Week 9: Modeling

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

Updated notes are here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Outline

- A collection of modeling techniques and tricks
- Big picture: We transform variables mostly for two reasons:
  1. Making the assumptions of models more plausible (typically involving the outcome variable)
  2. Presentation and interpretation (typically involving the explanatory variables)
- The most important part is that you understand that some transformation of variables imply that parameters are interpreted in a different way
- And by now you should recall that if the **parameters have a different meaning, so does the the null of the Wald test**

Centering

- We have seen several times that the meaning of the intercept is not that useful
- In the college GPA model:

  $colgpa_i = \beta_0 + \beta_1 hsgpa_i + \beta_2 act_i + \epsilon_i$
- The intercept is an **extrapolation**: average college GPA for those with a high school GPA of zero and ACT score of zero
- We can make the intercept more useful by **centering** the predictors at some value, usually the average (but it could be any value)
- For example, the average HS GPA is 3.4 and the average ACT score is 24

Centering

- We create two new variables:

```
 reg colgpa hsgpa act
...
 ------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsgpa |   .4534559   .0958129     4.73   0.000     .2640047    .6429071
         act |    .009426   .0107772     0.87   0.383    -.0118838    .0307358
       _cons |   1.286328   .3408221     3.77   0.000      .612419    1.960237
------------------------------------------------------------------------------
* Create new vars
 gen hsgpa_c = hsgpa - 3.4
gen act_c    = act - 24
```

- The model becomes:

  $$colgpa_i = \gamma_0 + \beta_1(hsgpa - 3.4)_i + \beta_2(act - 24)_i + \epsilon_i$$

- Now γ_0 has a different meaning: it is the average college GPA for those of average HS GPA and average ACT scores

Centering

- Nothing else has changed in the model
  $colgpa_i = \gamma_0 + \beta_1(hsgpa - 3.4)_i + \beta_2(act - 24)_i + \epsilon_i$
- You can rewrite as
  $colgpa_i = (\gamma_0 - \beta_1 3.4 - \beta_2 24) + \beta_1 hsgpa_i + \beta_2 act_i + \epsilon_i$
- In other words, the **interpretation of the coefficients (other than intercept)** for *hsgpa* and *act* is the same

```
reg colgpa hsgpa_c act_c

      Source |       SS           df       MS      Number of obs   =       141
-------------+----------------------------------   F(2, 138)       =     14.78
       Model |  3.42365514         2  1.71182757   Prob > F        =    0.0000
    Residual |  15.9824443       138  .115814814   R-squared       =    0.1764
-------------+----------------------------------   Adj R-squared   =    0.1645
       Total |  19.4060994       140  .138614996   Root MSE        =    .34032

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hsgpa_c |   .4534559   .0958129     4.73   0.000     .2640047    .6429071
       act_c |    .009426   .0107772     0.87   0.383    -.0118838    .0307358
       _cons |   3.054302   .0287056   106.40   0.000     2.997542    3.111062
------------------------------------------------------------------------------
```

Centering

- By the way, we also know that 3.05 is the **unconditional expectation** of *colgpa* (why?)
- Centering is also helpful presenting interactions of continuous variables:

  $colgpa_i = \gamma_0 + \gamma_1 hsgpa\_c + \gamma_2 act\_c + \gamma_3 hsgpa\_c * act\_c + \epsilon_i$
- Remember, with continuous variables, interactions are not so easy to interpret; easier with indicator variables, but that's what centering is doing in a sense
- Now, for example, γ_1 is the change in average college GPA for a small change in HS GPA for students with *average ACT scores*. Similar interpretation for γ_2
- Same as $\frac{\partial E[colgpa]}{\partial hsgpa} = \beta_1 + \beta_3(act = 24)$ in the uncentered interacted model

Centering

- Note how main effects change and intercept change (but not interaction)

```
* Centered
reg colgpa c.hsgpa_c##c.act_c
------------------------------------------------------------------------------------
           colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+-----------------------------------------------------------------
          hsgpa_c |   .4330686   .0967374     4.48   0.000      .241777    .6243603
            act_c |   .0103553   .0107684     0.96   0.338    -.0109384    .0316491
                  |
c.hsgpa_c#c.act_c |   .0485297   .0361728     1.34   0.182    -.0229995    .1200588
                  |
            _cons |   3.039022   .0308056    98.65   0.000     2.978106    3.099938
------------------------------------------------------------------------------------
* Uncentered
reg colgpa c.hsgpa##c.act
--------------------------------------------------------------------------------
       colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+-----------------------------------------------------------------
        hsgpa |   -.731644   .8884938    -0.82   0.412    -2.488579    1.025291
          act |  -.1546457    .122766    -1.26   0.210    -.3974069    .0881156
              |
c.hsgpa#c.act |   .0485297   .0361728     1.34   0.182    -.0229995    .1200588
              |
        _cons |   5.278084   2.994696     1.76   0.080    -.6437207    11.19989
--------------------------------------------------------------------------------
. di _b[hsgpa] + _b[c.hsgpa#c.act]*24
.43306863
```

Centering

- Not sure why centering is not used more often
- If you have **interactions of continuous variables, centering should be your first thought**
- **Be careful making predictions** with a centered model
- For example, if you want to predict college GPA for those with HS GPA of 3, you need to plug in -0.4, not 3
- Not a big deal. Make predictions using the uncentered model. Centering is done for **presentation and interpretation**
- Next semester, when you cover **regression discontinuity**, you will see that centering is useful because you want to interpret a parameter at one particular point (the cut-off point), so you center at that point

Changing scales

- We can of course change the scale of variables and we should expect that statistical inference will remain the same; parameter interpretation will change
- For example, we saw that age increases income in the GSS dataset:
- $realrinc_i = \beta_0 + \beta_1 age_i + \epsilon$
- We interpret β_1 as the change in average real income for a one year increase in age
- But that's not the most useful way to measure age. A one year increase is not that meaningful. A ten year increase would be perhaps more useful
- In this simple model, we could just calculate the increase for 10 years. It's $10 * \beta_1$; or we could **recode age in decades**

Rescaling

- Recoding age in decades; nothing other than the coefficient for age changes

```
qui reg realrinc age
est sto m1

gen aged = age/10

qui reg realrinc aged
est sto m2

est table m1 m2, star stats(N r2 ll F)
```

```
----------------------------------------------
    Variable |      m1              m2
-------------+--------------------------------
         age |  454.82891**
        aged |                   4548.2891**
       _cons |  12852.508        12852.508
-------------+--------------------------------
           N |       1186             1186
          r2 |  .00674942        .00674942
          ll | -15009.357       -15009.357
           F |  8.0456147        8.0456145
----------------------------------------------
      legend: * p<0.05; ** p<0.01; *** p<0.001
```

Rescaling

- What about if the model is number of children on income?

```
reg children realrinc

      Source |       SS           df       MS      Number of obs   =     1,189
-------------+----------------------------------   F(1, 1187)      =      0.96
       Model |  2.24240941         1  2.24240941   Prob > F        =    0.3268
    Residual |  2765.37071     1,187  2.32971416   R-squared       =    0.0008
-------------+----------------------------------   Adj R-squared   =   -0.0000
       Total |  2767.61312     1,188  2.32964067   Root MSE        =    1.5263

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    realrinc |   5.71e-07   5.82e-07     0.98   0.327    -5.71e-07    1.71e-06
       _cons |   1.668625   .0481168    34.68   0.000     1.574222    1.763029
------------------------------------------------------------------------------
```

- Not a great way of seeing the effect of income on the number of children. The coefficient of realinc is close to zero
- **Don't ever do this**. It takes less than 10 seconds to recode a variable and **you risk making a reviewer angry**

Rescaling

- Better, but still not great. At three decimals, the coefficient is 0.000; you could sill make it better by expressing the change by 10K or 5K increments (as we did for age)
- Again, this is purely to help with presentation

```
gen incthou = realrinc/1000

reg children incthou

      Source |       SS           df       MS      Number of obs   =     1,189
-------------+----------------------------------   F(1, 1187)      =      0.96
       Model |  2.24240939         1  2.24240939   Prob > F        =    0.3268
    Residual |  2765.37071     1,187  2.32971416   R-squared       =    0.0008
-------------+----------------------------------   Adj R-squared   =   -0.0000
       Total |  2767.61312     1,188  2.32964067   Root MSE        =    1.5263

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     incthou |   .0005713   .0005823     0.98   0.327    -.0005712    .0017137
       _cons |   1.668625   .0481168    34.68   0.000     1.574222    1.763029
------------------------------------------------------------------------------
```

"Beta" coefficients

- Consider adding education to the model

```
. reg children incthou educ

      Source |       SS           df       MS      Number of obs   =     1,189
-------------+----------------------------------   F(2, 1186)      =     21.14
       Model |   95.277586         2   47.638793   Prob > F        =    0.0000
    Residual |  2672.33553     1,186  2.25323401   R-squared       =    0.0344
-------------+----------------------------------   Adj R-squared   =    0.0328
       Total |  2767.61312     1,188  2.32964067   Root MSE        =    1.5011

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     incthou |   .0014315   .0005881     2.43   0.015     .0002777    .0025853
        educ |  -.0997509   .0155237    -6.43   0.000     -.130208   -.0692939
       _cons |   3.020157   .2155896    14.01   0.000     2.597178    3.443137
------------------------------------------------------------------------------
```

- We can't compare the **magnitude** of the coefficients to determine how important they are in explaining the outcome; after all, we just saw that we can change the size of the coefficients by changing the scale
- One trick is to express the coefficients in the **same scale**

Beta coefficients

- We have seen before that we can **standardize** a variable by subtracting its mean and dividing by the standard deviation $z_i = \frac{x_i - \bar{x}}{\sigma}$
- Then z will have a mean of zero and standard deviation of 1
- The idea behind a regression with so-called **beta coefficients** (yes, not the best name) is to standardize **all** variables
- The main advantage is that the **size of the coefficients tell you how important a variable is** in terms of it effect on the outcome because
- Now all of them are measured in the same scale and **a small change is a 1 standard deviation**. If linear, a 1 standard deviation change

Beta coefficients

- Example: **hedonic** pricing. What is the effect of pollution on housing prices?

```
* Get Wooldridge data
bcuse hprice2

reg price nox crime rooms dist stratio

      Source |       SS           df       MS      Number of obs   =       506
-------------+----------------------------------   F(5, 500)       =    174.47
       Model |  2.7223e+10         5  5.4445e+09   Prob > F        =    0.0000
    Residual |  1.5603e+10       500  31205611.6   R-squared       =    0.6357
-------------+----------------------------------   Adj R-squared   =    0.6320
       Total |  4.2826e+10       505    84803032   Root MSE        =    5586.2

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nox |  -2706.433   354.0869    -7.64   0.000    -3402.114   -2010.751
       crime |   -153.601   32.92883    -4.66   0.000    -218.2969   -88.90504
       rooms |   6735.498   393.6037    17.11   0.000     5962.177    7508.819
        dist |  -1026.806   188.1079    -5.46   0.000    -1396.386    -657.227
     stratio |  -1149.204   127.4287    -9.02   0.000    -1399.566   -898.8422
       _cons |   20871.13   5054.599     4.13   0.000     10940.26       30802
------------------------------------------------------------------------------
```

- *nox* is a measure of nitrogen oxide in the air over each community

Beta coefficients

- We can standardize variables "by hand" or use the egen command

```
* Hand (i.e. the hand of Stata)
qui sum price
gen zprice = (price - r(mean))/r(sd)

* Easier, use the egen function std() for all variables
foreach var of varlist price nox crime rooms dist stratio {
   egen z`var'=std(`var')
}

* Regress
reg zprice znox zcrime zrooms zdist zstratio
...
------------------------------------------------------------------------------
      zprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        znox |   -.340446   .0445411    -7.64   0.000    -.4279568   -.2529352
      zcrime |  -.1432828   .0307168    -4.66   0.000    -.2036327   -.0829328
      zrooms |   .5138878   .0300302    17.11   0.000      .454887    .5728887
       zdist |  -.2348385   .0430217    -5.46   0.000    -.3193642   -.1503129
    zstratio |  -.2702799   .0299698    -9.02   0.000    -.3291622   -.2113976
       _cons |   6.61e-09   .0269672     0.00   1.000    -.0529829    .0529829
------------------------------------------------------------------------------
```

- Now we can compare the size of the coefficients. And: a one st dv increase in nox decreases price by 0.34 st dvs

Beta coefficients

- Note that the intercept is zero
- **We also standardized price**; otherwise changes would be in the price scale
- Stata has a **beta** option for regress

```
reg price nox crime rooms dist stratio, beta

      Source |       SS           df       MS      Number of obs   =       506
-------------+----------------------------------   F(5, 500)       =    174.47
       Model |  2.7223e+10         5  5.4445e+09   Prob > F        =    0.0000
    Residual |  1.5603e+10       500  31205611.6   R-squared       =    0.6357
-------------+----------------------------------   Adj R-squared   =    0.6320
       Total |  4.2826e+10       505    84803032   Root MSE        =    5586.2

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
         nox |  -2706.433   354.0869    -7.64   0.000                 -.340446
       crime |   -153.601   32.92883    -4.66   0.000                -.1432828
       rooms |   6735.498   393.6037    17.11   0.000                 .5138878
        dist |  -1026.806   188.1079    -5.46   0.000                -.2348385
     stratio |  -1149.204   127.4287    -9.02   0.000                -.2702799
       _cons |   20871.13   5054.599     4.13   0.000                        .
------------------------------------------------------------------------------
```

- Replaces CIs for Beta coefficients

Beta coefficients

- Could also leave price in the original scale

```
qui reg price nox crime rooms dist stratio
est sto ori
qui reg zprice znox zcrime zrooms zdist zstratio
est sto m1
qui reg price znox zcrime zrooms zdist zstratio
est sto m2
est table ori m1 m2, stats(N r2)
-----------------------------------------------------
    Variable |    ori           m1            m2
-------------+---------------------------------------
         nox | -2706.4326
       crime | -153.60097
       rooms |  6735.4983
        dist | -1026.8063
     stratio | -1149.2038
        znox |              -.34044602   -3135.1184
      zcrime |              -.14328275   -1319.4702
      zrooms |               .51388784    4732.3191
       zdist |              -.23483854   -2162.5943
    zstratio |              -.27027989   -2488.9686
       _cons |  20871.127    6.608e-09     22511.51
-------------+---------------------------------------
           N |        506          506          506
          r2 |   .6356658    .63566579    .63566579
-----------------------------------------------------
```

Related but a digression

- Here we standardized so we can compare the contribution of some variables
- But we could standardize any of them so the parameter can be interpreted as change in 1 standard deviation
- Sometimes the measurement units do not mean much so it's helpful to think about the relevant units
- For example, if a predictor is a depression scale, what does it mean a unit change? Would 10 points be better?

Beta coefficients; back to those children

```
reg children incthou educ, beta

      Source |       SS           df       MS      Number of obs   =     1,189
-------------+----------------------------------   F(2, 1186)      =     21.14
       Model |   95.277586         2  47.638793   Prob > F        =    0.0000
    Residual |  2672.33553     1,186  2.25323401   R-squared       =    0.0344
-------------+----------------------------------   Adj R-squared   =    0.0328
       Total |  2767.61312     1,188  2.32964067   Root MSE        =    1.5011

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
     incthou |   .0014315   .0005881     2.43   0.015                 .0713254
        educ |  -.0997509   .0155237    -6.43   0.000                -.1882889
       _cons |   3.020157   .2155896    14.01   0.000                        .
------------------------------------------------------------------------------
```

- Education is more important than income but in the original scale $.0997509/.0014315 = 69.68$. We know this is meaningless
- Instead: $.1882889/\ .0713254 = 2.64$
- A linear model is not the best here; the **number of children is not normally distributed** (Poisson or negative binomial would be better)

Splines

- Linear relationships are easy to estimate and easy to interpret
- Splines are a way to divide relationships that are non-linear **into linear pieces connected by "knots"**
- They are fairly useful to a) accommodate non-linearities
- And b) great for testing changes in trends; used more commonly in longitudinal data
- **WARNING**: The **coding of splines can be utterly confusing** and there is more than one way of doing it (so careful if you google)

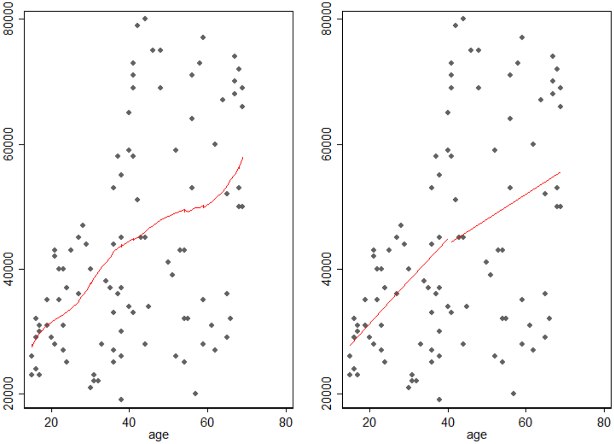
Example

- Example data from Stata on income vs age

```
* Get data
webuse mksp1

* See trend with lowess
lowess income age, gen(linc)
scatter income age || line linc age, color(red) sort

* Estimate separate models for before and after 40
scatter income age || lfit linc age if age <=40 || ///
               lfit linc age if age > 40
```

- What about if we wanted to test that the slope before 40 is the same as the slope after 40?
- If we estimated two models (just like in the graph above) we get an estimate of before and after 40, but not a statistical test

Seeing trends



- Note that the two linear pieces are not connected

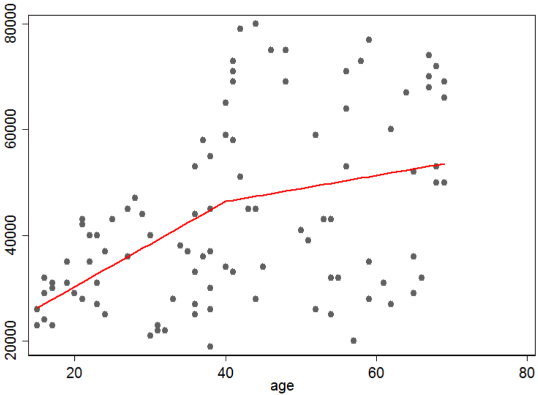
Splines

- We will use splines to model two lines joined by a knot at 40
  $income = \beta_0 + \beta_1 age + \beta_2 (age - k)_+ + \epsilon$
- The $(x)_+$ is called a **truncated line function** and is defined as being equal to x if x is positive and zero otherwise. k is the knot. In this example, $k = 40$ and $x = age - 40$
- It's similar to centering but we now make $(age - k)_+ = 0$ when age ≤ 40
- Again: $(age - k)_+$ will be equal to age (centered) if older than 40 and zero if less than 40
- The only difficult part about splines is to get the coding right, the rest is (relatively) easy

Splines

```
* Create truncated function
gen     aget = age - 40
replace aget = 0 if age <= 40

* Estimate model
reg income age aget
predict inchat

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(2, 97)        =     15.69
       Model |  7.1445e+09         2  3.5722e+09   Prob > F        =    0.0000
    Residual |  2.2078e+10        97   227605048   R-squared       =    0.2445
-------------+----------------------------------   Adj R-squared   =    0.2289
       Total |  2.9222e+10        99   295173333   Root MSE        =     15087

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   805.6953   210.8633     3.82   0.000       387.19    1224.201
        aget |  -559.8229   340.9786    -1.64   0.104    -1236.571    116.9253
       _cons |   14208.38   6680.308     2.13   0.036     949.8135    27466.94
------------------------------------------------------------------------------

* Plot
scatter income age || line inchat age, color(red) sort ///
    legend(off) saving(spli.gph, replace)
graph export sli.png, replace
```

Splines



- The two linear pieces are now connected
- **Important digression**: Is the model right? Probably not. We should compare it to others. **We just made it so because we ASSUMED a break at forty**

Understanding the model

- We estimated the model: $income = \beta_0 + \beta_1 age + \beta_2 (age - 40)_+$
- If age ≤ 40 the model is A) : $income = \beta_0 + \beta_1 age$
- If age > 40 the model is: $income = \beta_0 + \beta_1 age + \beta_2 (age - 40)$
- Same as centering, so if age > 40 the model is B):
  $income = (\beta_0 - \beta_2 * 40) + (\beta_1 + \beta_2) age$
- Compare A) and B). When are they going to be the same?
- If $\beta_2 = 0$, then the slope before and after is the same
- Note that β_2 is the *incremental* change in slope
- The trick of using the truncated function is that it allowed us the possibility of a different slope after 40

Testing if slope is the same before and after 40

- From the comparison of A) and B) it's clear that if we test the null $H_0 : \beta_2 = 0$ we are testing whether the slopes are the same before and after 40
- If we reject the null, then there is a change, which can be positive or negative
- From the output above, we do not reject the null: $p = 0.104$ so there is not enough evidence to suggest that there is a change in slope after 40
- See Stata's **mkspline** command for more ways of using splines;
- You can make cubic splines, assuming two or more non-linear lines with a knot

Paper example

From: **Incidence and Mortality of Hip Fractures in the United States**

JAMA. 2009;302(14):1573-1579. doi:10.1001/jama.2009.1462

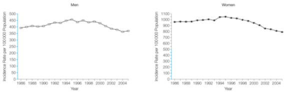


Figure Legend:

Data are based on a 20% sample of Medicare claims; error bars indicate 95% confidence intervals. P < .001 for a change in trend ir
1995. Regions of y-axes that are in blue indicate incidence rate of 0 to 500 per 100 000 population.

- Sample sizes were huge so not a lot of need of a test but...
- Very useful and flexible to test changes in trends, including a before and after policy change (with the caveat that the causal inference could be complicated)

Suggestions

- 1) Always verify that you coded splines correctly. Plot predicted values (this is generic example. Always plot predicted values to verify you code things correctly)
- 2) Write down the model for before and after the knot (remember the truncated function changes at the knot)
- 3) You can of course combine splines with interactions (homework)

Summary

- We will see more modeling issues next class
- This is important and the key is to understand the meaning of the parameters
- Once you get the meaning, hypothesis testing and modeling is easier
- More next class...