

Quantitative Research Methods: Introduction to correlation and regression

Brendan Halpin, Sociology, University of Limerick

January 2018



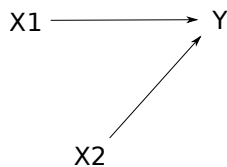
Multidimensional causality

- Regression analysis never proves causal relationships, but it "thinks" in causal terms
- To use it we need to understand causal relationships: what process generates the data we see, and what can regression tell us about it.
- Start by considering the relationship between variables and patterns of association



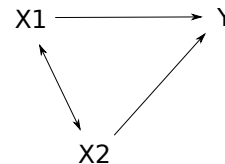
3-variable pictures

- Let's consider patterns of causality and association between three variables, X1 and X2, and Y
- If X1 and X2 are not correlated with each other, their separate effects on Y more or less just add up



Correlated X variables

- But if X1 and X2 are correlated, things can get funny:

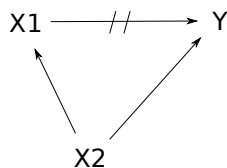


- In particular, if we measure the effect of one X without taking account of the other we will likely over-estimate it



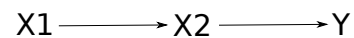
Spurious association

- X1 may have an association with Y, implying a causal relationship
- But if X2 affects both X1 and Y the relationship between X1 and Y may be **spurious**



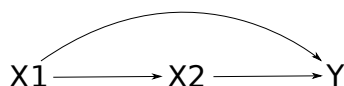
Indirect effects

- Where there is a time-order (X1 before X2), we may see direct and indirect effects
- X1 may affect X2, which affects Y, but not affect Y directly
- Thus there is association between X1 and Y without a direct causal effect



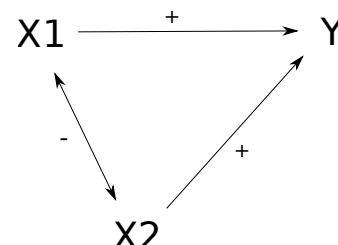
Direct and indirect effects

- However, it is possible for both direct and indirect effects to be present at the same time



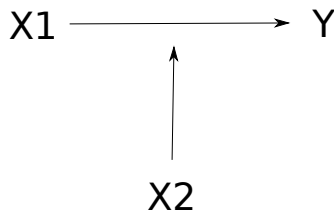
Suppression

- Where X1 and X2 have positive effects on Y, but a negative correlation, or different effects on Y with a positive correlation, the association between X1 and Y may be **suppressed**
- That is, it may be invisible if we don't take account of X2



Interactions

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



Multiple explanatory variables

- Regression analysis can be extended to the case where there is more than one explanatory variable – multivariate regression
- This allows us to estimate the net simultaneous effect of many variables, and thus to begin to disentangle more complex relationships
- Interpretation is relatively easy: each variable gets its own slope coefficient, standard error and significance
- The slope coefficient is the effect on the dependent variable of a 1 unit change in the explanatory variable, *while taking account of the other variables*

Example

- Example: domestic work time may be affected by gender, and also by paid work time: competing explanations – one or the other, or both could have effects
- We can fit bivariate regressions:

$$DWT = a + b \times PaidWork$$

or

$$DWT = a + b \times Female$$

- We can also fit a single multivariate regression

$$DWT = a + b \times PaidWork + c \times Female$$

Dichotomous variables

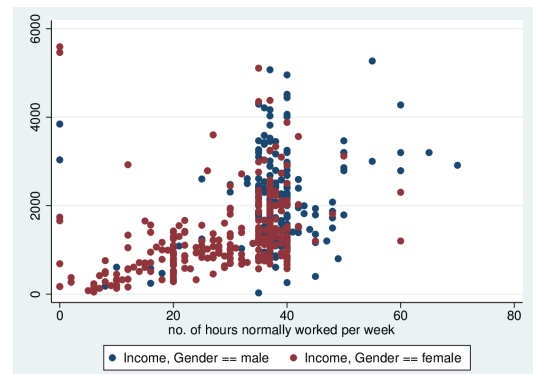
- We deal with gender in a special way: this is a *binary* or *dichotomous* variable – has two values
- We turn it into a yes/no or 0/1 variable – e.g., female or not
- If we put this in as an explanatory variable a *one-unit change in the explanatory variable* is the difference between being male and female
- Thus the *c* coefficient we get in the $DWT = a + b \times PaidWork + c \times Female$ regression is the net change in predicted domestic work time for females, once you take account of paid work time.
- The *b* coefficient is then the net effect of a unit change in paid work time, once you take gender into account.

Income, hours and gender

```
. corr Income Gender Hours
(obs=506)
```

	Income	Gender	Hours
Income	1.0000		
Gender	-0.3280	1.0000	
Hours	0.3638	-0.4360	1.0000

Income, hours and gender



T-test: Income by gender

```
. ttest Income, by(Gender)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
male	437	1618.348	59.11677	1235.809	1502.159 1734.537
female	531	992.1805	40.82127	940.6625	911.9892 1072.372
combined	968	1274.861	36.23219	1127.281	1203.759 1345.964
diff		626.1674	70.00484		488.7883 763.5465

```
diff = mean(male) - mean(female)          t = 8.9446
Ho: diff = 0                             degrees of freedom = 966
Ha: diff < 0                             Ha: diff != 0
Pr(T < t) = 1.0000                       Pr(|T| > |t|) = 0.0000
                                           Pr(T > t) = 0.0000
```

Regression: Just hours

```
. reg Income Hours
```

Source	SS	df	MS	Number of obs	=	506
Model	86947928.8	1	86947928.8	F(1, 504)	=	76.86
Residual	570128215	504	1131206.78	Prob > F	=	0.0000
				R-squared	=	0.1323
				Adj R-squared	=	0.1306
				Root MSE	=	1063.6

Income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Hours	37.82204	4.314061	8.77	0.000	29.34628 46.2978
_cons	449.7435	150.1722	2.99	0.003	154.703 744.7841

Regression: Hours and binary gender

```
. reg Income Hours i.Gender
```

Source	SS	df	MS	Number of obs	=	506
Model	110236231	2	55118115.6	F(2, 503)	=	50.70
Residual	546839912	503	1087156.88	Prob > F	=	0.0000
Total	657076144	505	1301140.88	R-squared	=	0.1678
				Adj R-squared	=	0.1645
				Root MSE	=	1042.7

Income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Hours	28.33857	4.699451	6.03	0.000	19.1056 37.57155
Gender female	-478.4214	103.3684	-4.63	0.000	-681.5084 -275.3344
_cons	1022.139	192.2717	5.32	0.000	644.3844 1399.893

Regression: for men only

```
. reg Income Hours if Gender==1
```

Source	SS	df	MS	Number of obs	=	232
Model	8009519.02	1	8009519.02	F(1, 230)	=	5.36
Residual	343845612	230	1494980.92	Prob > F	=	0.0215
Total	351855131	231	1523182.38	R-squared	=	0.0228
				Adj R-squared	=	0.0185
				Root MSE	=	1222.7

Income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Hours	24.61855	10.63597	2.31	0.022	3.662162 45.57495
_cons	1164.366	414.4901	2.81	0.005	347.6826 1981.049

Regression: for women only

```
. reg Income Hours if Gender==2
```

Source	SS	df	MS	Number of obs	=	274
Model	31772944.2	1	31772944.2	F(1, 272)	=	42.63
Residual	202744304	272	745383.469	Prob > F	=	0.0000
Total	234517248	273	859037.537	R-squared	=	0.1355
				Adj R-squared	=	0.1323
				Root MSE	=	863.36

Income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Hours	29.70376	4.549594	6.53	0.000	20.74687 38.66065
_cons	504.6153	140.3614	3.60	0.000	228.2824 780.9482

Regression: interaction

```
. reg Income c.Hours#i.Gender
```

Source	SS	df	MS	Number of obs	=	506
Model	110486228	3	36828742.8	F(3, 502)	=	33.82
Residual	546589915	502	1088824.53	Prob > F	=	0.0000
Total	657076144	505	1301140.88	R-squared	=	0.1681
				Adj R-squared	=	0.1632
				Root MSE	=	1043.5

Income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Hours	24.61855	9.076915	2.71	0.007	6.785132 42.45198
Gender female	-659.7502	392.3082	-1.68	0.093	-1430.518 111.0181
Gender#c.Hours female	5.085207	10.61255	0.48	0.632	-15.76529 25.9357
_cons	1164.366	353.7327	3.29	0.001	469.3865 1859.345

Regression: Direct and indirect 1

```
. reg ownscore fatherscore
```

Source	SS	df	MS	Number of obs	=	1,000
Model	13269.3853	1	13269.3853	F(1, 998)	=	53.50
Residual	247525.861	998	248.021905	Prob > F	=	0.0000
Total	260795.247	999	261.056303	R-squared	=	0.0509
				Adj R-squared	=	0.0499
				Root MSE	=	15.749

ownscore	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
fatherscore	.2370829	.032413	7.31	0.000	.1734773 .3006884
_cons	37.90861	1.672157	22.67	0.000	34.62726 41.18996

Regression: Direct and indirect 2

```
. reg education fatherscore
```

Source	SS	df	MS	Number of obs	=	1,000
Model	311.104929	1	311.104929	F(1, 998)	=	111.01
Residual	2797.00607	998	2.80261129	Prob > F	=	0.0000
Total	3108.111	999	3.11122222	R-squared	=	0.1001
				Adj R-squared	=	0.0992
				Root MSE	=	1.6741

education	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
fatherscore	.0363018	.0034455	10.54	0.000	.0295405 .0430631
_cons	1.295213	.1777516	7.29	0.000	.9464035 1.644023

Regression: Direct and indirect 3

```
. reg ownscore education
```

Source	SS	df	MS	Number of obs	=	1,000
Model	80742.8091	1	80742.8091	F(1, 998)	=	447.54
Residual	180052.437	998	180.413264	Prob > F	=	0.0000
Total	260795.247	999	261.056303	R-squared	=	0.3096
				Adj R-squared	=	0.3089
				Root MSE	=	13.432

ownscore	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
education	5.096871	.2409273	21.16	0.000	4.624089 5.569653
_cons	33.87079	.8556481	39.58	0.000	32.19171 35.54986

Regression: Direct and indirect 4

```
. reg ownscore education fatherscore
```

Source	SS	df	MS	Number of obs	=	1,000
Model	81453.7212	2	40726.8606	F(2, 997)	=	226.41
Residual	179341.525	997	179.881169	Prob > F	=	0.0000
Total	260795.247	999	261.056303	R-squared	=	0.3123
				Adj R-squared	=	0.3109
				Root MSE	=	13.412

ownscore	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
education	4.937369	.2535982	19.47	0.000	4.439722 5.435017
fatherscore	.0578475	.0290984	1.99	0.047	.0007463 .1149486
_cons	31.51367	1.461439	21.56	0.000	28.64582 34.38152