# Week 8: Qualitative predictors

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

# Outline

- Qualitative predictors, aka dummy variables, indicator variables, categorical variables
- ANOVA (the easier way)
- Effect coding and grand mean interpretation of intercept
- Contrasts
- Parameter interpretation
- Difference-in-difference models
- Interactions

# Big picture

- Qualitative variables are important in regression analysis because they provide more flexibility in modeling
- Sex, race, state, marital status, treatment group are all qualitative variables
- It's not uncommon to create categories from a continuous variable to make models a) easier to explain and b) relax the linearity assumption (for instance, age categories)
- Much of what we have learned so far applies to qualitative variables but we interpret models somewhat differently
- We often code qualitative variables as 0/1 but **this is not the only way of coding**

# Beauty dataset again

- Simplest case, two categories; for example, male or female
- We will define an indicator (aka dummy) variable that is equal to 1 if female and 0 if male
- **Random advice**: Name your variables in a way that makes clear which category is 1 and which is zero. Don't create a variable called "sex;" create a variable called female if female is 1 or male if male is 1
- We will estimate the model $wage_i = \beta_0 + \beta_1 female_i + \epsilon_i$

# Parameter interpretation

- Since *female* is a 1/0 variable, not a lot of sense to take the derivative (but you could, mechanically, although the derivative does not exist since you can't take the limit)
- $E[wage|female = 1] = \beta_0 + \beta_1$, so the sum of the coefficients is the **average wage for females**
- $E[wage|female = 0] = \beta_0$ is the average wage for males
- $E[wage|female = 1] - E[wage|female = 0] = \beta_0 + \beta_1 - \beta_0 = \beta_1$
- So $\beta_1$ is the **difference in average salaries between females and males**
- Note once again the difference between **sample and population**. What we just did will always be true in the sample. If we want to make statements at the population level, we need the *zero conditional mean* assumption again

# Stata output

- In the beauty dataset, the **average** female salary per hour is \$3 less than that of males

```
reg wage female

      Source |       SS           df       MS      Number of obs   =     1,260
-------------+----------------------------------   F(1, 1258)      =    137.04
       Model |  2686.38669         1  2686.38669   Prob > F        =    0.0000
    Residual |  24661.0525     1,258  19.6033803   R-squared       =    0.0982
-------------+----------------------------------   Adj R-squared   =    0.0975
       Total |  27347.4392     1,259  21.7215561   Root MSE        =    4.4276

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -3.069465   .2622068   -11.71   0.000    -3.583876   -2.555054
       _cons |   7.368823   .1542417    47.77   0.000     7.066223    7.671422
------------------------------------------------------------------------------
```

- Average salary for males: $\beta_0 = 7.368$. Average salary for females: $\beta_0 + \beta_1 = 7.368823 - 3.069465 = 4.29$
- The null for the Wald test is $H_0 : \beta_1 = 0$. If this is true, then we say that the average wage for males and females is the same: $\beta_0$

# Stata output

- Verify that it's the same as descriptive stats

```
tabstat wage, by(female) stats(N mean median sd min max)
Summary for variables: wage
     by categories of: female (=1 if female)

  female |         N      mean       p50        sd       min       max
---------+------------------------------------------------------------
       0 |       824  7.368823      6.41  4.592508      1.05     41.67
       1 |       436  4.299358      3.75  4.097392      1.02     77.72
---------+------------------------------------------------------------
   Total |      1260   6.30669       5.3  4.660639      1.02     77.72
----------------------------------------------------------------------
```

- Even if causality makes no sense in this study, the regression model is perfectly valid as a **descriptive** model

# Remember the old t-test for independent samples?

- Previous Wald test is the same as the stats 101 t-test for independent samples:

```
 ttest wage, by(female)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     824    7.368823    .1599876    4.592508    7.054791    7.682855
       1 |     436    4.299358    .1962295    4.097392    3.913682    4.685034
---------+--------------------------------------------------------------------
combined |   1,260     6.30669    .1312986    4.660639    6.049102    6.564279
---------+--------------------------------------------------------------------
    diff |            3.069465    .2622068                2.555054    3.583876
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  11.7063
Ho: diff = 0                                     degrees of freedom =     1258

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

- Same null (sign backwards) and same $t = 11.7063$

# More than one level

- It is fairly easy to incorporate more than one category
- Let' say that we are interested in the effect of experience on wage but for some reason we think that 0 and 10 years of experience are equivalent and we want to compare to more than 10 years to 30 and greater than 30 (note the different than missing part below)

```
* Create indicators
gen     expcat = 1 if exper >0   & exper <= 10
replace expcat = 2 if exper > 10 & exper <= 30
replace expcat = 3 if exper > 30 & exper ~= .

tab expcat

     expcat |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        424       33.76       33.76
          2 |        594       47.29       81.05
          3 |        238       18.95      100.00
------------+-----------------------------------
      Total |      1,256      100.00
```

# More than one level

- We could code indicator variables directly but I wanted to show you a handy way in Stata

```
tab expcat, gen(expcat)
sum expcat1-expcat3
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     expcat1 |      1,256    .3375796    .4730727          0          1
     expcat2 |      1,256    .4729299    .4994655          0          1
     expcat3 |      1,256    .1894904    .3920538          0          1
```

```
 list exper expcat1-expcat3 in 1/5

     +--------------------------------------+
     | exper   expcat1   expcat2   expcat3 |
     |--------------------------------------|
  1. |    30         0         1         0 |
  2. |    28         0         1         0 |
  3. |    35         0         0         1 |
  4. |    38         0         0         1 |
  5. |    27         0         1         0 |
     +--------------------------------------+
```

# Average salary by level of experience

- We can estimate the model
  $wage_i = \beta_0 + \beta_1 expcat2_i + \beta_2 expcat3_i + \epsilon_i$

```
reg wage expcat2 expcat3

      Source |       SS           df       MS      Number of obs   =     1,256
-------------+----------------------------------   F(2, 1253)      =     38.94
       Model |  1596.55872         2  798.279361   Prob > F        =    0.0000
    Residual |  25684.7847     1,253   20.498631   R-squared       =    0.0585
-------------+----------------------------------   Adj R-squared   =    0.0570
       Total |  27281.3434     1,255  21.7381222   Root MSE        =    4.5275

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     expcat2 |   2.217522    .287846     7.70   0.000     1.652809    2.782235
     expcat3 |   2.697418   .3667077     7.36   0.000     1.977989    3.416847
       _cons |   4.759599   .2198768    21.65   0.000     4.328232    5.190966
------------------------------------------------------------------------------
```

# Need to choose a reference category

- We leave one category out because otherwise we have perfect collinearity; if you don't do it, Stata will drop one

```
. reg wage expcat1 expcat2 expcat3
note: expcat1 omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =     1,256
-------------+----------------------------------   F(2, 1253)      =     38.94
       Model |  1596.55872         2  798.279361   Prob > F        =    0.0000
    Residual |  25684.7847     1,253   20.498631   R-squared       =    0.0585
-------------+----------------------------------   Adj R-squared   =    0.0570
       Total |  27281.3434     1,255  21.7381222   Root MSE        =    4.5275

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     expcat1 |          0  (omitted)
     expcat2 |   2.217522    .287846     7.70   0.000     1.652809    2.782235
     expcat3 |   2.697418   .3667077     7.36   0.000     1.977989    3.416847
       _cons |   4.759599   .2198768    21.65   0.000     4.328232    5.190966
------------------------------------------------------------------------------
```

# Parameter interpretation

- The intercept, $\beta_0$, is the average wage for individuals in the **reference** category 0-10
- The average salary for individuals with more than 30 years of experience is $2.70 higher than for those with 0-10 years of experience
- **Never forget (!!!!!)**: always a **comparison relative to the reference category**
- Get used to interpret models this way:

  $E[wage|expcat3 = 1] = \beta_0 + \beta_2$ and $E[wage|expcat1 = 1] = \beta_0$

  $E[wage|expcat3 = 1] - E[wage|expcat1 = 1] = \beta_2$
- So, comparisons to reference level

# Inference

- The Wald test $H_0 : \beta_2 = 0$. If not rejected, then people with over 30 years of experience make the same average salary than those with 0-10 years of experience
- The $F$ test is comparing the full model to the restricted model so the null hypothesis is that $\beta_2 = \beta_3 = 0$; the alternative is that at least one is not equal to zero
- **That's the ANOVA test**: if all coefficients are equal to zero, then the average wage is the same for all levels of experience
- ANOVA is a comparison of means when the number of groups $> 2$; an extension of the t-test. But its name, Analysis of **Variance**, comes fro the way the test was developed: as a comparison of (residual) variance (SSE)

# ANOVA should be equivalent to LRT (asymptotically)

- Do a LRT instead

```
qui reg wage expcat1 expcat2 expcat3
est sto full

qui reg wage if e(sample)
est sto red

lrtest red full
Likelihood-ratio test                                 LR chi2(2)  =     75.74
(Assumption: red nested in full)                      Prob > chi2 =    0.0000
```

- We need the e(sample) because the models have different sample sizes since there are missings in experience categories (Stata produces an error message)

# Digression ANOVA and parameter interpretation

- When you learn about ANOVA, it is usually presented as a model in which the intercept is supposed to represent the **grand mean** rather than the mean of the reference category as we just saw
- The grand mean in the previous example would be the average wage regardless of experience (or the unconditional mean)
- This is because coding indicator variables as 1/0 is not the only way of coding indicator variables
- Example: Data on cholesterol levels by age group. We want to test if the average cholesterol level is the same for all five age groups

# Cholesterol data

```
* Dummy variables coded as 0/1 - in the model the constant is cholesterol for 10-29 (180.51)

tabstat chol, by(agegrp) stats(N mean sd min max)
agegrp |         N      mean        sd       min       max
-------+--------------------------------------------------
 10-19 |        15  180.5198  9.959015  165.2215  204.7666
 20-29 |        15  188.7233  10.20568  170.6993  208.6496
 30-39 |        15  202.0608  10.38802  185.6186  220.5073
 40-59 |        15  210.6704   10.1015  196.3125  233.7877
 60-79 |        15   219.282  10.96153  196.7426  237.3754
-------+--------------------------------------------------
 Total |        75  200.2513  17.40287  165.2215  237.3754
----------------------------------------------------------

reg chol ageg2 ageg3 ageg4 ageg5
      Source |       SS           df       MS      Number of obs   =        75
-------------+----------------------------------   F(4, 70)        =     35.02
       Model |  14943.3997         4  3735.84993   Prob > F        =    0.0000
    Residual |  7468.21971        70  106.688853   R-squared       =    0.6668
-------------+----------------------------------   Adj R-squared   =    0.6477
       Total |  22411.6194        74  302.859722   Root MSE        =    10.329

------------------------------------------------------------------------------
        chol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ageg2 |   8.203575   3.771628     2.18   0.033     .6812991    15.72585
       ageg3 |   21.54105   3.771628     5.71   0.000     14.01878    29.06333
       ageg4 |   30.15067   3.771628     7.99   0.000      22.6284    37.67295
       ageg5 |   38.76221   3.771628    10.28   0.000     31.23993    46.28448
       _cons |   180.5198   2.666944    67.69   0.000     175.2007    185.8388
------------------------------------------------------------------------------
```

# Effect coding

- If in the age group of interest, code as 1, if not zero. Same as before. But the reference category is coded as -1. Constant now is the grand mean

```
gen age2029 = 0
replace age2029 = 1  if agegrp == 2
replace age2029 = -1 if agegrp == 1
gen age3039 = 0
replace age3039 = 1  if agegrp == 3
replace age3039 = -1 if agegrp == 1
...
. reg chol age2029 age3039 age4059 age6079

      Source |       SS           df       MS      Number of obs   =        75
-------------+----------------------------------   F(4, 70)        =     35.02
       Model |  14943.3997         4  3735.84993   Prob > F        =    0.0000
    Residual |  7468.21971        70  106.688853   R-squared       =    0.6668
-------------+----------------------------------   Adj R-squared   =    0.6477
       Total |  22411.6194        74  302.859722   Root MSE        =    10.329

------------------------------------------------------------------------------
        chol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age2029 |  -11.52793   2.385387    -4.83   0.000    -16.28543   -6.770423
     age3039 |   1.809552   2.385387     0.76   0.451    -2.947953    6.567057
     age4059 |   10.41917   2.385387     4.37   0.000     5.661668    15.17668
     age6079 |    19.0307   2.385387     7.98   0.000      14.2732    23.78821
       _cons |   200.2513   1.192694   167.90   0.000     197.8725      202.63
------------------------------------------------------------------------------
```

# Effect coding

- ANOVA test doesn't change
- Parameters are interpreted as deviations from the grand mean, which means that the interpretation of the Wald test does change
- Now the Wald test is testing if average cholesterol for each age group is different from the grand mean, not different from the reference category
- **Warning**: This only works for **balanced**; data in which all categories have the same number of observations. With unbalanced data, the intercept is no longer the grand mean
- Not the only way of coding; there are more schemes
- **Lesson**: **what you want to test drives how you code the data**

# Wage by experience agan using the anova command

- Stata of course has an anova command

```
anova wage expcat2 expcat3
                         Number of obs =      1,256    R-squared     =  0.0585
                         Root MSE      =    4.52754    Adj R-squared =  0.0570

                  Source | Partial SS         df         MS        F    Prob>F
              -----------+----------------------------------------------------
                   Model |  1596.5587          2   798.27936     38.94  0.0000
                         |
                 expcat2 |  1216.5794          1   1216.5794     59.35  0.0000
                 expcat3 |  1109.1269          1   1109.1269     54.11  0.0000
                         |
                Residual |  25684.785      1,253   20.498631
              -----------+----------------------------------------------------
                   Total |  27281.343      1,255   21.738122
test expcat2 expcat3

                  Source | Partial SS         df         MS        F    Prob>F
         ----------------+----------------------------------------------------
         expcat2 expcat3 |  1596.5587          2   798.27936     38.94  0.0000
                Residual |  25684.785      1,253   20.498631
```

# Wage by experience agan using the anova command

- The symbolic option helps you see what Stata is testing and how it's coded
- Type "help anova" for more details

```
 test expcat2 expcat3, symbolic
expcat2
          0  -r2
          1   r2
expcat3
          0  -r4
          1   r4
_cons         0
```

# A slide for Sue

- Your professor was doing was is called a "two-way factorial ANOVA." Time has three levels and group had two. Then there was an interaction
  $y = \beta_0 + \beta_1 group1 + \beta_2 time48 + \beta_3 time72 + \beta 4 group1 * time48 + \beta_5 group1 * time72 + \epsilon$

```
      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(5, 18)        =      8.02
       Model |  23.1264022         5  4.62528044   Prob > F        =    0.0004
    Residual |  10.3745437        18  .576363542   R-squared       =    0.6903
-------------+----------------------------------   Adj R-squared   =    0.6043
       Total |   33.500946        23  1.45656287   Root MSE        =    .75919

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.groupn |     .89375   .5368256     1.66   0.113    -.2340788    2.021579
             |
        time |
         48  |      -.817   .5368256    -1.52   0.145    -1.944829    .3108288
         72  |    -2.2905   .5368256    -4.27   0.000    -3.418329   -1.162671
             |
 groupn#time |
       1 48  |    -.69425   .7591861    -0.91   0.373    -2.289241    .9007408
       1 72  |     .23975   .7591861     0.32   0.756    -1.355241    1.834741
             |
       _cons |    6.06825   .3795931    15.99   0.000     5.270755    6.865745
------------------------------------------------------------------------------
```

# A slide for Sue

- Using ANOVA
- The test for the time row is a test of "main effects" for time. That is, whether the mean $y$ is the same for all times regardless of group
- You can do the using the F test comparing nested models . Because of small sample sizes, the LRT will be a bit different

```
anova y i.groupn i.time i.groupn#i.time

                         Number of obs =         24    R-squared     =  0.6903
                         Root MSE      =    .759186    Adj R-squared =  0.6043

                  Source | Partial SS         df         MS        F    Prob>F
              -----------+----------------------------------------------------
                   Model |   23.126402          5   4.6252804      8.02  0.0004
                         |
                  groupn |   3.3056104          1   3.3056104      5.74  0.0277
                    time |   18.879579          2   9.4397895     16.38  0.0001
             groupn#time |   .94121275          2   .47060637      0.82  0.4577
                         |
                Residual |   10.374544         18   .57636354
              -----------+----------------------------------------------------
                   Total |   33.500946         23   1.4565629
```

# A slide for Sue

- Replicate test

```
qui anova y i.groupn i.time i.groupn#i.time
test time
                  Source | Partial SS         df         MS         F    Prob>F
              -----------+----------------------------------------------------
                    time |   18.879579          2   9.4397895     16.38  0.0001
                Residual |   10.374544         18   .57636354
test time, symbolic
groupn
           0  0
           1  0
time
          24  -(r4+r5)
          48   r4
          72   r5
groupn#time
        0 24  -1/2 (r4+r5)
        0 48   1/2 r4
        0 72   1/2 r5
        1 24  -1/2 (r4+r5)
        1 48   1/2 r4
        1 72   1/2 r5
_cons         0
```

# Adding covariates

- Let's add experience (linearly) to the model:
  $wage_i = \beta_0 + \beta_1 female_i + \beta_2 exper_i + \epsilon_i$
- Same interpretation as before, with the addition that $\beta_1$ is the average difference in salaries of females versus males holding experience constant or after taking into account the effect of experience
- Let's say we hold exper constant at $exper = 10$
  $E[wage|female = 1; exper = 10] = \beta_0 + \beta_1 + \beta_2 * 10$ and
  $E[wage|female = 0; exper = 10] = \beta_0 + \beta_2 * 10$
- So $E[wage|female = 1] - E[wage|female = 0] = \beta_1$
- Same as before (this would be ANCOVA, by the way). The partialling out interpretation of adjusting still holds (try it)

# Factor syntax in Stata

- The other reason I created the expcat variable with 1, 2, 3 for each category of experience is because Stata has a convenient syntax so you don't have to create dummy variables
- It's called **factor variables**. For more, type **help fvvarlist** in Stata
- **Use it with caution**. I much prefer you do it the longer way (creating dummy variables) until you understand what you are doing
- But it does save time (once you understand what you are doing)

# Factor variables

```
 reg wage i.expcat
<...>
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      expcat |
          2  |   2.217522    .287846     7.70   0.000     1.652809    2.782235
          3  |   2.697418   .3667077     7.36   0.000     1.977989    3.416847
             |
       _cons |   4.759599   .2198768    21.65   0.000     4.328232    5.190966
------------------------------------------------------------------------------

* Change the reference category to second level
 reg wage ib2.expcat
<..>
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      expcat |
          1  |  -2.217522    .287846    -7.70   0.000    -2.782235   -1.652809
          3  |   .4798956   .3473301     1.38   0.167     -.201517    1.161308
             |
       _cons |   6.977121   .1857673    37.56   0.000     6.612672     7.34157
------------------------------------------------------------------------------
```

# Things to never forget

- **If you change the reference level you change the Wald test**
- What is the difference between the Wald test for the coefficient of expcat3 in the above models?
- In the first model the null is that the average for those with experience level of 3 (more than 30 years) is the same as the average salary of those with 0-10 years of experience. We reject that null
- In the second, we are comparing level 3 to level 2, the reference category. We do not reject that null
- Another way of creating different comparison with categorical variables is **contrasts** (more on that in one sec)

# Things to never forget II

- This has to be the **most common mistake** and source of confusion when using dummy variables
- How do we test if experience is related to wages?
- In the model $wage = \beta_0 + \beta_1 exper + \epsilon$ we would use the Wald test for exper (assumed to be linearly related to wage)
- But now we have $wage = \beta_0 + \beta_1 expercat2 + \beta_1 expercat3 + \epsilon$
- **We need to test all of them jointly**: $H_0 : \beta_1 = \beta_2 = 0$. If not rejected, the average wage is the same regardless of experience level
- In this simple model, that's the F test
- Think about a full and reduced model using either an F test or a LRT. Or a Wald test using the test command

# Several equivalent tests

```
 reg wage expcat2 expcat3

      Source |       SS           df       MS      Number of obs   =     1,256
-------------+----------------------------------   F(2, 1253)      =     38.94
       Model |  1596.55872         2  798.279361   Prob > F        =    0.0000
    Residual |  25684.7847     1,253   20.498631   R-squared       =    0.0585
-------------+----------------------------------   Adj R-squared   =    0.0570
       Total |  27281.3434     1,255  21.7381222   Root MSE        =    4.5275
<...>
test expcat2 expcat3
 ( 1)  expcat2 = 0
 ( 2)  expcat3 = 0
       F(  2,  1253) =   38.94
            Prob > F =    0.0000
qui reg wage expcat2 expcat3
est sto f
qui reg wage if e(sample)
est sto r
lrtest f r
 Likelihood-ratio test                                 LR chi2(2)  =     75.74
(Assumption: r nested in f)                          Prob > chi2 =    0.0000
* chi-squared = (numerator degrees of freedom) * F
di 2*38.94
* 77.88
```

# Contrasts

- This is one of those cultural issues in methods: if you were trained in economics, you have never heard about contrasts (ever, I have asked about 13 people so far)
- More common in psychology and stats and a bit old-fashioned. The researchers who still use ANOVA instead of regressions tend to use contrasts (or those who tend to use SAS, but this is anecdotal)
- In stats, you cover contrasts in design and analysis of experiments
- It involves linear combination of parameters (so to speak) to make comparisons
- Stata has a post-estimation command called contrasts that can be used to replicate what we did so far and much more (type "help contrast")

# Contrasts

```
qui reg wage i.expcat

contrast r.expcat
------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
      expcat |
   (2 vs 1)  |          1       59.35     0.0000
   (3 vs 1)  |          1       54.11     0.0000
      Joint  |          2       38.94     0.0000
             |
 Denominator |       1253
------------------------------------------------

contrast a.expcat
------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
      expcat |
   (1 vs 2)  |          1       59.35     0.0000
   (2 vs 3)  |          1        1.91     0.1673
      Joint  |          2       38.94     0.0000
             |
 Denominator |       1253
------------------------------------------------
```

# Let's bring beauty back

- We had to wait until now to actually measure the effect of looks on wages because looks was measured as a qualitative variable
- Check all p-values. Does beauty affect wages? Notice something odd?

```
tab looks, gen(look)
reg wage look2-look5

      Source |       SS           df       MS      Number of obs   =     1,260
-------------+----------------------------------   F(4, 1255)      =      2.58
       Model |  223.237407         4  55.8093518   Prob > F        =    0.0357
    Residual |  27124.2018     1,255  21.6129098   R-squared       =    0.0082
-------------+----------------------------------   Adj R-squared   =    0.0050
       Total |  27347.4392     1,259  21.7215561   Root MSE        =     4.649

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       look2 |   .7072643   1.347121     0.53   0.600    -1.935593    3.350122
       look3 |    1.88306   1.300948     1.45   0.148    -.6692133    4.435333
       look4 |   1.677802   1.312215     1.28   0.201    -.8965743    4.252179
       look5 |   2.766883   1.673336     1.65   0.098    -.5159616    6.049727
       _cons |   4.621538   1.289392     3.58   0.000     2.091937     7.15114
------------------------------------------------------------------------------
```

# Let's bring beauty back

- What changed now?

```
reg wage look1 look2 look4 look5
*same as
*reg wage ib3.looks

      Source |       SS           df       MS      Number of obs   =     1,260
-------------+----------------------------------   F(4, 1255)      =      2.58
       Model |  223.237407         4  55.8093518   Prob > F        =    0.0357
    Residual |  27124.2018     1,255  21.6129098   R-squared       =    0.0082
-------------+----------------------------------   Adj R-squared   =    0.0050
       Total |  27347.4392     1,259  21.7215561   Root MSE        =     4.649

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       look1 |   -1.88306   1.300948    -1.45   0.148    -4.435333    .6692133
       look2 |  -1.175796   .4267767    -2.76   0.006     -2.01307   -.3385211
       look4 |  -.2052577   .2988493    -0.69   0.492     -.791557    .3810416
       look5 |   .8838227   1.080489     0.82   0.414    -1.235941    3.003586
       _cons |   6.504598   .1730167    37.60   0.000     6.165164    6.844032
------------------------------------------------------------------------------
```

# Let's bring beauty back

- One problem with the previous model is that looks $= 1$ has very low sample size ($n = 13$); we would be better off comparing above average looks to the rest

```
 desc abvavg
...
abvavg          byte    %8.0g                 =1 if looks >=4
reg wage abvavg
      Source |       SS           df       MS      Number of obs   =     1,260
-------------+----------------------------------   F(1, 1258)      =      0.06
       Model |  1.19891395         1  1.19891395   Prob > F        =    0.8144
    Residual |  27346.2403     1,258  21.7378698   R-squared       =    0.0000
-------------+----------------------------------   Adj R-squared   =   -0.0008
       Total |  27347.4392     1,259  21.7215561   Root MSE        =    4.6624

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      abvavg |   .0670626   .2855582     0.23   0.814    -.4931602    .6272853
       _cons |   6.286306   .1574377    39.93   0.000     5.977436    6.595175
------------------------------------------------------------------------------
```

- Still, shaky evidence. Look at $R^2$. Clearly we need to explain more of the variance; let's ignore statistical significance and focus on coefficients

# Let's bring beauty back

- Adding female indicator

```
reg wage abvavg female

      Source |       SS           df       MS      Number of obs   =     1,260
-------------+----------------------------------   F(2, 1257)      =     68.76
       Model |  2696.97554         2  1348.48777   Prob > F        =    0.0000
    Residual |  24650.4636     1,257  19.6105518   R-squared       =    0.0986
-------------+----------------------------------   Adj R-squared   =    0.0972
       Total |  27347.4392     1,259  21.7215561   Root MSE        =    4.4284

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      abvavg |   .1994742   .2714608     0.73   0.463     -.333092    .7320404
      female |  -3.077489    .262482   -11.72   0.000     -3.59244   -2.562538
       _cons |   7.310966   .1732013    42.21   0.000      6.97117    7.650761
------------------------------------------------------------------------------
```

- Both are indicator variables, how do we interpret them?
- (Look at the p-value for abvage. What changed?)

# Interpretation

- The model is $wage_i = \beta_0 + \beta_1 abvavg_i + \beta_2 female_i + \epsilon_i$
- $\beta_0$ is the average wage for males rated as being of below average looks (holding sex constant)
- The other coefficients are interpreted as before, holding the other constant and in relation to their reference category
- $\beta_1$ is the average wage for those rated as having above average looks compared to those rated as having below average looks, holding the effect of sex constant (it is like a weighted average)
- Note other things: What is the average wage for females rated as having above average looks? It's $\beta_0 + \beta_1 + \beta_2$
- (It won't be exactly the same as using the summarize command but will be close. To get exact values, we need interactions)

# Expected value for above avg looks and females

- Not the same as summarize command but close enough

```
reg wage abvavg female
....
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      abvavg |   .1994742   .2714608     0.73   0.463     -.333092    .7320404
      female |  -3.077489    .262482   -11.72   0.000     -3.59244   -2.562538
       _cons |   7.310966   .1732013    42.21   0.000      6.97117    7.650761
------------------------------------------------------------------------------
di _b[_cons] + _b[abvavg] + _b[female]
4.4329506

sum wage if abvavg ==1 & female ==1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |        144    4.698264    6.436829       1.16      77.72
```

- In a second, we will get it exactly right

# Interactions

- When we did descriptive stats we saw that the effect of looks on wages was different for males and females
- In the model above, the **effect of looks on wage does not depend on sex**; it's the effect of above average looks once sex has been taken into account
- The partialling out interpretation still holds

```
reg wage female
predict wage_r, res
reg abvavg female
predict abvavg_r, res
reg wage_r abvavg_r
------------------------------------------------------------------------------
      wage_r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    abvavg_r |   .1994742   .2713529     0.74   0.462    -.3328799    .7318283
       _cons |   4.66e-09   .1247059     0.00   1.000    -.2446544    .2446544
------------------------------------------------------------------------------
```

- To make the effect of looks depend on sex, we need to add **interactions**

# Interactions

- A model with interactions is:

  $wage_i = \beta_0 + \beta_1 abvavg + \beta_2 female + \beta_3 abvavg * female_i + \epsilon_i$
- And here is where things get a bit complicated. **Never underestimate the power of interactions to get you all confused**
- **The easy part**: we are just making the effect of looks depend on sex, so the effect is different for males than females. Or the other way around, the effect of sex depends on looks
- The difficult part: interpreting the parameters and not getting things backwards because interactions go in both directions
- We will cover several strategies

# Interactions

- Female of above average looks: $\beta_0 + \beta_1 + \beta_2 + \beta_3$
- Female of below average looks: $\beta_0 + \beta_2$
- So (1) $\beta_1 + \beta_3$ is the difference in female average salaries for those of above average looks compared to those of below average looks
- Male of above average looks: $\beta_0 + \beta_1$
- Male of below average looks: $\beta_0$
- So (2) $\beta_1$ is the difference in male average salary for those of above average looks compared to those males of below average looks
- Both (1) and (2) are differences. And $\beta_3$ is the difference (1) - (2), so $\beta_3$ is a **difference of differences**: It is the **additional** effect of above average looks for females versus males

# Interactions

- The model again:

  $wage_i = \beta_0 + \beta_1 abvavg + \beta_2 female + \beta_3 abvavg * female_i + \epsilon_i$
- One mechanic way of remembering: $\beta_3$ is only "on" if both $abvavg = 1$ and $female = 1$; if either one is zero, $\beta_3$ is out of the picture
- So it's the incremental/additional effect of **above average looks for females versus males** as we just worked it out
- To make things more confusing, it is also the incremental or additional effect of being female for those with above average looks compared to those with below average looks
- That's how you can get easily confused
- About taking derivatives when you are not supposed to take derivatives... (don't tell anybody)

# Digression: difference-in-difference models

- Suppose you have data for a treatment group $T$ before and intervention and after an intervention, where *post* is a dummy variable equal to one if in the post period
- You run the following model (omitting subscripts):

  $y = \beta_0 + \beta_1 T + \beta_2 post + \beta_3 T * post + \epsilon$
- $\beta_1$ is the treatment effect in the pre-period. What is $\beta_3$?
- (Treated post-period - treated pre-period) $= \beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 - \beta_1 = \beta_2 + \beta_3$
- (Control post-period - control pre-period) $= \beta_0 + \beta_2 - \beta_0$
- So $\beta_3$ is (Treated post-period - treated pre-period) - (Control post-period - control pre-period)
- $\beta_3$ is a **difference-in-difference**. If $\beta_3 = 0$ then there is no treatment effect in the post-period

# Interactions

- The model with interactions: $\beta_3 = 0.62$;

```
gen abv_fem = abvavg*female
reg wage abvavg female abv_fem

      Source |       SS           df       MS      Number of obs   =     1,260
-------------+----------------------------------   F(3, 1256)      =     46.25
       Model |  2720.71294         3  906.904314   Prob > F        =    0.0000
    Residual |  24626.7262     1,256  19.6072661   R-squared       =    0.0995
-------------+----------------------------------   Adj R-squared   =    0.0973
       Total |  27347.4392     1,259  21.7215561   Root MSE        =     4.428

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      abvavg |  -.0256877   .3399345    -0.08   0.940    -.6925898    .6412143
      female |  -3.273637   .3172773   -10.32   0.000    -3.896088   -2.651185
     abv_fem |   .6213146   .5646815     1.10   0.271    -.4865083    1.729138
       _cons |   7.376273   .1830757    40.29   0.000     7.017106    7.735441
------------------------------------------------------------------------------
```

- Effect of looks depends on sex:
- For females: (1) $\hat{\beta}_1 + \hat{\beta}_3 = -.0256877 + .6213146 = .5956269$
- For males: (2) $\hat{\beta}_1 = -.0256877$

# Interactions: careful with tests

- Again, careful with tests. If we want to test if the effect of above average looks for *males* is significant, we can just look at the p-value for the coefficient of abvavg looks
- If we want to test if the effect is significant for females, we need to test $H_0 : \beta_1 = \beta_3 = 0$

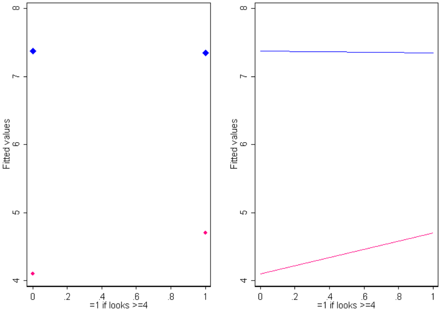
```
qui reg wage abvavg female abv_fem

 test abvavg abv_fem

 ( 1)  abvavg = 0
 ( 2)  abv_fem = 0

       F(  2,  1256) =    0.88
            Prob > F =    0.4170
```

- This seems fairly complicated but it **all starts with a clear understanding of the meaning of the model parameters**

# Interactions, graphically

- The more accurate graph is the one with dots: only four predicted values are possible but the lines help visualize the change

```
predict wagehat
scatter wagehat abvavg if female ==1, color(pink)  ||  scatter wagehat abvavg if female ==0, color(blue) ///
 legend(off) saving(int1.gph, replace)
line wagehat abvavg if female ==1, sort color(pink)  ||  line wagehat abvavg if female ==0, color(blue) ///
 legend(off) saving(int2.gph, replace)
graph combine int1.gph int2.gph
```



# In case you miss it...

- With a fully interacted model, we get four predicted means that are the same as creating a table with summary statistics by level of above average looks and sex:

```
 bysort abvavg female: sum wage
-> abvavg = 0, female = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |        585    7.376273    4.557269       1.05      38.86
-> abvavg = 0, female = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |        292    4.102637    2.149043       1.02      12.12
-> abvavg = 1, female = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |        239    7.350586    4.687264       1.46      41.67

-> abvavg = 1, female = 1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |        144    4.698264    6.436829       1.16      77.72
```

# Interactions and stratification

- What about if we estimated models separately for males and females? After all, we just saw that the effect is different for males and females

```
reg wage abvavg if female == 1
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      abvavg |   .5956269   .4167332     1.43   0.154    -.2234394    1.414693
       _cons |   4.102637   .2394948    17.13   0.000     3.631923    4.573351
------------------------------------------------------------------------------

reg wage abvavg if female == 0
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      abvavg |  -.0256877   .3527762    -0.07   0.942     -.718136    .6667605
       _cons |   7.376273   .1899917    38.82   0.000     7.003347    7.749199
------------------------------------------------------------------------------
```

- **Never forget**: The model with interactions is equivalent to a **stratified model**. If we had more covariates, say experience and education, you would need interactions between the female indicator and both experience and education (**triple highlight this**)

# A more complex model

- $wage_i = \beta_0 + \beta_1 abvavg + \beta_2 exper + \beta_3 female + \beta_4 abvavg * female_i + \beta_5 exper * female + \epsilon_i$
- Model for males (keeping other vars constant):
  $\beta_0 + \beta_1 \bar{abvavg} + \beta_2 \bar{exper}$
- Model for females (keeping other vars constant):
  $\beta_0 + \beta_1 \bar{abvavg} + \beta_2 \bar{exper} + \beta_3 + \beta_4 \bar{abvavg} + \beta_5 \bar{exper}$
- Testing $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ is testing whether there is *any* difference in models for females and males
- If there are, you may consider stratification. It is hard to present a **fully-interacted** model in a paper
- As I said before, never underestimate the power of interactions to get you all confused. Get used to do some math and make sure that you get the meaning of parameters right

# Digression

- As usual, language can be confusing and there is more than one way of teaching qualitative variables and interactions
- It is common to introduce indicator/dummy variables saying that adding an indicator variable is a model with different intercepts
- Adding interactions is a model with different intercepts and different slopes
- We cover the same already without using that kind of language (see last plot)
- I don't find that way of teaching very useful but it is a common way of introducing these ideas

# Interactions with continuous variables

- Not much changes but presentation tends to be a bit more difficult
- Suppose your model is: $y = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 age * educ + \epsilon$
- $\frac{\partial y}{\partial age} = \beta_1 + \beta_3 * educ$
- As before, the effect of *age* on *y* depends on the value of education
- You could present results choosing some meaningful values of education. For example 12, 16, 21 (high school, college, graduate school). For high school:
- $\frac{\partial y}{\partial age} = \beta_1 + \beta_3 * 12$
- A lot easier (for presentation) making education categories instead

# Other uses of indicator variables

- We saw that the relationship between wages and experience is better described by a curve than by a line
- $wage = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \epsilon$
- Similar to an interaction in the sense that there is no single effect of experience on wage; the change in average wages depends on the initial value of experience (take the derivative)
- By modeling experience by categories, we can take into account the non-linearity and make the model much easier to present
- Some statisticians **vehemently condemn** this practice but it's very common

# Graphically

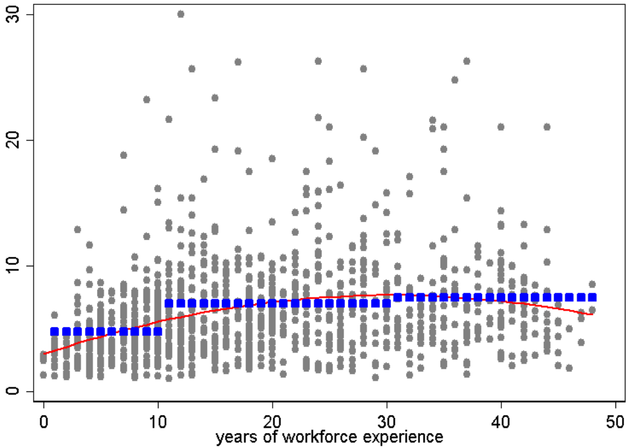
- Compare model with a quadratic term versus indicators for levels of experience

```
* Quadratic
reg wage c.exper##c.exper if wage < 30
predict wagehat2 if e(sample)

* Indicator variables
reg wage i.expcat
predict wagehatc if e(sample)

* Graph
scatter wage exper if wage < 30, color(gray) || line wagehat2 exper, color(red) sort ///
    || scatter wagehatc exper, color(blue) legend(off)
```

- Note the use of factor syntax to quickly create quadratic terms and interactions; I dropped high values of wages so the trend is easier to see

# Quadratic versus indicators

- We need to be careful when choosing categories



# Summary

- Qualitative variables or the categorization of a continuous variable adds a lot of flexibility in modeling
- Interpretation changes somewhat; never forget that everything is interpreted in relation to the reference category
- You change the reference category and the null of the Wald test changes
- ANOVA and ANCOVA are linear models, period
- Fully interacted and stratified models are equivalent
- **Be careful interpreting parameters**