Quadratic Terms, Connecting $(\mu_1 - \mu_2)$ to Regression, and an Economics Paper Illustrating Regression in Action

Lecture 23

Reading: "Quadratic Terms" (Quercus)

1

"Social Connectedness: Measurement, Determinants, and Effects"

ABSTRACT (excerpts): Social networks can shape many aspects of social and economic activity. Traditionally, the unavailability of large-scale and representative data on social connectedness has posed a challenge. We introduce a new measure of social connectedness at the US county level. Our Social Connectedness Index is based on friendship links on Facebook. It corresponds to the relative frequency of Facebook friendship links between every county-pair in the United States, and between every US county and every foreign country. Given Facebook's scale as well as the relative representativeness of Facebook's user body, these data provide the first comprehensive measure of friendship networks at a national level.

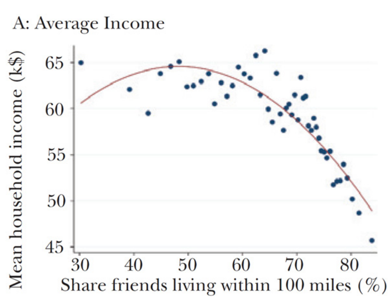
Bailey et al. (2018); https://www.aeaweb.org/articles?id=10.1257/jep.32.3.259

2

Figure 3: Network Concentrations and County-Level Characteristics

Figure 3 presents county-level binned scatterplots using the share of friends living within 100 miles and a number of socioeconomic outcomes.

The overall message is that counties where people have more concentrated social networks tend to have worse socioeconomic outcomes.



Notes: The red line shows the fit of a quadratic regression.

On average, they have lower income, lower education, higher teenage birth rate, and lower life expectancy.

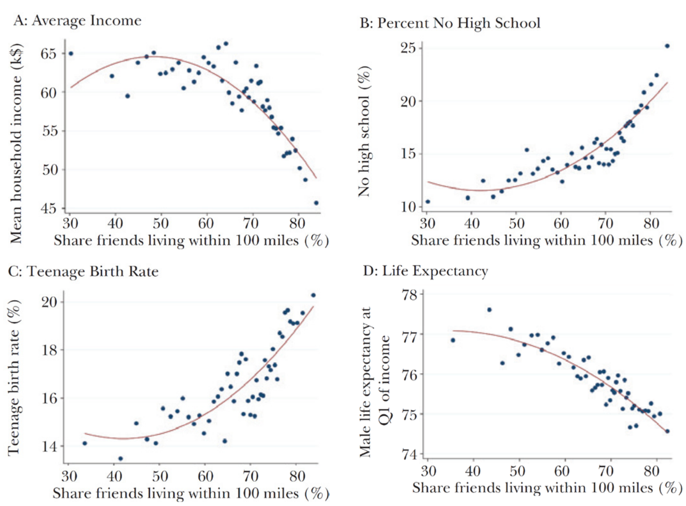
These correlations cannot be interpreted as causal.

3

Quadratic and Polynomials

- When non-linearity is non-monotonic try:
  - Quadratic: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + \cdots + \varepsilon$
  - Polynomial: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + \beta_m z + \cdots + \varepsilon$
  - When do we use these versus logarithms?
  - Careful when interpreting quadratic coefficients
    - You *cannot* hold x^2 constant while changing x
    - For $\hat{y} = b_0 + b_1 x + b_2 x^2$, the point estimate of the slope is $(b_1 + 2b_2 x)$. Note the slope varies with x.

4



5

Recall the diagnostic scatter plot of the residuals versus y-hat: we are hoping to see a cloud of dots with no clear pattern

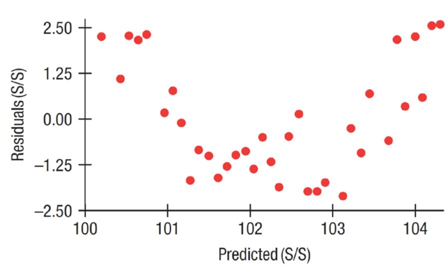


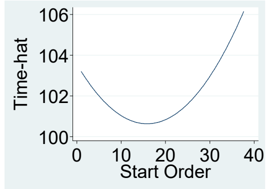
Figure 1 The residuals reveal a bend.

Dependent variable is: Time
R squared = 37.9% R squared (adjusted) = 36.0%
s = 1.577 with 35 − 2 = 33 degrees of freedom

Variable	Coeff	SE(Coeff)	*t*-ratio	P-Value
Intercept	100.069	0.5597	179	<0.0001
StartOrder	0.108563	0.0242	4.49	<0.0001

Table 1 Time to ski the women's downhill event at the 2002 Winter Olympics depended on starting position.

6



Dependent variable is: Time
R squared = 83.3% R squared (adjusted) = 82.3%
s = 0.8300 with 35 − 3 = 32 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	110.139	2	55.0694	79.9
Residual	22.0439	32	0.688871	

Variable	Coeff	SE(Coeff)	*t*-ratio	P-Value
Intercept	103.547	0.4749	218	<0.0001
StartOrder	−0.367408	0.0525	−6.99	<0.0001
StartOrder^2	0.011592	0.0012	9.34	<0.0001

Table 1: Correlates of Urban Air Pollution in China

	Dependent Variable: *log(PM10)*		
Explanatory Variables:	(1)	(2)	(3)
Log(GDP per capita)	-0.434 (0.129)	-0.424 (0.128)	-0.425 (0.128)
(Log(GDP per capita))^2	0.300 (0.075)	0.296 (0.074)	0.296 (0.074)
(Log(GDP per capita))^3	-0.0596 (0.0135)	-0.0592 (0.0134)	-0.0592 (0.0134)
Log(Population)	0.164 (0.014)	0.164 (0.014)	0.164 (0.014)
Log(Manuf. Share)	0.0498 (0.0397)	0.0450 (0.0396)	0.0478 (0.0394)
Log(Ave. Yrs. Schooling)	-0.918 (0.143)	-0.926 (0.142)	-0.923 (0.142)
Log(Rainfall)	-0.0987 (0.0347)	-0.0977 (0.0345)	-0.0980 (0.0345)
Log(Temperature Index)	0.391 (0.074)	0.394 (0.073)	0.393 (0.073)
Time Trend	-0.0316 (0.0031)	-	-0.0767 (0.0130)
(Time Trend)^2	-	-	0.0041 (0.0011)
Year Dummies	No	Yes	No
Constant	4.304 (0.428)	4.353 (0.425)	4.399 (0.426)
R^2	0.432	0.444	0.440
Observations	846	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.

Regression (1): Time Trend

```
      Source |       SS       df       MS              Number of obs =     846
-------------+------------------------------           F( 11,   834) =   57.56
       Model |  37.1271039    11  3.37519127           Prob > F      =  0.0000
    Residual |  48.9026999   834  .058636331           R-squared     =  0.4316
-------------+------------------------------           Adj R-squared =  0.4241
       Total |  86.0298038   845  .101810419           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
------------------------------------------------------------------------------
```

What does including a time trend control for?

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

Regression (2): Year Dummies

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

10

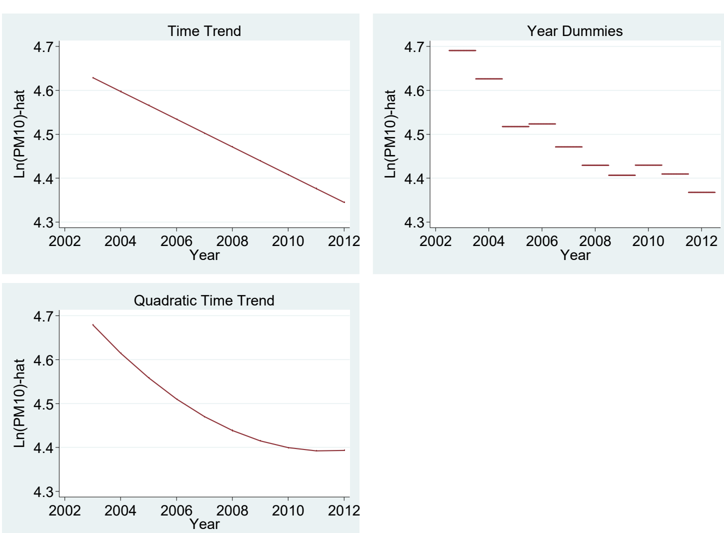
Regression (3): Quadratic Time Trend

Source	SS	df	MS
Model	37.8654782	12	3.15545652
Residual	48.1643256	833	.057820319
Total	86.0298038	845	.101810419

Number of obs	= 846
F(12, 833)	= 54.57
Prob > F	= 0.0000
R-squared	= 0.4401
Adj R-squared	= 0.4321
Root MSE	= .24046

ln_pm10	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_gdp_pc	-.4248166	.1277594	-3.33	0.001	-.6755847	-.1740485
ln_gdp_pc_2	.2962276	.0740302	4.00	0.000	.1509199	.4415353
ln_gdp_pc_3	-.059156	.0133827	-4.42	0.000	-.0854238	-.0328881
ln_pop	.1638634	.0136163	12.03	0.000	.137137	.1905897
ln_manu	.0477641	.0394457	1.21	0.226	-.0296606	.1251888
ln_edu	-.9234477	.1417355	-6.52	0.000	-1.201648	-.6452471
ln_rain	-.097978	.0344953	-2.84	0.005	-.165686	-.03027
ln_temp	.3933151	.0732961	5.37	0.000	.2494483	.5371818
longitude	-.0064097	.0014965	-4.28	0.000	-.009347	-.0034724
latitude	.0054001	.0040752	1.33	0.185	-.0025988	.0133989
trend	-.0767348	.0130054	-5.90	0.000	-.102262	-.0512076
trend_sq	.004085	.0011431	3.57	0.000	.0018413	.0063288
_cons	4.398518	.4257516	10.33	0.000	3.562846	5.23419

11



Multi-Dimensional Data & Fixed Effects

- A full set of fixed effects is common with multi-dimensional (e.g. panel) observational data
  - Idea: fixed effects can control for some lurking variables (e.g. differences across countries)
  - $y_{it} = \alpha + \beta x_{it} + \gamma_t + \delta_i + \varepsilon_{it}$
    - Where are the fixed effects in this model specification?
    - Kinds of lurking/confounding/omitted/unobserved variables these fixed effects can control for?

13

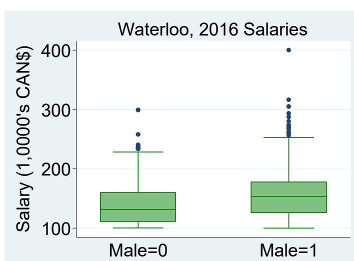
Connection: $(\mu_1 - \mu_2)$ & Regression

- Recall inference about $(\mu_1 - \mu_2)$ – the difference between population means for independent samples – from Chapter 14
  - Case 1 (general): Unequal variances (Section 14.2)
    - Use regression with a dummy for Group 1 (or 2) with *robust standard errors* to address *heteroscedasticity*
  - Case 2 (special): Assume $\sigma_1^2 = \sigma_2^2$ (Section 14.5)
    - Use regression with dummy assuming *homoscedasticity*
  - Control for other factors w/ multiple regression

14

Recall Lecture 18: 2017 ON Public Sector Disclosure of 2016 salaries for University of Waterloo employees

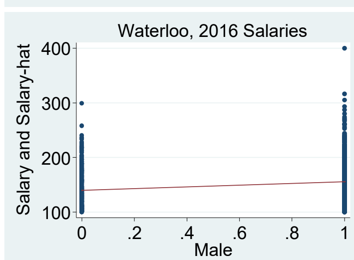
Sex	n	Mean	S.d.
F	416	\$139,743.09	\$33,740.99
M	941	\$155,359.54	\$36,962.36



OLS Results:

Salary-hat = 139.74 + 15.62*Male

R^2 = 0.0385, n = 1,357, s_e = 36.006



15

Regression, Assumes Homoscedasticity

```
.  regress salary male;

      Source |       SS       df       MS              Number of obs =    1357
-------------+------------------------------           F(  1,  1355) =   54.26
       Model |  70350.5619     1  70350.5619           Prob > F      =  0.0000
    Residual |   1756701.7  1355  1296.45882           R-squared     =  0.0385
-------------+------------------------------           Adj R-squared =  0.0378
       Total |  1827052.26  1356  1347.38367           Root MSE      =  36.006

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   15.61645   2.119961     7.37   0.000     11.45769    19.77521
       _cons |   139.7431   1.765358    79.16   0.000       136.28    143.2062
------------------------------------------------------------------------------
```

To test $H_0: (\mu_1 - \mu_2) = 0$ for Case 2 (specific) use t test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(155.35954 - 139.74309)}{\sqrt{\frac{1296.4588}{941} + \frac{1296.4588}{416}}} = \frac{15.61645}{2.119961} = 7.37$$

$$s_p^2 = \frac{(941-1)36.96236^2 + (416-1)33.74099^2}{941 + 416 - 2} = 1296.4588$$

16

Regression Addressing Heteroscedasticity w/ Robust S.E.'s

```
. regress salary male, robust;

Linear regression                                      Number of obs =    1357
                                                       F(  1,  1355) =   58.25
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0385
                                                       Root MSE      =  36.006

------------------------------------------------------------------------------
             |               Robust
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   15.61645   2.046117     7.63   0.000     11.60255    19.63035
       _cons |   139.7431   1.653518    84.51   0.000     136.4994    142.9868
------------------------------------------------------------------------------
```

To test $H_0: (\mu_1 - \mu_2) = 0$ for Case 1 (general) use t test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(155.35954 - 139.74309)}{\sqrt{\frac{36.96236^2}{941} + \frac{33.74099^2}{416}}} = \frac{15.61645}{2.046} = 7.63$$

17

The Economics of Cross-Border Travel

Abstract We model the decision to travel across an international border as a trade-off between benefits derived from buying a range of products at lower prices and the costs of travel. We estimate the model using microdata on Canada–United States travel. Price differences motivate cross-border travel; a 10% home appreciation raises the propensity to cross by 8% to 26%. The larger elasticity arises when the home currency is strong, a result predicted by the model. Distance to the border strongly inhibits crossings, with an implied cost of 87 cents per mile. Geographic differences can partially explain why American travel is less exchange rate responsive.

Chandra, Ambarish, Head, Keith, and Tappata, Mariano (2014) "The Economics of Cross-Border Travel." *Review of Economics and Statistics* 96.4, 648-661. Also, see "Readings" in portal.

18

Section 2.B: The Exchange Rate Elasticity of Cross-Border Travel

Excerpt (p. 650): Our first regression exercise is to determine the elasticity of cross-border trips with respect to the real exchange rate.

Our goal is establish simple data relationships to motivate the development of a model in the subsequent section of the paper.

We therefore work with a minimal specification. Denoting the number of cars that cross the border by n, and the real exchange rate by e, our specification is:

$$\ln(n_{it}) = \alpha + Month_t + Province_i + \eta_1 \ln(e_t) + \eta_2 post911_t + \eta_3 t + \eta_4 t^2 + \varepsilon_{it}$$

where i denotes a province and t denotes time (in months since January 1972).

19

$$\ln(n_{it}) = \alpha + Month_t + Province_i + \eta_1 \ln(e_t) + \eta_2 post911_t + \eta_3 t + \eta_4 t^2 + \varepsilon_{it}$$

Excerpt (p. 650): The month effects account for the strong seasonality in travel.

We add province fixed effects, as well as an indicator variable for the period following September 11, 2001 when border security was increased.

Finally, we add a linear and quadratic trend to capture secular effects such as population changes.

We estimate this equation separately for residents of each country. Therefore, for Canada, this regression models the number of cars returning from the US in a given province and month. For the US, it represents the cars that enter the corresponding Canadian province. (p. 5)

20

$$\ln(n_{it}) = \alpha + Month_t + Province_i + \eta_1 \ln(e_t) + \eta_2 post911_t + \eta_3 t + \eta_4 t^2 + \varepsilon_{it}$$

Excerpt (pp. 650 – 651): Implicit in the estimation of the above equation is the assumption that causation runs only from the real exchange rate to crossing decisions.

This assumption is defensible because demand for foreign currency created by US and Canadian cross-border shoppers is unlikely to be large enough to move the global foreign exchange markets.

To gain some perspective on relative magnitudes, Canadians spent \$4.2 billion in the US while Americans spent \$1.8 billon in Canada during the first quarter of 2010. This represents a mere 0.04% of the foreign exchange turnover involving the Canadian Dollar. (p. 6)

21

Nominal versus Real Exchange Rates

- Nominal Exchange Rate CAN/US
  - E.g. March 27, 2015 nominal CAN/US exchange rate (noon) is 1.2580: 1.00 USD = 1.26 CAN
- Real Exchange Rate CAN/US
  - p. 649 "We obtained monthly average data on the spot market exchange rate between the U.S. and Canadian currencies. Using data on monthly CPIs for both countries, we construct the Real Exchange Rate (RER) for each month."
    - "Why Real Exchange Rates?" by IMF researcher http://www.imf.org/external/pubs/ft/fandd/2007/09/pdf/basics.pdf

Table C.1. Summary Statistics: 1972 – 2010 (3276 province-months)

	Mean	SD	Median	Min	Max
Day Trips (1000 vehicles)					
U.S. Residents	114.7	211.4	42.7	1	1224.8
Canadian Residents	173.7	213.2	100.8	2.9	1192.9
Overnight Trips (1000 vehicles)					
U.S. Residents	41.7	71.9	14.4	0.5	519.1
Canadian Residents	42.8	51.6	18.3	1.1	346.4
Nominal ER (CAN/USD)	1.236	0.166	1.221	0.962	1.6
Real ER	1.007	0.127	0.99	0.814	1.333

7 Canadian provinces border U.S. * 39 years * 12 months = 3,276 province-months

Table 1. Regression of Log Crossings, 1972 – 2010

Length of stay:	Daytrip		Daytrip	
Residence:	U.S.	Canadian	U.S.	Canadian
$\ln(e)$ (CAN/USD)	1.24^{***} (0.17)	-1.62^{***} (0.24)	0.93^{***} (0.28)	-1.71^{***} (0.28)
$\ln(e) * [e > 1.09]$ (strong USD)			0.90^{**} (0.37)	0.54^{*} (0.33)
$\ln(e) * [e < 0.90]$ (strong CAN)			-0.87^{**} (0.34)	-0.87^{***} (0.24)
R^2	0.98	0.98	0.98	0.98

Notes: Newey-West standard errors in parentheses are robust to serial correlation out to 60 months. Significant at *10%, **5%, ***1%. An observation is a province-year-month. N = 3276. Regressions include month and province fixed-effects, a post 9/11 indicator, and trend variables.

What is the point estimate of the elasticity of day trips from the U.S. to Canada as the real exchange rate increases (i.e. U.S. dollar gets stronger) when the U.S. dollar is already strong?

Excerpt (p. 651): This section has uncovered four stylized facts of cross-border travel that should be features of a quantitative model of crossing decisions.

First, while there is always two-way movement across the border, there are large within- and between-year fluctuations.

Second, there is a robust relationship between exchange rates and travel: the stronger the currency in the country of residence, the more trips.

Third, elasticities are asymmetric. In absolute value Canadian residents have higher percentage responses to changes in the exchange rate.

Fourth, exchange rate elasticities are larger when the home currency is stronger.