Research Methods in Political Science I

8. Linear Regression (3)

Yuki Yanai

School of Law and Graduate School of Law

November 25, 2015

KOBE UNIVERSITY

Today's Menu

1. Assumptions of Linear Regression and Regression Diagnostics
   - Assumption of Linear Regression
   - Regression Diagnostics
2. Transformation of Variables
   - Linear Transformation
   - Centering
   - Correlation Coefficient and "Regression to the Mean"
   - Logarithmic Transformation
3. Presentation of the Results
   - What to Report
   - How to Present Your Results

Assumptions of OLS

1. **Estimated models are valid**
2. The response can be represented by a linear function of the predictors
3. Errors are independent: $\mathrm{Cor}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$
4. Errors are homoskedastic: $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all i
5. Errors are normally distributed: $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$

We can't test the assumptions!

Residual Plot

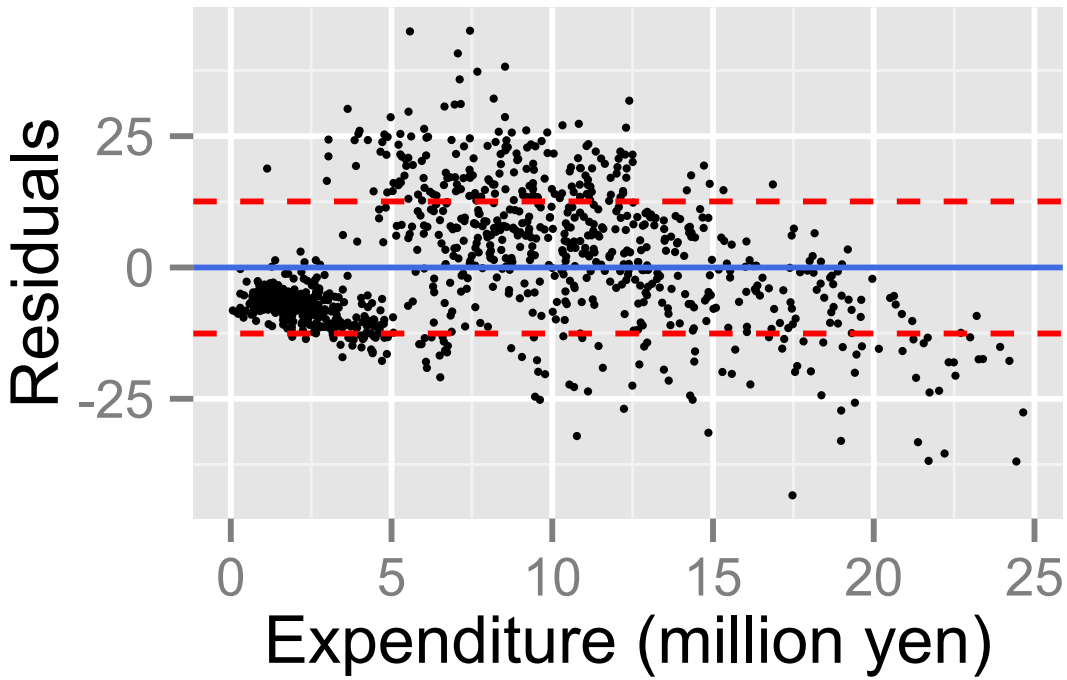

Figure: Residual plot of the model explaining the vote share by the expenditure. The red dashed line shows $\pm$ standard deviation. If we find some systematic pattern in a residual plot, it implies that the model has some problems.

Other Regression Diagnostics

- Inquiry of outliers: Studentized residuals
- Leverage: Hat values
- Influence of an observation: dfbeta, Cook's distance
- etc.
- Read the following books for more information
  - Cook, R. Dennis, and Sanford Weisberg. 1999. *Applied Regression Including Computing and Graphics.* New York: John Wiley: 334–369.
  - Fox, John. 1997. *Applied Regression, Linear Models, and Related Methods.* Thousand Oaks: SAGE:267–366.

Linear Transformation

- Transform some variables make the coefficient easier to interpret
- Transform by linear functions
- Does not change the content of the models

Scaling: Unit Conversion

Regression models explaining the vote share by the electoral expenditure can be represented as follows.

1. Expenditure measured by million yen

$$\text{vote share} = 7.7 + 3.1 \cdot \text{expenditure (million yen)} + \text{error}$$

2. Expenditure measured by yen

$$\text{vote share} = 7.7 + 0.0000031 \cdot \text{expenditure (yen)} + \text{error}$$

- Seemingly, the effect of the expenditure is larger in the first model than the second
- In fact, two equations show the same
- It is easier to interpret the first model than the second (why?)

Standardization by z Values

- We can fit the model using the z value (z score) of a variable
- The z value of a variable x is

$$z = \frac{x - \bar{x}}{u_x} = \frac{x - \text{mean of } x}{\text{unbiased sd of } x}$$

- Standardize all the predictors:
  - Estimated coefficient: the change of the response corresponding to the increase of the predictor by 1 unit, other thing equal
  - Intercept: predict value of the response when all the predictors take the mean

Other Standardization

- Another example: standardization of a 7-level Likert variable
  - 1: strongly disagree $\cdots$ 7: strongly agree $\rightarrow$ hard to interpret the coefficient
  - Standardization:

$$\frac{\text{point} - 4}{3}$$

    $\rightarrow -1$: strong disagree, 0: neither agree nor disagree, 1: strongly agree
  - Coefficient shows the difference between "strongly disagree" and "Neither..." or between "Neither..." and "strongly agree"

Interpretation of the Regression Intercepts

- Intercept: predicted values of the response when all the predictors are zero
- Some variables never take 0 $\rightarrow$ No meaningful interpretation of the intercept
- Zero is frequently the boundary value $\rightarrow$ Intercept shows an extreme of the data

$$\downarrow$$

- Center the predictors! (a kind of linear transformation, so the content of the regression model does not change)

Centering

1. Center variables to their sample means

$$\text{centered } x = x - \bar{x}$$

2. Centring by prior or common knowledge
   - E.g. Centering of female dummy: the female to male ratio should be 1 to 1

$$\text{c\_female} = \text{female} - 0.5$$

   - E.g. Centering of IQ: the mean of IQ should be 100

$$\text{c\_IQ} = \text{IQ} - 100$$

Intercept of the regression model of which all the predictors are centered:

predicted value of the response when all the predictors take the central values

Simple Regression with the Standardized Variables

Simple regression with the standardized variables x and y:
$y = a + bx + \varepsilon$

$$
\begin{aligned}
x &= \frac{x_{\text{raw}} - \bar{x}_{\text{raw}}}{u(x_{\text{raw}})} \\
y &= \frac{y_{\text{raw}} - \bar{y}_{\text{raw}}}{u(y_{\text{raw}})}
\end{aligned}
$$

- Intercept $a = 0$
- Slope $b \in [-1, 1]$: correlation coefficient of x and y:

$$|b| > 1 \Rightarrow \sigma_y > \sigma_x$$

Correlation Coefficients and the Coefficients of Simple Regression

Simple regression in general:

- σ_{xy}: covariance of x and y
- ρ: correlation coefficient of x and y

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- The slope of the regression line, b :

$$b = \rho \frac{\sigma_y}{\sigma_x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

Principal Components Line and Regression Line

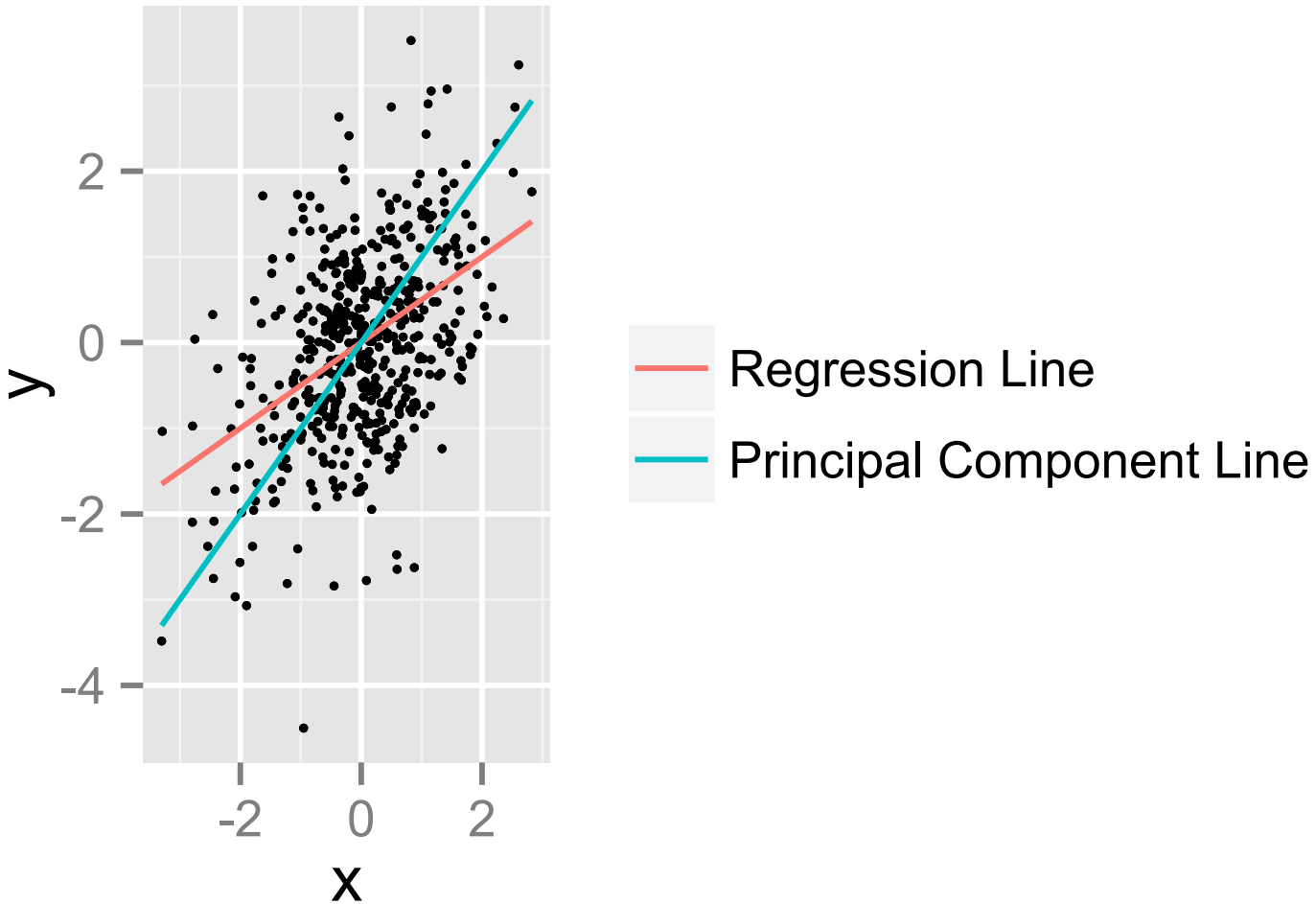

Figure: Standardized y and x: $\rho = 0.5$

Regression to the Mean

Compare the Regression Line to the Principal Component Line

- Principal Component Line
  - underestimates y when x is small
  - overestimates y when x is large
- Regression line goes through the center of data at any given x
- **Regression to the mean**: measure in standard deviation, $|\hat{y} - \bar{y}| < |x - \bar{x}|$
  - It does not tell that “any variables converge to the mean over time”
  - It states that the distance to the predicted values from its mean is smaller the distance to the observed value of the predictor form its mean.

Logarithm

- Logarithmic function: the inverse the exponential function
- For $x = a^p$, we call p "the base a logarithm of x" and write $p = \log_a x$
- Domain: $x > 0$
- E.g. Base 10 logarithm
  - When x increases as $1, 10, 100, \cdots = 10^0, 10^1, 10^2, \ldots,$
  - The base 10 logarithm of x increases as $0, 1, 2, \ldots$

  $\rightarrow$ We can consider the order of variables: easy to handle large numbers
- Frequently used base: e (exponential)

Log of x

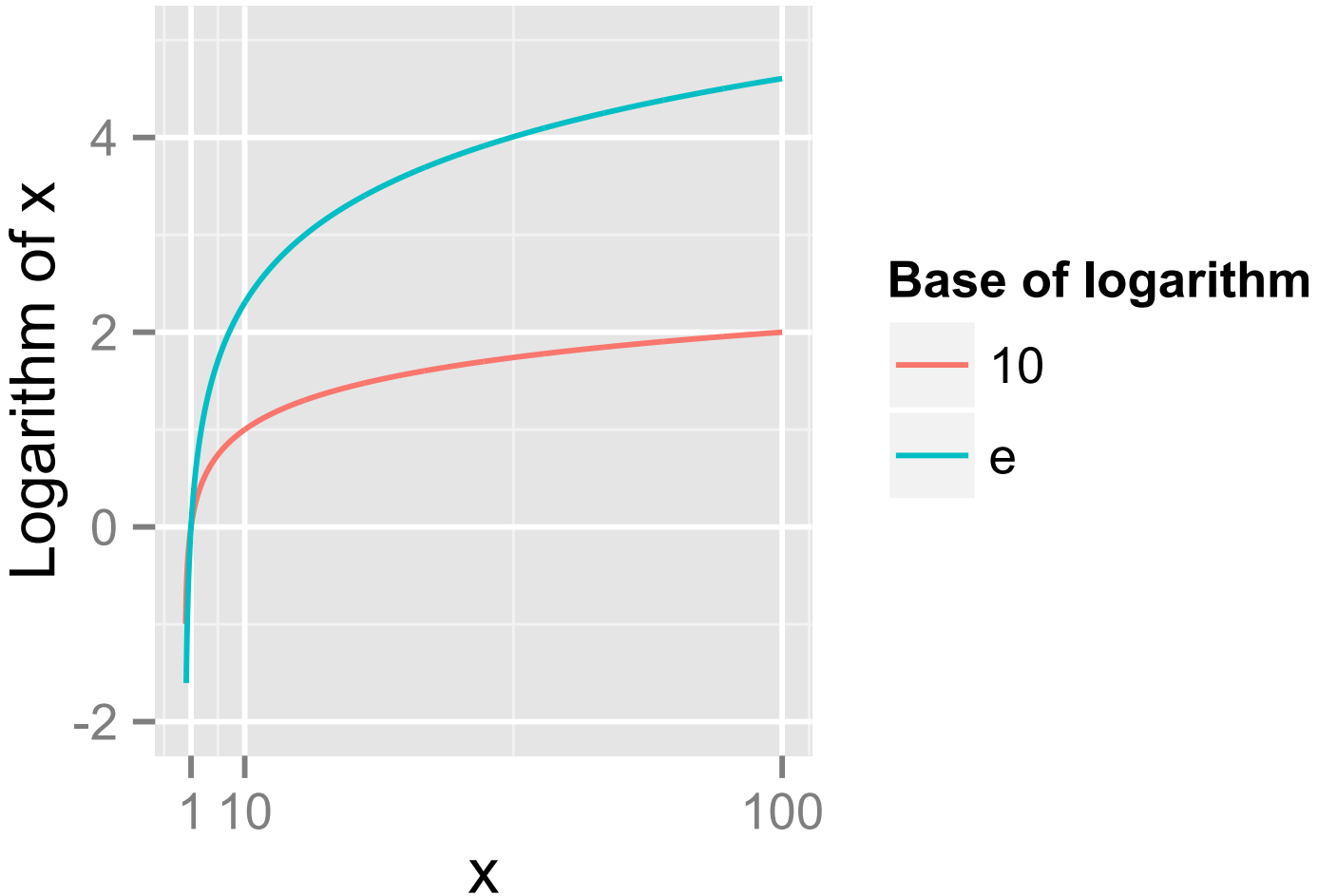

Figure: $\log_e x$ and $\log_{10} x$

Merits of Logarithmic Transformation

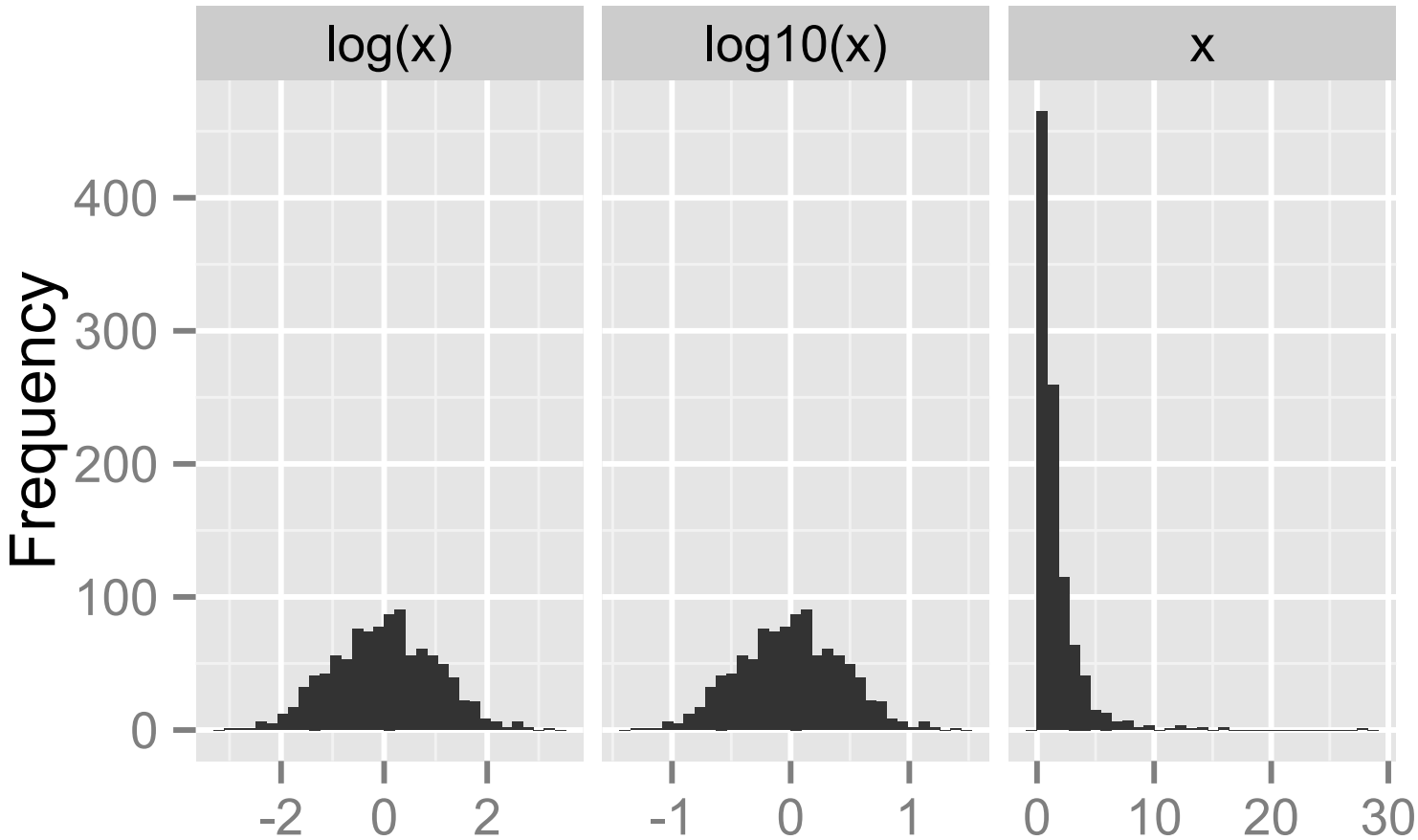

Figure: Distributions of log(x) ($= \log_e x$), log10(x) ($= \log_{10} x$), and x

Natural Logarithm: Base e Logarithm

- Natural logarithm of x: $\log_e(x) \rightarrow$ Writes $\ln(x)$ or simply $\log(x)$
- Easy to interpret the results with the natural logarithms of variables
- E.g. the response is the natural log

$$\log(y_i) = b_0 + 0.06x_i + \varepsilon_i$$

  - 1-unit increase of x increases $\log(y)$ by 0.06 units
  - percentage change of y by 1-unit increase of x is $(\exp(0.06) - 1) = 0.06 = 6\%$
  - 1-unit increase of x increases y by about 6% (0.06)
  - Coefficient 0.06: the change rate of y (this approximation works only for the values near 0)

Estimated Coefficient as the Change Rate: When the Response is the Natural Logarithm

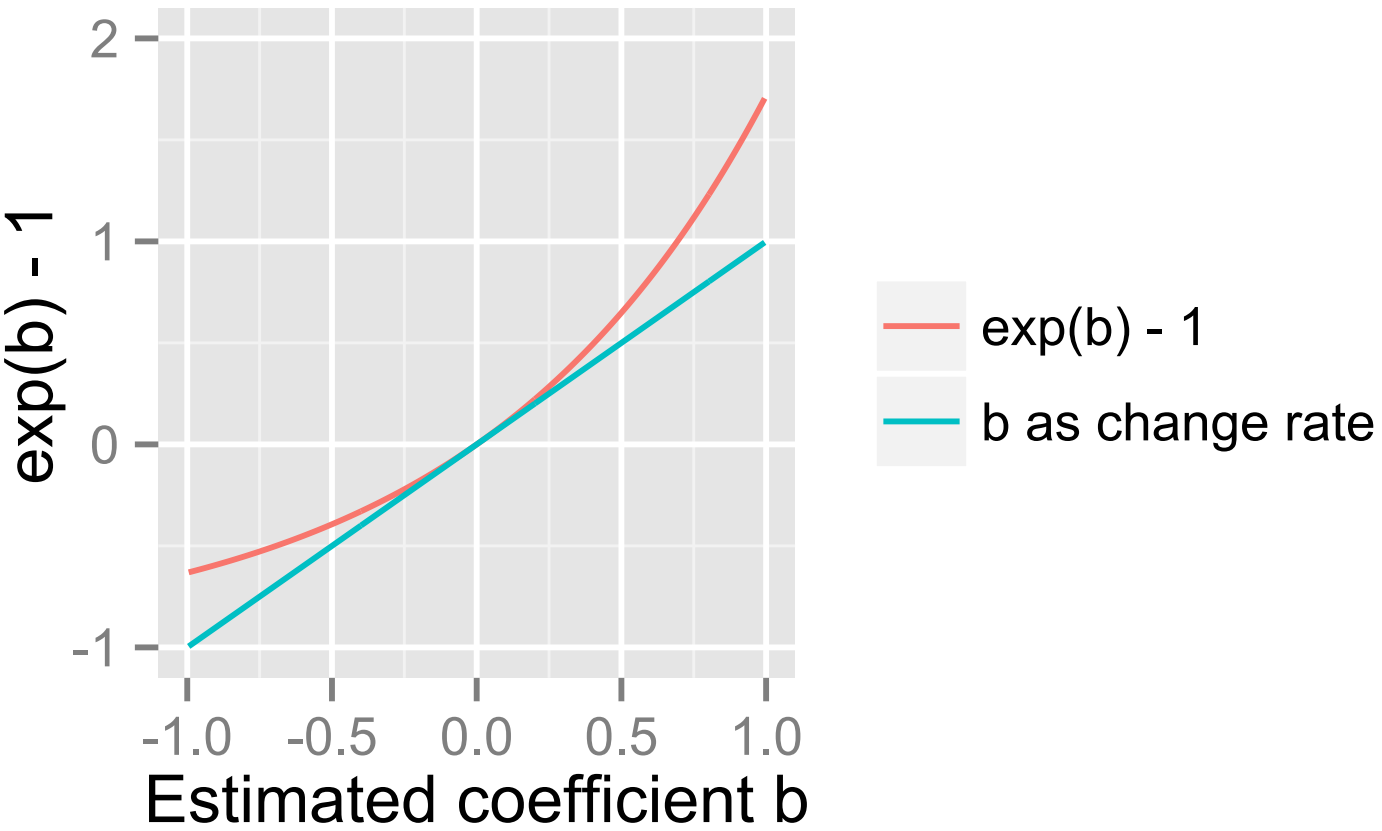

Figure: Coefficient as Change Rate

Natural Logarithms and Base 10 Logarithms

$$\log_{10} y_i = b_0 + 0.026 x_i + \varepsilon_i$$

- 1-unit increase of $x \rightarrow$ 0.026-unit increase of $\log_{10}(y)$
- 1-unit increase of $x \rightarrow 10^{0.026} - 1 = 0.06 =$6% increase of y
- Coefficient 0.026: hard to interpret as it is: we prefer natural logs

Interpretation of the Models with Logarithmic Variables

Response y	Predictor x	Meaning of b
y	x	1-unit increase of $x \to b$-unit increase of y
y	$\log(x)$	1% increase of $x \to b$-unit increase of y
$\log(y)$	x	1-unit increase of $x \to 100b$% increase of y
$\log(y)$	$\log(y)$	1% increase of $x \to 100b$% increase of y (elasticity)

Note: (2) When b is not near 0, we need to calculate $\exp(b) - 1$.

See Tufte, E. (1974) *Data Analysis for Politics and Policy*: pp.108–134 for detailed explanation.

Always Required Information

- Regression models: equations or words
- Detailed explanation of the response and predictor variables
- Estimated coefficients of the regression models
- Sample size and (adjusted-)R^2 (Not “R2”!)
- Values showing the estimation uncertainty (at least one)
  - Standard errors
  - t values
  - p values
- **Substantive interpretation of the results**

Reporting the Substantive Meaning (1)

E.g. $\widehat{\text{vote share}} = 7.7 + 3.1 \cdot \text{expenditure}$

- Bad: "the coefficient of the electoral expenditure is 3.1, and the effect is statistically significant."
  - What does "3.1" mean?
  - Why do we care the result? (Or so what?)

Reporting the Substantive Meaning (2)

E.g. $\widehat{\text{vote share}} = 7.7 + 3.1 \cdot \text{expenditure}$

- Explain the substantive meaning in sentences.
  - Meaning of "3.1": "Every additional million yen is expected increase the vote share by 1 percentage point."
  - Substantive significance: "The electoral expenditure ranges from ten thousand yen to 25 million yen, and the standard deviation is about five million yen. The increase of the expenditure by one standard deviation (e.g., from five million yen to 10 million yen) increases the vote share by about 15.5 percentage point. A fifteen point increase of the vote share could change the winner of the election. Thus, the effect is substantively significant.
    (These sentences are merely for explanation, and I do not argue these are true.)

How to Present

Use different methods depending on the situation

1. sentences + equations
2. sentences + tables (what to include in the tables?)
3. sentences + figures (what kinds of figures?)
   - Scatter plot with the regression line
   - Caterpillar plot

Criteria to choose the method

- **What you want to present to whom?**
- Complexity of your model: One (or a few) figure is enough?
- With or without interaction terms
- Rules of the journals
- etc.

Example: Equation (+ sentences)

$$\widehat{\text{vote}} = \underset{(0.69)}{7.91} + \underset{(1.23)}{18.10} \cdot \text{experience} + \underset{(0.12)}{1.85} \cdot \text{expenditure}$$

(Standard errors are in parentheses. The number of observations is 1,124, and the adjusted R^2 is 0.56)

Example: Table (+ sentences) (1)

Table: Estimation result by OLS: the response is the vote share (%)

Predictor	Estimated coefficient	Standard error
Intercept	7.91	0.69
MP Experience	18.10	1.23
Expenditure (million yen)	1.85	0.12
observations (n)	1124	
adjusted R^2	0.56	

Example: Table (+ sentences) (2)

Table: Estimation results by OLS: the response is vote share (%)

	Estimated coefficient	
Predictor	Model 3	Model 4
Intercept	7.91	−2.07
	(0.69)	(0.72)
MP Experience	18.10	45.91
	(1.23)	(1.58)
Expenditure (million yen)	1.85	4.87
	(0.12)	(0.16)
Experience × Expenditure		−4.76
		(0.21)
observations (n)	1,124	1,124
adjusted R^2	0.56	0.70

Note: Standard errors are in parentheses

Example: Scatter Plot (+ sentences)

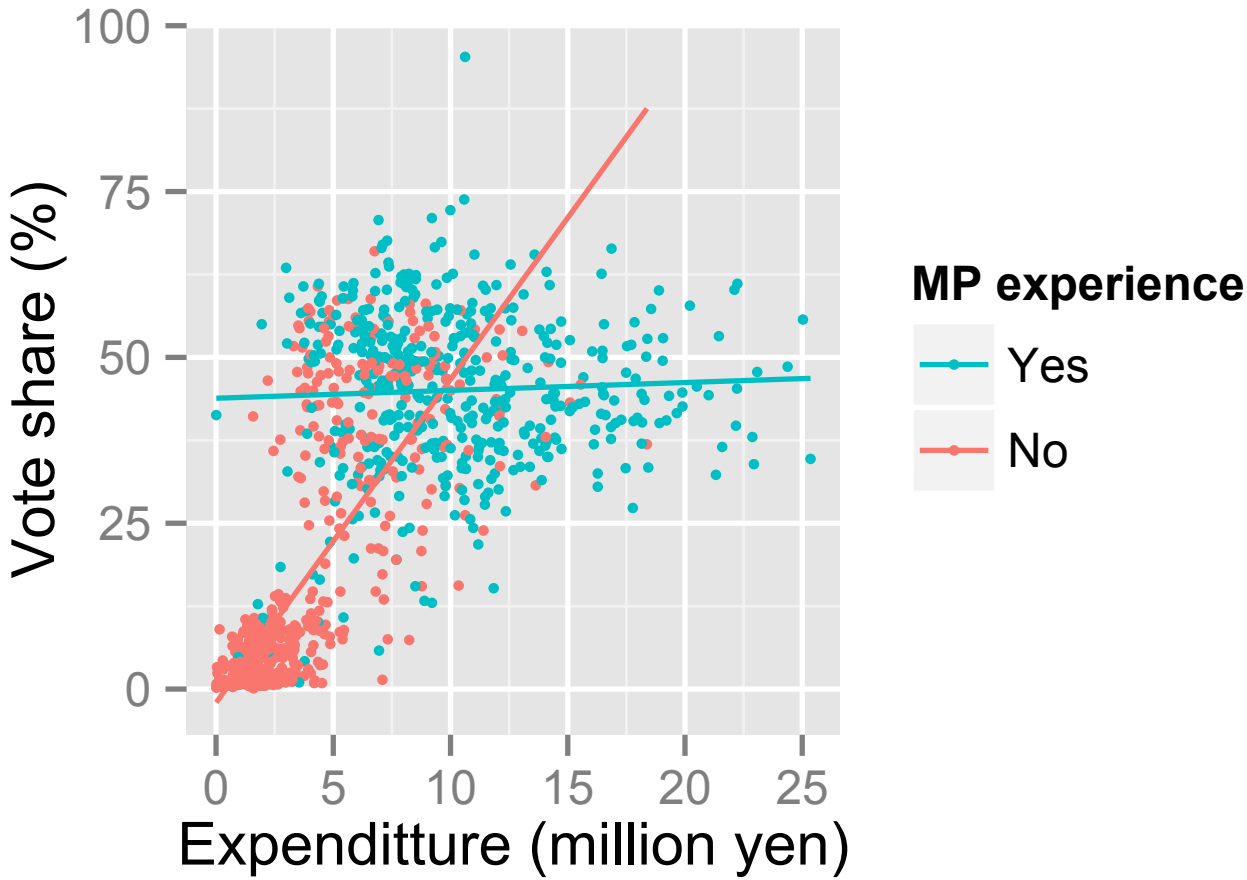

Figure: Explaining the vote share by the expenditure: fitting different lines by prior MP experience

Example: Caterpillar Plot (+ sentences)

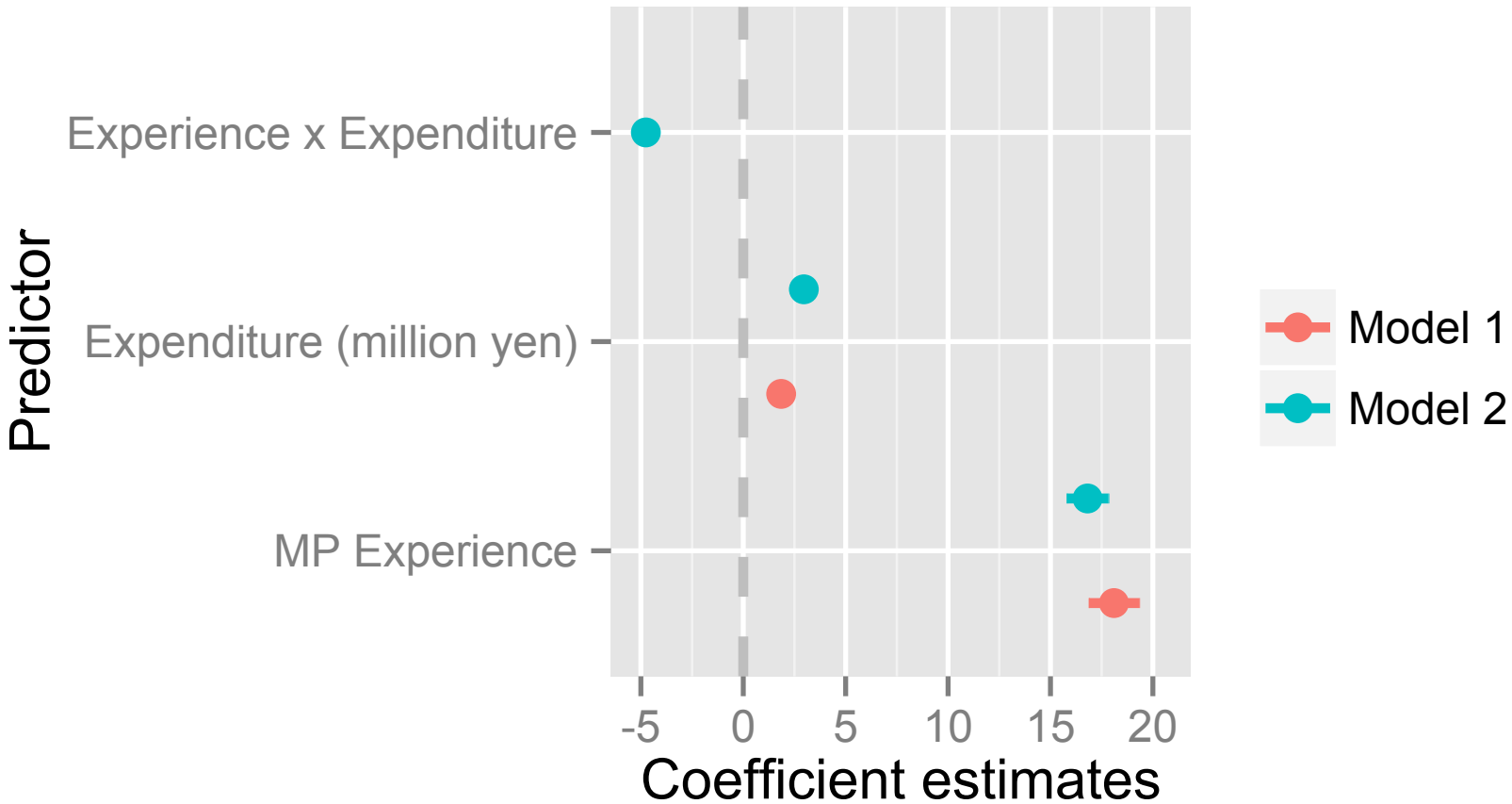

Figure: Coefficient estimates and 95% confidence intervals