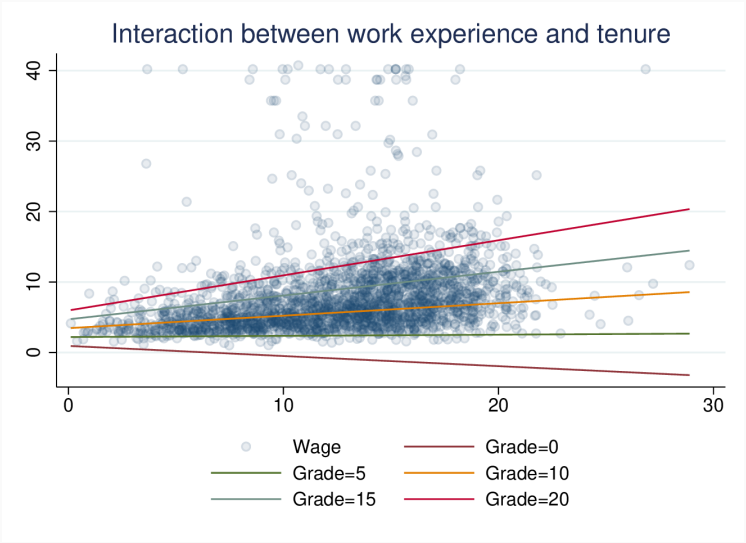
Source	SS	df	MS		
				Number of obs	= 2,244
				F(3, 2240)	= 133.83
Model	11301.2662	3	3767.08872	Prob > F	= 0.0000
Residual	63053.0643	2,240	28.1486894	R-squared	= 0.1520
				Adj R-squared	= 0.1509
Total	74354.3305	2,243	33.1495009	Root MSE	= 5.3055

wage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ttl_exp	-.143543	.1284932	-1.12	0.264	-.3955211	.1084352
grade	.2515455	.1315367	1.91	0.056	-.0064011	.5094921
c.ttl_exp#c.grade	.032074	.0099813	3.21	0.001	.0125005	.0516475
_cons	.933757	1.657647	0.56	0.573	-2.316929	4.184443

Without interaction, predictions for different levels of grade



With interaction



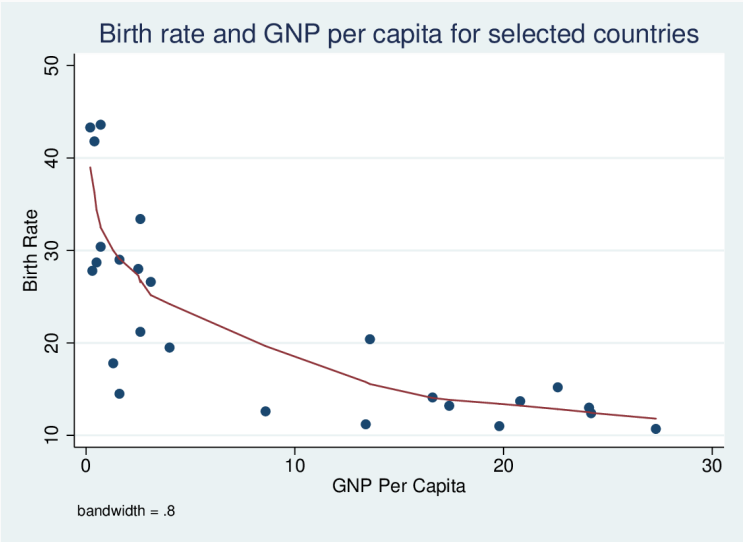
SO5032 Lecture 5

Non-linear linear regression

Birth rate and GNP example

```
do http://teaching.sociology.ul.ie/so5032/birth
sort gnp
label var bir "Birth Rate"
label var gnp "GNP Per Capita"
lowess bir gnp, title("Birth rate and GNP per capita for selected countries")
```

Nonlinear plot



Get linear relationship

```
reg bir gnp
```

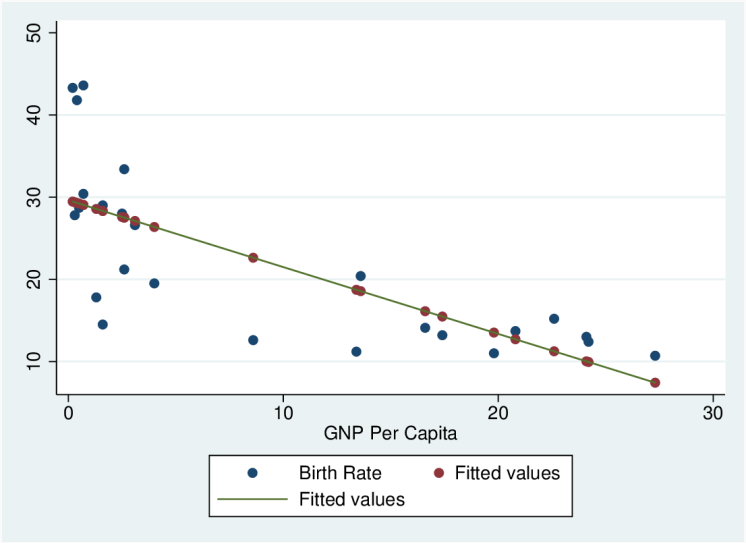
```
. reg bir gnp
```

Source	SS	df	MS	Number of obs	=	25
				F(1, 23)	=	27.52
Model	1450.2603	1	1450.2603	Prob > F	=	0.0000
Residual	1212.02523	23	52.696749	R-squared	=	0.5447
				Adj R-squared	=	0.5249
Total	2662.28552	24	110.928563	Root MSE	=	7.2593

bir	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gnp	-.8133082	.155033	-5.25	0.000	-1.134018	-.4925981
_cons	29.6227	2.037416	14.54	0.000	25.40798	33.83742

```
predict plin
scatter bir plin gnp|| line plin gnp
```

Linear plot


50
40
30
20
10
0
10
20
30
GNP Per Capita
Birth Rate
Fitted values
Fitted values

Quadratic

Linear regression doesn't fit well

Clearly, as GNP rises BIR falls, but the rate of fall declines

Let's try quadratic:

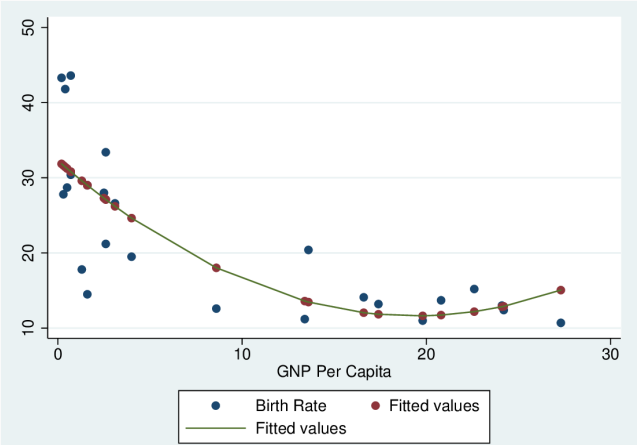
```
. reg bir c.gnp##c.gnp
```

Source	SS	df	MS	Number of obs	=	25
				F(2, 22)	=	18.39
Model	1665.82856	2	832.914278	Prob > F	=	0.0000
Residual	996.456968	22	45.2934985	R-squared	=	0.6257
				Adj R-squared	=	0.5917
Total	2662.28552	24	110.928563	Root MSE	=	6.73

bir	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gnp	-2.130192	.6205087	-3.43	0.002	-3.417048	-.8433351
c.gnp#c.gnp	.0549243	.0251762	2.18	0.040	.0027121	.1071366
_cons	32.27852	2.247195	14.36	0.000	27.61812	36.93892

Quatratic plot

```
predict pquad
scatter bir pquad gnp|| line pquad gnp
```



Let's try square root of GNP:

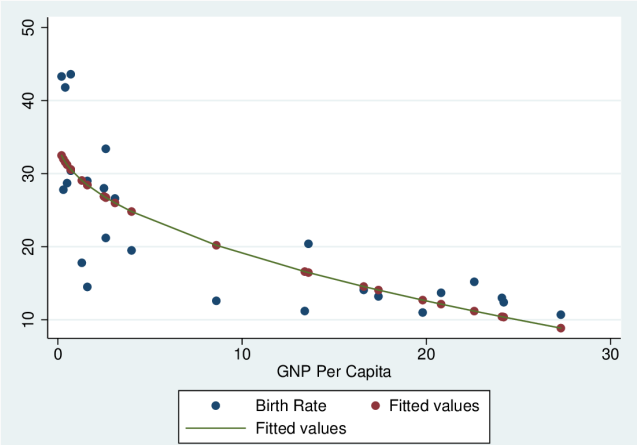
```
. gen sqg = sqrt(gnp)

. reg bir sqg
```

Source	SS	df	MS	Number of obs	=	25
				F(1, 23)	=	39.44
Model	1681.66084	1	1681.66084	Prob > F	=	0.0000
Residual	980.624685	23	42.6358559	R-squared	=	0.6317
				Adj R-squared	=	0.6156
Total	2662.28552	24	110.928563	Root MSE	=	6.5296

bir	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sqg	-4.945487	.7874579	-6.28	0.000	-6.574468	-3.316506
_cons	34.70314	2.391073	14.51	0.000	29.75683	39.64946

$\sqrt{GNP}$ plot

```
predict psqrt
scatter bir psqrt gnp|| line psqrt gnp
```



log(GNP)

Let's try the log of GNP:

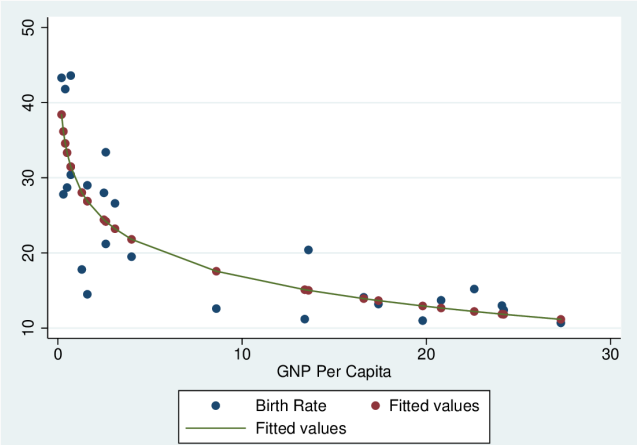
```
. gen lgg = log(gnp)

. reg bir lgg
```

Source	SS	df	MS	Number of obs	=	25
				F(1, 23)	=	54.84
Model	1875.68482	1	1875.68482	Prob > F	=	0.0000
Residual	786.600705	23	34.2000307	R-squared	=	0.7045
				Adj R-squared	=	0.6917
Total	2662.28552	24	110.928563	Root MSE	=	5.8481

bir	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lgg	-5.542152	.748362	-7.41	0.000	-7.090257	-3.994047
_cons	29.49466	1.53576	19.21	0.000	26.3177	32.67162

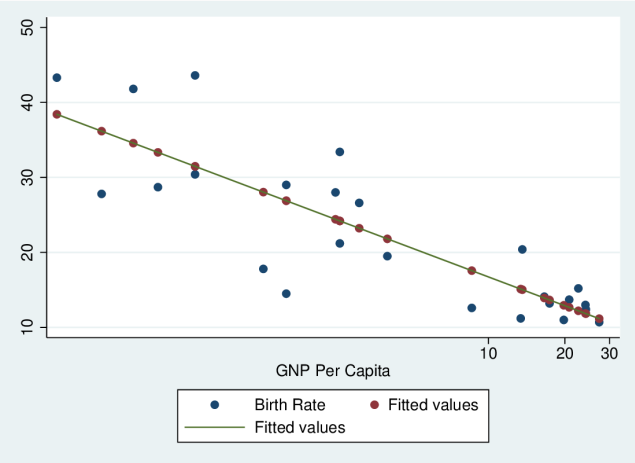
log(GNP) plot

```
predict plog
scatter bir plog gnp|| line plog gnp
```



Log-scale plot

```
scatter bir plog gnp, xscale(log)|| line plog gnp, xscale(log)
```



Square root and log compared

```
label var sqg "Sq Root GNP"
label var lg "Log of GNP"
scatter sqg lg gnp
```