July 21 - 24, 2021 | Garki Hospital Abuja | Resource Person

Research Methodology Boot Camp

with Epi Info Training

Dr. Adamu Onu
MBBS, FWACP (FM)
MS Epidemiology & Biostatistics
PhD Public Health (Epidemiology)

Target Audience Clinical Researchers, Post-Part 1 Residents, and Others

Important Information

- Limited slots are available on a first come, first served basis
- Laptop running Windows 10 required
- Organized as morning lecture sessions and afternoon hands on coaching sessions

Highlights

- Research Methodology
- Research Design
- Data Management
- Sample Size Calculations
- Test Statistics
- Interpretation of Results
- Report Writing
- Hands-on training sessions
- Statistical consulting sessions

For further details contact
Email: epimetrix@gmail.com
Phone: +234 803 474 9930

epimetrix
health systems
data for health decisions

Data Transformation

Purpose of transformations

- To linearize regression model.
- To stabilize variance (reduce heterogeneity of variance, “heteroscedasticity”).
- To normalize variables.
- Some transformations will serve more than one purpose.
    - For example, a transformation that linearizes a variable also normalizes it.

Reasons for data transformation

- Theoretical considerations.
- The dependent variable may have a probability distribution in which the mean is related to the variance.
- Empirical evidence from examination of the residuals.

Variables to be transformed

- The dependent variable can be transformed.
  - Note: This effects the relationship of the dependent variable with all of the predictor variables in the model.
- Individual predictor variables can be transformed.
- Both dependent and independent variables can be transformed

Major Drawbacks

- Interpretation of the regression involves transformed variables and not the original variables themselves.
- Relationship of the transformed variables to the original variables may be difficult or confusing.
- Transformation may not be able to rectify the problems in the original data and thus the regression analysis may still be suspect.

Data transformations

Log transformation	$y_i' = \ln(y_i), \quad y_i > 0$
Square root transformation	$y_i' = \sqrt{y_i}, \quad y_i \geq 0$
Square transformation	$y_i' = y_i^2$
Arcsin-Root transformation	$y_i' = \frac{1}{\sin}\sqrt{y_i}$
Poisson distribution	$y_i' = \sqrt{y_i}$, or $y_i' = \sqrt{y_i + 0.5}$, or $y_i' = \sqrt{y_i} + \sqrt{y_i + 0.5}$
Binomial distribution	$y_i' = \frac{1}{\sin}\sqrt{y_i}$
Negative binomial distribution	$y_i' = \frac{1}{\lambda} \cdot \frac{1}{\sin}(\lambda\sqrt{y_i})$, or $y_i' = \frac{1}{\lambda} \cdot \frac{1}{\sin}(\lambda\sqrt{y_i} + 0.5)$

Log Transformation

- To linearize regression model with consistently increasing slope (curvilinear upward).
- Stabilize variance when variance of residuals increases markedly with increasing *y*.
- To normalize *y* when distribution of residuals is positively skewed.

Square root transformation

- Used to stabilize variance when proportional to the mean of Y; especially when Y approximates a Poisson distribution.

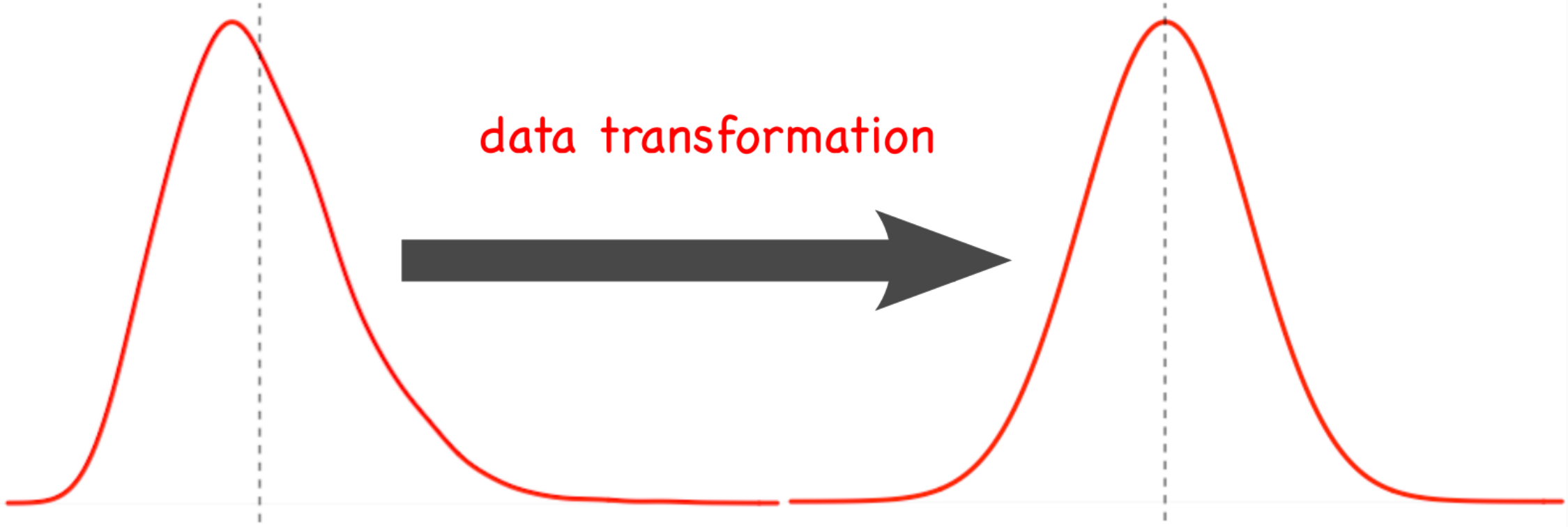
Square transformation

- Linearize when X vs Y is curvilinear downward, i.e., slope decreases as X increases
- Stabilize variance when it decreases with the mean of Y.
- Normalize Y when distribution of residuals is negatively skewed.

Arcsin-Root transformation

- Stabilize variance when *y* is a proportion or a rate

What to do if you can't figure out which transformation to use?



- Ladder of powers
- Box-Cox transformation

Ladder of powers

- The algorithm does each of the following transformations and tests for normality:
  - $y_i^3, y_i^2, y_i, \sqrt{y_i}, \ln(y_i), \frac{1}{\sqrt{y_i}}, \frac{1}{y_i}, \frac{1}{y_i^2}, \frac{1}{y_i^3}$
- In Stata the command is:

```
ladder y
gladder y
```

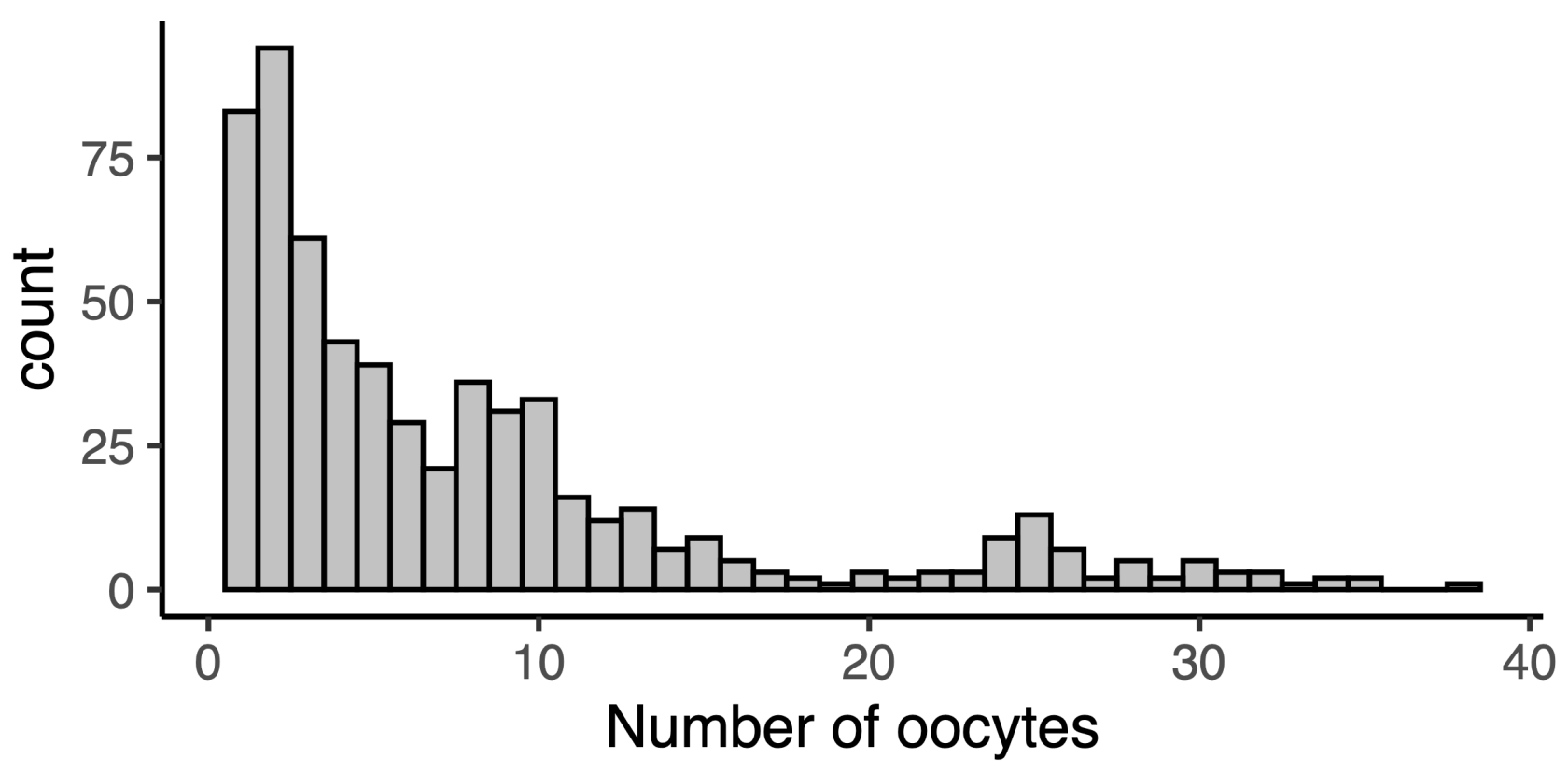
Box-Cox transformation

- Finds the value u for the transformation, $\frac{y_i^u - 1}{u}$, which normalizes the transformed variable.
- The values being transformed must be strictly positive, that is, greater than zero.
- In Stata the command is:

```
boxcox y, generate(newy)
```

Example

> Consider a regression model to determine the number of oocytes following a cycle of IVF treatment



It is apparent from the histogram that number of oocytes is right skewed and not normally distributed.

extract from the actual statistical report!

distributed. This is a deviation from one of the assumptions of linear regression models which requires the dependent variable to be normally distributed. Since the outcome variable is not normally distributed what do we then do? There are a number of approaches. We can try the following:

- Linear regression modeling with log-transformed outcome, or
- Generalized linear modeling with a Gamma distribution since the outcome variable is a right skewed non-negative distribution.

For this statistical analyses both linear regression models with log-transformed dependent variable and generalized linear modeling with a Gamma distribution were examined. However only the linear regression models are reported as they provided they better fit to the data.

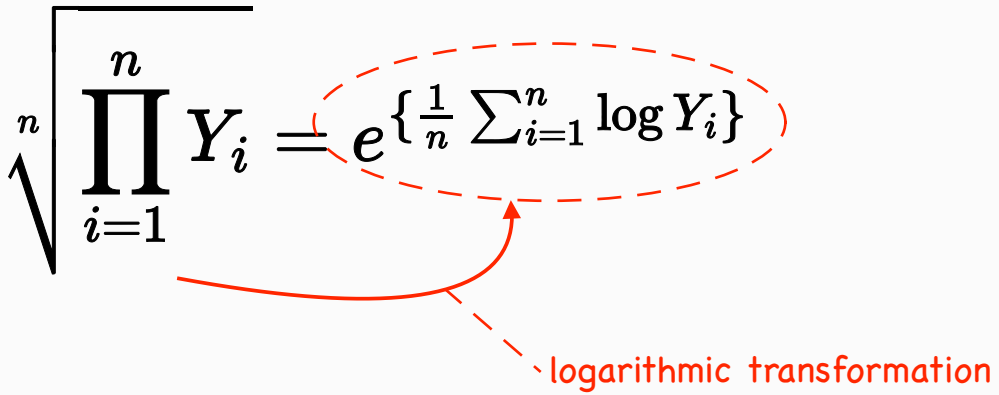
Log transformation of the outcome variable

- We can log-transform the outcome variable.
- This means that we are looking at a **multiplicative geometric mean** model.
- To refresh our memory the arithmetic mean is:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i$$

Geometric mean

- While the geometric mean is:



Geometric mean

- The linear model with the log-transformed outcome is a multiplicative geometric mean model
- Starting from the following:

$$E[\log(Y_i)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

- Thus:

$$\begin{aligned} Y_i &= e^{\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n\}} \\ &= e^{\beta_0} \times e^{\beta_1 X_1} \times e^{\beta_2 X_2} \times \cdots e^{\beta_n X_n} \end{aligned}$$

- This model is interpretable in terms of the change in the geometric mean – it assumes multiplicative effects on the original outcome by the predictors.

Linear regression with log-transformed outcome

	B	SE	Sig.
Intercept	2.369	0.071	***
Endometriosis	–0.234	0.073	**
Short protocol	–0.672	0.078	***
Cetrotide	0.054	0.119	
Zoladex	–0.888	0.080	***
N	605		
R^2	0.244		
R^2	0.239		

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

From the table above we see that the model can be written as

log(Number of oocytes) =

$$\beta_0 + \beta_1 \cdot \text{Endometriosis} + \beta_2 \cdot \text{Short Protocol} + \beta_3 \cdot \text{Cetrotide} + \beta_4 \cdot \text{Zoladex}$$

Which is:

$$2.37 - 0.23 \cdot \text{Endometriosis} - 0.67 \cdot \text{Short Protocol} + 0.05 \cdot \text{Cetrotide} - 0.888 \cdot \text{Zoladex}$$

Linear regression with log-transformed outcome

Interpretation

1. Exponentiate the coefficient
2. Subtract one from this number
3. Multiply by 100 to give the percent increase (or decrease) in the dependent variable for every one-unit increase in the independent variable:

- Thus for treatment protocol: $(e^{-0.672} - 1) \times 100 = -48.9$, i.e. a 48.9% decrease in number of oocytes compared to long protocol.