Analysis of Economics Data
Chapter 15: Regression with Transformed Variables

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 15: Regression with Transformed Variables

- Regression often involves variables that have been transformed
  - e.g. quadratics, natural logarithm, interactions (products of variables)
  - e.g. $\widehat{y}_i = b_1 + b_2 x_{2i} + b_3 x_{3i} + b_3 x_{2i} \times x_{3i}$.
- OLS estimation remains fine if model is still linear in coefficients $b_1, \ldots, b_k$.
- But interpreting results is more difficult when the model is nonlinear in the underlying variables
  - the marginal effect $\Delta\widehat{y}/\Delta x$ is no longer the slope coefficient
  - plus there are different ways to compute $\Delta\widehat{y}/\Delta x$
  - and if y is transformed then prediction of y becomes more difficult.

Outline

1. Example: Earnings, Gender, Education and Type of Worker
2. Marginal effects for Nonlinear Models
3. Quadratic Model and Polynomial Models
4. Interacted Regressors
5. Log-linear and Log-log models
6. Prediction from Log-linear and Log-log Models
7. Models with a Mix of Regressor Types

Datasets: EARNINGS_COMPLETE

15.1 Example: Earnings, Gender, Education, Worker Type

- Dataset EARNINGS_COMPLETE
  - 872 female and male full-time workers aged 25-65 years in 2000.

Variable	**Definition**	**Mean**	**Standard Deviation**	**Min**	**Max**
Earnings	Annual earnings in $	56369	51516	4000	504000
Age	Age in years	43.31	10.68	25	65
Gender	= 1 if female	0.433	0.496	0	1
Education	Years of schooling	13.85	2.88	0	20
d1 or dself	= 1 if self-employed	0.089	0.286	0	1
d2 or dpriv	=1 if private sector employee	0.760	0.427	0	1
d3 or dgovt	=1 if government sector employee	0.149	0.356	0	1
Agesq	*Age* squared	1989.7	935.7	625	4225
Educbyage	*Education* times *Age*	598.8	193.69	0	1260
Hours	Usual hours worked per week	44.34	8.50	35	99
Lnhours	Natural logarithm of *Hours*	3.78	0.16	3.56	4.60
Lnearnings	Natural logarithm of *Earnings*	10.69	0.68	8.29	13.13
n	872				

15.2 Marginal Effects for Nonlinear Models

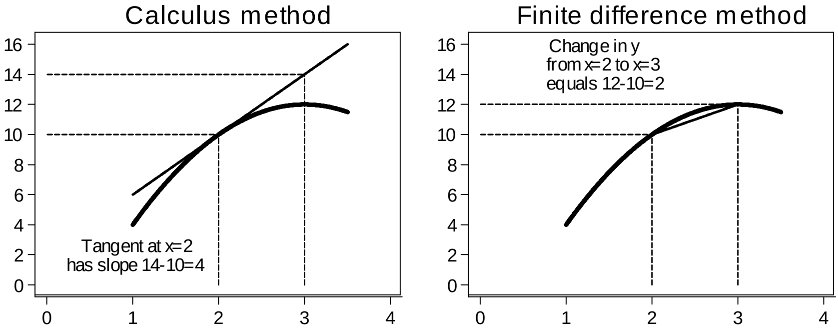
- Examples of nonlinear models
  - Quadratic: $\widehat{y} = b_1 + b_2 x + b_3 x^2$
  - Interactions: $\widehat{y} = b_1 + b_2 x + b_3 z + b_3 (x \times z)$
  - Natural logarithms: $\ln \widehat{y} = b_1 + b_2 x + b_3 z$.
- The **marginal effect (ME)** on the predicted value of y of a change in a regressor is

$$\mathrm{ME}_x = \frac{\Delta \widehat{y}}{\Delta x}.$$

- In nonlinear models we get different results depending on method
  - **calculus** method: use the derivative $d\widehat{y}/dx$ (for very small Δx)
  - **finite difference** methods: such as $\Delta x = 1$.

Calculus method versus Finite Difference Method

- Plotted curve is $y = 12 - 2 \times (x-3)^2$
  - calculus method at $x = 2 : \frac{dy}{dx} = 12 - 4x = 4$ at $x = 2$.
  - finite difference for $x = 2$ to $x = 3 : \Delta y = 12 - 10 = 2$.



AME, MEM and MER

- **Marginal effect** $\text{ME}_x = \Delta\widehat{y}/\Delta x$ varies with the level of x.
  - So what value of x do we evaluate at?
- 1. **Average** marginal effect (AME): evaluate for each i and average

$$\text{AME} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \text{ME}_i = \frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{\Delta\widehat{y}_i}{\Delta x_i}.$$

- 2. Marginal effect **at the mean** (MEM): evaluate ME at $x = \bar{x}$

$$\text{MEM} = \text{ME}|_{x=\bar{x}} = \left.\frac{\Delta\widehat{y}}{\Delta x}\right|_{x=\bar{x}}.$$

- 3. Marginal effect **at a representative value** (MER): evaluate ME at a representative value of x, say $x = x^*$

$$\text{MER} = \text{ME}|_{x=x^*} = \left.\frac{\Delta\widehat{y}}{\Delta x}\right|_{x=x^*}.$$

- Most often use AME, with ME_i evaluated using calculus methods.

Computation of Marginal Effects

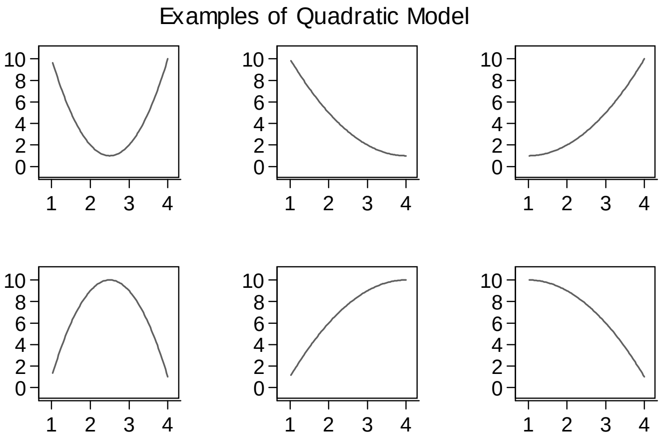
- Suppose $\text{ME}_x = 2x^2 + 3z^2$ so also depends on z.
- For AME evaluate for each individual and average
  - $\text{AME}_x = \frac{1}{n}\sum_{i=1}^{n}(2x_i^2 + 3z_i^2)$.
- For the MEM set all variables at their means
  - $\text{MEM}_x = 2\bar{x}^2 + 3\bar{z}^2$.
- For MER evaluate at a particular value x^* of x
  - with z taking the values for each individual $\text{MER}_x = 2(x^*)^2 + \frac{1}{n}\sum_{i=1}^{n} 3z_i^2$
  - or additionally specify a particular value z^* of z, so $\text{MER}_x = 2(x^*)^2 + 3(z^*)^2$.
- Some statistical packages provide post-estimation commands to calculate AME, MEM and MER
  - these additionally provide standard errors and confidence intervals for these estimates.

Nonlinear Models in Practice

- Several issues arise when the relationship is nonlinear.
- Estimation by OLS is possible if the coefficients in the model still appear linearly
  - e.g. $\mathrm{E}[y|x] = \beta_1 + \beta_2 \ln x$ is okay as linear in β_1 and β_2
  - e.g. $\mathrm{E}[y|x] = \exp(\beta_1 + \beta_2 x)$ is not okay as not linear in β_1 and β_2
- Direct interpretation of slope coefficients may not be possible
  - use marginal effects.
- Prediction of y problematic when y is transformed before regression
  - e.g. if $\mathrm{E}[\ln y|x] = \beta_1 + \beta_2 x$.
- Difficult to choose the appropriate nonlinear model
  - when can't do a scatter plot of several regressors.

15.3 Quadratic Model and Polynomial Models

- A **quadratic model** is the model $y = \beta_1 + \beta_2 x + \beta_3 x^2 + u$.
- The figure gives various examples
  - top row has $\beta_2 < 0$ and bottom row has $\beta_2 > 0$.



Marginal Effects for Quadratic Model

- Fitted quadratic model $\widehat{y} = b_1 + b_2 x + b_3 x^2$

$$\text{ME}_x = b_2 + 2b_3 x \text{ (using calculus methods).}$$

- The average marginal effect is

$$\begin{aligned} \text{AME} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(b_2 + 2b_3 x_i) \\ &= b_2 + 2b_3 \times \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i \\ &= b_2 + 2b_3 \bar{x}. \end{aligned}$$

Quadratic Example: Earnings and Age

- Regress *Earnings* (y) on *Age* (x), *Agesq* (x^2), and *Education* (z), with heteroskedastic-robust t-statistics in parentheses

$$\widehat{y} = \underset{(-4.02)}{-98620} + \underset{(2.86)}{3105x} - \underset{(-2.38)}{29.66\,x^2} + \underset{(8.94)}{5740z}, \quad R^2 = .1196, \quad n = 872,$$

- Quadratic term is warranted as for x^2 we have $|t| = 2.38 > t_{868;.025} = -1.963$.
- The turning point for the quadratic is at $x = -b_2/2b_3$
  - here at $Age = 3105/(2 \times (-29.66)) = 52.3$ years.
  - earnings on average increase to 52.3 years and then decline.
- ME$= 3105 - 29.66x - 29.66\Delta x$ by finite difference method
- ME$= 3105 - 59.32x$ using calculus method
- AME$= \frac{1}{n}\sum_{i=1}^{n}(3105 - 59.32x_i) = 3105 - 59.32\bar{x} = 3105 - 59.32 \times 43.31 = 536$

Polynomial Model

- A **polynomial model of degree** p includes powers of x up to x^p.
- The fitted model is

$$\widehat{y} = b_1 + b_2 x + b_3 x^2 + \cdots + b_{p+1} x^p.$$

- This model has up to $p - 1$ turning points.
- Determine polynomial order by progressively adding terms x^2, x^3, ...
  - until additional terms are no longer statistically significant.
- By calculus methods the marginal effect is

$$\text{ME} = b_2 + 2b_3 x + 3b_4 x^2 + \cdots + p b_{p+1} x^{p-1},$$

which again will vary with the point of evaluation x.

15.4 Interacted Regressors

- Example with $x \times z$ an interacted regressor is

$$y = \beta_1 + \beta_2 x + \beta_3 z + \beta_4 x \times z + u.$$

- **Estimation** is straightforward
  - create a variable xz, say, that equals $x \times z$
  - run OLS regression of y on an intercept, x, z and xz.
  - the fitted model (with $xz = x \times z$) is

$$\widehat{y} = b_1 + b_2 x + b_3 z + b_4 xz,$$

- **Interpretation** of regressors is more difficult.
- The marginal effect (ME) on $\widehat{y}$ of a change in x, holding z constant, depends on coefficients of both x and xz

$$\text{ME}_x = \frac{\Delta \widehat{y}}{\Delta x} = b_2 + b_4 z.$$

- To test statistical significance of x do joint F-test on variables x and xz : $H_0 : \beta_2 = 0, \beta_4 = 0$.

Interactions Example: Earnings, Education and Age

- OLS regression of *Earnings* on *Age* (x) and *Education* (z)
  - both variables are statistically significant at 5% (t stats in parentheses)

$$\widehat{y} = \underset{(-4.15)}{-46875} + \underset{(3.47)}{525}\, x + \underset{(9.06)}{5811}\, z, \quad R^2 = .115, \quad n = 872,$$

- Add *AgebyEduc* $(x \times z)$ as a regressor
  - now no regressors are statistically significant at 5%

$$\widehat{y} = \underset{(-0.94)}{-29089} + \underset{(0.18)}{127}\, x + \underset{(1.88)}{4515}\, z + \underset{(0.52)}{29.0}\, x \times z, \quad R^2 = .115, \quad n = 872,$$

- The marginal effect of one more year of schooling is

$$\text{ME}_{Ed} = 4515 + 29 \times Age.$$

  - So the returns to education increase as one ages.

Joint Hypothesis tests

- Individual coefficients are statistically insignificant at 5%
- But a joint test on *Age* (x) and *AgebyEduc* $(x \times z)$
  - a test of $H_0 : \beta_x = 0, \beta_{xz} = 0$ yields $F = 6.49$ with $p = 0.002$
  - so age remains highly statistically significant
  - similarly F-test for the two education regressors is $F = 43.00$ with $p = 0.000$.
- Why the difference between individual and joint tests?
- The interaction variable *AgebyEduc* is
  - quite highly correlated with *Age* $(\widehat{\rho} = 0.72)$
  - quite highly correlated with *Education* $(\widehat{\rho} = 0.64)$.
- When regressors are highly correlated with each other
  - individual contributions are measured much less precisely
  - here standard errors of *Age* and *Education* more than triple from 151 and 641 to 719 with inclusion of variable *AgebyEduc*.

15.5 Natural Logarithm Transformations

- Consider models with $\ln y$ and/or $\ln x$.
- Chapter 9 gave interpretation of coefficients
  - semi-elasticity in log-linear model
  - elasticity in log-log model.
- Now additionally consider **marginal effects** $\text{ME}_x = \Delta y / \Delta x$.
- For log-linear model $\ln y = b_1 + b_2 x$ use $\text{ME}_x = b_2 \widehat{y}$
  - reason: $\Delta \ln y / \Delta x = b_2$ but $\Delta \ln y \simeq \Delta y / y$
    so $(\Delta y / y) / \Delta x = b_2$ and on solving $\Delta y / \Delta x = b_2 y$
- Similarly for log-log model $\ln y = b_1 + b_2 \ln x$ use $\text{ME}_x = b_2 \widehat{y} / x$.

Log-linear Model

- OLS regression of ln(*Earnings*) on *Age* (x) and *Education* (z)
  - both variables are statistically significant at 5% (t stats in parentheses)

$$\widehat{\ln y} = \underset{(59.63)}{8.96} + \underset{(3.83)}{0.0078}x + \underset{(11.68)}{0.101}\,z, \quad R^2 = .190,$$

- One year of aging, controlling for education, is associated with a 0.78 percent $(= 100 \times 0.0078)$ increase in earnings.
- The marginal effect of aging is $0.0078\widehat{y}$
  - always positive and increases with age since $\widehat{y} \uparrow$ with age.
  - simplest to evaluate at $\bar{y}$, then MEM of a year of aging is a \$440 increase in earnings $(= 0.0078 \times 56369)$.

Log-log Models

- OLS regression of ln(*Earnings*) on ln(*Age*) (x) and *Education* (z)
  - both variables are statistically significant at 5% (t stats in parentheses)

$$\widehat{\ln y} = \underset{(24.23)}{8.01} + \underset{(4.21)}{0.346} \ln x + \underset{(11.67)}{0.100}\, z, \quad R^2 = .193,$$

- A one percent increase in age, controlling for education, is associated with a 0.346 percent increase in earnings.
- The marginal effect of aging is $0.346\widehat{y}/x$
  - always positive and increases with age since $\widehat{y} \uparrow$ with age.
  - simplest to evaluate at $\bar{y}$ and $\bar{x}$, then MEM of a year of aging is a $450 increase in earnings $(= 0.346 \times 56369/43.41)$.

15.6 Prediction from Log-linear and Log-log Models

- Consider log-linear model: $\widehat{\ln y} = b_1 + b_2 x + b_3 z$.
- A naive prediction in level is $\widehat{y} = \exp(\widehat{\ln y}) = \exp(b_1 + b_2 x + b_3 z)$.
- But this **underpredicts** due to retransformation bias (next page).
- Instead if errors were normal and homoskedastic predict y using

$$\widetilde{y} = \exp(s_e^2/2) \times \exp(\widehat{\ln y}).$$

- Here s_e is standard error of the regression for the $\ln y$ regression.
- Example: $s_e = 0.4$ (which is large for data on a log scale)
  - need to rescale by $\exp(s_e^2/2) = 1.215$

Retransformation Bias Correction

- Log-linear population model assumes $\mathrm{E}[u|x] = 0$ in

$$\ln y = \beta_1 + \beta_2 x + u.$$

- Taking the exponential on both sides: $y = \exp(\beta_1 + \beta_2 x + u)$.
- So the conditional mean of y given x is

$$\begin{aligned} \mathrm{E}[y|x] &= \mathrm{E}[\exp(\beta_1 + \beta_2 x + u)|x] \\ &= \exp(\beta_1 + \beta_2 x) \times \mathrm{E}[\exp(u)|x]. \end{aligned}$$

- Problem: We need to know $\mathrm{E}[\exp(u)|x]$.
  - in general $\mathrm{E}[\exp(u)|x] > 1$
  - $\mathrm{E}[\exp(u)|x] = \exp(\sigma_u^2/2)$ if $u|x \sim N(0, \sigma_u^2)$
    - i.e. normal homoskedastic errors.
  - then $\mathrm{E}[y|x] = \exp(\sigma_u^2/2)\exp(\beta_1 + \beta_2 x)$.

R-squared with Transformed Dependent Variable

- R^2 in regress y on x measures the fraction of the variation in y around $\bar{y}$ that is explained by the regressors.
- R^2 in regress $g(y)$ on x instead measures the fraction of the variation in $g(y)$ around $\overline{g(y)}$ that is explained by the regressors.
- So **meaningless** to compare R^2 across models with different transformations of the dependent variable.
- For right-skewed data R^2 is usually higher in models for $\ln y$ rather than y.
- For persistent time series right-skewed data R^2 is usually higher in models for y than for Δy.

15.7 Models with a Mix of Regressor Types

- Levels example with $R^2 = .206$, $n = 872$ is

$$\begin{aligned} \widehat{Earnings} &\\ = & \underset{(-5.38)}{-356631} - \underset{(-5.31)}{14330} \times Gender + \underset{(3.08)}{3283} \times Age - \underset{(-2.59)}{31.58} \times Agesq \\ & + \underset{(8.85)}{5399} \times Education + \underset{(1.07)}{9360} \times Dself - \underset{(-0.10)}{291} \times Dgovt \\ & + \underset{(4.34)}{69964} \times Lnhours, \end{aligned}$$

- Interpretation controlling for other regressors
  - ME of aging is $3283 - 63.16 \times Age$
  - Self-employed workers on average earn \$9,360 more than private sector workers (the omitted category)
    - ★ though this comparison is statistically insignificant at 5%
  - A 1% change in hours worked is associated with a \$699 increase in earnings.

Dependent Variable in Natural Logarithms

- Natural logarithms example with $R^2 = .206$, $n = 872$ is

$$\begin{aligned}
&\widehat{LnEarnings} \\
= &\underset{(6.89)}{4.459} - \underset{(-4.88)}{0.193} \times Gender + \underset{(3.55)}{0.0560} \times Age - \underset{(-2.99)}{0.000549} \times Agesq \\
&+\underset{(11.17)}{0.0934} \times Education - \underset{(-1.17)}{0.118} \times Dself + \underset{(1.53)}{0.070} \times Dgovt \\
&+\underset{(6.88)}{0.975} \times Lnhours
\end{aligned}$$

- Interpretation controlling for other regressors
  - women on average earn 19.3% less than men
  - earnings increase with age to 51.0 years $(= -.560/(2 \times (-.000549))$ and then decrease
  - Self-employed workers on average earn 11.8% less than private sector workers (the omitted category)
    - though this comparison is statistically insignificant at 5%
  - A 1% change in hours worked is associated with a 0.975% increase in earnings.

Some in-class Exercises

1. For $\widehat{y} = 2 + 3x + 4x^2$ for a dataset with $\bar{y} = 30$ and $\bar{x} = 2$ give the marginal effect of a one unit change in x. Hence give the AME.
2. For $\widehat{y} = 1 + 2x + 4d + 7d \times x$ for a dataset with $\bar{y} = 22$, $\bar{x} = 3$ and $\bar{d} = 0.5$ give the marginal effect of a one unit change in x. Hence give the AME.
3. For model $\ln y = \beta_1 + \beta_2 + u$ we obtain $\widehat{\ln y} = 1 + 2x$, $n = 100$, $s_e = 0.3$. Give an estimate of $\mathrm{E}[y|x]$.