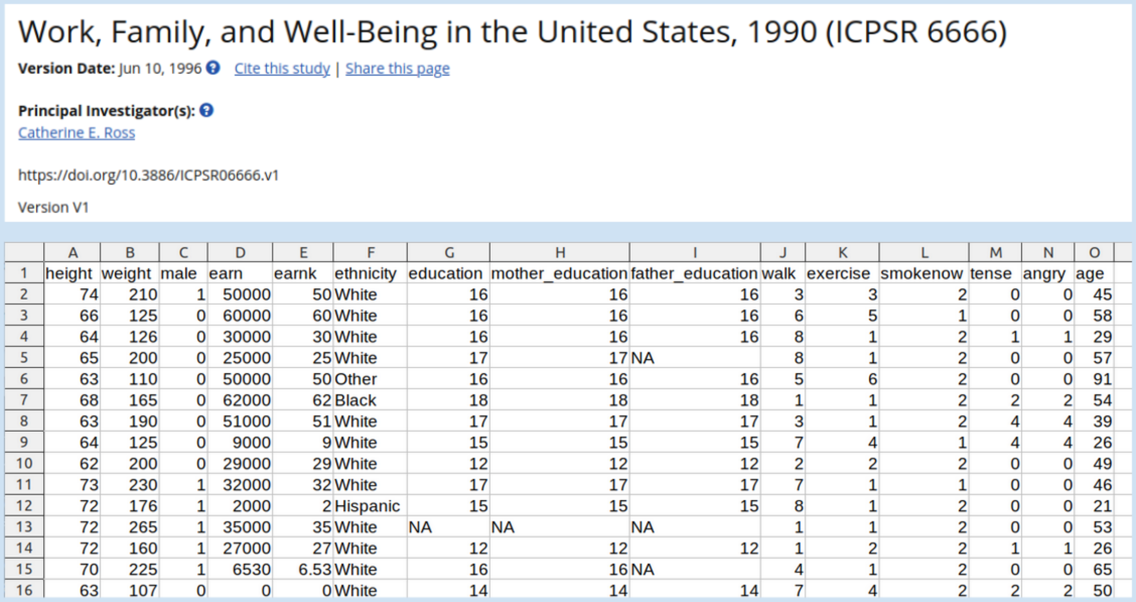
Today's Agenda

Extending the OLS Regression

1. Dichotomous predictors (Dummies)
2. 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

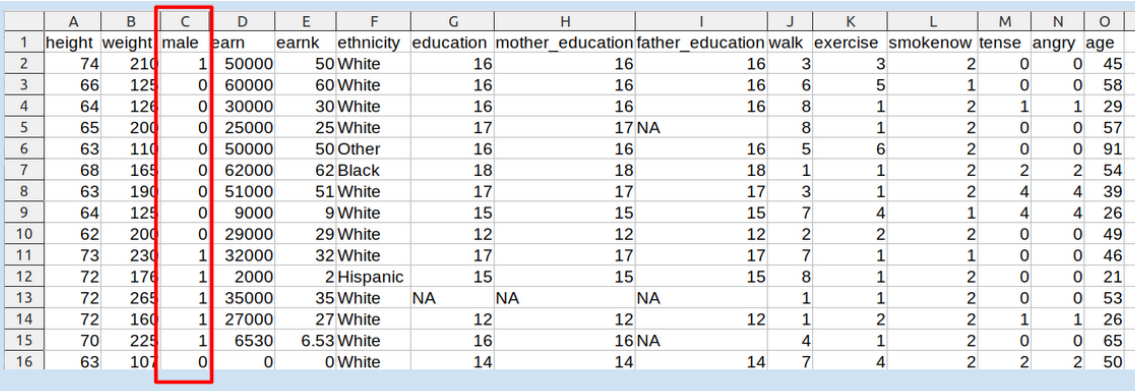
Principal Investigator(s):
Catherine E. Ross

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

Version V1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
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

Dichotomous Variables (e.g. Dummies)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
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

Dichotomous Variables (e.g. Dummies)

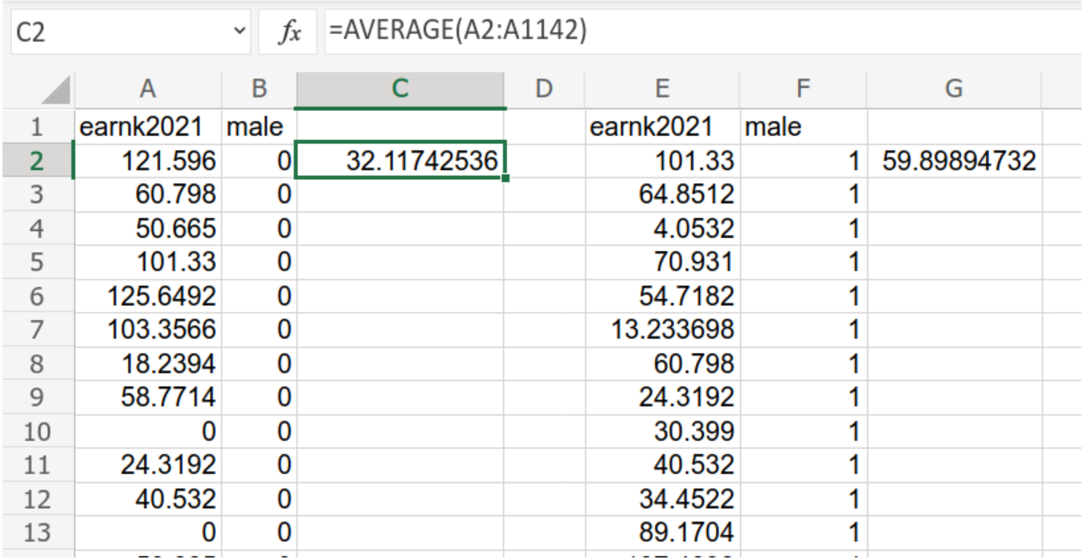
Is their evidence of a gender difference in earned income?

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

Dichotomous Variables (e.g. Dummies)

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 *fx* =AVERAGE(A2:A1142)

	A	B	C	D	E	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	

Dichotomous Variables (e.g. Dummies)

Is their evidence of a gender difference in earned income?

2. Fit an OLS regression of income on gender

Dichotomous Variables (e.g. Dummies)

- 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

- 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

$$\text{Income} = 32.12 + 27.78 \text{ x } (\text{Male})$$

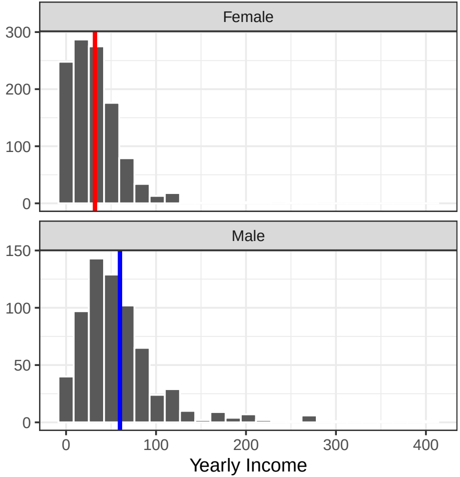
- Men = \$59.9k
- Women = \$32.1k

	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 x (Male)

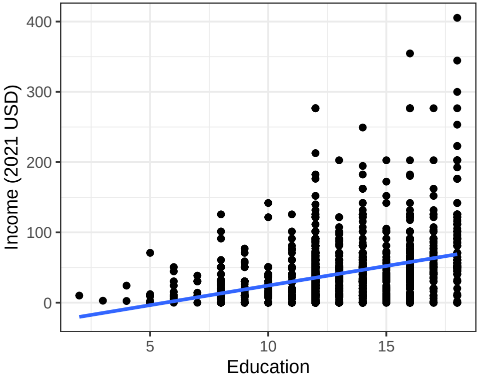
- P(Male = 1) = 32.12 + 27.78 x 1 = 59.9
- P(Male = 0) = 32.12 + 27.78 x 0 = 32.12

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 gender

	Income (Thousands USD)	
	(1)	(2)
Education	5.57^{*}	5.35^{*}
	(0.36)	(0.34)
Male		26.53^{*}
		(1.82)
Constant	-31.34^{*}	-38.28^{*}
	(4.89)	(4.65)
Observations	1,813	1,813
Adjusted R^2	0.11	0.21
Residual Std. Error	39.50 (df = 1811)	37.38 (df = 1810)
F Statistic	235.79^{*} (df = 1; 1811)	238.09^{*} (df = 2; 1810)
Note:		$^{*}p<0.05$

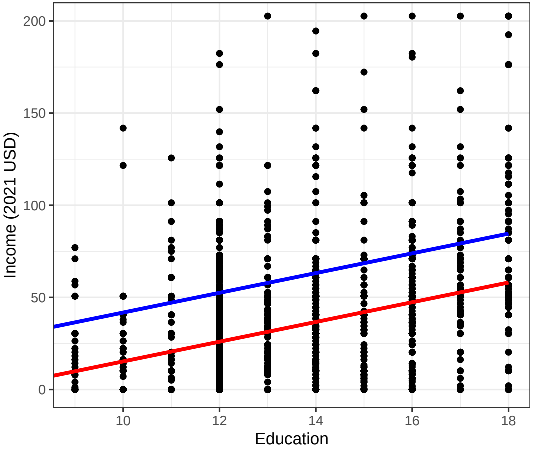
	Income (Thousands USD)
Education	5.35^{*}
	(0.34)
Male	26.53^{*}
	(1.82)
Constant	-38.28^{*}
	(4.65)
Observations	1,813
Adjusted R^2	0.21
Residual Std. Error	37.38 (df = 1810)
F Statistic	238.09^{*} (df = 2; 1810)
Note:	*p<0.05

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

1. Add a sheet
2. Column 1: The levels of education
3. Column 2: Model point estimates for a male across the levels of education
4. Column 3: Model point estimates for a female across the levels of education
5. 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

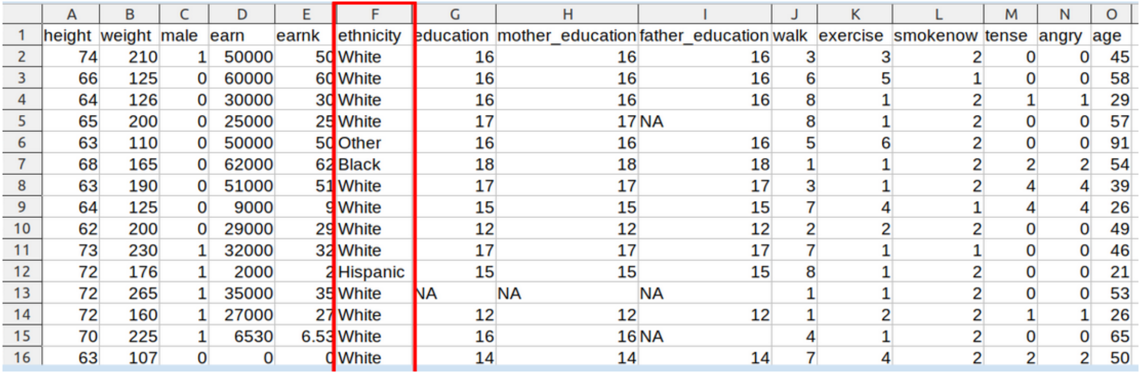


Dummy Variables in OLS Regressions

1. Point estimates produce the group means (with a significance test), and

2. The coefficient on the dummy moves the intercept, not the slope

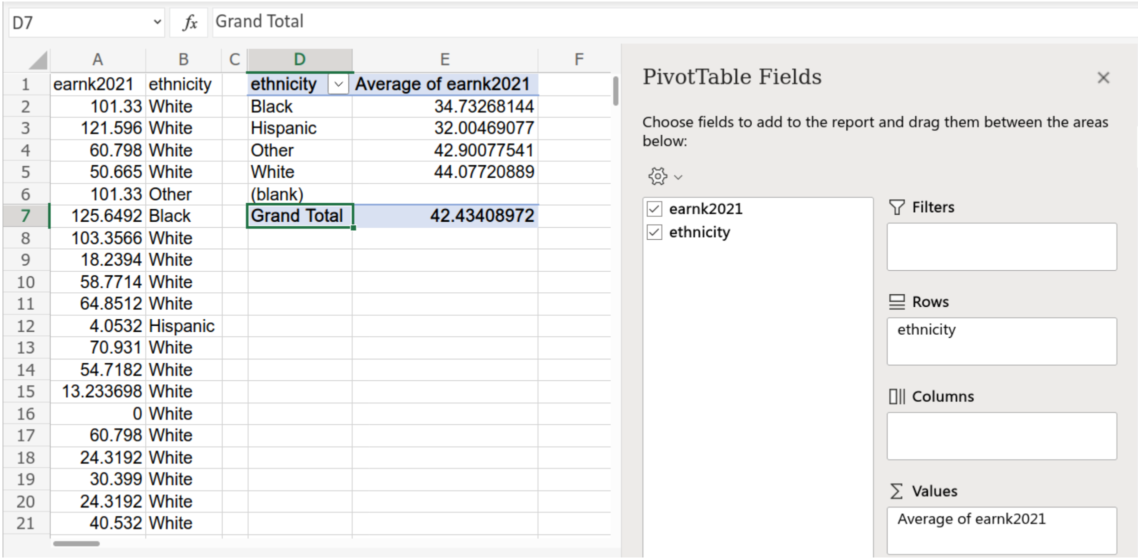
Categorical Variables



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
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

Categorical Variables

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



D7 ⌄ fx Grand Total

	A	B	C	D	E	F
1	earnk2021	ethnicity		ethnicity ⌄	Average of earnk2021	
2	101.33	White		Black	34.73268144	
3	121.596	White		Hispanic	32.00469077	
4	60.798	White		Other	42.90077541	
5	50.665	White		White	44.07720889	
6	101.33	Other		(blank)		
7	125.6492	Black		Grand Total	42.43408972	
8	103.3566	White				
9	18.2394	White				
10	58.7714	White				
11	64.8512	White				
12	4.0532	Hispanic				
13	70.931	White				
14	54.7182	White				
15	13.233698	White				
16	0	White				
17	60.798	White				
18	24.3192	White				
19	30.399	White				
20	24.3192	White				
21	40.532	White				

PivotTable Fields ×

Choose fields to add to the report and drag them between the areas below:

☑ earnk2021
☑ ethnicity

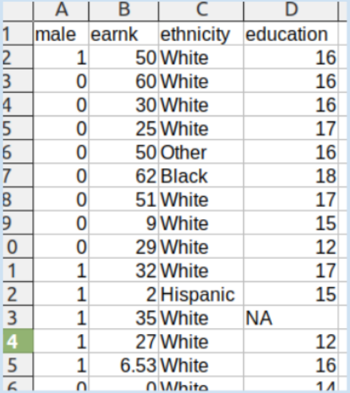
Filters

Rows: ethnicity

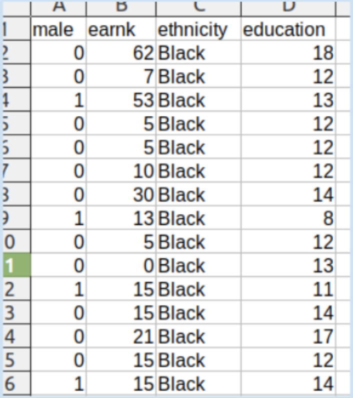
Columns

Values: Average of earnk2021

Using Categorical Variables in Excel: Make Dummies



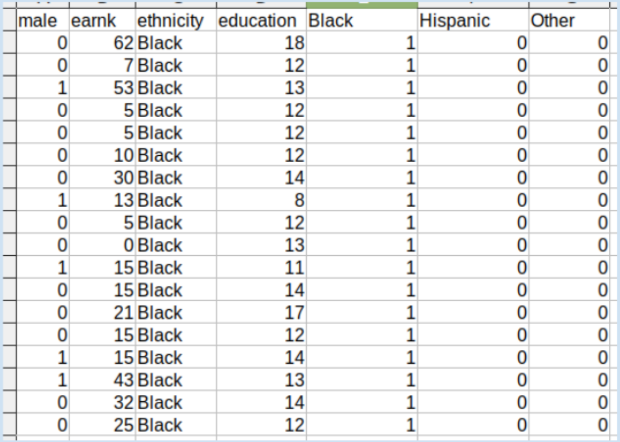
	A	B	C	D
1	male	earnk	ethnicity	education
2	1	50	White	16
3	0	60	White	16
4	0	30	White	16
5	0	25	White	17
6	0	50	Other	16
7	0	62	Black	18
8	0	51	White	17
9	0	9	White	15
0	0	29	White	12
1	1	32	White	17
2	1	2	Hispanic	15
3	1	35	White	NA
4	1	27	White	12
5	1	6.53	White	16
6	0	0	White	14



	A	B	C	D
1	male	earnk	ethnicity	education
2	0	62	Black	18
3	0	7	Black	12
4	1	53	Black	13
5	0	5	Black	12
6	0	5	Black	12
7	0	10	Black	12
8	0	30	Black	14
9	1	13	Black	8
0	0	5	Black	12
1	0	0	Black	13
2	1	15	Black	11
3	0	15	Black	14
4	0	21	Black	17
5	0	15	Black	12
6	1	15	Black	14

1. Sort data by categorical predictor

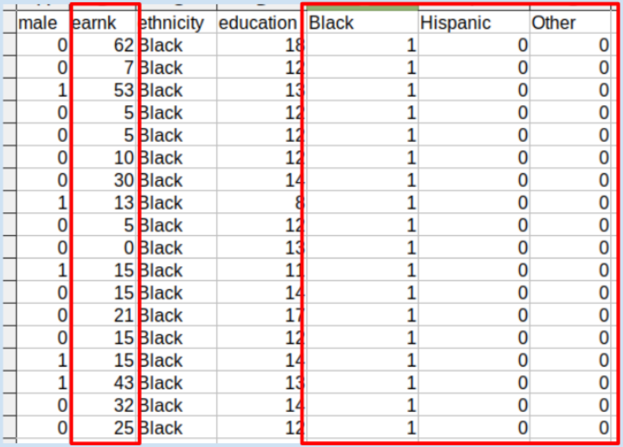
Using Categorical Variables in Excel: Make Dummies



male	earnk	ethnicity	education	Black	Hispanic	Other
0	62	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	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	15	Black	12	1	0	0
1	15	Black	14	1	0	0
1	43	Black	13	1	0	0
0	32	Black	14	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	62	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	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	15	Black	12	1	0	0
1	15	Black	14	1	0	0
1	43	Black	13	1	0	0
0	32	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	44.08^{*}
	(1.08)
Observations	1,815
Adjusted R^2	0.01
Residual Std. Error	41.83 (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

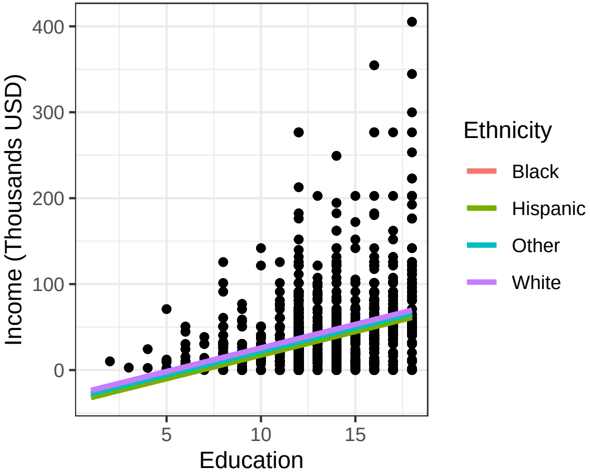
	Income (2021)
Ethnicity: Black	-9.34^{*}
	(3.30)
Ethnicity: Hispanic	-12.07^{*}
	(4.24)
Ethnicity: Other	-1.18
	(6.87)
Constant	44.08^{*}
	(1.08)
Observations	1,815
Adjusted R^2	0.01
Residual Std. Error	41.83 (df = 1811)
Note:	*p<0.05

$$\text{Income} = 44.08 + \text{-}9.34(\text{Black}) + \text{-}12.07(\text{Hispanic}) + \text{-}1.18(\text{Other})$$

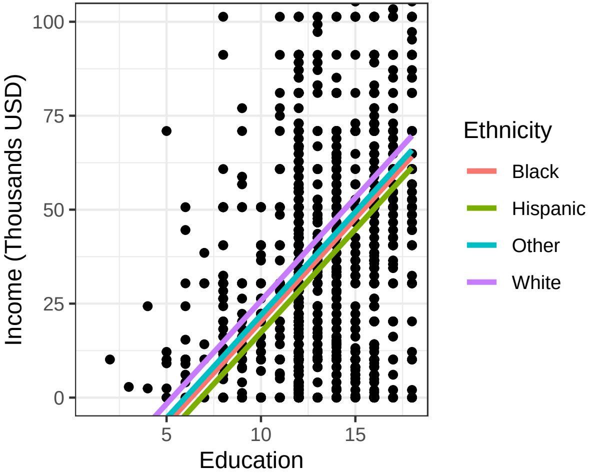
Categorical Variables in OLS Regressions

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)
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



	Income (2021)
Education	−5.47
	(3.13)
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