BTRY 6020: Module 2
Lecture 7: Transformations and Assumptions

Nayel Bettache

Spring 2025

Logistics

- End of Module 2 today
- Module 2 Assessment will be released today, due Feb 21 (Friday)

Recap

The population model we are trying to recover is

$$E(Y \mid \mathbf{X} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

Recap

The population model we are trying to recover is

$$E(Y \mid \mathbf{X} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

- Can include categorical variables by using dummy variables
  - Choose reference category
  - Include a binary variable for each other category
- Can include interactions to allow slope of one variable to depend on other variables
  - Covariates are product of other covariates
  - Always include main effect when including interactions

Transformations

Transformations

We transformed x by taking a square, but we can use other transformations

- Most common transform is $\log(y)$ transform
- Sometimes $1/y$ or $\sqrt{y}$ is also used
- Can transform covariates

$$E(Y \mid X = x) = b_0 + b_1 \log(x)$$

Transformations

We transformed x by taking a square, but we can use other transformations

- Most common transform is $\log(y)$ transform
- Sometimes $1/y$ or $\sqrt{y}$ is also used
- Can transform covariates
- Can transform dependent variable

$$E(\log(Y) \mid X = x) = b_0 + b_1 x$$

Transformations

We transformed x by taking a square, but we can use other transformations

- Most common transform is $\log(y)$ transform
- Sometimes $1/y$ or $\sqrt{y}$ is also used
- Can transform covariates
- Can transform dependent variable
- Can transform dependent variable and covariates

$$E(\log(Y) \mid X = x) = b_0 + b_1 \log(x)$$

Transformations

- Fitting a linear model with transformed data is conceptually the same
- Just "plug-in" transformed data
- Careful about interpretation!

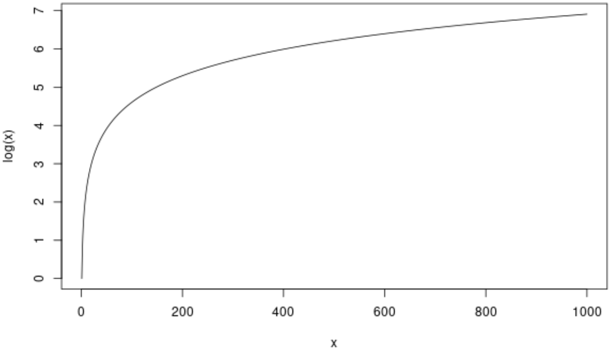
Properties of log

Key Properties of the Natural Logarithm:

$$\textbf{1. Definition:} \quad \log_e(x) = a \Leftrightarrow e^a = x$$

$$\textbf{2. Product Rule:} \quad \log_e(xy) = \log_e(x) + \log_e(y)$$

$$\textbf{3. Quotient Rule:} \quad \log_e(x/y) = \log_e(x) - \log_e(y)$$



Interpretation of Log-Transformed Covariates

Model:

$$Y_i = b_0 + b_1 \log(X_i) + \varepsilon_i$$

Expected Value:

$$E(Y_i \mid X_i = x) = b_0 + b_1 \log(x)$$

How does a 1% increase in X affect Y?

- Suppose $X_j = 1.01X_i$, meaning X_j is 1% larger than X_i.
- Difference in expectations:

$$E(Y_j \mid X_j = 1.01x) - E(Y_i \mid X_i = x) = b_1 \log(1.01) \approx 0.01 b_1$$

- Interpretation: For a 1% increase in X, the expected change in Y is approximately $b_1 \times 0.01$.

Interpreting a Log-Transformed Dependent Variable

Model:

$$\log(Y_i) = b_0 + b_1 X_i + \varepsilon_i$$

Exponentiating Both Sides:

$$Y_i = e^{b_0} \cdot e^{b_1 X_i} \cdot e^{\varepsilon_i}$$

Expected Value:

$$\begin{aligned} E(Y_i \mid X_i = x) &= e^{b_0} \cdot e^{b_1 x} \cdot E(e^{\varepsilon_i}) \\ &\neq e^{b_0} \cdot e^{b_1 x} \quad (\text{since } E(e^{\varepsilon_i}) \neq 1) \end{aligned}$$

Special Case: If $\varepsilon_i \sim N(0, \sigma^2)$, then:

$$E(e^{\varepsilon_i}) = e^{\sigma^2/2} \quad \Rightarrow \quad E(Y_i \mid X_i = x) = e^{b_0} \cdot e^{b_1 x} \cdot e^{\sigma^2/2}$$

Interpretation of b_1 in Log-Log Models

Model:

$$\log(Y_i) = b_0 + b_1 \log(X_i) + \varepsilon_i$$

Elasticity Interpretation:

- If X increases by 1

$$\frac{E(Y_j \mid X_j = 1.01X_i)}{E(Y_i \mid X_i)} = 1.01^{b_1}$$

- Percentage change:

$$100 \times (1.01^{b_1} - 1)$$

- Example: If $b_1 = 0.5$, a 1% increase in X leads to a **0.5% increase in Y**.

Comparison to Simple Linear Regression

Linear Model Assumptions

Linear regression model:

$$Y_i = b_0 + \sum_{j=1}^{p} b_j X_{i,j} + \varepsilon_i$$

Linear Model Assumptions

Linear regression model:

$$Y_i = b_0 + \sum_{j=1}^{p} b_j X_{i,j} + \varepsilon_i$$

Key Assumptions:

- **Linearity**: The relationship between Y and $\mathbf{X}$ is additive:

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + \sum_j b_j x_j$$

- **Independent Errors**: Errors across observations are uncorrelated:

$$\varepsilon_i \perp \varepsilon_k \quad \text{for } i \neq k$$

- **Error Independence from Covariates**: The error ε_i has mean zero and is independent of $\mathbf{X}_i$:

$$E(\varepsilon_i \mid \mathbf{X}_i) = 0$$

Linear Model Assumptions

Linear regression model:

$$Y_i = b_0 + \sum_{j=1}^{p} b_j X_{i,j} + \varepsilon_i$$

Key Assumptions:

- **Linearity**: The relationship between Y and $\mathbf{X}$ is additive:

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + \sum_j b_j x_j$$

- **Independent Errors**: Errors across observations are uncorrelated:

$$\varepsilon_i \perp \varepsilon_k \quad \text{for } i \neq k$$

- **Error Independence from Covariates**: The error ε_i has mean zero and is independent of $\mathbf{X}_i$:

$$E(\varepsilon_i \mid \mathbf{X}_i) = 0$$

Less critical assumption:

- **Normality (optional**: Sometimes, we assume $\varepsilon_i \sim N(0, \sigma^2)$ for inference (e.g., hypothesis testing, confidence intervals).

Model Assumptions: Linearity

Conditional Expectation:

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

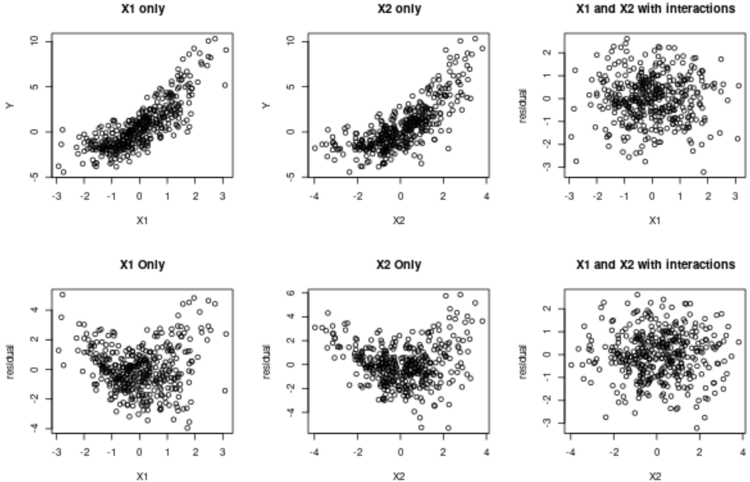
Key Points:

- A model may be **nonlinear in a single covariate** but still linear in multiple covariates.
- Examples:
  - **Simple linear regression** may not capture all relationships.
  - **Polynomial regression** (e.g., X^2 term) or **interaction terms** ($X_1 X_2$) can improve fit while still being **linear in parameters**.

Checking Linearity: Residuals

How to check for violations?

- **Plot residuals** vs. fitted values.
- **Look for patterns** (e.g., curves suggest nonlinearity).



Model Assumptions: Independent Errors

Definition: Errors across observations should be uncorrelated:

$$\varepsilon_i \perp \varepsilon_k \quad \text{for } i \neq k$$

Why is this important?

- **Unbiased estimates**: Coefficient estimates remain valid.
- **Invalid inference**: Standard errors and p-values may be incorrect.

Model Assumptions: Independent Errors

Definition: Errors across observations should be uncorrelated:

$$\varepsilon_i \perp \varepsilon_k \quad \text{for } i \neq k$$

Why is this important?

- **Unbiased estimates**: Coefficient estimates remain valid.
- **Invalid inference**: Standard errors and p-values may be incorrect.

Example: Time-Series Data

- If errors are correlated across time (e.g., stock prices), then:

$$E(\varepsilon_t \mid \varepsilon_{t-1}) \neq 0$$

- Solutions:
  - Include lagged variables (e.g., AR models).
  - Use robust inference.

Model Assumptions: Constant Error Variance (Homoscedasticity)

Definition: The error variance should be constant across observations:

$$\text{Var}(\varepsilon_i) = \sigma^2$$

Violations: Heteroscedasticity

- Occurs when **variance changes** with X.
- Common in income models: Variability in wages increases with experience.

Model Assumptions: Constant Error Variance (Homoscedasticity)

Definition: The error variance should be constant across observations:

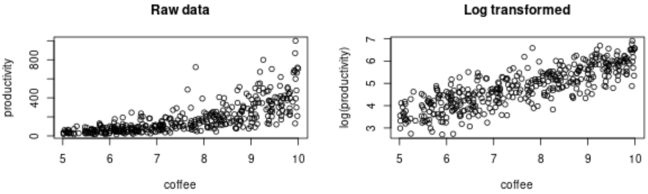
$$\text{Var}(\varepsilon_i) = \sigma^2$$

Violations: Heteroscedasticity

- Occurs when **variance changes** with X.
- Common in income models: Variability in wages increases with experience.

Detection:

- **Residual plot**: Plot residuals vs. fitted values.
- **Breusch-Pagan test**



Raw data   Log transformed

Assessing Explanatory Power: R^2

Decomposing Variance:

$$SS_{\text{Total}} = SS_{\text{Regression}} + SS_{\text{Error}}$$

Definition of R^2:

$$R^2 = 1 - \frac{SS_{\text{error}}}{SS_{\text{total}}} = \frac{SS_{\text{regression}}}{SS_{\text{total}}}$$

Assessing Explanatory Power: R^2

Decomposing Variance:

$$SS_{\text{Total}} = SS_{\text{Regression}} + SS_{\text{Error}}$$

Definition of R^2:

$$R^2 = 1 - \frac{SS_{\text{error}}}{SS_{\text{total}}} = \frac{SS_{\text{regression}}}{SS_{\text{total}}}$$

Key Interpretations:

- R^2 measures goodness-of-fit, not causality.
- A high R^2 does not mean a model is correct.
- Adjusted R^2 accounts for multiple predictors.

Wrap-up: Key Takeaways

- **Transformations**: Log transformations change interpretation.
- **Multiple covariates**: Improve flexibility and model fit.
- **Assumptions**: Linearity, independence, and homoscedasticity are critical.