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 arbitarily assign a dummy variable to decribe them.

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

TABLE 7.1

A Partial Listing of the Data in WAGE1.RAW

person	*wage*	*educ*	*exper*	*female*	*married*
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

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u.$$

where

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

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.

4 It is not possible to include all of the dummy alternatives in the same model

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

For example:

$$\begin{aligned} 1 &= female + male \\ female &= male + 1 \end{aligned}$$

or

$$\begin{aligned} 1 &= winter + spring + summer + fall \\ winter &= 1 - spring - summer - fall \end{aligned}$$

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

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

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

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3251146   .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 $married$.

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

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

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

Comments:

Consider a model which includes dummy variables for each gender/marital status combination– $marrmale$, $marrfem$ and $singfem$.

$$\log(wage) = \beta_0 + \delta_0 marrmale + \delta_1 marrfem + \delta_3 singfem + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u. \quad (8.1)$$

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

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

Comments:

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

$$\log(salary) = \beta_0 + \delta_0 top10 + \delta_1 r11\_25 + \delta_3 r26\_40 + \delta_4 r41\_60 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + u.$$

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.

```
. regress lsalary top10 r11_25 r26_40 r41_60 LSAT GPA llibvol lcost
```

Source	SS	df	MS
Model	9.16538532	8	1.14567316
Residual	1.2109665	127	.009535169
Total	10.3763518	135	.076861865

Number of obs	136
F(8, 127)	120.15
Prob > F	0.0000
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:

6 Interactions involving dummy variables

Case 1 Interactions among dummies

** We can use interactions among dummies to account for the effect of each combination of dummies as well:

A different way to estimate eq.(8.1) is

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0 female + \delta_1 married + \delta_3 female \cdot married + \beta_1 educ \\ &\quad + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u.\end{aligned}$$

where $female{\cdot}married = female \times married$.

```
. gen female_married = female*married

. regress lwage female married female_married educ exper expersq tenure tenursq

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

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.1103502   .0557421    -1.98   0.048     -.219859   -.0008414
     married |   .2126757   .0553572     3.84   0.000      .103923    .3214284
female_married |  -.3005931    .071767    -4.19   0.000    -.4415838   -.1596024
        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:

Case 2 Interaction between a dummy and a continuous variable

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

```
. gen female_educ = female*educ

. regress lwage female married female_educ educ exper expersq tenure tenursq

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  8,   517) =   51.35
       Model |   65.677852     8  8.2097315            Prob > F      =  0.0000
    Residual |  82.6518994   517  .159868277           R-squared     =  0.4428
-------------+------------------------------           Adj R-squared =  0.4342
       Total |  148.329751   525  .28253286            Root MSE      =  .39984

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2197774   .1675154    -1.31   0.190    -.5488721    .1093172
     married |   .0529779   .0407884     1.30   0.195    -.0271535    .1331092
 female_educ |  -.0056186   .0130532    -0.43   0.667    -.0312625    .0200254
        educ |   .0813472   .0085008     9.57   0.000     .0646469    .0980476
       exper |   .0268542    .005335     5.03   0.000     .0163733    .0373351
     expersq |  -.0005375   .0001124    -4.78   0.000    -.0007583   -.0003167
      tenure |   .0314803   .0068669     4.58   0.000     .0179898    .0449709
     tenursq |  -.0005792   .0002351    -2.46   0.014    -.0010412   -.0001173
       _cons |    .389629   .1186101     3.28   0.001     .1566119    .6226461
------------------------------------------------------------------------------
```

Comments:

7 Testing for Differences in Regression Functions across Groups

- Is it reasonable to believe that the population regression function that explains the dependent variable is the same across subsamples of populations?
- For example, is it reasonable to believe that the function that explains "GPA of college athlete" is the same for male and female students?

Consider

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

where
$cumgpa$ = $cummulative\ GPA$
sat = SAT score
$hsperc$ = high school rank percentile
$tothrs$ = total hours of college courses

- If we want to test whether "male" students and "female" students have the same values of $\beta_0, \beta_1, \beta_2, \beta_3$ we can estimate the following model

$$\begin{aligned} cumgpa &= \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc + \delta_2 female \cdot hsperc \\ &\quad + \beta_3 tothrs + \delta_3 female \cdot tothrs + u, \end{aligned}$$

and the null hypothesis would be

$$\begin{aligned} H_0 &: \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0 \\ H_a &: otherwise \text{ (at least one } \delta_j = 0) \end{aligned} \tag{8.2}$$

We can use the F-test to test this type of null hypothesis:

1. The restricted model (r)

```
. regress cumgpa sat hsperc tothrs

      Source |       SS       df       MS              Number of obs =     732
-------------+------------------------------           F(  3,   728) =   74.72
       Model |  168.533658     3  56.1778861           Prob > F      =  0.0000
    Residual |  547.364897   728  .751874858           R-squared     =  0.2354
-------------+------------------------------           Adj R-squared =  0.2323
       Total |  715.898555   731  .979341389           Root MSE      =  .86711

------------------------------------------------------------------------------
      cumgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sat |   .0009028   .0002079     4.34   0.000     .0004947    .0013109
      hsperc |  -.0063791   .0015678    -4.07   0.000    -.0094572   -.0033011
      tothrs |   .0119779   .0009314    12.86   0.000     .0101494    .0138064
       _cons |   .9291105   .2285515     4.07   0.000     .4804118    1.377809
------------------------------------------------------------------------------
```

2. The unrestricted model (ur)

```
. gen female_sat = female*sat

. gen female_hsperc = female*hsperc

. gen female_tothrs = female*tothrs

. regress cumgpa female sat female_sat hsperc female_hsperc tothrs female_tothrs

      Source |       SS       df       MS              Number of obs =     732
-------------+------------------------------           F(  7,   724) =   35.15
       Model |  181.589407     7  25.9413439           Prob > F      =  0.0000
    Residual |  534.309148   724   .73799606           R-squared     =  0.2537
-------------+------------------------------           Adj R-squared =  0.2464
       Total |  715.898555   731  .979341389           Root MSE      =  .85907

------------------------------------------------------------------------------
       cumgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       female |  -1.113638    .528539    -2.11   0.035     -2.15129   -.0759859
          sat |   .0006113    .000235     2.60   0.009     .0001499    .0010727
   female_sat |   .0011167      .0005     2.23   0.026     .0001351    .0020984
       hsperc |  -.0059675   .0017765    -3.36   0.001    -.0094551   -.0024798
female_hsperc |   .0000508   .0041025     0.01   0.990    -.0080035     .008105
       tothrs |   .0103004   .0010928     9.43   0.000     .0081549    .0124459
female_tothrs |   .0055599   .0020696     2.69   0.007     .0014968     .009623
        _cons |   1.213984   .2648281     4.58   0.000     .6940617    1.733907
------------------------------------------------------------------------------
```

Comments:

7.1 We can use the "Chow statistics" to test this type of hypothesis as well

- When there are many variables in the model, adding an interaction for every explanatory variable would make the regression analysis messy.

- In which case, we can use the "Chow test" or "Chow statistic" to test the hypothesis expressed in (8.2).

- Chow statistic is a type of F-statistic.

$$F = \frac{SSR_p - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k+1)]}{k+1},$$

where n is the total number of observations.

$SSR_p = SSR$ from the pooled model (include observations from both subsamples)

$SSR_1 = SSR$ from subsample 1
$SSR_2 = SSR$ from subsample 2

regress cumgpa sat hsperc tothrs if female == 0

Source	SS	df	MS
Model	89.6937042	3	29.8979014
Residual	390.619421	548	.712809162
Total	480.313125	551	.871711661

Number of obs	552
F(3, 548)	41.94
Prob > F	0.0000
R-squared	0.1867
Adj R-squared	0.1823
Root MSE	.84428

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.0006113	.000231	2.65	0.008	.0001576	.001065
hsperc	-.0059675	.0017459	-3.42	0.001	-.0093969	-.002538
tothrs	.0103004	.001074	9.59	0.000	.0081907	.0124101
_cons	1.213984	.2602697	4.66	0.000	.7027359	1.725233

regress cumgpa sat hsperc tothrs if female == 1

Source	SS	df	MS
Model	83.4816253	3	27.8272084
Residual	143.689727	176	.816418902
Total	227.171352	179	1.2691137

Number of obs	180
F(3, 176)	34.08
Prob > F	0.0000
R-squared	0.3675
Adj R-squared	0.3567
Root MSE	.90356

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.0017281	.0004642	3.72	0.000	.0008119	.0026442
hsperc	-.0059167	.0038895	-1.52	0.130	-.0135927	.0017594
tothrs	.0158603	.0018485	8.58	0.000	.0122122	.0195085
_cons	.1003465	.4810947	0.21	0.835	-.8491105	1.049803

Comments:

8 A Binary Dependent Variable (y variable): The Linear Probability Model

- So far, our Y variables are continuous.
- What if we are interested in explaining a qualitative Y variable (that is, Y is a dummy variable)?

Consider

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \\ E(y|\mathbf{x}) &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \end{aligned}$$

where $\mathbf{x}$ denotes all of the explanatory variables $(x_1, ..., x_k)$.

```
. regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6

      Source |       SS       df       MS              Number of obs =     753
-------------+------------------------------           F(  7,   745) =   38.22
       Model |  48.8080578     7  6.97257969           Prob > F      =  0.0000
    Residual |  135.919698   745  .182442547           R-squared     =  0.2642
-------------+------------------------------           Adj R-squared =  0.2573
       Total |  184.727756   752  .245648611           Root MSE      =  .42713

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0034052   .0014485    -2.35   0.019    -.0062488   -.0005616
        educ |   .0379953    .007376     5.15   0.000      .023515    .0524756
       exper |   .0394924   .0056727     6.96   0.000     .0283561    .0506287
     expersq |  -.0005963   .0001848    -3.23   0.001    -.0009591   -.0002335
         age |  -.0160908   .0024847    -6.48   0.000    -.0209686    -.011213
     kidslt6 |  -.2618105   .0335058    -7.81   0.000    -.3275875   -.1960335
     kidsge6 |   .0130122    .013196     0.99   0.324    -.0128935    .0389179
       _cons |   .5855192    .154178     3.80   0.000     .2828442    .8881943
------------------------------------------------------------------------------
```

where

$inlf$ = 1 if the woman reports working for a wage outside the home at some point during the year, zero otherwise.

$nwifeinc$ = husband's earnings (in thousands of dollars)

$educ$ = years of education

$exper$ = past years of labor market experience

age = age

$kidslt6$ = number of children less than 6 years old

$kidsage6$ = number of kinds between 6 - 18 years old