

Multiple Regression Analysis with Qualitative Information:

1 Outline

- Describing qualitative information
- Using a single dummy independent variable
- Using dummy variables for multiple categories
- Interactions involving dummy variables
- A binary dependent variable (Y variable): The linear probability model

2 Describing Qualitative Information

- "Female" and "Married" are qualitative variable.
- We arbitrarily assign a dummy variable to describe them.

$$female = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

$$married = \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise (of if single)} \end{cases}$$

TABLE 7.1

A Partial Listing of the Data in WAGE1.RAW

<i>person</i>	<i>wage</i>	<i>educ</i>	<i>exper</i>	<i>female</i>	<i>married</i>
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

3 Models with a single dummy independent variable

Consider

$$wage = \beta_0 + \delta_0 \underline{female} + \beta_1 \underline{educ} + u. \quad (1)$$

where

$$female = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

In this case, the δ_0 notation is used to highlight the interpretation of the parameters multiplying dummy variables. In other cases, we can use any notation that is the most convenient.

$$\begin{aligned} \textcircled{1} E(wage | female, educ) &= E(\beta_0 + \delta_0 female + \beta_1 educ + u | female, educ) \\ &= \beta_0 + \delta_0 female + \beta_1 educ + E(u | female, educ) \\ &= \beta_0 + \delta_0 female + \beta_1 educ \quad \downarrow = 0 \\ &\quad \text{(ass MLR1-4 holds)} \end{aligned}$$

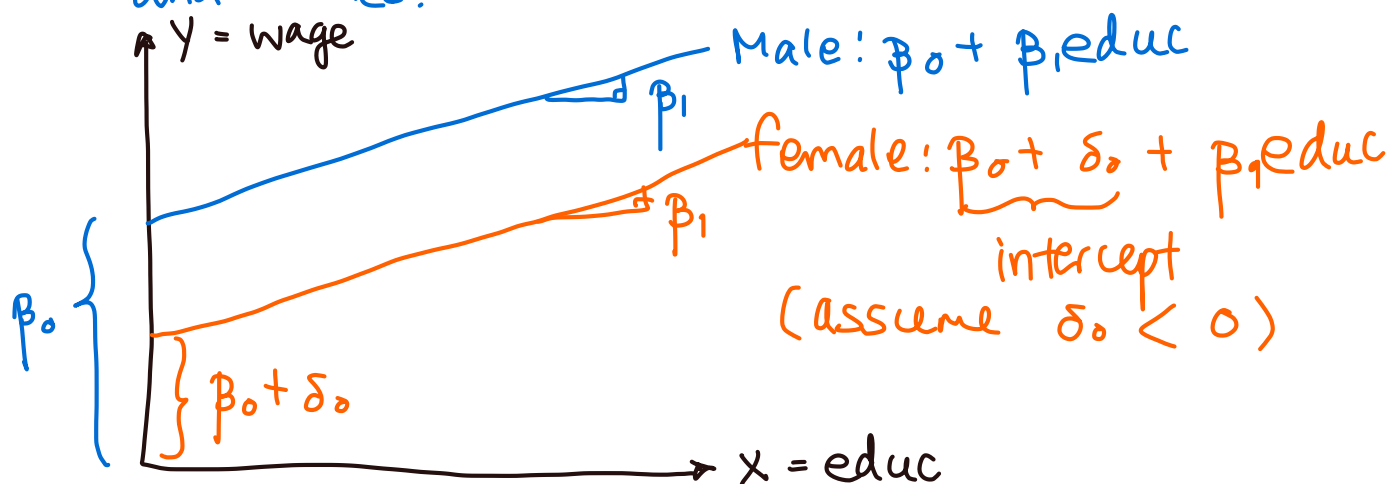
② Thus

$$\textcircled{f} : E(wage | female=1, educ) = \beta_0 + \delta_0(1) + \beta_1 educ = \beta_0 + \delta_0 + \beta_1 educ$$

$$\textcircled{m} : E(wage | female=0, educ) = \beta_0 + \delta_0(0) + \beta_1 educ = \beta_0 + \beta_1 educ$$

$$\begin{aligned} \delta_0 &= E(wage | female=1, educ) - E(wage | female=0, educ) \\ \text{or } \delta_0 &= E(wage | female, educ) - E(wage | male, educ) \end{aligned}$$

* given the same value of educ (same education level), δ_0 is the difference in the expected wage of females and males.



→ By the way we model this regression function, "female" is going to give a constant impact on wage, regardless of the level of educ.

4 It is not possible to include all of the dummy alternatives in the same model (as long as there is an intercept in the model).

- If we include all alternatives of a dummy variable in the same model, we will face the "perfect collinearity" problem.

For example:

$$\text{wage} = \beta_0 X_0 + \delta_1 \text{female} + \beta_2 \text{educ} + \delta_3 \text{male} + u$$

↑ intercept = 1

$$X_0 = X_1 + X_3$$

$$1 = \text{female} + \text{male}$$

$$\text{female} = \text{male} + 1$$

id.	female	male	X_0
1	1	0	1
2	1	0	1
3	0	1	1
4	0	1	1
...	...	...	...
99	...	...	...

or If there are "n" categories, we omit "1" category to avoid multi collinearity. $\text{winter} = \text{spring} + \text{summer} + \text{fall}$ ✓
 $\text{winter} = 1 - \text{spring} - \text{summer} - \text{fall}$ ✓

winter = $\begin{cases} 1 & \text{if winter} \\ 0 & \text{otherwise} \end{cases}$

spring = $\begin{cases} 1 & \text{if spring} \\ 0 & \text{otherwise} \end{cases}$
etc.

id	winter	spring	summer	fall	X_0
1	1	0	0	0	1
2	1	0	0	0	1
3	0	1	0	0	1
4	0	1	0	0	1
...	...	...	...	...	...
1	0	0	1	0	1
1	0	0	1	0	1
1	0	0	1	0	1

- At least one alternative has to be dropped. We treat the dropped alternative as the "BASE GROUP" or "BASELINE" or "BENCHMARK GROUP".

$\begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$ $\begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$

`. regress lwage female male married educ exper`
 note: male omitted because of collinearity

↑ intercept = β_0

↑ in this case, male

Source	SS	df	MS	Number of obs =	526
Model	54.3265253	4	13.5816313	F(4, 521) =	75.27
Residual	94.0032262	521	.180428457	Prob > F	= 0.0000
				R-squared	= 0.3663
				Adj R-squared	= 0.3614
Total	148.329751	525	.28253286	Root MSE	= .42477

Female workers are expected to have less wage compared to male workers.

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
female	-.251146	.0377061	-8.62	0.000	-.3991892 -.25104
male	0 (omitted)				
married	.1380145	.0411197	3.36	0.001	.0572338 .2187953
educ	.0872644	.0071554	12.20	0.000	.0732075 .1013213
exper	.0076213	.0015314	4.98	0.000	.0046129 .0106297
_cons	.4690918	.1040575	4.51	0.000	.264668 .6735156

5 Using dummy variables for multiple categories

Case 1 We can use many dummy variables in the same model

Consider a model which includes 2 dummy variables— *female* and *married*.

$$\log(\text{wage}) = \beta_0 + \delta_0 \text{female} + \delta_1 \text{married} + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \beta_4 \text{tenure} + \beta_5 \text{tenure}^2 + u.$$

$\begin{cases} 1 & \text{if female} \\ 0 & \text{if otherwise} \end{cases}$
 $\begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise} \end{cases}$

```
regress lwage female married educ exper expersq tenure tenursq
```

Source	SS	df	MS	Number of obs = 526		
Model	65.6482326	7	9.37831895	F(7, 518) = 58.76		
Residual	82.6815188	518	.159616832	Prob > F = 0.0000		
Total	148.329751	525	.28253286	R-squared = 0.4426		
				Adj R-squared = 0.4351		
				Root MSE = .39952		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-.2901838	.0361121	-8.04	0.000	-.3611279	-.2192396
married	.0529219	.0407561	1.30	0.195	-.0271456	.1329894
educ	.0791547	.0068003	11.64	0.000	.0657952	.0925143
exper	.0269535	.0053258	5.06	0.000	.0164907	.0374163
expersq	-.0005399	.0001122	-4.81	0.000	-.0007603	-.0003196
tenure	.0312962	.0068482	4.57	0.000	.0178426	.0447499
tenursq	-.0005744	.0002347	-2.45	0.015	-.0010355	-.0001134
_cons	.4177837	.0988662	4.23	0.000	.2235557	.6120116

$\hat{\beta}$

2) δ_1 measures the impact of be married.
(marriage premium) But since $|t| < 1.96$ or $p > 0.05$,
Comments: we do not reject H_0 of no impact.

1) δ_0 measures the expected difference between female & male workers given the same marital status and other factors.

$$\frac{\partial \log(\text{wage})}{\partial \text{female}} = \frac{1}{\text{wage}} \frac{d \text{wage}}{d \text{female}} = -0.29$$

female workers are expected to earn less than male workers by 29.02%, holding other factors the same.

$$100 \cdot \frac{1}{\text{wage}} \frac{d \text{wage}}{d \text{female}} = 100 \cdot -0.29$$

$$\frac{\% \Delta \text{wage}}{\partial \text{female}} = 29.02\%$$

	♀	♂
mar	marrfem	marmale
sim	singfem	singmale

Basecase

8. Multiple Regression Analysis with Qualitative Information: 85

Consider a model which includes dummy variables for each gender/marital status combination— marrmale, marrfem and singfem. (or singmale used as the basecase)

$$\log(\text{wage}) = \beta_0 + \delta_0 \text{marrmale} + \delta_1 \text{marrfem} + \delta_2 \text{singfem} + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \beta_4 \text{tenure} + \beta_5 \text{tenure}^2 + u. \quad (8.1)$$

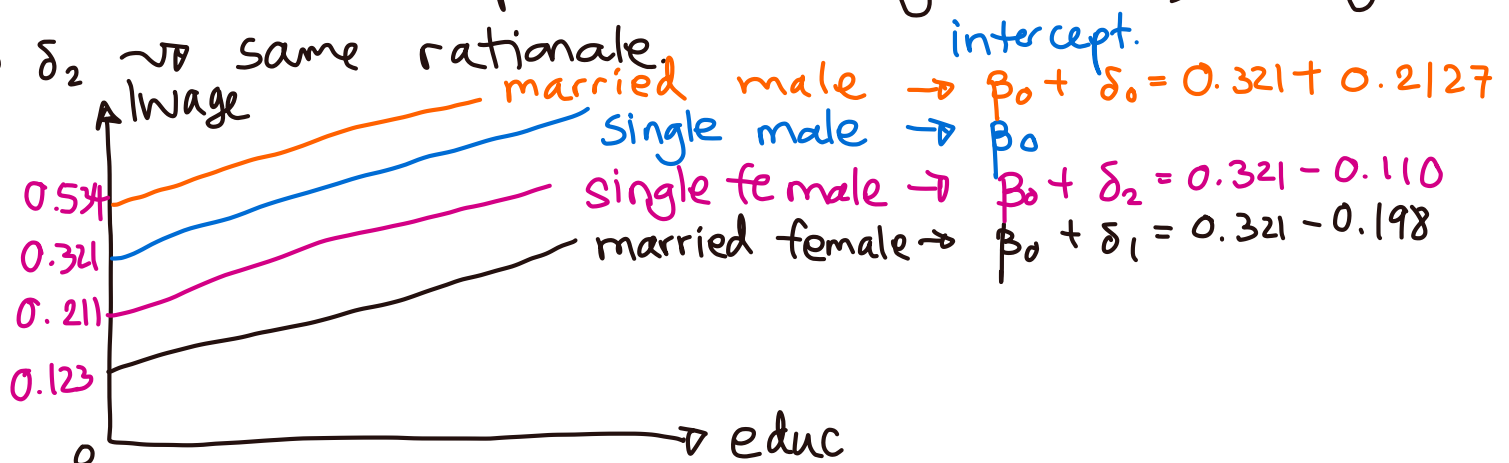
regress lwage marrmale marrfem singfem educ exper expersq tenure tenursq

Source	SS	df	MS	Number of obs = 526
Model	68.3617623	8	8.54522029	F(8, 517) = 55.25
Residual	79.9679891	517	.154676961	Prob > F = 0.0000
Total	148.329751	525	.28253286	R-squared = 0.4609
				Adj R-squared = 0.4525
				Root MSE = .39329

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
δ_0 marrmale	.2126757	.0553572	3.84	0.000	.103923 .3214284
δ_1 marrfem	-.1982676	.0578355	-3.43	0.001	-.311889 -.0846462
δ_2 singfem	-.1103502	.0557421	-1.98	0.048	-.219859 -.0008414
educ	.0789103	.0066945	11.79	0.000	.0657585 .092062
exper	.0268006	.0052428	5.11	0.000	.0165007 .0371005
expersq	-.0005352	.0001104	-4.85	0.000	-.0007522 -.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031 .0423719
tenursq	-.0005331	.0002312	-2.31	0.022	-.0009874 -.0000789
_cons	.3213781	.100009	3.21	0.001	.1249041 .5178521

Comments: This regression is not the same as the previous one. It uses "Single Male" as the base group. (The previous one use male & Single as 2 base groups)

- δ_0 measures the expected diff. in wage of married male as compared with single males, holding other factors constant.
- δ_1 measures the expected diff. in married female as compared with single males, holding —.
- δ_2 same rationale



Case 2 We can use dummy variables to represent multiple categories of a variable. Consider the relationship between law school rankings and starting salaries

$$\log(\text{salary}) = \beta_0 + \delta_0 \text{top10} + \delta_1 r11_25 + \delta_3 r26_40 + \delta_4 r41_60 + \beta_1 \text{LSAT} + \beta_2 \text{GPA} + \beta_3 \log(\text{libvol}) + \beta_4 \log(\text{cost}) + u.$$

where top10 , $r11_25$, $r26_40$, $r41_60$ would be equal to 1 when the variable rank falls into the appropriate range.

** Rank below 60 would be the base case.

* In many cases the "range of value" serve as a better explanatory variable than the "value" itself.
e.g. age may explain the model better if split into generations.

regress lsalary top10 r11_25 r26_40 r41_60 LSAT GPA llibvol lcost

Source	SS	df	MS	Number of obs =	136
Model	9.16538532	8	1.14567316	F(8, 127) =	120.15
Residual	1.2109665	127	.009535169	Prob > F =	0.0000
Total	10.3763518	135	.076861865	R-squared =	0.8833
				Adj R-squared =	0.8759
				Root MSE =	.09765

lsalary		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
top10	✓	.5393428	.053542	10.07	0.000	.4333927 .6452928
r11_25	✓	.4716199	.0390921	12.06	0.000	.3942637 .548976
r26_40	✓	.2790977	.0346972	8.04	0.000	.2104383 .3477571
r41_60	✓	.182382	.0283098	6.44	0.000	.126362 .238402
LSAT		.0060482	.0034919	1.73	0.086	-.0008616 .012958
GPA		.1305893	.0818678	1.60	0.113	-.0314122 .2925908
llibvol		.0725522	.0289213	2.51	0.013	.0153221 .1297824
lcost		.0249169	.0283224	0.88	0.381	-.031128 .0809619
_cons		8.363103	.4457314	18.76	0.000	7.481081 9.245125

Comments:

rank	top10	r11-25	r26-40	etc.
1	1	0	0	
2	1	0	0	
3	1	0	0	
⋮	⋮	⋮	⋮	
10	1	0	0	
11	0	1	0	
12	0	1	0	
⋮	⋮	⋮	⋮	
25	0	1	0	
26	0	1	0	
⋮	⋮	⋮	⋮	
40	0	0	1	
⋮	⋮	⋮	⋮	

- 1) δ_0 measures the difference in expected $\log(\text{salary})$ of a law-school graduate from a top 10 university compared to expected $\log(\text{salary})$ of those who graduated from the school ranked 61th and worse.
- 2) δ_1 use the same rationale.