9. Dummy Variable Regression Models

In the previous chapter, the dependent and independent variables in our multiple regression models have had **quantitative** meaning. For example, the salary of CEO,annual firm sales,return on equity in percent, and return on firm's stock. In each case the magnitude of the variable conveys useful information.

However, in the empirical work, we must also incorporate **qualitative factors** into regression models. The gender or race of an individual, the industry of a firm (manufacturing, retail, and so on), and the region in Thailand where a city is located (north, south, west, and so on) are all considered as the qualitative factors.

9.1 Describing Qualitative Information

Normally, qualitative factors often come in the form of binary information:

Example:

[1] A person is female or male or female.

[2] A firm offers a certain kind of employee pension plan or it does not.

[3] A farm is located nearby the dam or not.

All of these examples, the relevant information can be captured by defining a **binary variable** or a zero-one variable.

In econometrics, binary variables are most commonly called **dummy variables**, although this name is not especially descriptive.

In defining a dummy variable, we must decide which event is assigned the value one and which is assigned the value zero.

Question: Why do we use the the values zero and one to describe qualitative information?

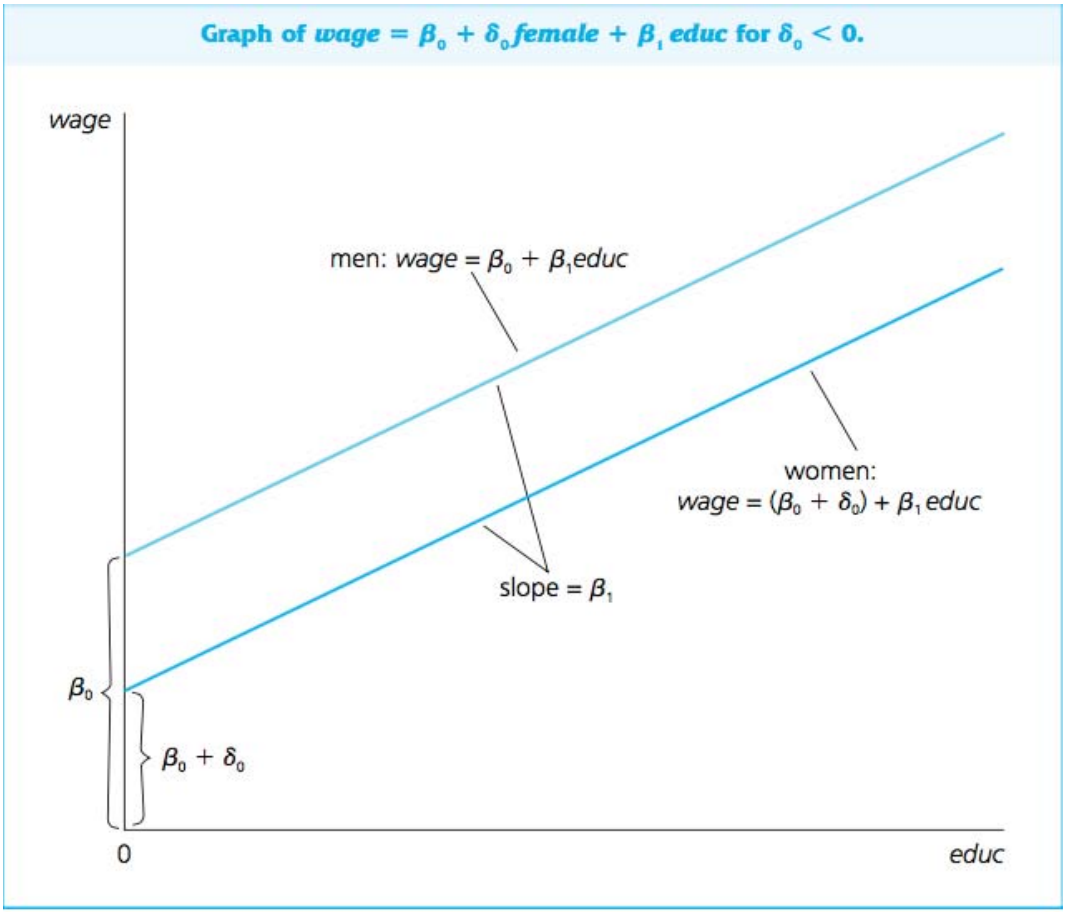
Answer: These values are arbitrary: any two different values would do. The real benefit of capturing qualitative information using zero-one variable is that it leads to regression models where the parameters have very natural interpretations.

9.1.1 A Single Dummy Independent Variable

Suppose we would like to estimate the following simple model of hourly wage determination:

$$wage_i = \beta_0 + \delta_0 \text{female} + \beta_1 \text{edu} + u_i$$

Figure 9.1: Graph of Wage



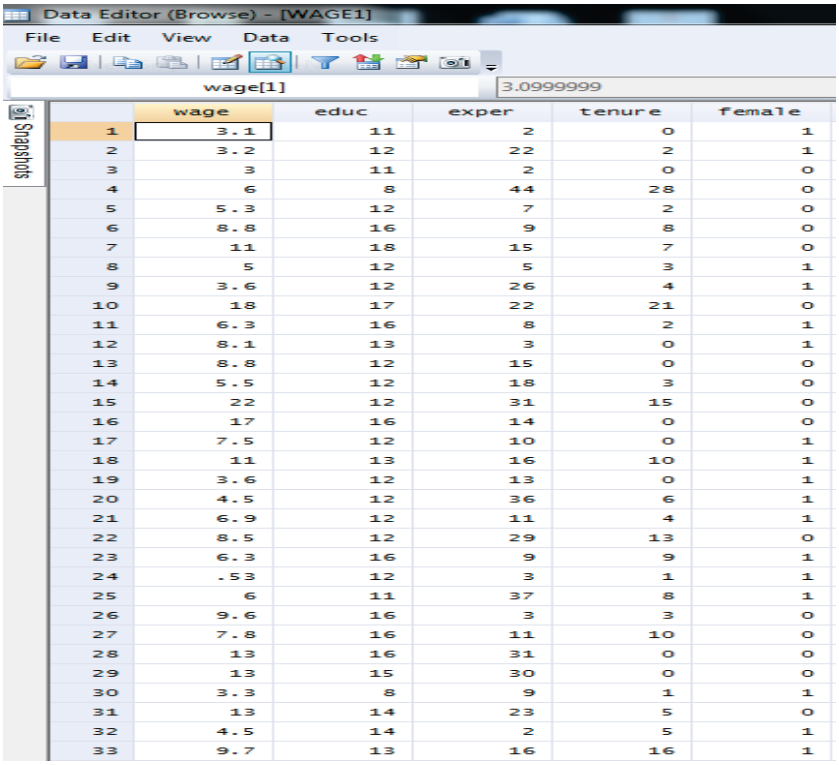
Now, we added more variables into the wage model. Taking males as the base group, a model that controls for experience and tenure in addition to education is

$$wage_i = \beta_0 + \delta_0 \text{female} + \beta_1 \text{edu} + \beta_2 \text{exper} + \beta_3 \text{tenure} + u_i$$

If edu, expert, and tenure are all relevant productivity characteristics, the null hypothesis of no difference between men and women (No wage discrimination) is:

In table 9.1, it represents the partial listing of the sample data of wage model. We see that Person 1 is female, Person 2 is female , Person 3 is male, and so on.

Table 9.1: A Partial Listing of the Wage Data.



	wage	educ	exper	tenure	female
1	3.1	11	2	0	1
2	3.2	12	22	2	1
3	3	11	2	0	0
4	6	8	44	28	0
5	5.3	12	7	2	0
6	8.8	16	9	8	0
7	11	18	15	7	0
8	5	12	5	3	1
9	3.6	12	26	4	1
10	18	17	22	21	0
11	6.3	16	8	2	1
12	8.1	13	3	0	1
13	8.8	12	15	0	0
14	5.5	12	18	3	0
15	22	12	31	15	0
16	17	16	14	0	0
17	7.5	12	10	0	1
18	11	13	16	10	1
19	3.6	12	13	0	1
20	4.5	12	36	6	1
21	6.9	12	11	4	1
22	8.5	12	29	13	0
23	6.3	16	9	9	1
24	.53	12	3	1	1
25	6	11	37	8	1
26	9.6	16	3	3	0
27	7.8	16	11	10	0
28	13	16	31	0	0
29	13	15	30	0	0
30	3.3	8	9	1	1
31	13	14	23	5	0
32	4.5	14	2	5	1
33	9.7	13	16	16	1

Table 9.2: The command function to estimate the wage model in STATA program

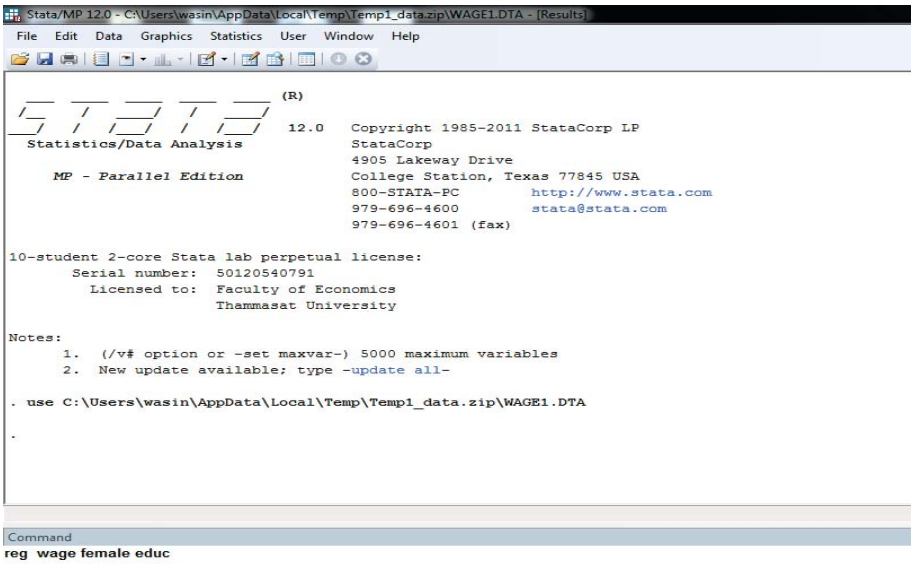


Table 9.3: $wage_i = \beta_0 + \delta_0 \textbf{female} + \beta_1 \textbf{edu} + \beta_2 \textbf{exper} + \beta_3 \textbf{tenure} + u_i$

```
. reg wage female educ exper tenure
```

Source	SS	df	MS
Model	2603.10658	4	650.776644
Residual	4557.30771	521	8.7472317
Total	7160.41429	525	13.6388844

Number of obs	= 526
F(4, 521)	= 74.40
Prob > F	= 0.0000
R-squared	= 0.3635
Adj R-squared	= 0.3587
Root MSE	= 2.9576

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-1.810852	.2648252	-6.84	0.000	-2.331109	-1.290596
educ	.5715048	.0493373	11.58	0.000	.4745802	.6684293
exper	.0253959	.0115694	2.20	0.029	.0026674	.0481243
tenure	.1410051	.0211617	6.66	0.000	.0994323	.1825778
_cons	-1.567939	.7245511	-2.16	0.031	-2.991339	-.144538

Example: the Hourly Wage Equation:

$$\begin{aligned} \widehat{\text{wage}} &= -1.5679 - 1.8109 \text{ female} + 0.5715 \text{ edu} + 0.025 \text{ exper} + 0.141 \text{ tenure} \\ &= (0.7246) \quad (0.2648) \quad (0.0493) \quad (0.0116) \quad (0.0212) \end{aligned} \tag{9.1}$$

$R^2 = 0.3635 \quad n = 526$

Interpret the model:

The intercept:

Table 9.4: $wage_i = \beta_0 + \delta_0 \textbf{female} + u_i$

```
reg wage female
```

Source	SS	df	MS	Number of obs =	526
				F(1, 524) =	68.54
Model	828.220467	1	828.220467	Prob > F =	0.0000
Residual	6332.19382	524	12.0843394	R-squared =	0.1157
				Adj R-squared =	0.1140
Total	7160.41429	525	13.6388844	Root MSE =	3.4763

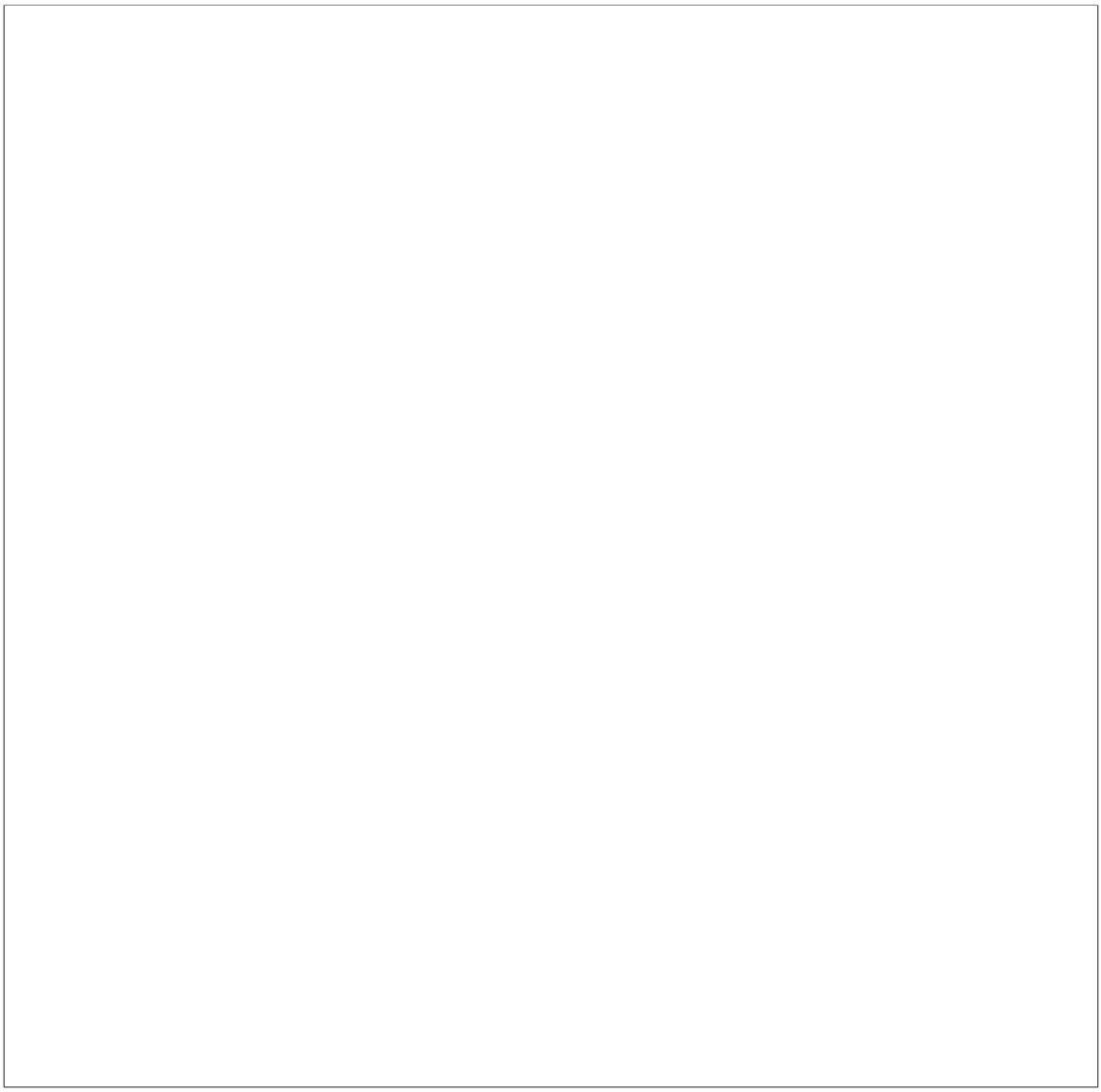
wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-2.51183	.3034092	-8.28	0.000	-3.107878	-1.915782
_cons	7.099489	.2100082	33.81	0.000	6.686928	7.51205

The coefficient on female

It is informative to compare the coefficient on female in the above equation to the estimate we get when all other explanatory variables are dropped from the equation:

$$\begin{aligned} \widehat{\text{wage}} &= 7.0995 - 2.5118 \ \text{female} \\ se &= (0.2100) \qquad (0.3034) \end{aligned} \tag{9.2}$$

$R^2 = 0.1157 \quad n = 526$



9.2 Interpreting Coefficients on Dummy Explanatory Variables When the Dependent Variable is log(y)

In this section, we will study a model that has the dependent variable appearing in logarithmic form, with one or more dummy variables approving as independent variables.
Question: How do we interpret the dummy variable coefficients in this case?

Answer: Not surprisingly, the coefficients have a percentage interpretation.

Let us reestimate the wage equation , using log (wage) as the dependent variable and adding quadratics in expert and tenure:

$$log(wage_i) = \beta_0 + \delta_0 female + \beta_1 edu + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u_i$$

The Stata result is shown in table 9.5.

Table 9.5:

```
reg lwage female educ exper expersq tenure tenursq
```

Source	SS	df	MS		
Model	65.3791009	6	10.8965168	Number of obs	= 526
Residual	82.9506505	519	.159827843	F(6, 519)	= 68.18
				Prob > F	= 0.0000
Total	148.329751	525	.28253286	R-squared	= 0.4408
				Adj R-squared	= 0.4343
				Root MSE	= .39978

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.296511	.0358055	-8.28	0.000	-.3668524	-.2261696
educ	.0801967	.0067573	11.87	0.000	.0669217	.0934716
exper	.0294324	.0049752	5.92	0.000	.0196585	.0392063
expersq	-.0005827	.0001073	-5.43	0.000	-.0007935	-.0003719
tenure	.0317139	.0068452	4.63	0.000	.0182663	.0451616
tenursq	-.0005852	.0002347	-2.49	0.013	-.0010463	-.0001241
_cons	.416691	.0989279	4.21	0.000	.2223425	.6110394

Example: Log Hourly Wage Equation:

$$\begin{aligned}\widehat{\log(\text{wage})} &= 0.4167 - 0.2965\ \text{female} + 0.0802\ \text{edu} + 0.0294\ exper - 0.0006\ exper^2 \\ &\quad (0.0989) \qquad (0.0358) \qquad (0.0068) \qquad (0.0050) \qquad (0.0001) \\ &\quad +0.0317\ tenure - 0.0006\ tenure^2 \\ &\quad (0.0068) \qquad (0.0002)\end{aligned} \tag{9.3}$$

$R^2 = 0.4408 \quad n = 526$

Interpret the model:

The coefficient on female

9.3 Using Dummy Variables for Multiple Categories

We can use several dummy independent variables in the same equation. For example, we could add the dummy variable **married** to the wage model.

The previous model:

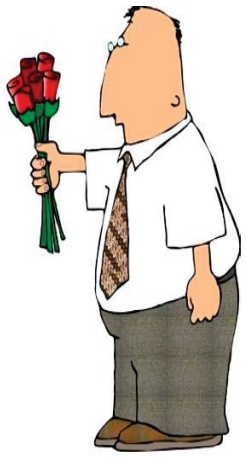
$$log(wage_i) = \beta_0 + \delta_0 female + \beta_1 edu + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u_i$$

Now, Let us estimate a model that allows for wage differences among four groups:

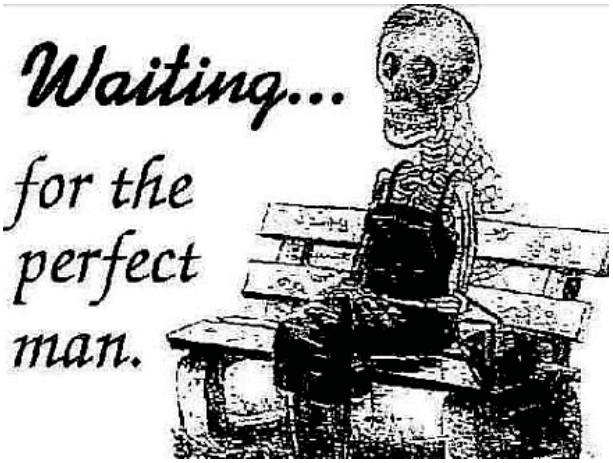
[1.] Married Men

[2] Married Women

[3] Single Men



[4] Single Women



To do this, we must select a base group:

Now, we need to define dummy variables for each of the remaining groups.

Therefore, our model is:

$$log(wage_i) = \beta_0 + \delta_0 marrmale + \delta_1 marrfem + + \delta_2 singfem +$$

$$\beta_1 edu + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u_i$$

We of course drop the dummy variable (female). (Why?)

Table 9.5:

```
reg lwage marrmale marrfem singfem educ exper expersq tenure tenursq
```

Source	SS	df	MS		
Model	68.3617623	8	8.54522029	Number of obs	= 526
Residual	79.9679891	517	.154676961	F(8, 517)	= 55.25
				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]
marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
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

$$\begin{aligned}\widehat{\log(\text{wage})} &= 0.3214 + 0.2127\ \text{marrmale} - 0.1983\ \text{marrfem} - 0.1104\ sing fem \\ &\quad (0.1000)\qquad (0.0554)\qquad (0.0578)\qquad (0.0557) \\ &\quad +0.0789\ edu + 0.0268\ exper - 0.0005\ exper^2 + 0.0291\ tenure - 0.0005\ tenure^2 \\ &\quad (0.0067)\qquad (0.0268)\qquad (0.0001)\qquad (0.0068) + \qquad (0.0002)\end{aligned} \tag{9.4}$$

$R^2 = 0.4609 \quad n = 526$

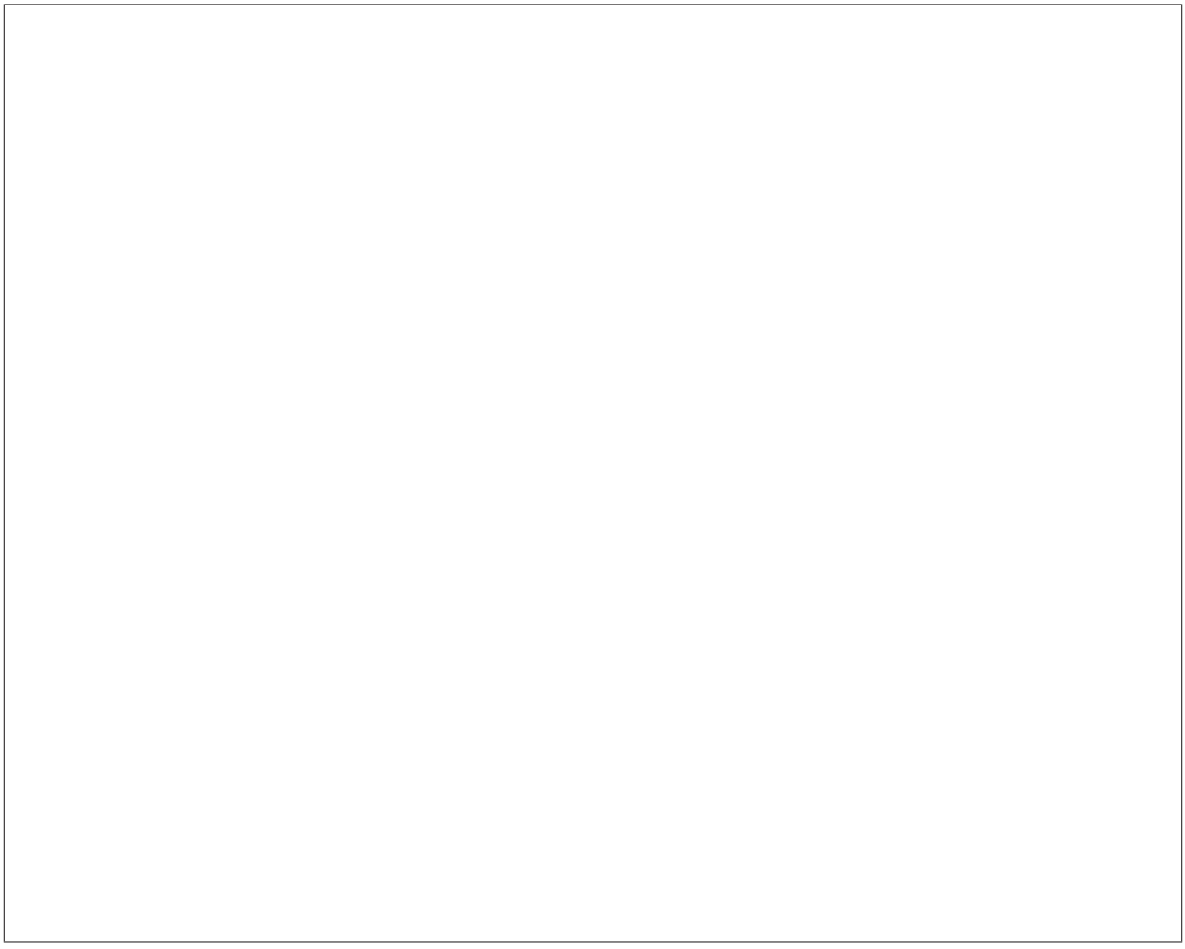
Interpret the model:

9.3.1 Interactions Involving Dummy Variables

8.5.1 The Interactions Among Dummy Variables:

We can recast the model by adding an **interaction term** between female and married to the model where female and married appear separately. This allows the marriage premium to depend on gender. The estimated model with the female-married interaction term is :

$$\begin{aligned}\widehat{\log(\text{wage})} &= 0.321 - 0.110\ \text{female} + 0.213\ \text{married} \\ &\quad (0.100) \qquad (0.056) \qquad (0.055) \\ &\quad +0.301\ \ female \cdot married + ..., \\ &\quad (0.072)\end{aligned} \tag{9.5}$$



8.5.2 The interaction between Dummy Variable/s and Explanatory Variable/s: the Allowing for the Different Slopes

There are also occasions for interacting dummy variables with explanatory variables that are not dummy variables to allow for a **difference in slope.**

To see the interaction between female and edu, we can rewrite the model as follow:

$$wage_i = \beta_0 + \delta_0 \text{female} + \beta_1 \text{edu} + \delta_1 female \cdot edu + u_i$$

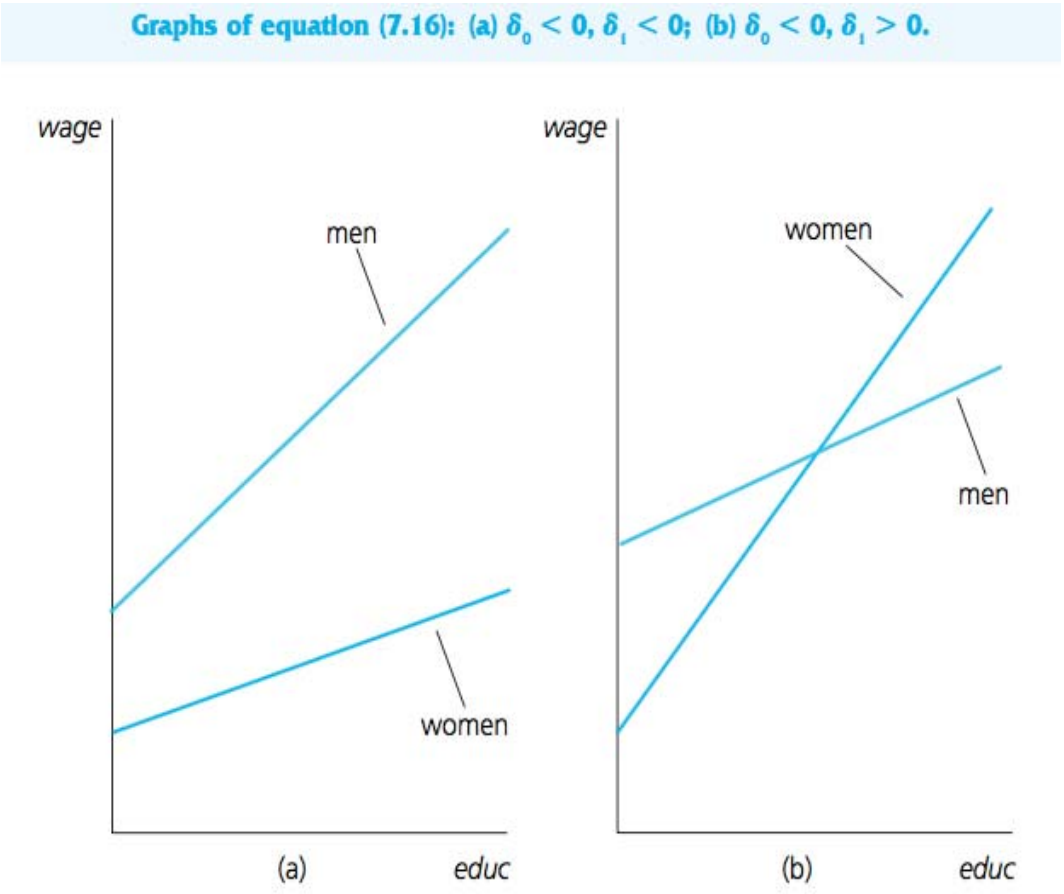
Men Group we plug female =0

Therefore:

Women Group we plug female =1

Therefore:

Figure 9.2: Graph of the Wage Model with an Interaction between female and education



Example

Table 9.5:

```
gen femed = female*educ

reg lwage female educ femed exper expersq tenure tenursq
```

Source	SS	df	MS	Number of obs =	526
				F(7, 518) =	58.37
Model	65.4081534	7	9.34402192	Prob > F =	0.0000
Residual	82.921598	518	.160080305	R-squared =	0.4410
				Adj R-squared =	0.4334
Total	148.329751	525	.28253286	Root MSE =	.4001

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.2267886	.1675394	-1.35	0.176	-.5559289	.1023517
educ	.0823692	.0084699	9.72	0.000	.0657296	.0990088
femed	-.0055645	.0130618	-0.43	0.670	-.0312252	.0200962
exper	.0293366	.0049842	5.89	0.000	.019545	.0391283
expersq	-.0005804	.0001075	-5.40	0.000	-.0007916	-.0003691
tenure	.0318967	.006864	4.65	0.000	.018412	.0453814
tenursq	-.00059	.0002352	-2.51	0.012	-.001052	-.000128
_cons	.388806	.1186871	3.28	0.001	.1556388	.6219732

$$\begin{aligned}\widehat{\log(\text{wage})} &= 0.3889 - 0.2268\ \text{female} + 0.082\ \text{edu} - 0.0056\ female \cdot edu \\ &\quad (0.1187)\qquad (0.1675)\qquad (0.0085)\qquad (0.0131) \\ &\quad +0.0293\ exper - 0.0006\ exper^2 + 0.0319\ tenure - 0.00059\ tenure^2 \\ &\quad (0.0050)\qquad (0.0001)\qquad (0.0069) + \qquad (0.0002)\end{aligned} \tag{9.6}$$

$R^2 = 0.4410 \quad n = 526$