Week 5: Multiple Linear Regression II

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

- Adjusted R^2
- More on testing hypotheses in linear models

R^2 versus R_a^2 (adjusted)

- We saw before that the **goodness of fit** of a linear regression can be measured by $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- This is equivalent to $[cor(\hat{y}, y)]^2$
- We can still use this measure but when we **compare models that have different number of predictors** it is better to to take into account the number of predictors
- In the linear model, R^2 will always increase (or not decrease) when we add more parameters, regardless of whether they are relevant or not
- The "adjusted" (for the number of parameters) model is $R_a^2 = 1 - \frac{\frac{SSE}{(n-p-1)}}{\frac{SST}{(n-1)}}$
- Note that the more parameters we estimate the larger is p and the more SSE is penalized

Example

- Stata shows these quantities in the ANOVA table

```
. reg colgpa hsgpa male skipped

      Source |       SS           df       MS      Number of obs   =       141
-------------+----------------------------------   F(3, 137)       =     13.30
       Model |  4.37665441         3  1.4588848    Prob > F        =    0.0000
    Residual |   15.029445       137  .109703978   R-squared       =    0.2255
-------------+----------------------------------   Adj R-squared   =    0.2086
       Total |  19.4060994       140  .138614996   Root MSE        =    .33122

. di 1-.109703978/.138614996
.20857064
```

- But why an extra parameter reduces SSE? This is because $SST = SSR + SSE$, so $SSE = SST - SSR$. SST is not going to change (it's the unexplained, observed variance) but the more variables we add to the model the more we can "explain" with the regression, so SSR will tend to go down
- As usual, **remember the context**: we are talking about the vanilla linear model. **This is not true in non-linear models like logit or probit**. Adding more variables could make the model worse

Example

- Add (literally) random noise to the regression

```
gen noise = uniform()
qui reg colgpa hsgpa male skipped
est sto m1
reg colgpa hsgpa male skipped noise
------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsgpa |   .4710294   .0898855     5.24   0.000     .2932754    .6487834
        male |   .0409904   .0584738     0.70   0.484    -.0746451    .1566259
     skipped |   -.080972   .0265302    -3.05   0.003    -.1334371   -.0285069
       noise |  -.0026149   .0984739    -0.03   0.979     -.197353    .1921232
       _cons |   1.521224   .3209863     4.74   0.000     .8864544    2.155994
------------------------------------------------------------------------------
est sto m2
est table m1 m2, star stats(N r2 r2_a) b(%7.3f)
----------------------------------------
    Variable |     m1           m2
-------------+--------------------------
       hsgpa |   0.471***     0.471***
        male |   0.041        0.041
     skipped |  -0.081**     -0.081**
       noise |               -0.003
       _cons |   1.520***     1.521***
-------------+--------------------------
           N |     141          141
          r2 |   0.226        0.226
        r2_a |   0.209        0.203
----------------------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001
```

A couple of things to notice

- The parameter for noise is not significant, which makes sense
- None of the other coefficients were affected at all because noise is not correlated to any of them (verify)
- The R_a^2 went down, which is somewhat reassuring
- R^2 did not change at 3 decimals (actual numbers are 0.225530 vs 0.225534)
- **One more time**: Remember the context. This is true in linear models. In other models adding irrelevant variables may make the fit of the model *worse*

Small digression

- What if we add random noise that is **correlated to one of the covariates**?

```
gen noise2 = skipped*noise + rnormal(0,5)
             |   noise2  skipped   colgpa    hsgpa     male
-------------+---------------------------------------------
      noise2 |   1.0000
     skipped |   0.2573   1.0000
      colgpa |  -0.1022  -0.2618   1.0000
       hsgpa |  -0.0357  -0.0897   0.4146   1.0000
        male |   0.0422   0.2010  -0.0765  -0.2075   1.0000
qui reg colgpa hsgpa male skipped
est sto m1
qui reg colgpa hsgpa male skipped noise2
est sto m2
est table m1 m2, star stats(N r2 r2_a) b(%7.6f)
------------------------------------------
    Variable |     m1            m2
-------------+----------------------------
       hsgpa | 0.471037***   0.470484***
        male | 0.040885      0.040586
     skipped | -0.080895**    -0.078127**
      noise2 |               -0.002154
       _cons | 1.519816***   1.521428***
-------------+----------------------------
           N |      141           141
          r2 | 0.225530      0.226446
        r2_a | 0.208571      0.203694
------------------------------------------
  legend: * p<0.05; ** p<0.01; *** p<0.001
```

Hypotheses testing

- Nothing much has changed respect to Wald tests but now the degrees of freedom for the t-student are different
- For confidence intervals
- $\hat{\beta}_j \pm t_{(n-p-1,\alpha/2)} se(\hat{\beta}_j)$
- We need to take into account that we are now estimating p+1 parameters. $t_{(n-p-1,\alpha/2)}$ is still close to 2 and with large samples closer to 1.96 (as the z from the standard normal)
- We could do the same simulations we did before because we know that $\hat{\beta}_j$ distributes normal
- If we wanted to do simulations to do tests or probabilities about **two or more parameters** at the same time, we need to consider their covariance

Simulating from multivariate normals

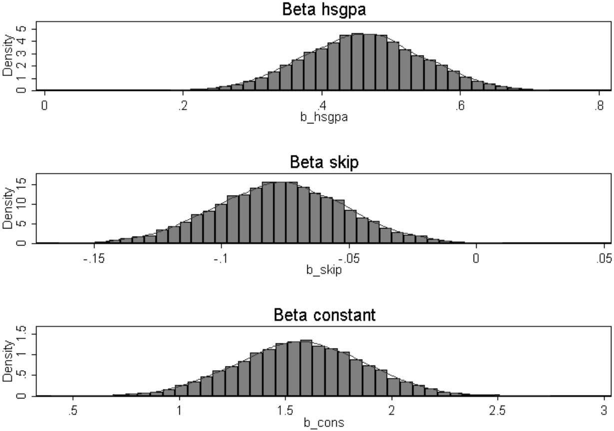
- It used to be a bit of a hassle to do this simulation but Stata now has a command to do it

```
qui reg colgpa hsgpa skipped

* Save coefficients and variance-covariance matrix
matrix M = e(b)
matrix V = e(V)
clear
* won't delete matrices
matrix list M
matrix list V
* Simulate 10,000 draws from multivariate normal with mean M and var-covar V
drawnorm b_hsgpa b_skip b_cons, n(10000) cov(V) means(M)
sum
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     b_hsgpa |     10,000    .4604699     .088031   .1144291    .807568
      b_skip |     10,000   -.0771271    .0258411  -.1671696    .017728
      b_cons |     10,000    1.573506    .3040728   .3955161   2.747055
. corr
             |  b_hsgpa   b_skip   b_cons
-------------+---------------------------
     b_hsgpa |   1.0000
      b_skip |   0.0837   1.0000
      b_cons |  -0.9918  -0.1727   1.0000
```

Simulating from multivariate normals

- Each β has a marginal normal too



Simulating from multivariate normals

- What is the probability that $b\_hsgpa > 0.4$ and $b\_skipped < -0.05$?

```
count if b_hsgpa > 0.4 & b_skip < -0.05
 6,410

di 6410/10000
0.641
```

- Fairly likely
- Remember, we need to take into account their **joint probability**

Comparing nested models

- Models are said to be **nested** if one can be obtained as a special case of the other
- a) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is nested within b) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
- If $\beta_3 = 0$, the we can obtain a) from b)
- Two non-nested models:
- a) $y = \beta_0 + \beta_1 x_1 + \beta_2 y$ is NOT nested within b) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
- We often call the smaller model the **reduced or restricted** model and the larger model the **full model**
- Lot's of theory behind the above statements; it's a paradigm for doing statistical tests. We will learn about this after we learn Maximum Likelihood Estimation (MLE)

Comparing nested models

- The intuition for comparing nested models is fairly simple: we will compare their SSEs
- Recall that SSE is the sum of squares of the **residuals**, which gives a measure of the variance **not explained** by our model
- Comparing SSE is similar to comparing R_a^2. We are essentially trying to figure out what improvement in R_a^2 is good enough (is the improvement due to chance?)
- Define SSE(RM) as the sum of square of the residuals of the reduced model and SSE(FM) as the sum of square of the residuals of the full model
- We will use the ratio $F = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$

Comparing nested models

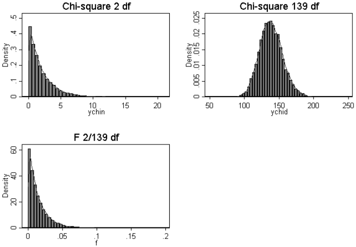
- $F = \frac{[SSE(RM)-SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$
- The above expression is just the proportion of **unexplained** variance between the reduced and full model relative to the full model
- We just divided by the degrees of freedom to take into account the parameters estimated. The parameters in the full model are $p + 1$ while the parameters of the reduced model are denoted by k
- What is the sign of $[SSE(RM) - SSE(FM)]$?
- The smaller F the more convinced we should be that the full model is not that great. We are estimating more parameters but not reducing the unexplained variation

Null hypotheses

- When comparing models, our **null hypothesis is that the reduced model is adequate**
- The alternative is that the full model is adequate
- By now you should remember that the ratio of SSEs distributes F with some degrees of freedom
- We reject the null if $F \geq F(p + 1 - k, n - p - 1; \alpha)$
- $F(p + 1 - k, n - p - 1; \alpha)$ is the critical value
- Note that p+1-k is just the number of additional parameters in the full model

Digression about χ^2 and F distributions

```
set obs 10000
gen ychin = rchi2(2)
gen ychid = rchi2(139)
gen f = ychin/ychid
```



- See a pattern here? Chi-square (χ^2) with df 139 converges to normal
- Why is χ^2 positive? Rejection for F is the tail on right, so large values of F will be likely to be rejected; also, $F = (t - \mathit{student})^2$

Test all parameters are equal to zero

Reduced model: $colgpa = \beta_0$

Full model: $colgpa = \gamma_0 + \gamma_1 hsgpa + \gamma_2 skipped$

- Recall that the null is that the reduced model is adequate
- Since the reduced model is just the mean of colgpa, then $SSE = SST$
- This test is essentially testing $\gamma_1 = \gamma_2 = 0$
- In words, all parameters p are simultaneously equal to zero

F test "by hand"

- Stata stores SSE in a temporary variable called e(rss)

```
qui reg colgpa
* ereturn list
scalar sse_r = e(rss)
qui reg colgpa hsgpa skipped
scalar sse_f = e(rss)
di ((sse_r - sse_f)/2)/(sse_f/(141-3))
 19.77258
di invFtail(2,138,0.05)
3.0617157
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
```

- It matches the regression output: 19.77
- Note that the critical value is usually around 3, larger for smaller samples (see Stata code for this class)

Digression: Be curious

- How is the rejection region affected by sample size in an F-test?

```
 forvalues i = 10(10)300 {
      di `i' "     "  invFtail(2,`i',0.05)
 }
10     4.102821
20     3.4928285
30     3.3158295
40     3.231727
50     3.1826099
60     3.1504113
70     3.1276756
80     3.1107662
90     3.097698
100     3.0872959
110     3.0788195
120     3.0717794
130     3.0658391
140     3.0607595
150     3.0563663
160     3.0525291
170     3.0491486
...
```

- Why? Remember what happened to the t-student critical value when the sample size increases

But we can also use the test command

- The test command is quite flexible

```
qui reg colgpa hsgpa skipped

test hsgpa skipped
 ( 1)  hsgpa = 0
 ( 2)  skipped = 0

       F(  2,   138) =   19.77
            Prob > F =    0.0000

* Remember this: shortcut for
 test _b[hsgpa] = _b[skipped] = 0

 ( 1)  hsgpa - skipped = 0
 ( 2)  hsgpa = 0

       F(  2,   138) =   19.77
            Prob > F =    0.0000
```

- You can do much more with test but don't forget the logic of the test
- Terminology and software can be confusing; the above F-test is a Wald test

More

- Your textbook has more examples that you can easily do with the test command
- They are extensions of the idea of comparing reduced and full models
- Remember too that the theory about the Wald test is not for testing one parameter but rather a linear combination of parameters. Some examples:

```
test hsgpa = skipped

 ( 1)  hsgpa - skipped = 0

       F(  1,   138) =   36.18
            Prob > F =    0.0000

 test hsgpa + skipped =1

 ( 1)  hsgpa + skipped = 1

       F(  1,   138) =   43.69
            Prob > F =    0.0000
```

Summary

- The idea of partitioning the variance and using $SSE = \sum(y_i - \hat{y}_i)^2$ as a measure of the variation in y not explained by the model leads to a general method for comparing models
- The models must be nested
- We want our models to be **parsimonious** ("unwilling to spend money or use resources; stingy or frugal; sparing, restrained")
- We haven't covered the inferential theory but it all starts with the assumption of normally distributed iid error terms
- Next, we will cover **maximun likelihood estimation** and will show that we can use the likelihood function in a similar way we used SSE or SSR
- For the Nth time: the advantage of focusing on MLE is that the method applies to many other models, not just linear regression