Multiple Regression 2

Adam Okulicz-Kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Monday 2nd March, 2015 17:51

outline

misc

intuition

testing hypotheses

adj Rsq, stata output

dummies and interactions

outline

misc

intuition

testing hypotheses

adj Rsq, stata output

dummies and interactions

when final?

- 5/5 or 5/12 ?
- let's vote

more data

- http://www.stateoftheusa.org/blog.php
- http://www.stateoftheusa.org/content/health-measures-for-the-develo.php
- http://www.stateoftheusa.org/content/fbi-report-violent-crime-down.php
- http://www.stateoftheusa.org/content/economy-seen-as-prompting-cohabitation.php
- http://stateoftheusa.org/content/measuring-economic-well-being.php
- http://www.stateoftheusa.org/content/report-hispanics-outlive-other-american.php

a short note on collinearity

- collinearity/multicollinearity simply means correlation among RHS vars.
- don't do anything about it
- the problem of collinearity is that CI are wider
- but this is the nature of the data...
- ... not a problem with your model
- conceptually it is the same problem as "micronumerosity" (wider CI)

a short note on academic research

- ◇ have a research idea: a problem/question/hypothesis
- ◇ read about it, mostly peer reviewed articles (literature review)
  - write literature review
- ◇ find data that has variables that can be used to test your hypotheses
  - write about your data and show des stats
- ◇ build your model based on literature AND your research idea
  - write about your model and defend it robustness/contribution/novelty
- ◇ interpret your results and discuss them

ps4

- ps4 is almost like a mini paper
- you will do des stats, build a model, defend it and interpret it
- again, if you have questions – email us

outline

misc

intuition

testing hypotheses

adj Rsq, stata output

dummies and interactions

you can do a lot with multiple regression

- you can test complex hypotheses
- the most interesting are interactions
  - you can test interesting hypotheses
  - and contribute to the literature
- remember, world is always more complicated than your model
  - interactions are a great way to get closer to the real complexity

outline

misc

intuition

testing hypotheses

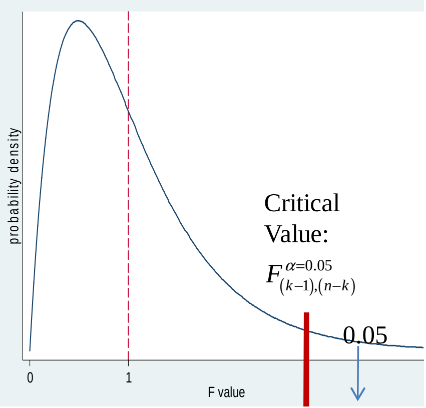
adj Rsq, stata output

dummies and interactions

F-test

- $F = \frac{\text{explained variation per regressor}}{\text{Residual variation per degree of freedom}} = \frac{ESS/(k-1)}{RSS/(n-k)}$
- $F = \frac{\sum(\hat{Y}_i - \bar{Y})^2/(k-1)}{\sum e_i^2/(n-k)}$
- $F = \frac{\frac{ESS}{TSS}/(k-1)}{\frac{RSS}{TSS}/(n-k)} = \frac{R^2/(k-1)}{1-R^2/(n-k)}$

F-test

- $H_o : \beta_2 = \beta_3 = ... = \beta_k = 0$
- H_A : At least one $\beta \neq 0$
- 

- assuming that the Null is true, the expected value of F is 1

F-test for restrictions

- UR: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$
- R: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + [0] X_{4i} + [0] X_{5i} + u_i$
- $H_0 : \beta_4 = \beta_5 = 0$
- H_A : at least one $\beta \neq 0$
- $F = \frac{ESS_U - ESS_R / m}{RSS_U / (n-k)}$ $\quad \frac{\text{m=\# of restrictions}}{\text{k=\#of betas (incl interecept) in UR}}$
- critical F: $(m, n-k)$
- blackboard: draw a real example like in exam
- dofile:F

chow test (F-test)

- chow test is just an F-test that tests stability of betas across groups
  - e.g.: men vs women; black vs white; before 2000 vs after 2000
- first, run a model and get RSS – it will be your RSS_R
- second, run the same model for each group separately and get:
  - $RSS_U = RSS_{male} + RSS_{female}$
- $F = \frac{(RSS_R - RSS_U)/k}{RSS_U/(n-2k)}$
- dofile:chow

testing equality of betas

- $H_0 : \beta_2 = \beta_3$ or $\beta_2 - \beta_3 = 0$
- $H_A : \beta_2 \neq \beta_3$ or $\beta_2 - \beta_3 \neq 0$
- $t = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - (\beta_2 - \beta_3)}{s_{(\hat{\beta}_2 - \hat{\beta}_3)}}$
- $var(A - B) = var(A) + var(B) - 2cov(A, B)$
- $s_{(\hat{\beta}_2 - \hat{\beta}_3)} = \sqrt{var(\hat{\beta}_2) + var(\hat{\beta}_3) - 2cov(\hat{\beta}_3, \hat{\beta}_3)}$

var-cov matrix of betas

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$\hat{\beta}_1$	$\mathrm{var}\left(\hat{\beta}_1\right) = s^2_{\hat{\beta}_1}$		
$\hat{\beta}_2$	$\mathrm{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)$	$\mathrm{var}\left(\hat{\beta}_2\right) = s^2_{\hat{\beta}_2}$	
$\hat{\beta}_3$	$\mathrm{cov}\left(\hat{\beta}_1, \hat{\beta}_3\right)$	$\mathrm{cov}\left(\hat{\beta}_2, \hat{\beta}_3\right)$	$\mathrm{var}\left(\hat{\beta}_3\right) = s^2_{\hat{\beta}_3}$

- assuming that the Null is true, the expected value of F is 1
- dofile: vce

outline

misc

intuition

testing hypotheses

adj Rsq, stata output

dummies and interactions

adj Rsq

- $R^2 = 1 - \frac{RSS}{TSS}$ $\quad adj.R^2 = \bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - \frac{s^2}{s_Y^2}$
- regular R^2 always increases when new variables added, even if they are just noise
- Adj. R^2 "corrects" for degrees of freedom
- can decline, or even become negative
- widely used, but not vary useful
- neither accurate as a description nor a valid test statistic for some hypothesis
- don't use it
- if you see it ignore it and complain
- if you are concerned about the significance of a variable or variables, look to t and F tests

stata output

. **regress** $Y\ X_2\ X_3\ \dots\ X_k$, [beta]

Source	SS	df	MS
Model	$ESS=\sum\left(\hat{Y}_i-\bar{Y}\right)^2$	$k-1$	$\frac{ESS}{k-1}$
Residual	$RSS=\sum e_i^2$	$n-k$	$s^2=\frac{RSS}{n-k}$
Total	$TSS=\sum\left(Y_i-\bar{Y}\right)^2$	$n-1$	$s_Y^2=\frac{TSS}{n-1}$

Number of obs	=	n
$\mathrm{F}(1,n-2)$	=	$F=\frac{ESS/(k-1)}{RSS/(n-k)}$
Prob > F	=	p value for the model
R-squared	=	$R^2=1-\frac{RSS}{TSS}$
Adj R-Squared	=	$\bar{R}^2=1-\frac{RSS/(n-k)}{TSS/(n-1)}$
Root MSE	=	s

Y	Coef.	Std.Err.	t	P>|t|	[95% Conf.	Interval]	[Beta]
X_2	$\hat{\beta}_2$	$s_{\hat{\beta}_2}$	$\hat{\beta}_2/s_{\hat{\beta}_2}$	$\mathrm{H}_0:\beta_2=0$	$\hat{\beta}_2-t_{0.025}s_{\hat{\beta}_2}$	$\hat{\beta}_2+t_{0.025}s_{\hat{\beta}_2}$	$\hat{\beta}_2\left(s_{X_2}/s_Y\right)$
X_3	$\hat{\beta}_3$	$s_{\hat{\beta}_3}$	$\hat{\beta}_3/s_{\hat{\beta}_3}$	$\mathrm{H}_0:\beta_3=0$	$\hat{\beta}_3-t_{0.025}s_{\hat{\beta}_3}$	$\hat{\beta}_3+t_{0.025}s_{\hat{\beta}_3}$	$\hat{\beta}_2\left(s_{X_3}/s_Y\right)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	$\hat{\beta}_k$	$s_{\hat{\beta}_k}$	$\hat{\beta}_k/s_{\hat{\beta}_k}$	$\mathrm{H}_0:\beta_k=0$	$\hat{\beta}_k-t_{0.025}s_{\hat{\beta}_k}$	$\hat{\beta}_k+t_{0.025}s_{\hat{\beta}_k}$	$\hat{\beta}_2\left(s_{X_k}/s_Y\right)$
$\_cons$	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$	$\hat{\beta}_1/s_{\hat{\beta}_1}$	$\mathrm{H}_0:\beta_1=0$	$\hat{\beta}_1-t_{0.025}s_{\hat{\beta}_1}$	$\hat{\beta}_1+t_{0.025}s_{\hat{\beta}_1}$	.

outline

intuition

- dummies and interactions are fun !
- this is one of the most interesting things in regression
- you can test some interesting hypotheses
- and you can contribute to the literature

what is it?

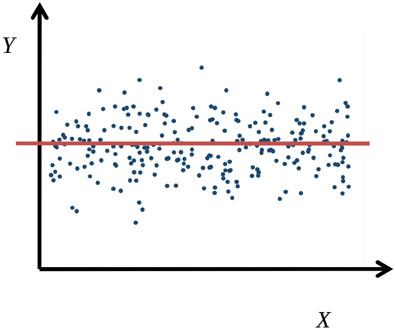
- Indicator variables identify nominal or ordinal characteristics, such as gender, race, region, religion, or education (measured as highest degree attained).
- "Dummy" variables are indicator variables that are binary indicators of a specific attribute - you either have the attribute or you do not.

what is it?

- Dummy variables are almost always coded 1 if the condition is true and 0 otherwise, which greatly simplifies interpretation.
- Dummies can be used to create separate intercepts and/or slopes for subgroups of the sample within one regression.
- Coefficients on dummy variables must always be interpreted relative to a "base case," i.e. a reference group.

regression on a constant only

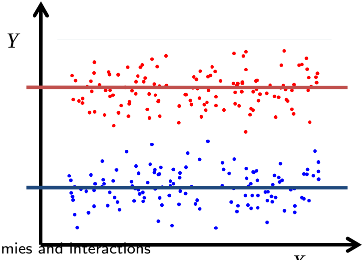
- $\hat{\beta}_2 = 0$
- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = \bar{Y}$



-

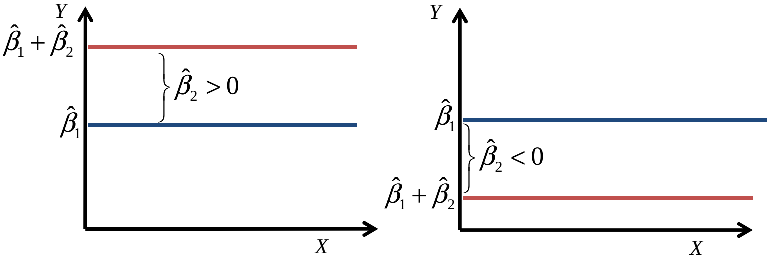
now add a dummy

- $Y_i = \beta_1 + \beta_2 female_i + u_i$
- if $female_i = 1 \quad \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2(1) = \hat{\beta}_1 + \hat{\beta}_2$
  - $E[Y|female = 1] = \hat{\beta}_1 + \hat{\beta}_2$
- if $female_i = 0 \quad \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2(0] = \hat{\beta}_1$
  - $E[Y|female = 0] = \beta_1$
- hence, β_2 is the difference between $\bar{Y}$ for males and females



schematic

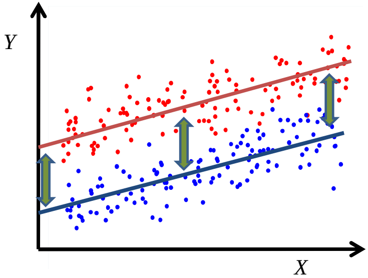
- $Y_i = \beta_1 + \beta_2 female_i + u_i$
- $\hat{\beta}_2 = \bar{Y}_{female} - \bar{Y}_{male}$



-

and add a continuous var

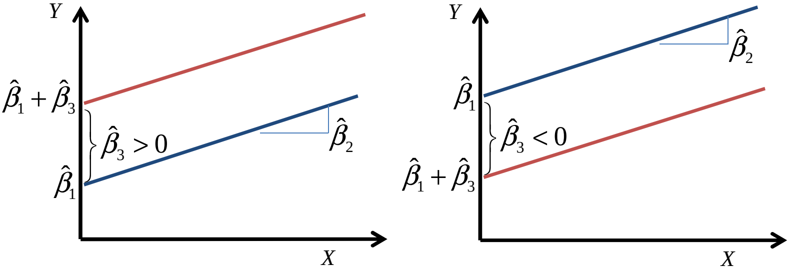
- $Y_i = \beta_1 + \beta_2 X_i + \beta_3 female_i + u_i$
- if $female_i = 1 \quad \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3(1) = (\hat{\beta}_1 + \hat{\beta}_3) + \hat{\beta}_2 X_i$
- if $female_i = 0 \quad \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3(0) = (\hat{\beta}_1) + \hat{\beta}_2 X_i$



-

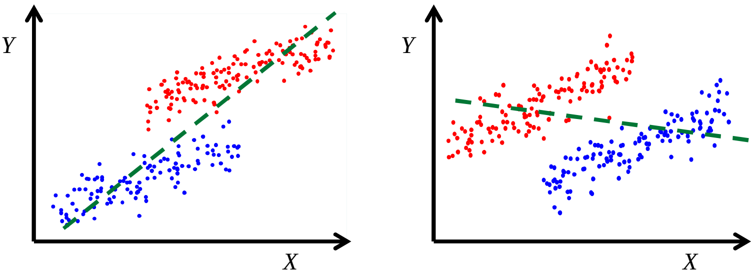
schematic

◇ $Y_i = \beta_1 + \beta_2 X_i + \beta_3 female_i + u_i$



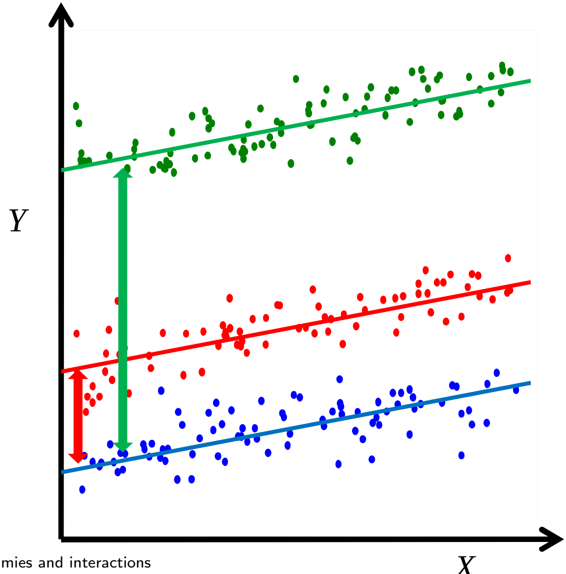
◇

bias from omitting a dummy...



ordinal variables

◇ omit one category (base case)

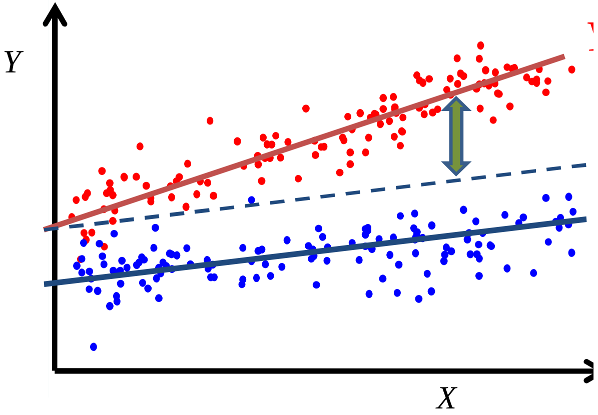


choosing the base case

- don't let the software pick for you!
- usually the largest category is best, but it depends what comparisons you want to highlight (coefficients and t tests are relative to base case)
- think about what hypotheses you are most interested in
- remember that a different base case can change which coefficients are significant
- make your choice(s) clear in your tables and text

continuous/dummy interactions

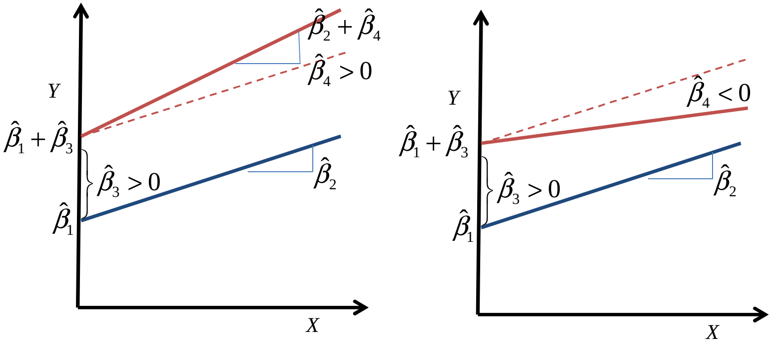
- $Y_i = \beta_1 + \beta_2 X_i + \beta_3 female_i + \beta_4 female_i * X_i + u_i$



-

schematic

- $Y_i = \beta_1 + \beta_2 X_i + \beta_3 female_i + \beta_4 female_i * X_i + u_i$



-

interaction of dummies

- If there is an interaction effect between two variables, the effect of one variable depends on the level of the other.
- For example, the effect of marriage on wage depends on gender.
- Interactions go both ways. The effect of gender depends on marital status.

interaction of dummies

- $Y_i = \beta_1 + \beta_2 female + \beta_3 married + \beta_4 female * married + u_i$

	Male	Female	Gender Difference
Unmarried	$\hat{\beta}_1$	$\hat{\beta}_1 + \hat{\beta}_2$	$\hat{\beta}_2$
Married	$\hat{\beta}_1 + \hat{\beta}_3$	$\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4$	$\hat{\beta}_2 + \hat{\beta}_4$
Effect of Marriage	$\hat{\beta}_3$	$\hat{\beta}_3 + \hat{\beta}_4$	$\hat{\beta}_4$

-

example

```
. table married female, c(mean wage) row col f(%7.2f)
----------------------------------
          |        Gender
  Married |   male  female   Total
----------+-----------------------
       no |   8.35    8.26    8.31
      yes |  10.88    7.68    9.40
          |
    Total |   9.99    7.88    9.02
----------------------------------
```

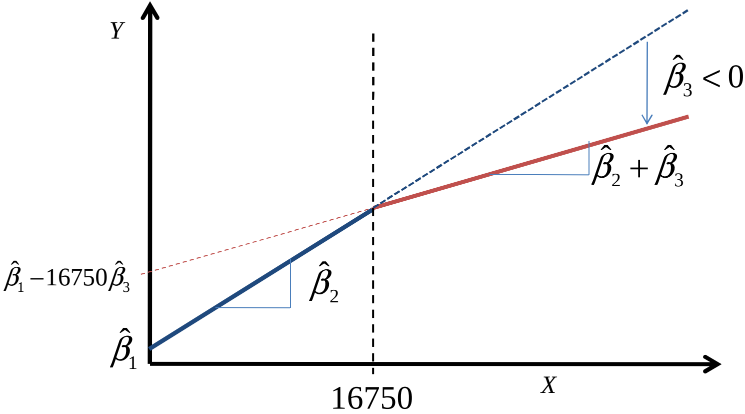
```
. gen femxmar = female*married
. reg wage female married femxmar
```

	wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
$\hat{\beta}_2$	female	-.0951892	.7350367	-0.13	0.897	-1.539132	1.348754
$\hat{\beta}_3$	married	2.521222	.6120814	4.12	0.000	1.318819	3.723626
$\hat{\beta}_4$	femxmar	-3.09704	.9072785	-3.41	0.001	-4.879344	-1.314737
$\hat{\beta}_1$	_cons	8.354677	.4936728	16.92	0.000	7.384882	9.324473

piece-wise linear

- Suppose you want to estimate the effect of income on rent paid. The coefficient tells how much of each additional dollar they allocate to rent.
  - $rent_i = f(income) = \beta_1 + \beta_2 income + u_i$
- However, you suspect that those with a higher tax rate may allocate less per dollar of income to rent (since they have less of that income to spend).
  - create a dummy $D = 1$ if $income > 16,750$; 0 otherwise
  - $rent_i = f(income, taxrate) =$
    $\beta_1 + \beta_2 income + \beta_3[D * (income_i - 16,750)] + u_i$

piece-wise linear



◇

◇ the model forces the lines to connect (no gap), because the new rate only applies to dollars of income above the cut point

dummy practice

- ◇ instead of the dofile, see the links on the website for the code
- ◇ let's especially focus on the dummy variables
- ◇ we'll do it in the class if we have time...

interactions of continuous variables

- $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 (X_{2i} X_{3i}) + u_i$
- $\frac{\Delta Y_i}{\Delta X_{2i}} = \beta_2 + \beta_4 X_{3i}$
- $\frac{\Delta Y_i}{\Delta X_{3i}} = \beta_3 + \beta_4 X_{2i}$

the marginal effect of x_2 depends on the level of x_3

X_{2i}	X_{3i}	$\hat{Y}_i$	$\Delta\hat{Y}_i$
5	0	10	+2
6	0	12	
5	10	65	+7
6	10	72	
50	10	380	+7
51	10	387	

$$\hat{Y}_i = 0 + 2X_{2i} + 3X_{3i} + 0.5\left(X_{2i}X_{3i}\right)$$

$$\begin{aligned}\hat{Y}_i &= 2X_{2i} + 3(0) + 0.5X_{2i}(0)\\ &= 2X_{2i}\end{aligned}$$

$$\begin{aligned}\hat{Y}_i &= 2X_{2i} + 3(10) + 0.5X_{2i}(10)\\ &= 2X_{2i} + 5X_{2i}\\ &= 7X_{2i}\end{aligned}$$

note: add 30 in the last equation to get $\hat{Y}_i$

practice

- let's practice in stata using ucla regression webbook
- let's think of the regression of wage on educ and female
- let's write down equation and test with stata using wages data from the last class...