PPOL502-01, Spring 2016
Course Notes #10: Indicator Variables as Explanatory Variables

---

I. WHAT ARE INDICATOR (DUMMY) VARIABLES, AND WHY USE THEM?

- Up until now, we've been using only interval-ratio variables for the independent and dependent variables in regression models.

- This may have seemed particularly restrictive to you, and in fact it is. Regression models *are* uniquely suited (compared to Chi-Square tests and other test statistics) to specify the relationship between an interval-ratio explanatory variable and an interval-ratio dependent variable.

- Regression models can also accommodate a number of different kinds of variables. One type of variable that is particularly common, and that will prove to be particularly useful are **indicator variables** (also called *binary* variables, *dummy* variables, *zero/one* variables).

  We will usually use the term "indicator" because it INDICATES a particular category of interest (examples in a moment).

- Constructs often operationalized with indicators include: gender, race/ethnicity, whether a person has a degree (e.g., high school). For example:

Variable name	**= 1 if…**	**= 0 if….**
MALE	male	female
FEMALE	female	male
HSGRAD	graduated from high school	did not graduate from high school

- These are nominal or ordinal variables that contain *qualitative* information. Thus, there is no *required* coding of the categories, that is, which category is assigned the value "1"

  ** *Note1*: ***Naming convention****:* give the indicator variable the name of the category with the "1" value (e.g., male or female in the previous table). This way, you'll never forget how the variable is defined. It also makes it much easier for new users of the data.

  ** *Note2:* ***Which numbers to use?*** Assigning values of "1" and "0" instead of (for example) "1" and "2" makes interpretation of these variables much easier in regression and in interpreting means. The actual value of the variables was not important for running Chi-Square tests – only the fact that separate categories were identified.

  ** *Note3*: The mean of an indicator variable coded as "0" or "1" value equals the proportion of the observations that are in the "1" category.

- How to identify indicator variables in a data set that someone else has put together? Using the Stata command *summarize,*
  -- Look at the minimum and maximum values: are they 0 and 1?
  -- Do the variable names look like they are indicator variables?
  -- Always double-check the codebook to confirm that you are right!

Example:

```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      faminc |      1388    29.02666    18.73928         .5         65
      cigtax |      1388    19.55295    7.795598          2         38
    cigprice |      1388     130.559    10.24448      103.8      152.5
       bwght |      1388    118.6996    20.35396         23        271
    fatheduc |      1192    13.18624    2.745985          1         18
-------------+--------------------------------------------------------
    motheduc |      1387    12.93583    2.376728          2         18
      parity |      1388    1.632565    .8940273          1          6
        male |      1388    .5208934    .4997433          0          1
       white |      1388    .7845821    .4112601          0          1
        cigs |      1388    2.087176    5.972688          0         50
-------------+--------------------------------------------------------
      lbwght |      1388    4.760031    .1906622   3.135494   5.602119
    bwghtlbs |      1388    7.418723    1.272123     1.4375    16.9375
       packs |      1388    .1043588    .2986344          0        2.5
     lfaminc |      1388    3.071271    .9180645  -.6931472   4.174387
```

What do the data look like? (for example)…..

```
. list bwght male white motheduc in 1/5

     +--------------------------------+
     | bwght   male   white   motheduc |
     |--------------------------------|
  1. |   109      1       1         12 |
  2. |   133      1       0         12 |
  3. |   129      0       0         12 |
  4. |   126      1       0         12 |
  5. |   134      1       1         12 |
     +--------------------------------+
```

From BWGT codebook:

```
 faminc                  1988 family income, $1000s
 bwght                   birth weight, ounces
 fatheduc                father's yrs of educ
 motheduc                mother's yrs of educ
 male                    =1 if male child
 white                   =1 if white
 cigs                    cigs smked per day while preg
 packs                   packs smked per day while preg
```

II. INDICATOR VARIABLES IN REGRESSION: SIMPLE EXAMPLES

Example 1:

```
. reg bwght male

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =    7.27
       Model |  2998.87965     1  2998.87965           Prob > F      =  0.0071
    Residual |   571612.84  1386  412.419077           R-squared     =  0.0052
-------------+------------------------------           Adj R-squared =  0.0045
       Total |   574611.72  1387  414.283864           Root MSE      =  20.308

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |     2.94235   1.091149     2.70   0.007     .8018673    5.082832
       _cons |   117.1669   .7875145   148.78   0.000     115.6221    118.7118
------------------------------------------------------------------------------
```

- Statistical significance: The statistical significance of the indicator variable coefficient is tested exactly the same as any other coefficient. The coefficient on MALE is highly statistically significant: $p = 0.0071$.

- Interpretation of MALE coefficient estimate: Male babies weigh 2.94 ounces more at birth, on average, *than do female babies.*

- Idea of "turning on" the indicator variable: When a variable is equal to zero, the coefficient in front of it does not matter in making predictions. When MALE=0, the baby is female. What is your best prediction of the birthweight for female babies?

  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 \qquad BWGT - hat = 117.17 + 2.94 MALE$

  When an indicator variable is equal to 1, the coefficient in front of the variable adds to the prediction value: When MALE=1, the baby is male. What's your best prediction of the birthweight

  for male babies?

  for female babies?

- Idea of a baseline category (sometimes called the "omitted category" or "reference category": You'll notice that there is no FEMALE variable in the regression above.

  * So, where are the females? They are in the intercept: Remember that the intercept is the predicted value of Y when all the X variables are equal to zero.

* The indicator variable MALE gives the *marginal difference*, or "effect," of being MALE (on average), *compared to being female*. **The category to which an indicator is compared (i.e., the category that is reflected in the intercept) is the baseline category.**

* How do you decide which category to use as the "baseline category?" The choice of baseline category is arbitrary: you just need to keep straight which category is the baseline category. (a rule of thumb is to use the most frequently-occurring category as the baseline category, however the research question may prompt using another category)

-- For example, above we could have created a variable called FEMALE and included it in the regression *instead of* MALE).

```
. recode male (1=0) (0=1) , gen(female)
(1388 differences between male and female)

. reg bwght female

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =    7.27
       Model |  2998.87965     1  2998.87965           Prob > F      =  0.0071
    Residual |   571612.84  1386  412.419077           R-squared     =  0.0052
-------------+------------------------------           Adj R-squared =  0.0045
       Total |   574611.72  1387  414.283864           Root MSE      =  20.308

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -2.94235   1.091149    -2.70   0.007    -5.082832   -.8018673
       _cons |   120.1093   .7552665   159.03   0.000     118.6277    121.5909
------------------------------------------------------------------------------
```

Using this new regression, what's your best prediction of the birthweight

for male babies?

for female babies?

-- What Gauss-Markov assumption would be violated if we were to include *both* variables FEMALE and MALE in the regression?

- Another fun fact is that when you include only one indicator variable in a simple regression, the coefficient simply gives you the *difference of means* between the two groups:

  From example above, if we were to use the *summarize* command and then run a ttest:

```
. bys male: summarize bwght

-------------------------------------------------------------------------------
-> male = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       bwght |       665    117.1669    20.32805         30        271

-------------------------------------------------------------------------------
-> male = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       bwght |       723    120.1093    20.28974         23        192


. ttest bwght , by(male)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     665    117.1669     .788288    20.32805    115.6191    118.7148
       1 |     723    120.1093    .7545836    20.28974    118.6278    121.5907
---------+--------------------------------------------------------------------
combined |    1388    118.6996     .546329    20.35396    117.6278    119.7713
---------+--------------------------------------------------------------------
    diff |            -2.94235    1.091149               -5.082832   -.8018673
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.6966
Ho: diff = 0                     degrees of freedom =     1386

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0035         Pr(|T| > |t|) = 0.0071          Pr(T > t) = 0.9965
```

In other words, the following hypothesis tests are equivalent:

T-test approach:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

Regression approach with one indicator X variable:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Example 2:

```
. reg bwght white

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =   22.71
       Model |  9263.73189     1  9263.73189            Prob > F      =  0.0000
    Residual |  565347.988  1386  407.898981            R-squared     =  0.0161
-------------+------------------------------           Adj R-squared =  0.0154
       Total |   574611.72  1387  414.283864           Root MSE      =  20.197

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   6.284029   1.318626     4.77   0.000     3.697311    8.870747
       _cons |   113.7692   1.167994    97.41   0.000      111.478    116.0605
------------------------------------------------------------------------------
```

- What is the baseline category (hint: look back at variable descriptions)?

- Is the coefficient on WHITE statistically significant?

- Interpret the coefficient on WHITE:

- What is the predicted birthweight for white babies?

  for the baseline category, _____ babies?

- What is the average birthweight in this sample for white babies?

  for the baseline category, _____ babies?

III. INTERPRETING INDICATOR VAR COEFFICIENTS IN ln(Y) MODELS

Example 3:

```
. reg lbwght male

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =    5.71
       Model |  .206958881     1  .206958881           Prob > F      =  0.0170
    Residual |  50.2133747  1386  .036228986           R-squared     =  0.0041
-------------+------------------------------           Adj R-squared =  0.0034
       Total |  50.4203336  1387  .036352079           Root MSE      =  .19034

------------------------------------------------------------------------------
      lbwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0244431   .0102269     2.39   0.017     .0043813     .044505
       _cons |   4.747299    .007381   643.18   0.000      4.73282    4.761778
------------------------------------------------------------------------------
```

- Interpretation of MALE indicator coefficient: Males are predicted to weigh 2.4 percent more at birth, on average, than females ($p = 0.0170$).

- General interpretation of indicator coefficients in predicting ln(Y): [*indicator variable group]* is predicted to [*relevant verb*] $(100 * \hat{\beta})\%$ [*more/less*] than the [*baselinegroup]*

- Note: when the coefficient estimate indicates a relatively "small" percentage change (e.g., as above – say less than 10 percent), then the interpreting the coefficient as above is not far off.

  -- However, when the percentage change indicated is relatively high, then you should calculate an *exact percentage difference* in Y, using the following formula (note: this formula can be used for coefficients of any size – it just produces relatively different estimates when the effect is larger):

  Exact percentage difference in Y for indicator coefficients = $100 * [e^{\hat{\beta}_k} - 1]$

  e.g.: for the male coefficient above: $100 * [e^{0.02444} - 1] = 100 * [1.02474 - 1] = 2.47\%$

IV. INDICATOR VARS WITH OTHER INT/RAT VARS IN THE REGRESSION

- Up until now, we've only included indicator variables in the regression. How about if we include a single indicator, plus a single int/rat variable?

Example 4:

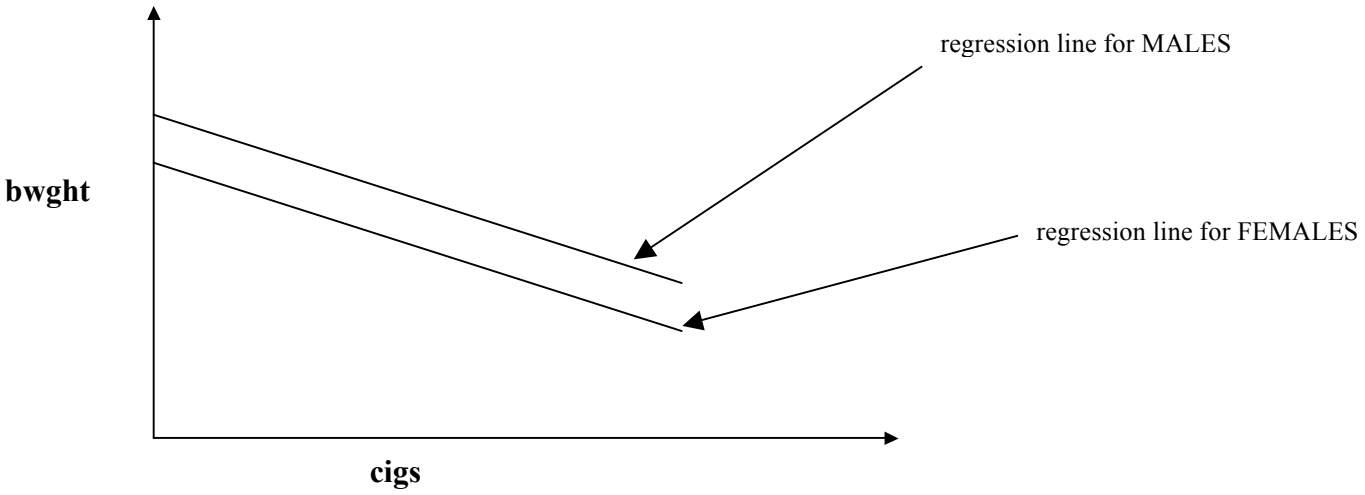
```
. reg bwght male cigs

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  2,  1385) =   19.90
       Model |  16053.1711     2  8026.58556           Prob > F      =  0.0000
    Residual |  558558.549  1385  403.291371           R-squared     =  0.0279
-------------+------------------------------           Adj R-squared =  0.0265
       Total |   574611.72  1387  414.283864           Root MSE      =  20.082

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   2.939342   1.079007     2.72   0.007     .8226776    5.056007
        cigs |  -.5136516   .0902821    -5.69   0.000     -.690756   -.3365473
       _cons |   118.2406   .8012893   147.56   0.000     116.6687    119.8124
------------------------------------------------------------------------------
```

- Interpretation of MALE coefficient estimate: *Holding constant the number of cigarettes smoked per day while pregnant*, male babies weigh 2.94 ounces more at birth, on average, than do female babies (p =0.0065).

- What's going on here? (also see Wooldridge p. 228 for another picture)

  -- Think back to when we just included a single indicator var in the equation: we got a difference of means in Y (i.e., moving up or down the Y-axis)

  -- Here, the indicator variable just shifts the intercept for the indicator group (here males) for the regression line of BWGHT on CIGS) (rough picture):



- This type of model says that for a given level of mom's cigarette smoking, male and female's birthweight is always predicted to differ by _____ ounces, on average.

- This idea can be extended to including many other variables in the equation (e.g., family income, education of the mom and dad, etc.) – it is just not as easy to depict in a graph. But the basic idea is the same: *Holding constant the other Xs in the model*….

V. CONCEPTUAL SETS OF INDICATOR VARIABLES (MUTUALLY EXCLUSIVE CATEGORIES)

- What if the concept or construct you're interested may have more than one category? Often, there are many different "correct" ways to code the variable; how you do it will depend on the research question you're interested in, and prior ways of measuring this variable.

- The interpretation of each coefficient will depend on your definition of other categories (which in turn defines the baseline category).

Example from the NLSY97:

FAMSTR. The relationship of the youth to the parent figure(s)/guardian(s) he or she lives with currently

1 both biological parents
2 two parents, biological mother
3 two parents, biological father
4 biological mother only
5 biological father only
6 adoptive parent(s)
7 foster parent(s)
8 no parents, grandparents
9 no parents, other relatives
10 anything else

```
. tab famstr, missing

     family |
  structure |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      4,371       48.65       48.65
          2 |        989       11.01       59.66
          3 |        213        2.37       62.03
          4 |      2,560       28.50       90.53
          5 |        311        3.46       93.99
          6 |         90        1.00       94.99
          7 |         42        0.47       95.46
          8 |        202        2.25       97.71
          9 |        137        1.52       99.23
         10 |         63        0.70       99.93
          . |          6        0.07      100.00
------------+-----------------------------------
      Total |      8,984      100.0
```

- Question: Would you want to simply include the variable FAMSTR in a regression?

- How might you construct a variable (or variables) to reflect family structure?

```
. recode famstr     (1   = 1 "both biological parents")     ///
>                   (2 3 = 2 "two parents, one biological") ///
>                   (4 5 = 3 "one biological parent")       ///
>                   (6/max = 4 "other parents")             ///
>                   , generate(famstr_reduced)
(3618 differences between famstr and famstr_reduced)
.
. tab famstr_reduced, gen(fam_)
.
. rename fam_1 bothbio
. rename fam_2 twobiol
. rename fam_3 onebiol
. rename fam_4 paroth
.
. list famstr famstr_reduced twobiol onebiol paroth in 1/10

     +-------------------------------------------------------------------+
     | famstr                famstr_reduced   twobiol   onebiol   paroth |
     |-------------------------------------------------------------------|
  1. |      1       both biological parents         0         0        0 |
  2. |      2   two parents, one biological         1         0        0 |
  3. |      4         one biological parent         0         1        0 |
  4. |      4         one biological parent         0         1        0 |
  5. |      1       both biological parents         0         0        0 |
     |-------------------------------------------------------------------|
  6. |      4         one biological parent         0         1        0 |
  7. |      4         one biological parent         0         1        0 |
  8. |      1       both biological parents         0         0        0 |
  9. |      1       both biological parents         0         0        0 |
 10. |      1       both biological parents         0         0        0 |
     +-------------------------------------------------------------------+
.
. bys famstr_reduced: sum piatm

-------------------------------------------------------------------------------
-> famstr_reduced = both biological parents

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       piatm |      2915    100.3125    19.12293         55        142

-------------------------------------------------------------------------------
-> famstr_reduced = two parents, one biological

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       piatm |       826    94.69492    18.99637         55        141

-------------------------------------------------------------------------------
-> famstr_reduced = one biological parent

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       piatm |      2011    93.21929    19.19723         55        138

-------------------------------------------------------------------------------
-> famstr_reduced = other parents

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       piatm |       372    87.82796    19.63817         55        136

-------------------------------------------------------------------------------
```

```
Dependent Variable: piatm: youth piat math score

. reg piatm twobiol onebiol paroth

      Source |       SS       df       MS              Number of obs =    6124
-------------+------------------------------           F(  3,  6120) =   85.95
       Model |  94675.7106     3  31558.5702           Prob > F      =  0.0000
    Residual |  2247152.69  6120  367.181812           R-squared     =  0.0404
-------------+------------------------------           Adj R-squared =  0.0400
       Total |   2341828.4  6123  382.464217           Root MSE      =  19.162

------------------------------------------------------------------------------
       piatm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     twobiol |  -5.617606   .7553098    -7.44   0.000    -7.098279   -4.136933
     onebiol |  -7.093228   .5554724   -12.77   0.000    -8.182149   -6.004306
      paroth |  -12.48456   1.054993   -11.83   0.000    -14.55272   -10.41641
       _cons |   100.3125   .3549125   282.64   0.000     99.61677    101.0083
------------------------------------------------------------------------------
```

- Interpret the coefficient on TWOBIO1:
- Interpret the coefficient on ONEBIO1:
- Interpret the coefficient on PAROTH:
- We can test the null:

  H0: $\delta_{twobio1} = \delta_{onebio1} = \delta_{paroth}$

Test allequal Results for Dependent Variable piatm

```
. test twobiol=onebiol=paroth

 ( 1)  twobiol - onebiol = 0
 ( 2)  twobiol - paroth = 0

       F(  2,  6120) =   16.91
            Prob > F =    0.0000
```

- Or the null:

  H0: $\delta_{twobio1}=0$ *and* $\delta_{onebio1}=0$ *and* $\delta_{paroth}=0$

```
Test allzero Results for Dependent Variable piatm
. test twobiol=onebiol=paroth=0

 ( 1)  twobiol - onebiol = 0
 ( 2)  twobiol - paroth = 0
 ( 3)  twobiol = 0

       F(  3,  6120) =   85.95
            Prob > F =    0.0000
```

VI. CONCEPTUAL SETS OF INDICATOR VARIABLES: MUTUALLY-EXCLUSIVE CATEGORIES (ANOTHER EXAMPLE)

```
. reg bwght clthalf1 chpack1 cgepack1

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  3,  1384) =   12.15
       Model |  14748.9928     3  4916.33092           Prob > F      =  0.0000
    Residual |  559862.727  1384  404.525092           R-squared     =  0.0257
-------------+------------------------------           Adj R-squared =  0.0236
       Total |   574611.72  1387  414.283864           Root MSE      =  20.113

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    clthalf1 |   -6.75088   2.705283    -2.50   0.013    -12.05778   -1.443982
     chpack1 |    -8.9473   2.337639    -3.83   0.000      -13.533   -4.361602
    cgepack1 |  -10.55456    2.39534    -4.41   0.000    -15.25345   -5.855669
       _cons |   120.0612   .5865014   204.71   0.000     118.9107    121.2118
------------------------------------------------------------------------------
```

- What's the baseline category for the C* variables (*C* refers to all variable with variable names that start with C)*?

- Are the C* indicators mutually exclusive?

- Interpretation of CLTHALF1: babies whose mothers smoked less than half a pack of cigarettes a day while they were pregnant weighed 6.75 ounces less at birth, on average, than babies whose mothers did not smoke at all while they were pregnant (p =0.0127).

- Interpretation of CHPACK1:

- Interpretation of CGEPACK1:

- What other hypothesis tests might you be interested in?

```
test1: test clthalf1=chpack1=cgepack1;

. test clthalf1=chpack1=cgepack1

 ( 1)  clthalf1 - chpack1 = 0
 ( 2)  clthalf1 - cgepack1 = 0

       F(  2,  1384) =    0.59
            Prob > F =    0.5572
```

- The above result suggests that the effects on birthweight are not different of smoking less than half a pack, between half and one, and one or more packs a day while pregnant. So

maybe it is just smoking anything at all that matters! We might also then wonder about the following test:

```
test2: test clthalf1=0, chpack1=0, cgepack1=0;
```

```
. test clthalf1=chpack1=cgepack1=0

 ( 1)  clthalf1 - chpack1 = 0
 ( 2)  clthalf1 - cgepack1 = 0
 ( 3)  clthalf1 = 0

       F(  3,  1384) =   12.15
            Prob > F =    0.0000
```

- This result suggests that smoking ***does*** matter. So how about if we just include a single indicator variable in the regression that indicates whether the person smoked AT ALL?

Example:

```
. reg bwght smoker

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =   35.31
       Model |  14275.6609     1  14275.6609           Prob > F      =  0.0000
    Residual |  560336.059  1386  404.282871           R-squared     =  0.0248
-------------+------------------------------           Adj R-squared =  0.0241
       Total |   574611.72  1387  414.283864           Root MSE      =  20.107

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |  -8.914998   1.500258    -5.94   0.000    -11.85802   -5.971977
       _cons |   120.0612   .5863258   204.77   0.000      118.911    121.2114
------------------------------------------------------------------------------
```

- Interpretation of SMOKER coeff: Babies of mothers who smoke at *one or more cigarettes a day* during pregnancy weigh 8.915 ounces less at birth, on average, compared to babies whose mothers don't smoke at all ($p < 0.0001$)

- We predict that babies whose mothers smoke at *least one cigarette a day* during pregnancy weigh (120.06122-8.915) = 111.146 ounces at birth.

  -- This is a weighted average of the 3 "groups" included in this "smoker" category (see means by group below):

$$\left(\left(\frac{58}{212} * 113.3103\right) + \left(\frac{79}{212} * 111.114\right) + \left(\frac{75}{212} * 109.507\right)\right) = 111.146$$

```
. bys smoker: sum bwght
```

```
For nonsmoking moms
-------------------------------------------------------------------------------
-> smoker = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       bwght |      1176    120.0612    20.26849         23        271

For smoking moms
-------------------------------------------------------------------------------
-> smoker = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       bwght |       212    111.1462    19.18141         50        159


. * For moms smoking less than 1/2 pack
. sum bwght if clthalf1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       bwght |        58    113.3103    18.93961         50        153


. * For moms smoking 1/2 or more but less than one pack
. sum bwght if chpack1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       bwght |        79    111.1139    18.86252         68        159


. * For moms smoking a pack or more
. sum bwght if cgepack1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       bwght |        75    109.5067    19.78286         60        151
```

- What if we think back to the pack breakdown, and include only one of the indicators?

Example:

```
. reg bwght cgepack1

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =   16.35
       Model |  6700.25259     1  6700.25259           Prob > F      =  0.0001
    Residual |  567911.467  1386  409.748533           R-squared     =  0.0117
-------------+------------------------------           Adj R-squared =  0.0109
       Total |   574611.72  1387  414.283864           Root MSE      =  20.242

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cgepack1 |   -9.71801   2.403203    -4.04   0.000    -14.43232   -5.003702
       _cons |   119.2247   .5586327   213.42   0.000     118.1288    120.3205
------------------------------------------------------------------------------
```

- What's the baseline category for CGEPACK1 ?

- Interpretation of CGEPACK1:

- What do you think the intercept reflects?

- Do you think this model is sensible? Why or why not?

VII. CONCEPTUAL SETS OF INDICATOR VARIABLES (OVERLAPPING CATEGORIES)

- We could have defined cigarette smoking differently, where the categories are *not* mutually exclusive.:

Example:

```
. reg bwght c*2

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  3,  1384) =   12.15
       Model |  14748.9928     3  4916.33092           Prob > F      =  0.0000
    Residual |  559862.727  1384  404.525092           R-squared     =  0.0257
-------------+------------------------------           Adj R-squared =  0.0236
       Total |   574611.72  1387  414.283864           Root MSE      =  20.113

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    clthalf2 |   -6.75088   2.705283    -2.50   0.013    -12.05778   -1.443982
     chpack2 |  -2.196421   3.477807    -0.63   0.528    -9.018764    4.625922
    cgepack2 |  -1.607257   3.242567    -0.50   0.620    -7.968134    4.753619
       _cons |   120.0612   .5865014   204.71   0.000     118.9107    121.2118
------------------------------------------------------------------------------
```

- What's the baseline category for the C* variables?

- Are the C* indicators mutually exclusive?

- Interpretation of CLTHALF2: Babies whose mothers smoked *up to a half pack* of cigarettes a day while they were pregnant weighed 6.75 ounces less at birth, on average, than babies whose mothers did not smoke at all while they were pregnant (p =0.0127).

- Interpretation of CHPACK2: Babies whose mothers smoked *½ up to 1 pack* a day did not weigh a statistically significant different amount, on average, compared to babies whose mothers who smoked only up to half a pack a day while pregnant.

- Interpretation of CGEPACK2:

- Ho tests: `test1: test clthalf2=chpack2=cgepack2;`

```
. test clthalf2=chpack2=cgepack2

       F(  2,  1384) =    0.81
            Prob > F =    0.4471
```

`test2: test clthalf2=0, chpack2=0, cgepack2=0;`

```
. test clthalf2=chpack2=cgepack2=0

       F(  3,  1384) =   12.15
            Prob > F =    0.0000
```

```
cd "C:\...\Stata datasets"

capture: log close
log using notes10.txt, text replace
set more off

/*******************************************************************************
Birthweight Data
*******************************************************************************/

clear
use bwght.dta
* drop if missing(motheduc) | missing(fatheduc)

*** Investigate Data
summarize
list bwght male white motheduc in 1/5
describe faminc bwght fatheduc motheduc male white cigs packs

*** Example 1
reg bwght male

gen female = 1-male
reg bwght female

bys male: summarize bwght
ttest bwght , by(male)

*** Example 2
reg bwght white


*** Example 3
reg lbwght male

*** Example 4
reg bwght male cigs

/*******************************************************************************
NLSY Data
*******************************************************************************/
clear
use pitnew.dta

tab famstr, missing
recode famstr (1   = 1 "both biological parents")     ///
              (2 3 = 2 "two parents, one biological") ///
              (4 5 = 3 "one biological parent")       ///
              (6/max = 4 "other parents")             ///
              , generate(famstr_reduced)

tab famstr_reduced, gen(fam_)

rename fam_1 bothbio
rename fam_2 twobiol
rename fam_3 onebiol
rename fam_4 paroth
```

```
list famstr famstr_reduced twobiol onebiol paroth in 1/10

bys famstr_reduced: sum piatm

reg piatm twobiol onebiol paroth
test twobiol=onebiol=paroth
test twobiol=onebiol=paroth=0

/*******************************************************************************
Birthweight Data
*******************************************************************************/
clear
use bwght.dta
gen female = 1-male

* create new indicator variables
gen clthalf1=0
gen chpack1 =0
gen cgepack1=0

* change values
replace clthalf1 =1 if packs>0    & packs<0.5
replace chpack1  =1 if packs>=0.5 & packs<1
replace cgepack1 =1 if packs>=1

* correct for any missing values of packs
foreach var of varlist c*1{
      replace `var'=. if missing(packs)
}

* create new indicator variables
gen clthalf2=0
gen chpack2 =0
gen cgepack2=0

* change values
replace clthalf2 =1 if packs>0
replace chpack2  =1 if packs>=0.5
replace cgepack2 =1 if packs>=1

* correct for any missing values of packs
foreach var of varlist c*2{
      replace `var'=. if missing(packs)
}

gen smoker=0
replace smoker=1 if cigs>0
replace smoker = . if missing(cigs)

gen frstbrn=0
replace frstbrn=1 if parity==1
replace frstbrn=. if missing(parity)

* create interaction variables
gen whcigs = white*cigs
gen whmale = white*male
```

```
* Regressions

reg bwght clthalf1 chpack1 cgepack1
test clthalf1=chpack1=cgepack1
test clthalf1=chpack1=cgepack1=0

reg bwght smoker


* Summary statistics by smoker group
bys smoker: sum bwght

* For moms smoking less than 1/2 pack
sum bwght if clthalf1
* For moms smoking 1/2 or more but less than one pack
sum bwght if chpack1
* For moms smoking a pack or more
sum bwght if cgepack1

* Regressions
reg bwght cgepack1

reg bwght c*2
test clthalf2=chpack2=cgepack2
test clthalf2=chpack2=cgepack2=0

reg bwght male cigs

log close
```