Week 5: Multiple Linear Regression

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

- Multiple linear regression estimation
- Algebraic properties, redux
- Matrix form
- Interpretation: "controlling," "adjusting," "taking account" other factors

Big picture

- We seldom run a regression model with just one predictor but for pedagogical reasons is always easier to understand the basics of regression with only one variable
- When modeling, it's useful to start with just one predictor/covariate of interest and then add variables
- We often want to describe relationships "controlling" for other factors
- Today, we will see different ways of understanding what it means to "control" or "adjust" for other variables and to "hold them constant" or "taking them into account"
- Note, once again, how language can be confusing. All are words to describe the **same idea** but you can **understand that idea in different ways**, hence the many terms

Estimation

- Estimation of parameters is the same as before. We want to minimize SSE:
- $S(\beta_0 ... \beta_p) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - ... - \hat{\beta}_p x_{ip})^2$
- Now, we will have $p + 1$ first order conditions (p parameters not counting the intercept) and a system of linear equations with $p + 1$ unknowns (I dropped the ^ because it's easier to write)

  $\frac{\partial SSE}{\partial \beta_0} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip}) = 0$

  $\frac{\partial SSE}{\partial \beta_1} = \sum_{i=1}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip}) = 0$

  ...

  $\frac{\partial SSE}{\partial \beta_p} = \sum_{i=1}^{n} x_{ip}(y_i - \beta_0 - \beta_1 x_i - ... - \beta_p x_{ip}) = 0$

Estimation

- Note that the **algebraic properties** of OLS will hold here too: the sum of the residuals will be zero and every covariate x will be uncorrelated to the residual (remember: in the **sample**, not the population)
- **Whenever you see a system of linear equations that need to be solved you should think about linear algebra**
- That's pretty much the *origin* of linear algebra. Liner algebra was developed (invented?) as shortcut to deal with multiple linear equations
- So no surprise here that we could write the regression model as $Y = X\beta + \epsilon$, where Y is a nx1 vector and X is an nx$(p+1)$ matrix, also know as the **design matrix**. β is also a vector of $p+1$ elements. ϵ is also a vector

Estimation

- From your textbook:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

- Again, just a way of writing a long set of equations in a very simple way. Minimizing the sum of squares can be written as:

  $S(\beta) = \epsilon'\epsilon = (Y - X\beta)'(Y - X\beta)$

Estimation

- The minimization problem will lead to
  $(X'X)\beta = X'Y$ (we could write $\hat{\beta}$ here)
- If the inverse of $(X'X)$ exists, then we can **left multiply** by $(X'X)^-1$ to get
- $\hat{\beta} = (X'X)^{-1}X'Y$, where $\hat{\beta}$ is a vector of $p + 1$ coefficients. **And that's it**. That gives us the OLS solution
- **Again**: We used linear algebra to more easily solve a complicated set of linear equations but the logic is the same as with simple linear regression
- Before, we were trying to draw the line that passes as close as possible to all points. We are still trying to do so but now the **line is a plane in multiple dimensions**
- If you know linear algebra well, the intuition behind the solution is that the optimal betas are an orthogonal projection. This **means nothing** if you don't know linear algebra well

Inference

- As before, we made no assumptions about the distribution of ϵ. We just solved a system of linear equations
- We again do need to assume that the error distributes normal idd $N(0, \sigma^2)$ to **figure out what is the distribution of the estimated coefficients** β
- We will find equivalent results about the distribution of $\hat{\beta}_j$ and the standard errors $se(\hat{\beta}_j)$ with the difference that now we have to think about **multivariate normals**
- The vector $\hat{\beta}$ has a (p+1)-variate normal distribution with mean vector β and **variance-covariance** matrix $\sigma^2(X'X)^{-1}$. Your textbook defines $C = (X'X)^{-1}$ so the variance-covariance matrix is $\sigma^2 C$
- If you are confused, please review joint distributions and the multivariate normal. See Wooldridge and `https://en.wikipedia.org/wiki/Multivariate_normal_distribution`

Inference

- **Important**: We can still use the intuition from the SLR model because each $\hat{\beta}_j$ has a **marginal** distribution that is normal with $se(\hat{\beta}_j) = \hat{\sigma}\sqrt{c_{jj}}$, where c_{jj} is the j diagonal element of C
- In other words, similar to SLR, we can do Wald test as with the SLR model
- We can do simulations about each β as we did in SLR
- **BUT**: since the $\hat{\beta}_j$ have a (p+1)-variate normal distribution, they also have a covariance (the off-diagonal elements of $\sigma^2 C$)
- If we are going to do simulation about **multiple parameters**, we do need to take into account their **covariance** (correlation) when doing simulations

College grades again

- We now add a variable called skipped, which has the average number of lectures missed per week:

  $$colgpa_i = \beta_0 + \beta_1 hsgpa_i + \beta_2 skipped_i + \epsilon_i$$

```
reg colgpa hsgpa skipped
      Source |       SS           df       MS      Number of obs   =       141
-------------+----------------------------------   F(2, 138)       =     19.77
       Model |  4.32237812         2  2.16118906   Prob > F        =    0.0000
    Residual |  15.0837213       138  .109302328   R-squared       =    0.2227
-------------+----------------------------------   Adj R-squared   =    0.2115
       Total |  19.4060994       140  .138614996   Root MSE        =    .33061

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsgpa |   .4588039   .0876908     5.23   0.000     .2854125    .6321952
     skipped |  -.0774349   .0257646    -3.01   0.003    -.1283792   -.0264905
       _cons |   1.579167   .3033669     5.21   0.000     .9793183    2.179015
------------------------------------------------------------------------------
```

- What is the null of the F(2,138) test? What is the null of the Wald tests?

Variance-covariance matrix

- Stata saves the variance-covariance matrix

```
 ereturn list
scalars:
                  e(N) =  141
               e(df_m) =  2
     ...
macros:
            e(cmdline) : "regress colgpa hsgpa skipped"
          e(estat_cmd) : "regress_estat"
matrices:
                  e(b) :  1 x 3
                  e(V) :  3 x 3
matrix list e(V)
symmetric e(V)[3,3]
              hsgpa     skipped       _cons
  hsgpa   .00768968
skipped   .00020257   .00066381
  _cons  -.02637928  -.00140361    .0920315
di sqrt(.00768968)
.08769082
 * note it matches the SE of hsgpa
```

- The covariance of the coefficient for hsgpa and skipped is .00020257. We typically don't use this information in empirical work but we would need it to make simulations about the betas

Interpretation

- We can interpret the parameters as before with a slight modification: The average college GPA increases by 0.45 points for one point increase in high school GPA **holding the number of classes skipped constant**
- The average college GPA decreases 0.08 points for an additional class skipped during the semester **holding high school GPA constant**
- The average college GPA for a student with a GPA of zero who didn't skip class during the semester is 1.57 (an extrapolation, of course)
- **The difficult part is, what exactly do we mean by "holding constant" high school GPA or classes skipped?**
- The language comes from **experiments** in which investigators can keep one condition **unchanged (constant)** while changing another

Holding constant?

- One way in which holding constant (*ceteris paribus*) is clear is that the the coefficient of a variable is the **partial change** when the other variables are not changing:

  $colgpa_i = \beta_0 + \beta_1 hsgpa_i + \beta_2 skipped_i + \epsilon_i$
- **Total** change: $\Delta \hat{colgpa} = \hat{\beta}_1 \Delta hsgpa + \hat{\beta}_2 \Delta skipped$
- For a small **partial** change: $\frac{\partial \hat{colgpa}}{\partial hsgpa} = \hat{\beta}_1$
- (I'm simplifying the notation but what is changing is $E[colgpa|X]$)
- The confusing part is that it sounds like we conducted an experiment is which we were able to study students that, say, skip no classes but have different hsgpa. Another way: we found students who skip two classes but who have different hsgpa...
- With observational data, we can't hold other factors constant in this more **literal meaning**, but we are trying to **replicate** that situation with a regression analysis

Partialling out interpretation (**read this slide slowly and carefully**)

- Another common way of describing a model is to say that the coefficient of hsgpa is the effect on average colgpa after "**adjusting**" or "**taking into account**" the effect of skipping classes
- Our original model is: $colgpa_i = \beta_0 + \beta_1 hsgpa_i + \beta_2 skipped_i + \epsilon_i$

  1) Estimate $hsgpa = \gamma_0 + \gamma_1 skipped + \mu$

  This is the effect of skipping classes on *high school* GPA

  2) Calculate the residuals of the above regression:

  $\hat{\mu}_i = hsgpa_i - \hat{hsgpa}_i$

  The residuals are the part of high school GPA that is not explained by skipping classes

  3) Estimate the regression $colgpa = \alpha_0 + \alpha_1 \hat{\mu}_i + \lambda$

  In this model, we used the unexplained (by skipped) part of hsgpa to explain colgpa. Or, after we take into account the effect of skipping classes, what is the partial effect of hsgpa on colgpa?
- **We will find that** $\hat{\alpha}_1 = \hat{\beta}_1$

Convince yourself

- We estimated: $colgpa_i = 1.57 + 0.4588 hsgpa_i - 0.077 skipped_i$

```
reg hsgpa skipped
 ...
------------------------------------------------------------------------------
       hsgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     skipped |  -.0263437   .0248204    -1.06   0.290     -.075418    .0227306
       _cons |    3.43048   .0379318    90.44   0.000     3.355482    3.505478
------------------------------------------------------------------------------
predict hs_r, r
reg colgpa hs_r
...
------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hs_r |   .4588039   .0911463     5.03   0.000     .2785914    .6390164
       _cons |   3.056738   .0289395   105.63   0.000     2.999519    3.113956
------------------------------------------------------------------------------
```

Not the only way of partialling out

- Your textbook presents another way of understanding partialling out

```
* partialling out (Chatterjee)
qui reg colgpa skipped
predict colres, r
* The part of colgpa that is not explained by skipped

qui reg hsgpa skipped
predict hsgres, r
* The part of hsgpa not explained by skipped

reg colres hsgres
...
------------------------------------------------------------------------------
      colres |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hsgres |   .4588039   .0873748     5.25   0.000     .2860484    .6315594
       _cons |   1.11e-09    .027742     0.00   1.000    -.0548508    .0548508
------------------------------------------------------------------------------
```

- We removed the effect of skipped from both colgpa and hsgpa. Then we regressed both unexplained (by skipped) parts. So the effect of hsgres on colres is the effect once skipped has been taken into account
- Note that now the standard error matches the original standard error

Useful to stop and think about hsgpa and skipped

- Let's compare the regressions

```
qui reg colgpa hsgpa
est sto m1
qui reg colgpa hsgpa skipped
est sto m2
est table m1 m2, star stats(N r2_a)  b(%7.3f)
----------------------------------------
    Variable |     m1           m2
-------------+--------------------------
       hsgpa |   0.482***     0.459***
     skipped |               -0.077**
       _cons |   1.415***     1.579***
-------------+--------------------------
           N |     141          141
        r2_a |   0.166        0.211
----------------------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001
```

- We reject the null that skipped=0. Also, note that the coefficient of hsgpa went down, from 0.48 to 0.46
- Some of the effect of hsgpa is **accounted/explained** for skipping classes. Look at the regression of hsgpa and skipped above. Students who skip more classes have a lower high school GPA

When are they going to be the same?

- If skipped and hsgpa were completely uncorrelated, then in the regression, reg hsgpa skipped, the coefficient for skipped would be zero and the variance of the residuals would be the same as the unexplained variance
- So adjusting would not change the coefficient for hsgpa
- **This is why in a clinical trial adjusting for other variables doesn't change the estimation of treatment effects. Treatment and other factors are independent**
- Remember, if two variables are independent, their correlation/covariance is zero

Confounder, mediator, or modifier?

- Skipping classes is a *counfounder* of the relationship between high school GPA and college GPA: skipped is related to both
- But we need to know more about the subject matter (mechanism) to figure out if skipping classes is a **mediator**. It's not that hsgpa causes lower colgpa, it's that low hsgpa causes skipping classes which in turns causes lower colgpa (probably not)
- Note that we have implicitly **assumed** that skipping classes is not a **modifier**. In our model, the effect of hsgpa on colgpa does not depend on skipping classes
- We can check for it by adding an interaction:
  $colgpa_i = \beta_0 + \beta_1 hsgpa_i + \beta_2 skipped_i + \beta_3 hsgpa * skipped + \epsilon_i$
- Now $\frac{\partial colgpa}{\partial hsgpa} = \beta_1 + \beta_3 skipped$
- So now there is not a single effect of hsgpa (nonlinear). Instead, the effect of hsgpa on colgpa depends on class skipped

Summary

- Most of what we learned in SLR easily extends to MLR
- We interpret coefficients in a different way and in MLR. We need to adjust, control for, hold something constant, take into account... Many ways of saying the same
- Under the assumption of iid errors that distribute normal, the estimated coefficients have a multvariate normal distribution with normal marginal distributions
- The key with MRL is to gain intuition about what it means holding other factors constant or adjusting for other factors
- Next, inference and comparing models using F tests