# Biostatistics 200

## Review & Short Topics

# Interaction terms

- Always include each main effect as well as the interaction term
- Center your continuous variables
- Is one continuous and the other categorical?
  - This will divide the data into (2 or more) strata and then regress Y on the continuous variable
- Are they both categorical?
  - This just divides the data into strata for a regression of Y on any other variables (=ANOVA)
- Are both the variables continuous?
  - Statistically an interaction is just multiplying them together but how to interpret this?

# Using the Class dataset

Gender almost predicts the amount of cash you are carrying

Height predicts the amount of cash you are carrying

```
. reg cash  sex

      Source |       SS           df       MS      Number of obs   =       104
-------------+----------------------------------   F(1, 102)       =      3.59
       Model |  7088.56199         1  7088.56199   Prob > F        =    0.0609
    Residual |  201316.202       102  1973.68825   R-squared       =    0.0340
-------------+----------------------------------   Adj R-squared   =    0.0245
       Total |  208404.764       103  2023.34722   Root MSE        =    44.426

------------------------------------------------------------------------------
        cash |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   16.76473   8.846201     1.90   0.061    -.7816665    34.31113
       _cons |   14.84379   13.24092     1.12   0.265     -11.4195    41.10709
------------------------------------------------------------------------------

. reg cash height

      Source |       SS           df       MS      Number of obs   =       103
-------------+----------------------------------   F(1, 101)       =      4.90
       Model |  9622.91663         1  9622.91663   Prob > F        =    0.0291
    Residual |  198222.331       101  1962.59734   R-squared       =    0.0463
-------------+----------------------------------   Adj R-squared   =    0.0369
       Total |  207845.248       102  2037.69851   Root MSE        =    44.301

------------------------------------------------------------------------------
        cash |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   2.304966   1.040942     2.21   0.029     .2400159    4.369915
       _cons |  -115.5096   69.81001    -1.65   0.101    -253.9939    22.97464
------------------------------------------------------------------------------
```

# Create new variable = Height*Sex

```
. reg cash height sex htsex
```

| Source | SS | df | MS | | | |
|---|---|---|---|---|---|---|
| Model | 15184.9449 | 3 | 5061.64829 | Number of obs | = | 103 |
| Residual | 192660.303 | 99 | 1946.06367 | F(3, 99) | = | 2.60 |
| | | | | Prob > F | = | 0.0563 |
| Total | 207845.248 | 102 | 2037.69851 | R-squared | = | 0.0731 |
| | | | | Adj R-squared | = | 0.0450 |
| | | | | Root MSE | = | 44.114 |

| cash | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| height | 3.885287 | 1.873133 | 2.07 | 0.041 | .1685854 | 7.601989 |
| sex | 250.2801 | 170.7333 | 1.47 | 0.146 | -88.49177 | 589.052 |
| htsex | -3.821873 | 2.521609 | -1.52 | 0.133 | -8.825293 | 1.181547 |
| _cons | -222.5126 | 130.7697 | -1.70 | 0.092 | -481.9881 | 36.9629 |

In a t-test comparing cash by sex we see that on average Males = $48.37 and Females = $31.61 )p = 0.061 and for height Males = 69.7 in, Females = 64.94 (p<0.001)

In my model, Males=0 and Females=1

Equation says 3.89*Ht + 250*sex – 3.82*Htsex – 222.5 = Cash

# Multinomial Logit (sometimes ordered)

- Multinomial Logit strategy - Contrast outcomes with a common reference point
    - Similar to conducting a series of 2-outcome logit models comparing pairs of categories
    - The “reference category” is like the reference group when using dummy variables in regression
- It serves as the contrast point for all analyses  (use mlogit in Stata)
- Example: Multinomial Logit strategy for types of graduate schools
    - Analysis of 5 categories yields 4 tables of results:
        - No grad school vs. MA
        - No grad school vs. MBA
        - No grad school vs. Prof’l school
        - No grad school vs. PhD.

# Multinomial Logit Example:  Family Vacation

- Mode of Travel.  Reference category = Train

Large families less likely to take bus (vs. train)

```
. mlogit mode income familysize

Multinomial logistic regression                   Number of obs   =        152
                                                  LR chi2(4)      =      42.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -138.68742                       Pseudo R2       =     0.1332

------------------------------------------------------------------------------
        mode |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Bus          |
      income |   .0311874   .0141811     2.20   0.028     .0033929    .0589818
 family size |  -.6731862   .3312153    -2.03   0.042    -1.322356   -.0240161
       _cons |  -.5659882    .580605    -0.97   0.330    -1.703953    .5719767
-------------+----------------------------------------------------------------
Car          |
      income |    .057199   .0125151     4.57   0.000     .0326698    .0817282
 family size |   .1978772   .1989113     0.99   0.320    -.1919817    .5877361
       _cons |  -2.272809   .5201972    -4.37   0.000    -3.292377   -1.253241
------------------------------------------------------------------------------
(mode==Train is the base outcome)
```

Note:  It is hard to directly compare Car vs. Bus in this table

# Multinomial Logit Example: Family Vacation

- Mode of Travel. Reference category = Car

```
. mlogit mode income familysize, base(3)

Multinomial logistic regression                   Number of obs   =        152
                                                  LR chi2(4)      =      42.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -138.68742                       Pseudo R2       =     0.1332

------------------------------------------------------------------------------
        mode |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Train        |
      income |   -.057199   .0125151    -4.57   0.000    -.0817282   -.0326698
 family size |  -.1978772   .1989113    -0.99   0.320    -.5877361    .1919817
       _cons |   2.272809   .5201972     4.37   0.000     1.253241    3.292377
-------------+----------------------------------------------------------------
Bus          |
      income |  -.0260117   .0139822    -1.86   0.063    -.0534164     .001393
 family size |  -.8710634   .3275472    -2.66   0.008    -1.513044   -.2290827
       _cons |   1.706821   .6464476     2.64   0.008      .439807    2.973835
------------------------------------------------------------------------------
(mode==Car is the base outcome)
```

Here, the pattern is clearer: Wealthy & large families use cars

# Multinomial Logit - Stata Notes: mlogit

- Dependent variable: any categorical variable
  - Don’t need to be positive or sequential
  - Ex: Bus = 1, Train = 2, Car = 3
    - Or: Bus = 0, Train = 10, Car = 35
- Base category can be set with option:
  - mlogit mode income familysize, baseoutcome(3)
- Exponentiated coefficients called “relative risk ratios”, rather than odds ratios
  - mlogit mode income familysize, rrr

# MLogit Example: Car vs. Bus vs. Train

- Exponentiated coefficients: relative risk ratios

```
Multinomial logistic regression                   Number of obs   =        152
                                                  LR chi2(4)      =      42.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -138.68742                       Pseudo R2       =     0.1332

------------------------------------------------------------------------------
        mode |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Train        |
      income |   .9444061   .0118194    -4.57   0.000     .9215224    .9678581
  familysize |   .8204706   .1632009    -0.99   0.320     .5555836    1.211648
-------------+----------------------------------------------------------------
Bus          |
      income |   .9743237   .0136232    -1.86   0.063     .9479852    1.001394
  familysize |   .4185063   .1370806    -2.66   0.008     .2202385    .7952627
------------------------------------------------------------------------------
(mode==Car is the base outcome)
```

exp(-.057)=.94. Interpretation is just like odds ratios… BUT comparison is with reference category.

# Multinomial Logit: Predicted Probabilities

- You can predict probabilities for each case
  - Each outcome has its own probability (they add up to 1)

```
. predict predtrain predbus predcar if e(sample), pr

. list predtrain predbus predcar
```

| | predtrain | predbus | predcar |
|---|---|---|---|
| 1. | .3581157 | .3089684 | .3329159 |
| 2. | .448882 | .1690205 | .3820975 |
| 3. | .3080929 | .3106668 | .3812403 |
| 4. | .0840841 | .0562263 | .8596895 |
| 5. | .2771111 | .1665822 | .5563067 |
| 6. | .5169058 | .279341 | .2037531 |
| 7. | .5986157 | .2520666 | .1493177 |
| 8. | .3080929 | .3106668 | .3812403 |
| 9. | .0934616 | .1225238 | .7840146 |
| 10. | .6262593 | .1477046 | .2260361 |

This case has a high predicted probability of traveling by car

This probabilities are pretty similar here…

# Multinomial Logit: Predicted Probabilities

- Similar to the logit command, you can show how probabilies change across independent variables
  - Manually compute mean of predicted probabilities
  - Other variables will be left “as is” (ceteris paribus) unless you set them manually before you use “predict”

```
. mean predcar, over(familysize)

----------------------------
        Over |          Mean
-------------+--------------
predcar      |
           1 |      .2714656
           2 |      .4240544
           3 |      .6051399
           4 |      .6232910
           5 |      .8719671
           6 |      .8097709
```

Probability of using car increases with family size

Note: Values bounce around because other vars are not set to common value.

Note 2: Again, scatter plots aid in summarizing such results

# Multinomial Logit Assumptions

- Can also look at *ologit* (ordered outcomes) – similar setup in Stata
- Multinomial logit is designed for outcomes that are **not complexly interrelated**
- Critical assumption:  Independence of Irrelevant Alternatives (IIA)
    - Odds of one outcome versus another should be **independent** of other alternatives
    - Problems often come up when dealing with individual choices…aren't most of our choices somewhat interrelated?

# Power & Sample Size in Regression

- Sample size will depend on the number of predictors
- Power will depend on sample size AND the magnitude of the effect (i.e.; the size of the coefficients)
- In simple linear regression with 1 predictor we can determine this by comparing the slope (=coefficient on b) to a slope of 0 or ???
- In multiple regression you can find the sample size by knowing the number of predictors & the $R^2$ value
- Using the Power & Sample Size menu in Stata