Lecture 11

Annoucements

- PS3 due on 03/08/2020

Outline

- Interaction effects
- Non-linear Models

Can we test whether the effect of being female is different depending on marital status?

- $y = \beta_0 + \beta_1 Female + \beta_2 Married + \beta_3 Female * Married + u$
- Suppose we estimated this model:
- $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} Female + \widehat{\beta_2} Married + \widehat{\beta_3} Female * Married$

- H_0: $\beta_3 = 0$
- H_0: $\beta_3 \neq 0$

- t=$\frac{\widehat{\beta_3} - 0}{se(\widehat{\beta_3})}$, we can use t-statistic to test the hypothesis that the effect of being a female is different for married people and non-married people

```
. reg earnings Female Married inter_Female_Married
```

Source	SS	df	MS	Number of obs	=	17,870
				F(3, 17866)	=	969.40
Model	1.8132e+12	3	6.0441e+11	Prob > F	=	0.0000
Residual	1.1139e+13	17,866	623494086	R-squared	=	0.1400
				Adj R-squared	=	0.1398
Total	1.2953e+13	17,869	724863544	Root MSE	=	24970

earnings	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Female	-5781.058	641.0552	-9.02	0.000	-7037.588	-4524.528
Married	17651.91	597.0708	29.56	0.000	16481.59	18822.22
inter_Female_Married	5476.871	791.8643	6.92	0.000	3924.741	7029.002
_cons	36653.81	488.2928	75.07	0.000	35696.71	37610.91

The effect of being a female is “statistically significantly” different for married people and non-married people?

Interactions between independent variables: interactions of dummy variable and continuous variable

- Example: education -> earnings
- Perhaps the effects of education on earnings can be different for different gender
- We assumed “constant” effect of education so far
- We want to allow for different effect of being education depending on gender
- How can we do this?

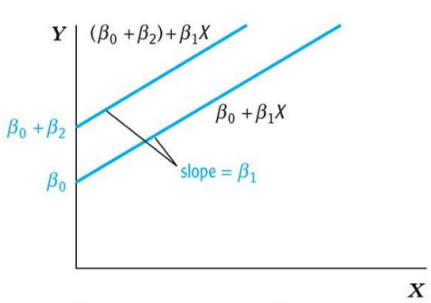
Interactions between independent variables:
interactions of dummy variable and continuous variable

- $y = \beta_0 + \beta_1 Educ + \beta_2 Female + \beta_3 Educ * Female + u$
- E(y|Female = 0) = $\beta_0 + \beta_1 Educ$
- E(y|Female = 1) = $\beta_0 + \beta_2 + (\beta_1 + \beta_3) Educ$

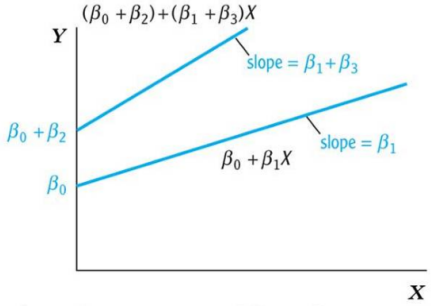
- The effect of education can be different for males and females

- The intercept difference is β_2
- Now we also have slope difference, which is β_3

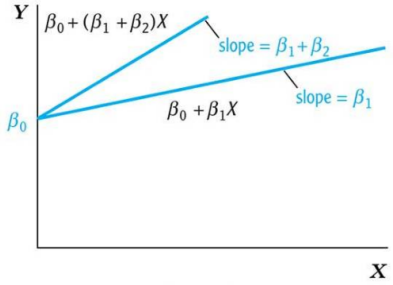
dummy-continuous interactions, ctd.



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

Can we test whether the slopes are different for different gender?

- $y = \beta_0 + \beta_1 Educ + \beta_2 Female + \beta_3 Educ * Female + u$
- Suppose we estimated this model:
- $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} Educ + \widehat{\beta_2} Female + \widehat{\beta_3} Educ * Female$

- H_0: β_3 = 0
- H_0: $\beta_3 \neq 0$

- t=$\frac{\widehat{\beta_3}-0}{se(\widehat{\beta_3})}$, we can use t-statistic to test the hypothesis that the effect of being a female is different for married people and non-married people

Can we test whether the intercepts are different for different gender?

- $y = \beta_0 + \beta_1 Educ + \beta_2 Female + \beta_3 Educ * Female + u$
- Suppose we estimated this model:
- $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} Educ + \widehat{\beta_2} Female + \widehat{\beta_3} Educ * Female$

- H_0: $\beta_2 = 0$
- H_0: $\beta_2 \neq 0$

- t=$\frac{\widehat{\beta_2}-0}{se(\widehat{\beta_2})}$, we can use t-statistic to test the hypothesis that the effect of being a female is different for married people and non-married people

```
. reg earnings Female educ inter_Female_educ
```

Source	SS	df	MS	Number of obs	=	17,870
				F(3, 17866)	=	1075.84
Model	1.9819e+12	3	6.6062e+11	Prob > F	=	0.0000
Residual	1.0971e+13	17,866	614055642	R-squared	=	0.1530
				Adj R-squared	=	0.1529
Total	1.2953e+13	17,869	724863544	Root MSE	=	24780

earnings	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Female	-4028.901	1936.579	-2.08	0.038	-7824.783	-233.0192
educ	3900.419	99.75824	39.10	0.000	3704.883	4095.955
inter_Female_educ	98.64415	140.3479	0.70	0.482	-176.4513	373.7396
_cons	-4421.657	1380.96	-3.20	0.001	-7128.472	-1714.841

Can you reject the null that the intercepts for females and males are the same?
Can you reject the null that the slopes for females and males are the same?

Interactions between independent variables: interactions of two continuous variables

- Example: student-teacher ratio (STR) -> test score
- Perhaps the effect of STR on test score might be different depending on the share of English learners in the class
- We assumed “constant” effect of being STR so far
- We want to allow for different effect of STR depending on the share of English learners
- How can we do this?

Interactions between independent variables: interactions of two continuous variables

- $y = \beta_0 + \beta_1 STR + \beta_2 PctEL + \beta_3 STR * PctEL + u$

- The effect of STR can be different for different PctEL (%)

- How can we interpret this?
- Suppose we have:
- $\hat{y}$ = 686.3 – 1.12*STR* – 0.67*PctEL* + .0012(*STR*PctEL*),
  (11.8) (0.59) (0.37) (0.019)
- The effect of STR on test score is negative. However, as PctEL increases the effect size of STR decreases.
- What would be the effect of STR when PctEL = 20?
- What would be the effect of STR when PctEL = 100?

Interactions between independent variables: interactions of two continuous variables

- $y = \beta_0 + \beta_1 STR + \beta_2 PctEL + \beta_3 STR * PctEL + u$

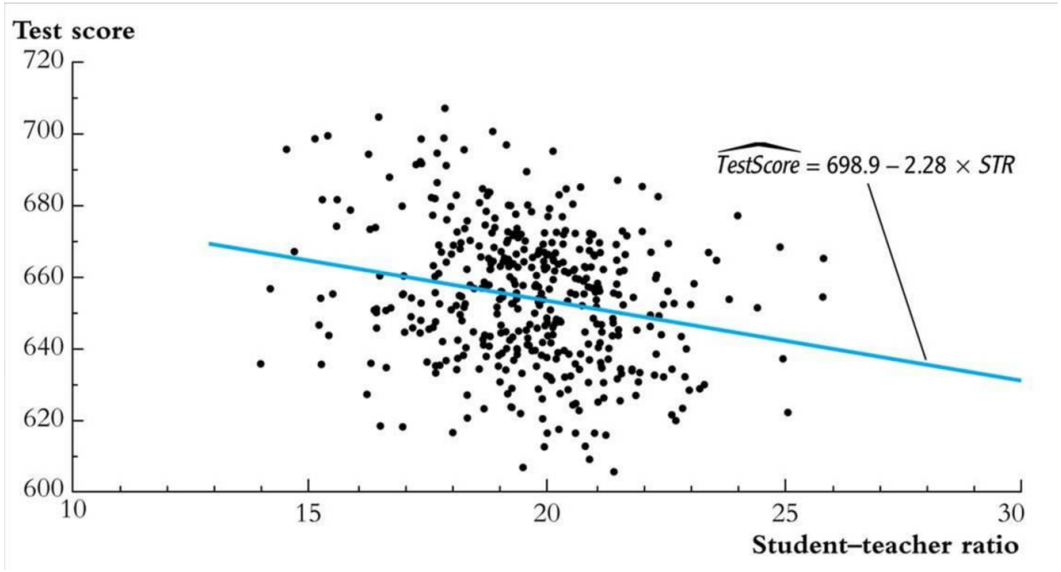
- The effect of STR can be different for different PctEL (%)

- How can we interpret this?
- Suppose we have:
- $\hat{y}$ = 686.3 – 1.12*STR* – 0.67*PctEL* + .0012(*STR*PctEL*),
  (11.8) (0.59) (0.37) (0.019)
- The effect of STR on test score is negative. However, as PctEL increases the effect size of STR decreases.
- What would be the effect of STR when PctEL = 20? -1.12STR+0.0012(STR*20)=-1.096STR
- What would be the effect of STR when PctEL = 100? -1.12STR+0.0012(STR*100)=-1STR

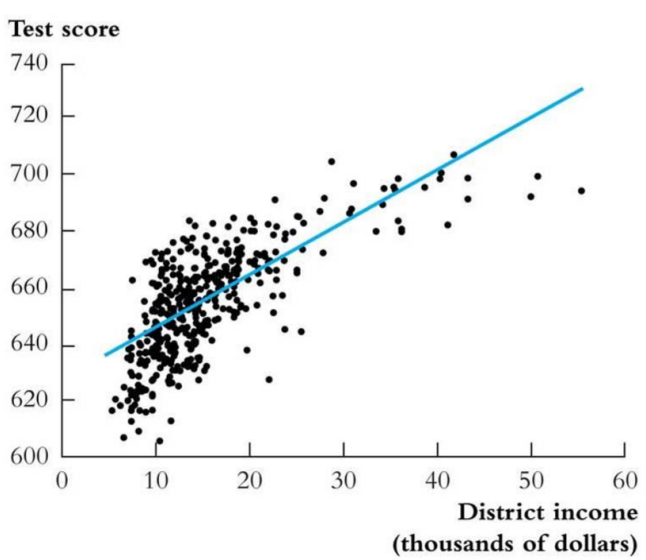
Nonlinear regression functions

- Everything so far has been linear in the *X*'s
- But the linear approximation is not always a good one
- The multiple regression framework can be extended to handle regression functions that are nonlinear in one or more *X*.

The *TestScore* – *STR* relation looks linear (maybe)…



But the *TestScore* – *Income* relation looks nonlinear...



Example: the *TestScore* – *Income* relation

- $Income_i$ = average district income in the i^{th} district (thousands of dollars per capita)

- Quadratic specification:
- $TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + u_i$

- Cubic specification:
- $TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + \beta_3 (Income_i)^3 + u_i$

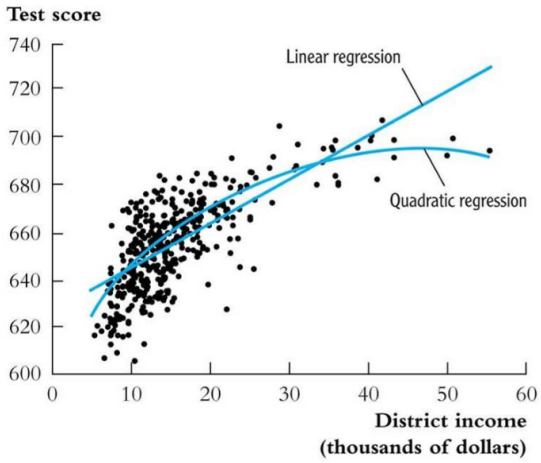
Estimation of the quadratic specification in STATA

```
generate avginc2 = avginc*avginc;       Create a new regressor
reg testscr avginc avginc2, r;

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  428.52
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5562
                                                      Root MSE      =  12.724

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   3.850995   .2680941    14.36   0.000      3.32401    4.377979
     avginc2 |  -.0423085   .0047803    -8.85   0.000     -.051705   -.0329119
       _cons |   607.3017   2.901754   209.29   0.000     601.5978    613.0056
------------------------------------------------------------------------------
```

Test the null hypothesis of linearity against the alternative that the regression function is a quadratic….

Interpreting the estimated regression function:

- $\widehat{TestScore}_i$ = 607.3+ 3.85$Income_i$ + -0.0423$(Income_i)^2$



Interpreting the estimated regression function, ctd:

- (b) Compute "effects" for different values of *X*

- Predicted change in *TestScore* for a change in income from $5,000 per capita to $6,000 per capita:
- $\Delta Testscore$ =607.3 + 3.85*6 – 0.0423*6^2

  – (607.3 + 3.85*5 – 0.0423*5^2) = 3.4

- Predicted "effects" for different values of *X*:

Change in Income ($1000 per capita)	Δ testscore
from 5 to 6	3.4
from 25 to 26	1.7
from 45 to 46	0.0

- The "effect" of a change in income is greater at low than high income levels (perhaps, a declining marginal benefit of an increase in school budgets?)

Estimation of a cubic specification in STATA

```
gen avginc3 = avginc*avginc2;              Create the cubic regressor
reg testscr avginc avginc2 avginc3, r;

Regression with robust standard errors                Number of obs =     420
                                                      F(  3,   416) =  270.18
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5584
                                                      Root MSE      =  12.707

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   5.018677   .7073505     7.10   0.000     3.628251    6.409104
     avginc2 |  -.0958052   .0289537    -3.31   0.001    -.1527191   -.0388913
     avginc3 |   .0006855   .0003471     1.98   0.049     3.27e-06    .0013677
       _cons |    600.079   5.102062   117.61   0.000     590.0499     610.108
------------------------------------------------------------------------------
```

- Testing the null hypothesis of linearity, against the alternative that the population regression is quadratic and/or cubic, that is, it is a polynomial of degree up to 3:

- H_0: pop'n coefficients on *Income*^{2} and *Income*^{3} = 0
- H_1: at least one of these coefficients is nonzero.

```
test avginc2 avginc3;  Execute the test command after running the regression

 ( 1)  avginc2 = 0.0
 ( 2)  avginc3 = 0.0

       F(  2,   416) =   37.69
       Prob > F =    0.0000
```

- The hypothesis that the population regression is linear is rejected at the 1% significance level against the alternative that it is a polynomial of degree up to 3.

TABLE 8.3 Nonlinear Regression Models of Test Scores

Dependent variable: average test score in district; 420 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student–teacher ratio (STR)	−1.00** (0.27)	−0.73** (0.26)	−0.97 (0.59)	−0.53 (0.34)	64.33** (24.86)	83.70** (28.50)	65.29** (25.26)
STR^2					−3.42** (1.25)	−4.38** (1.44)	−3.47** (1.27)
STR^3					0.059** (0.021)	0.075** (0.024)	0.060** (0.021)
% English learners	−0.122** (0.033)	−0.176** (0.034)					−0.166** (0.034)
% English learners ≥ 10%? (Binary, $HiEL$)			5.64 (19.51)	5.50 (9.80)	−5.47** (1.03)	816.1* (327.7)	
$HiEL \times STR$			−1.28 (0.97)	−0.58 (0.50)		−123.3* (50.2)	
$HiEL \times STR^2$						6.12* (2.54)	
$HiEL \times STR^3$						−0.101* (0.043)	
% Eligible for subsidized lunch	−0.547** (0.024)	−0.398** (0.033)		−0.411** (0.029)	−0.420** (0.029)	−0.418** (0.029)	−0.402** (0.033)
Average district income (logarithm)		11.57** (1.81)		12.12** (1.80)	11.75** (1.78)	11.80** (1.78)	11.51** (1.81)
Intercept	700.2** (5.6)	658.6** (8.6)	682.2** (11.9)	653.6** (9.9)	252.0 (163.6)	122.3 (185.5)	244.8 (165.7)

F-Statistics and *p*-Values on Joint Hypotheses							
(a) All *STR* variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (< 0.001)	4.96 (< 0.001)	5.91 (0.001)
(b) $STR^2, STR^3 = 0$					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)
(c) $HiEL \times STR$, $HiEL \times STR^2$, $HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\overline{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

- Do we have enough evidence that there's a non-linear effect of STR on average test score?

F-Statistics and *p*-Values on Joint Hypotheses							
(a) All *STR* variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (< 0.001)	4.96 (< 0.001)	5.91 (0.001)
(b) $STR^2, STR^3 = 0$					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)
(c) $HiEL \times STR$, $HiEL \times STR^2$, $HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\overline{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

- Do we have enough evidence that there's a non-linear effect of STR on average test score? Yes because in (5), (6), (7), we all reject the joint hypothesis that the coefficients of squared term and cubic terms are zero.

Review for quizzes

- Know how to interpret interaction terms.
- Know how to check for non-linear effects