Today's Agenda

Extending the OLS Regression

- Dichotomous predictors (Dummies)
- Categorical predictors

Dataset: Ross (1990)

Justin Leinaweaver (Spring 2022)

Work, Family, and Well-Being in the United States, 1990 (ICPSR 6666)

Version Date: Jun 10, 1996 @ Cite this study | Share this page

Principal Investigator(s): 3

Catherine E. Ross

https://doi.org/10.3886/ICPSR06666.v1

Version V1

	Α	В	С	D	E	F	G	Н	l l	J	K	L	M	N	0
1	height	weight	male	earn	earnk	ethnicity	education	mother_education	father_education	walk	exercise	smokenow	tense	angry	age
2	74	210	1	50000	50	White	16	16	16	3	3	2	0	C	45
3	66	125	0	60000	60	White	16	16	16	6	5	1	0	C	58
4	64	126	0	30000	30	White	16	16	16	8	1	. 2	1	. 1	29
5	65	200	0	25000	25	White	17	17	NA	8	1	. 2	0	0	57
6	63	110	0	50000	50	Other	16	16	16	5	6	2	0	0	91
7	68	165	0	62000	62	Black	18	18	18	1	. 1	. 2	2	2	54
8	63	190	0	51000	51	White	17	17	17	3	1	. 2	4	. 4	39
9	64	125	0	9000	9	White	15	15	15	7	4	1	4	. 4	26
10	62	200	0	29000	29	White	12	12	12	2	2	2	0	0	49
11	73	230	1	32000	32	White	17	17	17	7	1	1	0	0	46
12	72	176	1	2000	2	Hispanic	15	15	15	8	1	. 2	0	C	21
13	72	265	1	35000	35	White	NA	NA	NA	1	. 1	. 2	0	C	53
14	72	160	1	27000	27	White	12	12	12	1	. 2	2	1	. 1	26
15	70	225	1	6530	6.53	White	16	16	NA	4	1	. 2	0	0	65
16	63	107	0	0	0	White	14	14	14	7	4	. 2	2	2	50

	Α	В	C	D	E	F	G	Н		1	J	K	L	M	N	0
1	height	weight	male	earn	earnk	ethnicity	education	mother_education	father	education	walk	exercise	smokenow	tense	angry	age
2	74	210	1	50000	50	White	16	16		16	3	3	2	0	0	45
3	66	125	0	60000	60	White	16	16		16	6	5	1	0	0	58
4	64	126	0	30000	30	White	16	16		16	8	1	2	1	1	29
5	65	200	0	25000	25	White	17	17	NA		8	1	2	0	0	57
6	63	110	0	50000	50	Other	16	16		16	5	6	2	0	0	91
7	68	165	0	62000	62	Black	18	18		18	1	1	2	2	2	54
8	63	190	0	51000	51	White	17	17		17	3	1	2	4	4	39
9	64	125	0	9000	9	White	15	15		15	7	4	1	4	4	26
10	62	200	0	29000	29	White	12	12		12	2	2	2	0	0	49
11	73	230	1	32000	32	White	17	17		17	7	1	1	0	0	46
12	72	176	1	2000	2	Hispanic	15	15		15	8	1	2	0	0	21
13	72	265	1	35000	35	White	NA	NA	NA		1	1	2	0	0	53
14	72	160	1	27000	27	White	12	12		12	1	2	2	1	1	26
15	70	225	1	6530	6.53	White	16	16	NA		4	1	2	0	0	65
16	63	107	0	0	0	White	14	14		14	7	4	2	2	2	50

Is their evidence of a gender difference in earned income?

- 1. Calculate the mean income for each gender
 - Men = ?
 - Women = ?

Is their evidence of a gender difference in earned income?

- 1. Calculate the mean income for each gender
 - Men = \$59.9k
 - Women = \$32.1k

C2 f_x =AVERAGE(A2:A1142)

	Α	В	С	D	Е	F	G	
1	earnk2021	male			earnk2021	male		
2	121.596	0	32.11742536		101.33	1	59.89894732	
3	60.798	0			64.8512	1		
4	50.665	0			4.0532	1		
5	101.33	0			70.931	1		
6	125.6492	0			54.7182	1		
7	103.3566	0			13.233698	1		
8	18.2394	0			60.798	1		
9	58.7714	0			24.3192	1		
10	0	0			30.399	1		
11	24.3192	0			40.532	1		
12	40.532	0			34.4522	1		
13	0	0			89.1704	1		

Is their evidence of a gender difference in earned income?

2. Fit an OLS regression of income on the male dummy

• Men = \$59.9k

• Women = \$32.1k

	Income (Thousands USD)
Male	27.78*
	(1.93)
Constant	32.12*
	(1.18)
Observations	1,815
Adjusted R ²	0.10
Residual Std. Error	39.77 (df = 1813)
F Statistic	206.76* (df = 1; 1813)
Note:	*n<0.05

• Men = \$59.9k

• Women = \$32.1k

	Income (Thousands USD)
Male	27.78*
	(1.93)
Constant	32.12*
	(1.18)
Observations	1,815
Adjusted R ²	0.10
Residual Std. Error	39.77 (df = 1813)
F Statistic	206.76* (df = 1; 1813)
Note:	*p<0.05

 $Income = 32.12 + 27.78 \times (Male)$

• Men = \$59.9k

• Women = \$32.1k

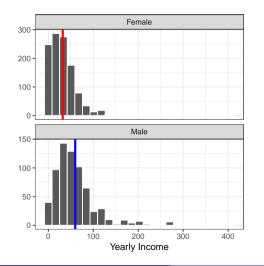
	Income (Thousands USD)
Male	27.78*
	(1.93)
Constant	32.12*
	(1.18)
Observations	1,815
Adjusted R ²	0.10
Residual Std. Error	39.77 (df = 1813)
F Statistic	206.76* (df = 1; 1813)
Note:	*p<0.05

Income = $32.12 + 27.78 \times (Male)$

•
$$P(Male = 1) = 32.12 + 27.78 \times 1 = 59.9$$

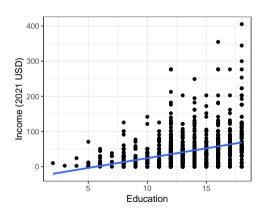
•
$$P(Male = 0) = 32.12 + 27.78 \times 0 = 32.12$$

Gender Differences in Income?



	Income (Thousands USD)
Male	27.78*
	(1.93)
Constant	32.12*
	(1.18)
Observations	1.815
Adjusted R ²	0.10
Residual Std. Error	$39.77 \; (df = 1813)$
F Statistic	206.76* (df = 1; 1813)
Note:	*p<0.05

Gender Differences in Income?



	Income (Thousands USD)
Education	5.57*
	(0.36)
Constant	-31.34*
	(4.89)
Observations	1,813
Adjusted R ²	0.11
Residual Std. Error	39.50 (df = 1811)
F Statistic	235.79* (df = 1; 1811)
Note:	*p<0.05

Dummy Variables in OLS Regressions

Regress earnings (2021) on education and the male dummy

		Income (Thousands USD)	
	(1)	(2)	(3)
male	27.78*		26.53*
	(1.93)		(1.82)
education		5.57*	5.35*
		(0.36)	(0.34)
Constant	32.12*	-31.34*	-38.28*
	(1.18)	(4.89)	(4.65)
Observations	1,815	1,813	1,813
Adjusted R ²	0.10	0.11	0.21
Residual Std. Error	39.77 (df = 1813)	39.50 (df = 1811)	37.38 (df = 1810)
F Statistic	206.76* (df = 1; 1813)	235.79* (df = 1; 1811)	238.09* (df = 2; 1810)

Note:

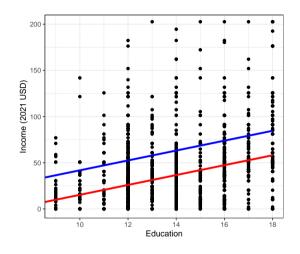
Income (Thousands USD)
5.35*
(0.34)
26.53*
(1.82)
-38.28*
(4.65)
` ,
1,813
0.21
37.38 (df = 1810)
238.09* (df = 2; 1810)
*p<0.05

Make a marginal effects plot of education with separate lines for each gender

Make a marginal effects plot of education with separate lines for each gender

- Add a sheet
- Column 1: The levels of education
- Oclumn 2: Model point estimates for a male across the levels of education
- Column 3: Model point estimates for a female across the levels of education
- Highlight all three columns, insert a scatterplot and polish it

Education	Male	Female
9	36.4	9.9
10	41.8	15.2
11	47.1	20.6
12	52.4	25.9
13	57.8	31.3
14	63.1	36.6
15	68.5	42
16	73.8	47.3
17	79.2	52.7
18	84.6	58



Dummy Variables in OLS Regressions

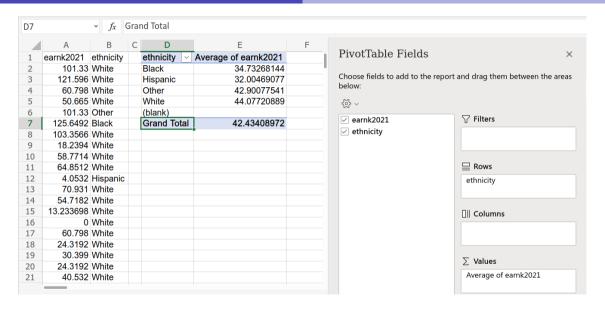
- Point estimates produce the group means (with a significance test), and
- The coefficient on the dummy moves the intercept, not the slope

Categorical Variables

	Α	В	С	D	E	F	G	Н		1	J	K	L	M	N	0
1	height	weight	male	earn	earnk	ethnicity	education	mother_educati	n fathe	r_education	walk	exercise	smokenow	tense	angry	age
2	74	210	1	50000	50	White	16		.6	16	3	3	2	0	0	45
3	66	125	0	60000	60	White	16		.6	16	6	5	1	. 0	0	58
4	64	126	0	30000	30	White	16		.6	16	8	1	. 2	1	1	29
5	65	200	0	25000	25	White	17		7 NA		8	1	2	0	0	57
6	63	110	0	50000	50	Other	16		.6	16	5	6	2	0	0	91
7	68	165	0	62000	62	Black	18		.8	18	1	1	2	2	2	54
8	63	190	0	51000	51	White	17		.7	17	3	1	2	4	4	39
9	64	125	0	9000	9	White	15		.5	15	7	4	1	4	4	26
10	62	200	0	29000	29	White	12		2	12	2	2	2	0	0	49
11	73	230	1	32000	32	White	17		.7	17	7	1	. 1	. 0	0	46
12	72	176	1	2000	2	Hispanic	15		.5	15	8	1	2	0	0	21
13	72	265	1	35000	35	White	NA	NA	NA		1	1	. 2	0	0	53
14	72	160	1	27000	27	White	12		.2	12	1	2	2	1	1	26
15	70	225	1	6530	6.53	White	16		6NA		4	1	. 2	0	0	65
16	63	107	0	0	C	White	14		4	14	7	4	2	2	2	50

Categorical Variables

Ethnicity	Income
Black	\$34.73k
Hispanic	\$32k
Other	\$42.9k
White	\$44.08k



Using Categorical Variables in Excel: Make Dummies

education

	A	В	C	D
1	male	earnk	ethnicity	education
2	1		White	16
3	0		White	16
4	0		White	16
5	0		White	17
6	ő		Other	16
7	0		Black	18
В	0		White	17
9	ő		White	15
0	0		White	12
1	1		White	17
2	1		Hispanic	15
2	1		White	NA
4	1		White	12
	1		White	16
5	1		White	10

Sort data by categorical predictor

Using Categorical Variables in Excel: Make Dummies

male earnk ethnicity education Black Hispanic Other			-	_	-			- ,
0 7 Black 12 1 0 0 0 1 5 Black 12 1 0 0 0 0 5 Black 12 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	L	male	earnk	ethnicity	education	Black	Hispanic	Other
1 53 Black 13 1 0 0 0 5 Black 12 1 0 0 0 5 Black 12 1 0 0 0 10 Black 12 1 0 0 0 30 Black 14 1 0 0 1 13 Black 8 1 0 0 0 5 Black 12 1 0 0 0 0 Black 13 1 0 0 1 15 Black 11 1 0 0 0 15 Black 14 1 0 0 0 21 Black 17 1 0 0		0	62	Black	18	1	0	0
0 5 Black 12 1 0 0 0 0 0 0 10 Black 12 1 0 0 0 0 0 0 10 Black 14 1 0 0 0 0 1 1 13 Black 8 1 0 0 0 0 0 5 Black 12 1 0 0 0 0 0 1 1 15 Black 13 1 0 0 0 1 1 15 Black 11 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0	7	Black	12	1	0	0
0 5 Black 12 1 0 0 0 0 0 0 10 Black 12 1 0 0 0 0 10 Black 14 1 0 0 0 1 1 13 Black 8 1 0 0 0 0 5 Black 12 1 0 0 0 1 1 15 Black 11 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		1	53	Black	13	1	0	0
0 10 Black 12 1 0 0 0 1 1 3 Black 14 1 0 0 0 0 1 1 13 Black 8 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0	5	Black	12	1	0	0
0 30 Black 14 1 0 0 0 1 1 13 Black 8 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0	5	Black	12	1	0	0
1 13 Black 8 1 0 0 0 5 Black 12 1 0 0 0 0 Black 13 1 0 0 1 15 Black 11 1 0 0 0 15 Black 14 1 0 0 0 21 Black 17 1 0 0	С	0	10	Black	12	1	0	0
0 5 Black 12 1 0 0 0 0 Black 13 1 0 0 0 1 1 15 Black 11 1 0 0 0 0 0 0 15 Black 14 1 0 0 0 0 0 21 Black 17 1 0 0	Е	0	30	Black	14	1	0	0
0 0 Black 13 1 0 0 0 1 1 15 Black 11 1 0 0 0 0 0 0 15 Black 14 1 0 0 0 0 0 21 Black 17 1 0 0	С	1	13	Black	8	1	0	0
1 15 Black 11 1 0 0 0 15 Black 14 1 0 0 0 21 Black 17 1 0 0		0	5	Black	12	1	0	0
0 15 Black 14 1 0 0 0 21 Black 17 1 0 0	С	0	0	Black	13	1	0	0
0 21 Black 17 1 0 0	Г	1	15	Black	11	1	0	0
	С	0	15	Black	14	1	0	0
0 15 Black 12 1 0 0		0	21	Black	17	1	0	0
	С	0	15	Black	12	1	0	0
1 15 Black 14 1 0 0		1	15	Black	14	1	0	0
1 43 Black 13 1 0 0	С	1	43	Black	13	1	0	0
0 32 Black 14 1 0 0		0	32	Black	14	1	0	0
0 25 Black 12 1 0 0		0	25	Black	12	1	0	0

2. Create dummies for each category (omit baseline)

Using Categorical Variables in Excel: Make Dummies

male	earnk	ethnicity	education	Black	Hispanic	Other
0		Black	18	1	0	0
0	7	Black	12	1	0	0
1	53	Black	13	1	0	0
0	5	Black	12	1	0	0
0	5	Black	12	1	0	0
0	10	Black	12	1	0	0
0		Black	14	1	0	0
1		Black	8	1	0	0
0		Black	12	1	0	0
0		Black	13	1	0	0
1		Black	11	1	0	0
0		Black	14	1	0	0
0		Black	17	1	0	0
0		Black	12	1	0	0
1		Black	14	1	0	0
1		Black	13	1	0	0
0		Black	14	1	0	0
0	25	Black	12	1	0	0

Regress income on the three dummy predictors

Ethnicity	Income
Black	\$34.73k
Hispanic	\$32k
Other	\$42.9k
White	\$44.08k

	Income (2021)	
Ethnicity: Black	-9.34*	
	(3.30)	
Ethnicity: Hispanic	-12.07^{*}	
	(4.24)	
Ethnicity: Other	-1.18	
	(6.87)	
Constant	44.00*	
Constant	44.08*	
	(1.08)	
Observations	1 015	
	1,815	
Adjusted R ²	0.01	
Residual Std. Error	$41.83 \; (\mathrm{df} = 1811)$	
F Statistic	4.96* (df = 3; 1811)	
Note:	*p<0.05	

Ethnicity	Income
Black	\$34.73k
Hispanic	\$32k
Other	\$42.9k
White	\$44.08k

	Income (2021)
Ethnicity: Black	-9.34*
,	(3.30)
Ethnicity: Hispanic	-12.07*
	(4.24)
Ethnicity: Other	-1.18
	(6.87)
Constant	44.08*
	(1.08)
Observations	1,815
Adjusted R ²	0.01
Residual Std. Error	41.83 (df = 1811)
Note:	*p<0.05

Income = 44.08 + -9.34(Black) + -12.07(Hispanic) + -1.18(Other)

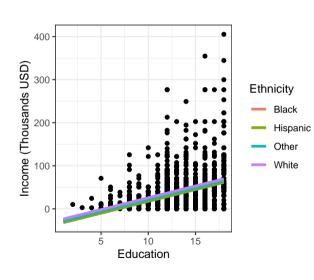
Categorical Variables in OLS Regressions

Regress earnings (2021) on education and the ethnicity dummies

		Income (2021)	
	(1)	(2)	(3)
Education	5.57*		5.49*
	(0.36)		(0.36)
Ethnicity: Black		-9.34*	-5.47
		(3.30)	(3.13)
Ethnicity: Hispanic		-12.07*	-8.43*
		(4.24)	(4.01)
Ethnicity: Other		-1.18	-3.80
		(6.87)	(6.48)
Constant	-31.34*	44.08*	-29.11*
	(4.89)	(1.08)	(4.97)
Observations	1,813	1,815	1,813
Adjusted R ²	0.11	0.01	0.12
Residual Std. Error	39.50 (df = 1811)	41.83 (df = 1811)	39.46 (df = 1808)
F Statistic	235.79* (df = 1; 1811)	4.96* (df = 3; 1811)	60.85* (df = 4; 1808)

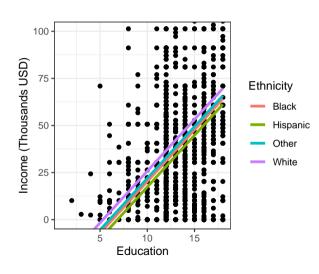
Note:

*p<0.05



	Income (2021)
Education	-5.47
	(3.13)
	(31-3)
Ethnicity: Black	-8.43*
•	(4.01)
	, ,
Ethnicity: Hispanic	-3.80
	(6.48)
Ethnicity: Other	5.49*
	(0.36)
Constant	-29.11*
	(4.97)
Observations	1,813
Adjusted R ²	0.12
Residual Std. Error	39.46 (df = 1808)
F Statistic	$60.85^* (df = 4; 1808)$
Note:	*p<0.05

Note: *p<0.05



	Income (2021)
Education	-5.47
	(3.13)
	(3.13)
Ethnicity: Black	-8.43*
	(4.01)
	(1.01)
Ethnicity: Hispanic	-3.80
	(6.48)
	(51.15)
Ethnicity: Other	5.49*
•	(0.36)
	()
Constant	-29.11*
	(4.97)
	()
Observations	1,813
Adjusted R ²	0.12
Residual Std. Error	39.46 (df = 1808)
F Statistic	60.85^* (df = 4; 1808)
	,
Note:	*p<0.05