POLS/CS&SS 503: Advanced Quantitative Political Methodology

TRANSFORMATIONS

May 5, 2015

Jeffrey B. Arnold

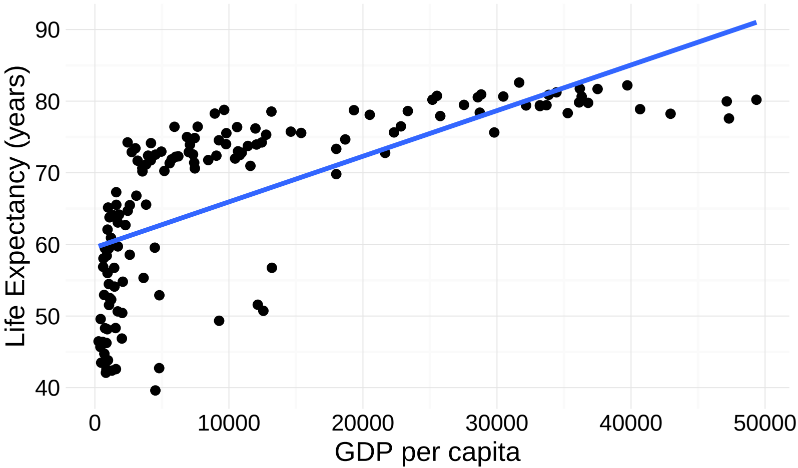
UNIVERSITY of WASHINGTON
DEPARTMENT OF POLITICAL SCIENCE

CENTER for STATISTICS and the SOCIAL SCIENCES

Overview

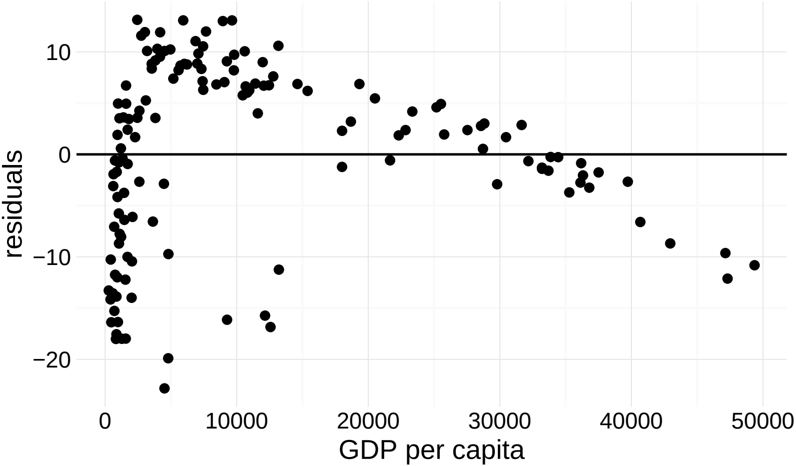
Residuals and Misspecification

Life Expectancy (years) on GDP per capita (2007)



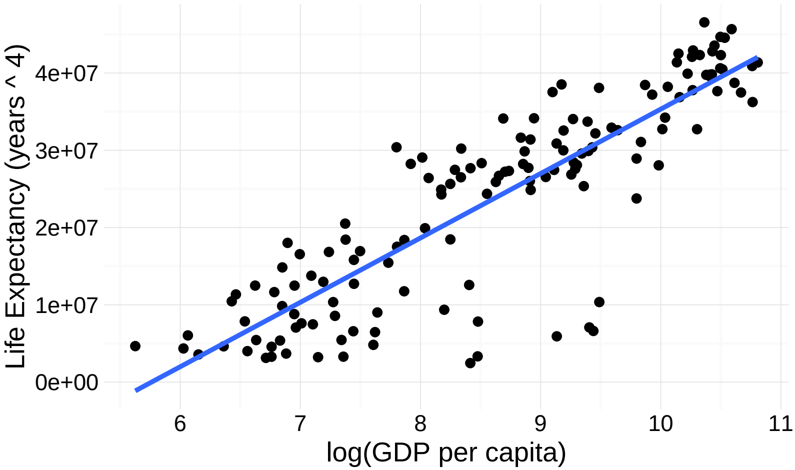
Residuals and Misspecification

Residuals of Life Expectancy (years) on GDP per capita (2007)



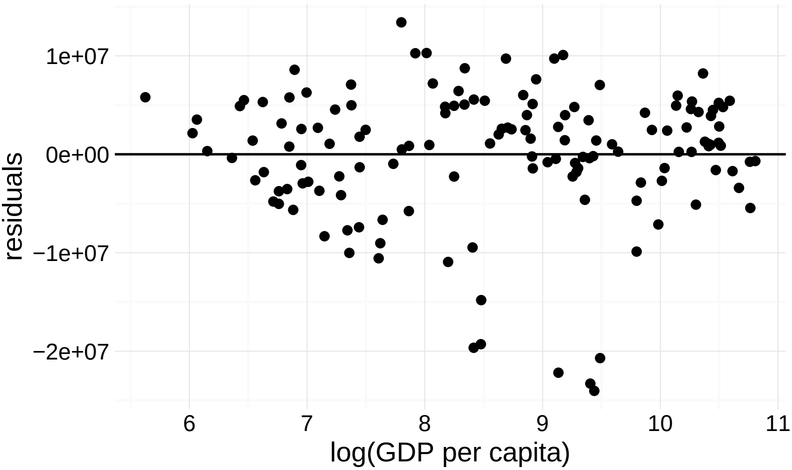
Residuals and Misspecification

Life Expectancy (years ^{4}) on log GDP per capita (2007)



Residuals and Misspecification

Residuals of Life Expectancy (years ^{4}) on log GDP per capita (2007)



Interpreting Logarithms

How would you interpret the following?

- GDP per $\text{cap}_i = \alpha + \beta \log(\text{school})_i$
- log GDP per $\text{cap}_i = \alpha + \beta(\text{school})_i$
- log GDP per $\text{cap}_i = \alpha + \beta \log(\text{school})_i$

Linearizing Functions

Can you linearize these functions by taking the logarithms of both sides?

Exponential

$$y_i = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i}$$

Yes

$$\log y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

Gravity Equation

$$\text{trade}_{ij} = \frac{\alpha \text{GDP}_i^{\beta_1} \text{GDP}_j^{\beta_2}}{\delta d_{ij}^{\beta_3}}$$

Yes

$$\log \text{trade}_{ij} = (\log \alpha + \log \delta) + \beta_1 \log \text{GDP}_i + \beta_2 \text{GDP}_j - \beta_3 d_{ij}$$

Cobb-Douglas Production Function

$$y = \alpha x_1^{\beta} x_2^{\gamma}$$

Yes

$$\log y = \log \alpha + \beta \log x_1 + \delta \log x_2$$

CES Production Function

$$y = \alpha(\delta x_1^\rho + (1-\delta)x_2^\rho)^{\gamma/\rho}$$

No

$$\log y = \log \alpha + (\gamma/\rho)\log(\delta x_1^\rho + (1-\delta)x_2^\rho)$$

Can't simplify $\log(\delta x_1^\rho + (1-\delta)x_2^\rho)$.

Close to 0, $\log(1 + x) \approx x$

Why can diff in logs be interpreted as a %$\triangle$

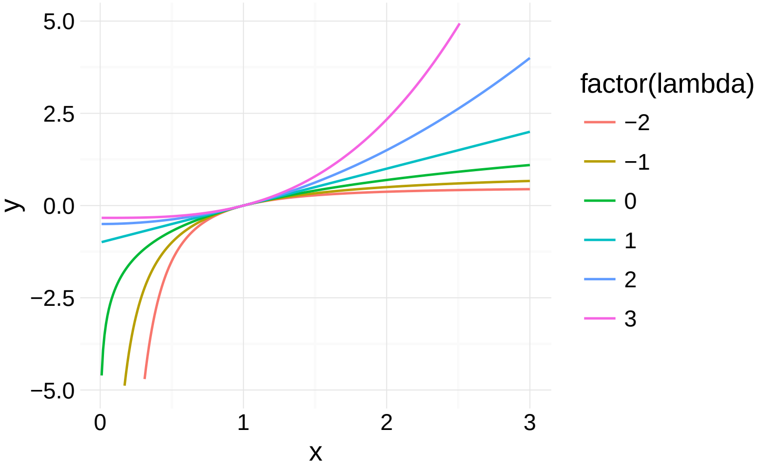
Note: $\log(1+r) \approx r$ when r small
Then,

$$\begin{aligned} \log(x) - \log(x(1+r)) &= \log(1+r) \approx r \\ &= \%\Delta x/100 \end{aligned}$$

This property only holds for the natural logarithm.

Box-Cox Family of Transformations

```
## Warning: Removed 95 rows containing missing values (geom_path).
```



Plot for $\lambda = 0.25, 0.5, 0, 2, 4, 8$ for $x = (0, 4]$

Box-Cox Family of Transforms

$$\begin{cases} f(x,\lambda) = \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ f(x,\lambda) = \log x & \text{if } \lambda = 0 \end{cases}$$

- Can solve for λ to transform x to be symmetric.
- **car** function: `powerTransform`, `bcTransform`.
- In regression: If know λ can transform y or x.

Logarithms and Power Transformations

Linear Transformations of Regressions

Transforming Dependent Variable

- Do not change the fit (R^2, SSE) of OLS
- Can be useful (sometimes) for interpretation

Linear Transformations of Regression

Scalar Multiplication

$$y = \alpha + \beta x_i + \epsilon$$

Multiplying x_i by a just changes the slope to βa

$$y = \alpha + (\beta a)x_i + \epsilon$$

Linear Transformations of Regression

Scalar Addition

$$y = \alpha + \beta x_i + \epsilon$$

Adding a constant c to x_i

$$y = \alpha + \beta(x_i + c) + \epsilon$$

Standardized Coefficients / Regressors

$$y = \alpha + \beta_0 + \beta_1 \frac{x_i - \bar{x}}{\mathsf{SD}\left(x\right)} + \epsilon_i$$

- Can be useful for default interpretation (controversial)
- But about same as comparing $x + \mathsf{SD}(x)$ post-estimation.
- Bad for skewed variables, binary variables?
- Transform regressors, not functions of regressors.
- Gelman: Continuous: divide by $2\,\mathsf{SD}(x)$; Binary: center at mean.
- No need for them for default interpretation. With computational power, simulations better.
- Very important to standardize X in machine learning applications, or anywhere with complicated optimization problems.

Logarithms and Power Transformations

Linear Transformations of Regressions

Transforming Dependent Variable

Logit Transformation

- Suppose $Y \in (0, 1)$
- The logit transformation $\tilde{y} = \text{log}(y/(1-y))$,

$$\text{log}\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 x_1 + \cdots + \epsilon$$

- What if original data included 0s or 1s
- Not a “logit model”, linear regression with logit transformed response variable