Dummy Variables and Interaction Terms

Lecture 22

Reading: Sections 19.3, 21.1 – 21.3, "Waterloo 2016 Salary Disparities" (Optional: "Standardized Residuals," "Influence Measures" pp. 737-9)

1

Dummy Variables in Regression

- Dummy variable: Captures qualitative information with 2 possible values: 0 or 1
  - Also called: indicator variables, fixed effects
  - Allows inclusion of categorical/nominal variables
  - Example: Does sex affect wages even if we control for years of education?
    - $wage$ (dollars per hour)
    - $educ$ (years of education)
    - fem (= 1 if female; = 0 if male)

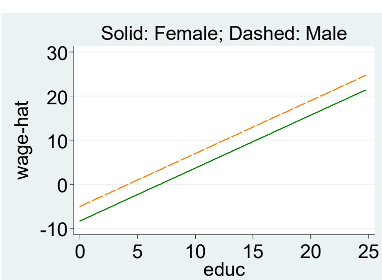
Why not name the dummy variable sex?

2

Wage Regression

Model: $wage_i = \alpha + \beta educ_i + \delta fem_i + \varepsilon_i$

Results: $\widehat{wage}_i = \underset{(3.6)}{-5.0} + \underset{(0.5)}{1.2} educ_i - \underset{(1.1)}{3.3} fem_i$



Answers causal research question?

Is difference in wages statistically significant after accounting for education?

$H_0: \delta = 0$
$H_1: \delta \neq 0$

$$t = \frac{-3.3}{1.1} = -3$$

After controlling for years of education, hourly wages for females are \$3.30 lower on average than for males.

3

Omitted Category (Reference Group)

- Omitted category (aka reference group): The category that is *not* included as a dummy
  - The regular constant term (intercept) picks up the constant value for the omitted category
    - What is omitted category in the wage regression: $\widehat{wage} = -5.0 + 1.2educ - 3.3fem$?
    - What if we switched the omitted category?
  - Coefficient estimates on dummy variables are *relative to* the omitted category ("baseline")

4

What If More Than 2 Categories?

- To include a categorical variable, the number of dummy vars is *one less than* number of unique categories (one will be reference cat.)
  - E.g. To fully control for occupation with 40 occupational categories requires 39 dummies
  - E.g. Zheng and Kahn (2017) from DACM A.2
    - PM10 – conc. of particulate matter – from 2003 to 2012 (10 years) and across cities (85 Chinese cites)
    - How to control for changes over time across all cities?

Which kind of data: cross sectional, time series, or panel?

5

Table 1: Correlates of Urban Air Pollution in China

	Dependent Variable: *log(PM10)*	
Explanatory Variables:	(1)	(2)
Log(GDP per capita)	-0.434 (0.129)	-0.424 (0.128)
(Log(GDP per capita))^2	0.300 (0.075)	0.296 (0.074)
(Log(GDP per capita))^3	-0.0596 (0.0135)	-0.0592 (0.0134)
Log(Population)	0.164 (0.014)	0.164 (0.014)
Log(Manufacturing Share)	0.0498 (0.0397)	0.0450 (0.0396)
Log(Average Years of Schooling)	-0.918 (0.143)	-0.926 (0.142)
Log(Rainfall)	-0.0987 (0.0347)	-0.0977 (0.0345)
Log(Temperature Index)	0.391 (0.074)	0.394 (0.073)
Time Trend	-0.0316 (0.0031)	-
Year Dummies	No	Yes
Constant	4.304 (0.428)	4.353 (0.425)
R^2	0.432	0.444
Observations	846	846

Note: The latitude and longitude of each city are controlled for in each column. Standard errors in parentheses. Four cities are missing PM10 data in 2003.

6

Regression (1): Time Trend

Source	SS	df	MS
Model	37.1271039	11	3.37519127
Residual	48.9026999	834	.058636331
Total	86.0298038	845	.101810419

Number of obs =	846
F(11, 834) =	57.56
Prob > F =	0.0000
R-squared =	0.4316
Adj R-squared =	0.4241
Root MSE =	.24215

ln_pm10	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_gdp_pc	-.4340424	.1286315	-3.37	0.001	-.6865218	-.1815629
ln_gdp_pc_2	.2998217	.0745439	4.02	0.000	.153506	.4461375
ln_gdp_pc_3	-.0595622	.0134763	-4.42	0.000	-.0860137	-.0331107
ln_pop	.1638094	.0137121	11.95	0.000	.1368952	.1907236
ln_manu	.0498194	.0397189	1.25	0.210	-.0281413	.1277801
ln_edu	-.9182325	.1427245	-6.43	0.000	-1.198374	-.638091
ln_rain	-.0987354	.0347372	-2.84	0.005	-.1669181	-.0305527
ln_temp	.3907443	.0738079	5.29	0.000	.2458731	.5356154
longitude	-.0063736	.001507	-4.23	0.000	-.0093315	-.0034157
latitude	.005419	.0041039	1.32	0.187	-.0026361	.0134741
trend	-.0316037	.003127	-10.11	0.000	-.0377415	-.025466
_cons	4.303665	.4279114	10.06	0.000	3.463755	5.143575

A time trend measures passage of time: the variable trend above equals 1 for 2003, 2 for 2004, …, and 10 for 2012.

Regression (2): Year Dummies

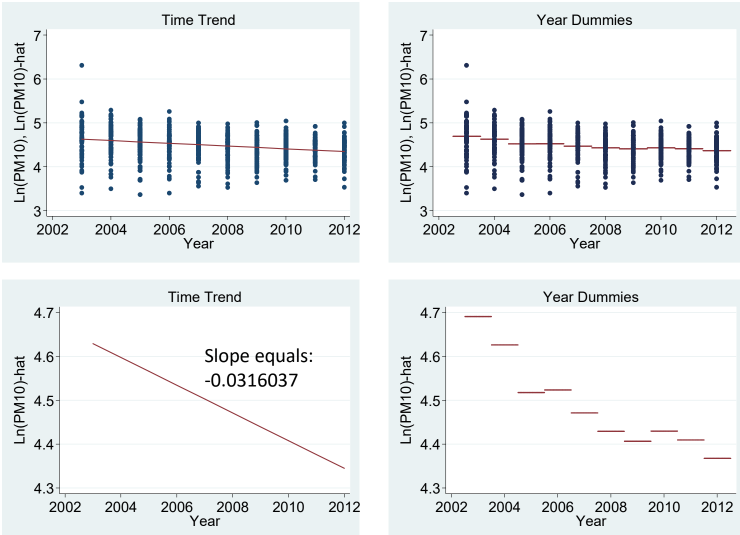
Source	SS	df	MS
Model	38.2139593	19	2.01126101
Residual	47.8158446	826	.057888432
Total	86.0298038	845	.101810419

Number of obs =	846
F(19, 826) =	34.74
Prob > F =	0.0000
R-squared =	0.4442
Adj R-squared =	0.4314
Root MSE =	.2406

ln_pm10	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_gdp_pc	-.4241961	.1278504	-3.32	0.001	-.675146	-.1732461
ln_gdp_pc_2	.2961769	.0740776	4.00	0.000	.1507745	.4415793
ln_gdp_pc_3	-.0591624	.0133912	-4.42	0.000	-.0854471	-.0328776
ln_pop	.1636883	.0136248	12.01	0.000	.1369451	.1904316
ln_manu	.0449651	.0396028	1.14	0.257	-.0327688	.122699
ln_edu	-.9262087	.1419217	-6.53	0.000	-1.204778	-.6476391
ln_rain	-.0976617	.0345163	-2.83	0.005	-.1654117	-.0299116
ln_temp	.393586	.0733424	5.37	0.000	.2496265	.5375455
longitude	-.0064208	.0014975	-4.29	0.000	-.0093601	-.0034814
latitude	.0054305	.0040779	1.33	0.183	-.0025738	.0134347
yr_2004	-.0648882	.0373851	-1.74	0.083	-.1382692	.0084929
yr_2005	-.1731407	.0374578	-4.62	0.000	-.2466644	-.0996171
yr_2006	-.1673246	.0375447	-4.46	0.000	-.2410188	-.0936304
yr_2007	-.2196464	.0376449	-5.83	0.000	-.2935372	-.1457555
yr_2008	-.2616172	.0377134	-6.94	0.000	-.3356426	-.1875919
yr_2009	-.2840717	.0381066	-7.45	0.000	-.3588689	-.2092744
yr_2010	-.2611697	.0382683	-6.82	0.000	-.3362843	-.1860551
yr_2011	-.2812865	.0382972	-7.34	0.000	-.3564577	-.2061153
yr_2012	-.3232032	.0386962	-8.35	0.000	-.3991577	-.2472486
_cons	4.35313	.425458	10.23	0.000	3.518023	5.188236

Interpreting Coefficients on Time

- In Reg. (1), coefficient on trend is -.0316037***
  - After controlling for GDP per capita, population, manufacturing share, average education, rainfall, temperature, latitude, and longitude, PM10 concentrations on average declined by 3.2 percent annually in Chinese cities between 2003 and 2012.
- In Reg. (2), coefficient on yr_2006 is -.1673246***
  - After controlling for GDP per capita, population, manufacturing share, average education, rainfall, temperature, latitude, and longitude, Chinese cities in 2006 had PM10 concentrations that were 16.7 percent lower on average compared to 2003.



To plot Ln(PM10)-hat against time, plugged in mean values for all other variables.

Table 2
A Simple International Education Production Function: A Least-Squares Regression
(dependent variable is student's mathematics test score)

Excerpt: 2016 article in *Journal of Economic Perspectives* "The Importance of School Systems: Evidence from International Differences in Student Achievement." DOI: 10.1257/jep.30.3.3

(PISA score: mean ~500)	*Coefficient*	*Standard error*
Family Background		
Age (years)	17.825***	(3.160)
Female	−14.733***	(1.639)
Preprimary education (more than 1 year)	6.832***	(2.428)
School starting age	−3.869*	(2.030)
Grade repetition in primary school	−54.579***	(4.734)
Grade repetition in secondary school	−33.726***	(6.702)
Grade		
7th grade	−47.003***	(10.051)
8th grade	−19.213*	(10.242)
9th grade	−6.772	(6.896)
11th grade	−3.275	(5.236)
12th grade	11.949*	(6.398)
… there are many more explanatory variables		
Constant	116.126**	(51.774)
Students	219,794	
Schools	8,245	
Countries	29	
R^2 (at student level)	0.340	

For *Grade*, what is the omitted category (i.e. reference group)? How to interpret "–47.003***"?

Which kind of data are these?

y-variable?

x-variables?

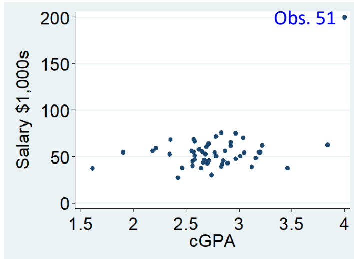
Which are dummy variables?

Outliers & Their Impact

- Outliers: Observations substantially different from the bulk of data
  - Incorrect data entry, confusing question, non-sampling errors or valid data point illustrating extreme situation
- Textbook distinguishes *leverage* and *influential*
- Outliers can affect slope estimate, R^2, and s.e.'s
  - If outlier has large residual, it pulls line towards itself
    - OLS minimizes SSE
    - (Large residual)^2 = ridiculously huge
  - If outlier close to line, makes R^2 higher and s.e. lower (maybe a lot)

Finding & Dealing with Outliers

- Find with graphs (scatter & histograms) & summary statistics
- Investigate outliers
  - Report results with and without outlier(s), hoping they are robust
    - If keep outlier must say why it is valid
    - If drop outlier must show it is invalid



- What can we do? Keep it, drop it, or include a dummy variable for it

13

If Keep Outlier (obs. 51)

```
. regress salary cGPA

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  1,    49) =   12.00
       Model |  5474.43281     1  5474.43281           Prob > F      =  0.0011
    Residual |  22355.0309    49  456.225119           R-squared     =  0.1967
-------------+------------------------------           Adj R-squared =  0.1803
       Total |  27829.4637    50  556.589273           Root MSE      =  21.359

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cGPA |   25.96476   7.495564     3.46   0.001     10.90186    41.02766
       _cons |  -16.53706   20.93776    -0.79   0.433    -58.61305    25.53894
------------------------------------------------------------------------------
```

14

If Drop Outlier (obs. 51)

```
. regress salary cGPA if dummy_obs51==0

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    0.93
       Model |  123.340729     1  123.340729           Prob > F      =  0.3385
    Residual |   6333.8788    48  131.955808           R-squared     =  0.0191
-------------+------------------------------           Adj R-squared = -0.0013
       Total |  6457.21953    49   131.77999           Root MSE      =  11.487

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cGPA |   4.334865     4.4837     0.97   0.338    -4.680219    13.34995
       _cons |   40.47529   12.39228     3.27   0.002     15.55894    65.39165
------------------------------------------------------------------------------
```

15

If Include a Dummy for the Outlier

```
. regress salary cGPA dummy_obs51

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  2,    48) =   81.45
       Model |  21495.5849     2  10747.7924           Prob > F      =  0.0000
    Residual |  6333.8788     48  131.955808           R-squared     =  0.7724
-------------+------------------------------           Adj R-squared =  0.7629
       Total |  27829.4637    50  556.589273           Root MSE      =  11.487

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cGPA |   4.334865     4.4837     0.97   0.338    -4.680219    13.34995
 dummy_obs51 |   142.1852   12.90393    11.02   0.000     116.2402    168.1303
       _cons |   40.47529   12.39228     3.27   0.002     15.55894    65.39165
------------------------------------------------------------------------------
```

How do the coefficient on cGPA and the intercept compare with simply dropping observation 51 from the analysis?

What about the R^2?

16

Interaction Terms

- Interaction term: A variable that is the product (multiplication) of two variables

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

  – How to interpret $(\beta_1 + \beta_3 x_2)$? $(\beta_2 + \beta_3 x_1)$?
- Eg: Test research hypothesis that education is more important for women wrt earnings:

  $$wage = \alpha + \beta educ + \delta fem + \gamma fem * educ + \varepsilon$$

  - If your research hypothesis is true what do you expect about the parameter gamma?

17

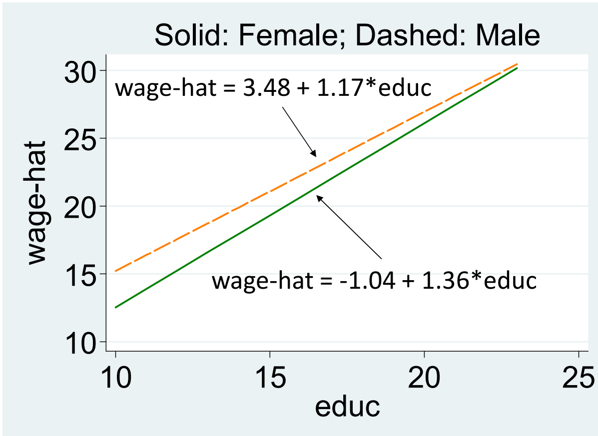
Wage Regression

```
. regress wage educ female femXeduc;

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  3,   996) =  926.32
       Model |  12205.9118     3  4068.63728           Prob > F      =  0.0000
    Residual |  4374.70952   996  4.39227864           R-squared     =  0.7362
-------------+------------------------------           Adj R-squared =  0.7354
       Total |  16580.6214   999  16.5972186           Root MSE      =  2.0958

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.173098   .0363003    32.32   0.000     1.101864    1.244332
      female |  -4.514158     .74495    -6.06   0.000     -5.97601   -3.052307
    femXeduc |   .1832757   .0497587     3.68   0.000     .0856318    .2809197
       _cons |   3.477994   .5404336     6.44   0.000     2.417475    4.538513
------------------------------------------------------------------------------
```

How to interpret these results?

18

Meaning of Interaction Effects



Alternate Wage Regression

```
. regress wage educ male maleXeduc;

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  3,   996) =  926.32
       Model |  12205.9118     3  4068.63728           Prob > F      =  0.0000
    Residual |  4374.70952   996  4.39227864           R-squared     =  0.7362
-------------+------------------------------           Adj R-squared =  0.7354
       Total |  16580.6214   999  16.5972186           Root MSE      =  2.0958

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.356374   .0340326    39.86   0.000      1.28959    1.423158
        male |   4.514158     .74495     6.06   0.000     3.052307     5.97601
   maleXeduc |  -.1832757   .0497587    -3.68   0.000    -.2809197   -.0856318
       _cons |  -1.036165   .5127202    -2.02   0.044      -2.0423   -.0300288
------------------------------------------------------------------------------
```


Another Alternate Specification

```
. regress wage female femXeduc maleXeduc;

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  3,   996) =  926.32
       Model |  12205.9118     3  4068.63728           Prob > F      =  0.0000
    Residual |  4374.70952   996  4.39227864           R-squared     =  0.7362
-------------+------------------------------           Adj R-squared =  0.7354
       Total |  16580.6214   999  16.5972186           Root MSE      =  2.0958

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -4.514158     .74495    -6.06   0.000     -5.97601   -3.052307
    femXeduc |   1.356374   .0340326    39.86   0.000      1.28959    1.423158
   maleXeduc |   1.173098   .0363003    32.32   0.000     1.101864    1.244332
       _cons |   3.477994   .5404336     6.44   0.000     2.417475    4.538513
------------------------------------------------------------------------------
```

While with this specification you can see the slope for males and females directly. The disadvantage is that the statistical tests are NOT whether there is a difference in slope between males and females, but rather whether each differs from zero.

Yet Another Alternate Specification

```
. regress wage educ if female==1;

      Source |       SS       df       MS              Number of obs =     517
-------------+------------------------------           F(  1,   515) = 1525.90
       Model |  6976.8312     1  6976.8312             Prob > F      =  0.0000
    Residual | 2354.71791   515  4.57226779            R-squared     =  0.7477
-------------+------------------------------           Adj R-squared =  0.7472
       Total | 9331.54911   516  18.0843975            Root MSE      =  2.1383

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.356374   .0347229    39.06   0.000     1.288158     1.42459
       _cons |  -1.036165     .52312    -1.98   0.048    -2.063876    -.008453
------------------------------------------------------------------------------
```

One more option, which is less powerful, but yet very popular, is to simply run separate regressions for each sex.

This yields the same lines as shown in the original graph, but cannot test for statistically significant differences by sex.

22

And the Regression for Just Males

```
. regress wage educ if female==0;

      Source |       SS       df       MS              Number of obs =     483
-------------+------------------------------           F(  1,   481) = 1092.28
       Model | 4587.09535     1  4587.09535            Prob > F      =  0.0000
    Residual | 2019.99161   481  4.19956676            R-squared     =  0.6943
-------------+------------------------------           Adj R-squared =  0.6936
       Total | 6607.08696   482  13.7076493            Root MSE      =  2.0493

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.173098    .035495    33.05   0.000     1.103354    1.242843
       _cons |   3.477994   .5284448     6.58   0.000     2.439648    4.516339
------------------------------------------------------------------------------
```

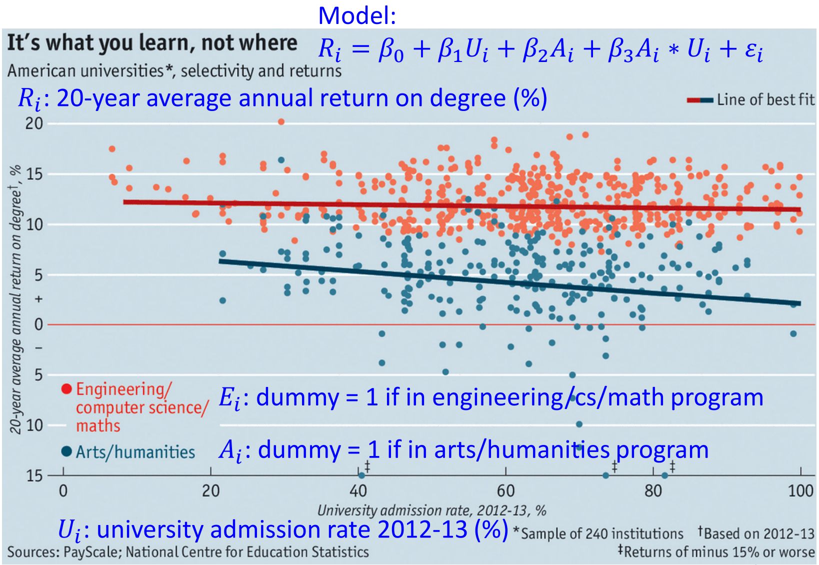
23

"The log-on degree" *The Economist*, March 14, 2015

"A new report from PayScale, a research firm, calculates the returns to a college degree. Its authors compare the career earnings of graduates with the present-day cost of a degree at their alma maters, net of financial aid. College is usually worth it, but not always, it transpires. And what you study matters far more than where you study it." (p. 30)

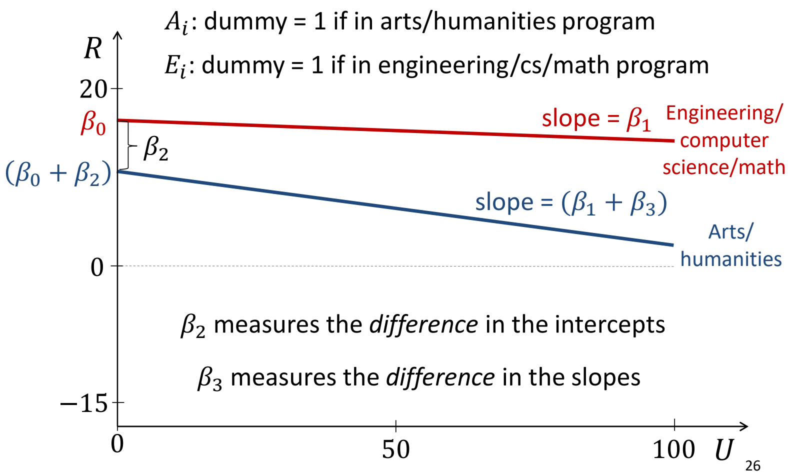
"Engineers and computer scientists do best, earning an impressive 20-year annualised return of 12% on their college fees (the S&P 500 yielded just 7.8%). Engineering graduates from run-of-the-mill colleges do only slightly worse than those from highly selective ones." (p. 30)

24



25

Model: $R_i = \beta_0 + \beta_1 U_i + \beta_2 A_i + \beta_3 A_i * U_i + \varepsilon_i$



26

Article cont'd…

"Business and economics degrees also pay well, delivering a solid 8.7% average return. Courses in the arts or the humanities offer vast spiritual rewards, of course, but less impressive material ones. Some yield negative returns. An arts degree from the Maryland Institute College of Art had a hefty 20-year net negative return of $92,000, for example." (p. 30)

Let R be the 20-year average annual return on a degree (%) and U the university admission rate, 2012-2013 (%), E an indictor for Engineering/computer science/maths, and A an indicator for Arts/humanities. Which model specification fits with the figure?

Cool interactive chart:
http://www.economist.com/blogs/graphicdetail/2015/03/daily-chart-2

27