Functional forms

- Why is this important?
- OLS can be used for relationships that are not strictly linear in *x* and *y* by using nonlinear functions of *x* and *y* – will still be linear in the parameters
  - Can take the natural log of *x*, *y* or both
  - Can use quadratic forms of *x*
  - Can use interactions of *x* variables
- How do you interpret β_1?
  - $ln(y) = \beta_0 + \beta_1 ln(x) + u$
  - $ln(y) = \beta_0 + \beta_1 x + u$
  - $y = \beta_0 + \beta_1 ln(x) + u$
- Why would you use log models?

Functional forms: Quadratics

- How about when we have a model of the form?:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$$

- Can we interpret β_1 alone as the change in *y* with respect to *x*? Do we need to take into account β_2 as well?
- Let's see this case:

$$wage_i = \beta_0 + \beta_1 exp_i + u_i \quad vs. wage_i = \beta_0 + \beta_1 exp_i + \beta_2 exp_i^2 + u_i$$

- Which one is more realistic? Interpretation?

Functional forms: Adding in qualitative Xs

- How about when we add in dummies/binaries and dichotomous Xs?
  - Example: Is there a relationship between wage and race?

- Write down the model specification and interpret the following results (next slide)
- Remember that there is a benchmark or baseline group. What does it happen if there is not?

Interaction terms: Application

```
reg wage female##nonwhite union education exper
```

```
      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  6,  1282) =  102.39
       Model |  26016.8082     6  4336.13471           Prob > F      =  0.0000
    Residual |  54293.0165  1282  42.3502469           R-squared     =  0.3240
-------------+------------------------------           Adj R-squared =  0.3208
       Total |  80309.8247  1288  62.3523484           Root MSE      =  6.5077

------------------------------                  ---------
           wage |      Coef.                     P>|t|
----------------+-------------                  ---------
       1.female |  -3.240148                     0.000
     1.nonwhite |  -2.158525                     0.004
                |
female#nonwhite |   1.095371                     0.022
                |
          union |   1.115044                     0.028
      education |   1.370113                     0.000
          exper |    .165856                     0.000
          _cons |  -7.088725                     0.000
------------------------------                  ---------
```

(result just for illustration)

Interaction terms

- Interacting dummy variables is like subdividing the group
  - Example: have dummies for gender (female or male), as well as for educational attainment (HS incomplete, HS, and BS)
    - How many categories do you have? What's the base group?
    - Write the model (at home):

- Interactions can also include continuous variables.

Goodness of fit and selection of regressors

- Remember that R^2 cannot decrease (usually increase) when you add more Xs to the model
  - adj-R^2 takes this into account
- Don't think only on the adj-R^2; think about the theory you are testing (and also common sense)
- Let's answer the following questions
  - What happens if we include variables in our specification that don't belong?
  - What if we exclude a variable from our specification that does belong?

Goodness of fit and selection of regressors: Ex 1

- Restaurant new location for a established restaurant chain based on volume of customers (Y), using competition (N), population (P) and average household income (I) from nearby places (from Stundenmund, 2016)

	est1	est2
	b/t	b/t
N	-9074.674***	-1487.344
	(-4.42)	(-0.84)
P	0.355***	
	(4.88)	
I	1.288*	2.322**
	(2.37)	(3.50)
Constant	102192.428***	84438.590***
	(7.98)	(5.19)
N	33.000	33.000
r2	0.618	0.305
r2_a	0.579	0.258

* p<0.05, ** p<0.01, *** p<0.001

- What did it happen with the Adj-R^2?
- What did it change?
- Why?

Goodness of fit and selection of regressors: Ex 2

- What if we add an irrelevant and uncorrelated variable? (e.g. A: the last 3 digits of the street address)

	est1	est2
	b/t	b/t
N	-9074.674***	-9099.163***
	(-4.42)	(-4.46)
P	0.355***	0.348***
	(4.88)	(4.80)
I	1.288*	1.239*
	(2.37)	(2.29)
A		10.993
		(1.16)
Constant	102192.428***	97877.397***
	(7.98)	(7.39)
N	33.000	33.000
r2	0.618	0.636
r2_a	0.579	0.584

* p<0.05, ** p<0.01, *** p<0.001

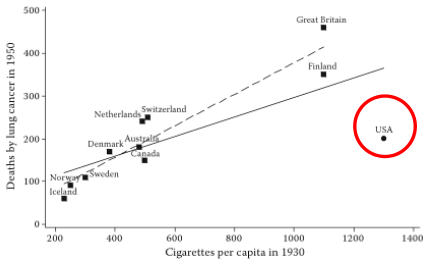
- What did it happen with the R^2 and Adj-R^2?
- What did it change?
- Why?

Goodness of fit and selection of regressors: Ex 3

- Demand for gasoline using petroleum consumption by U.S. State (PCON), using urban highway miles with the state (UHM), gasoline tax rates (TAX) and number of motor vehicles (REG) (from Stundenmund, 2016)
- Write down the model with the expected sign of each coefficient
- Board: estimation with and without UHM
- What did it happen with the Adj-R^2?
- What did it happen with REG?
- Does it help to know that REG and UHM are extremely correlated?

(beyond) LRM: Some common problems with the data

- Non-linear relationships (you can use data transformation - we talked about this)
- Extrapolating beyond the data
- Missing data (common problem beyond LRM)
  - Example: We want to predict the effect of voter turnout in the last elections on people's donations. We find a significant positive relationship, but only when political ideology is included. The problem is that this latter variable wasn't answer by 80% of the people in the dataset.
  - Other common solutions
- Outliers and influence point. Example:



	Mod1	Mod2
cig	0.228**	0.369***
	(3.27)	(8.00)
Constant	67.561	9.139
	(1.38)	(0.32)
N	11	10
r2_a	0.493	0.875
aic	130.353	105.698

Models with binary dependent variable

- What does it happen when there is a binary/dichotomous/dummy dependent variable regression models?
  - This is part of the "qualitative response" regression models involving nominal scale dependent variables that you will see in other courses (e.g. ordinal and multinomial models)
- Examples:
  - Effect of when a person started to smoke on whether he/she smokes at the age of 50
  - Effect of a marketing technique on whether people buy a product
- What are the problems of the linear probability model (LPM) using OLS?
- You will see logit and probit models in other courses (perhaps in one of our review session)

Summary of the last two Parts

- Empirical research methods. Why do we care?
- Data collections and type of measures
- Ethics
- Visualization
- Linear regression models (review)
  - Interpretation
  - Assumptions
  - Model specification errors
  - Garbage in, garbage out