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

Logistic Regression

Objectives

1. When to use logistic regression
2. Properties of logistic regression
3. Interpreting logistic regression results
4. Model building
5. Dummy variables
6. Interaction of variables

Logistic regression

Logistic regression is one of the most commonly used statistical methods in medical research

Key reasons to use logistic regression

- To identify variables that are significant predictors of outcome independent of the effects of other variables
- To determine if a specific variable is related to outcome while controlling for the effect of confounding variables

Regression techniques

- Predict the value of one variable based upon the value of other variables
- Develop an equation which will predict a dependent variable (Y) given a value for an independent variable (X)
- Multiple independent variables X_1, X_2, X_3, … X_n may affect an outcome Y

Outcome variable

- Dependent variable Y (outcome)
- If outcome variable is continuous, use multiple linear regression: $\hat{Y} = \alpha + b_1X_1 + b_2X_2 + b_3X_3 + \ldots b_nX_n$
- If outcome variable is dichotomous (binary), multiple linear regression will not work. Need model that will permit only two values of Y.

Examples of binary outcome

- Mortality
- Disease outcome
- Event occurrence

> For logistic regression, a dichotomous variable is coded as a 0 or 1.

> A value of 1 indicates the presence of disease or an event (1 = yes, 0 = no).

Use of proportions

- Proportion is the descriptive statistic used for dichotomous variables.
- Proportion represents the mean of 0s and 1s for the sample.
- Proportion also represents the probability of drawing a subject with a 1 from the sample.
- P = probability of an outcome = proportion of 1s
- Permitted values of P range from 0 to 1
- In regression, P represents the probability of an outcome based on values of the independent variables (X_1, X_2, X_3, … X_n)

Prediction of male sex based on height

- Prediction of male sex based on height
- y-axis is *p* = proportion of 1s (male) at any given height
- Regression line is non- linear: none of the fall on the regression line.
- They all fall on 0 or 1 (male = 1, female = 0)

Problems with linear regression

Inadmissible values

- If you use linear regression, the predicted values will become greater than one and less than zero if you move far enough on the x-axis.
- Such values are theoretically inadmissible.

Problems with linear regression

Lack of constant variance

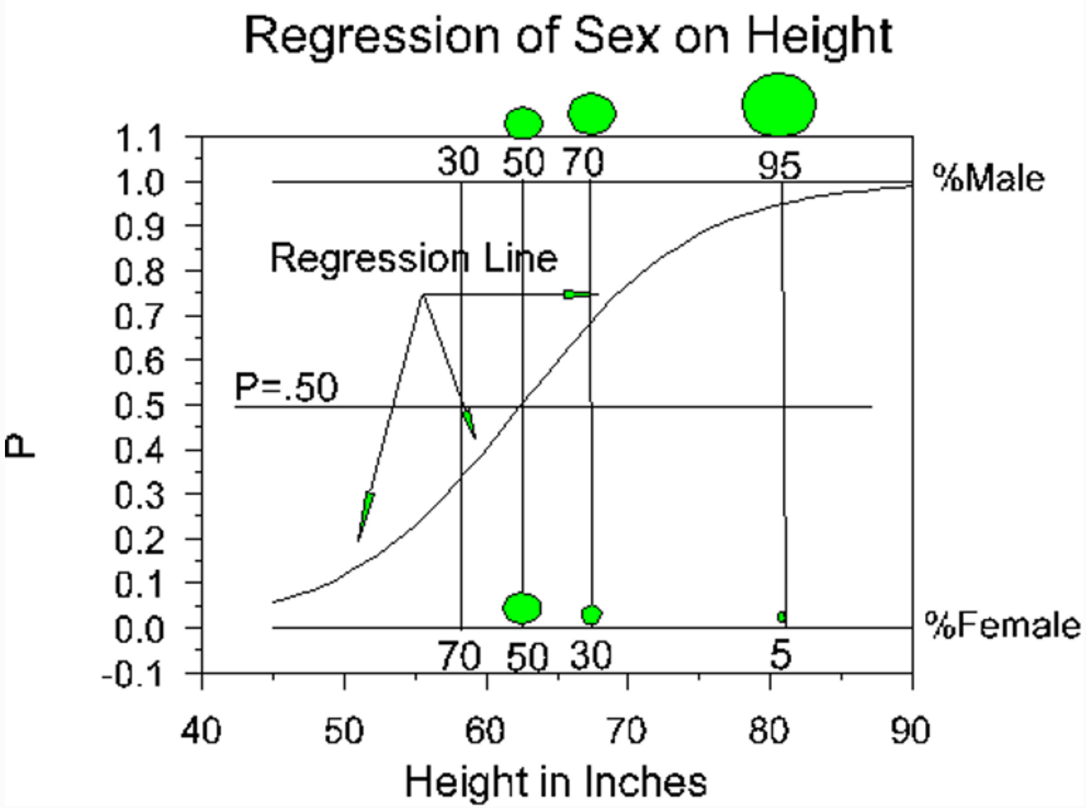
- One assumption of regression is that the variance of y is constant across values of x (homoscedasticity).
- This cannot be the case with a binary variable, because the variance is $P(1 - P)$.
- When 50 percent of the people are 1s, then the variance is .25 — its maximum value.
- At more extreme values, the variance decreases.
- When $P = .10$, the variance is $.1 \times .9 = .09$, so as $P \rightarrow 1$ or 0, the variance $\rightarrow 0$

Problems with linear regression

Lack of normal distribution of predicted values

- Significance testing of the regression coefficients (*b*) rests upon the assumption that errors of prediction ($Y - Y'$) are normally distributed.
- Because *Y* only takes the values 0 and 1, this assumption is pretty hard to justify

Equation



- Equation of regression line is nonlinear, using the base of the natural logarithm *e*
- Since the relation between X and P is nonlinear, *b* does not have a straightforward interpretation as it does in ordinary linear regression

$$P = \frac{1}{1 + e^{-a+bX}}$$

Odds

$$\text{Odds} = \frac{\text{proportion with outcome}}{\text{proportion without outcome}} = \frac{\text{P}}{\text{1 - P}}$$

- P = proportion of 1s, 1 – P = proportion of 0s

Example:

- Probability of male at given height is 0.9
- Odds of being male = 0.9/0.1 = 9 to 1 = 9
- Odds of being female = 0.1/0.9 = 1 to 9 = 0.11
- Asymmetry (9 vs 0.11) is difficult to interpret, because the odds of being male should be the opposite of odds of being female

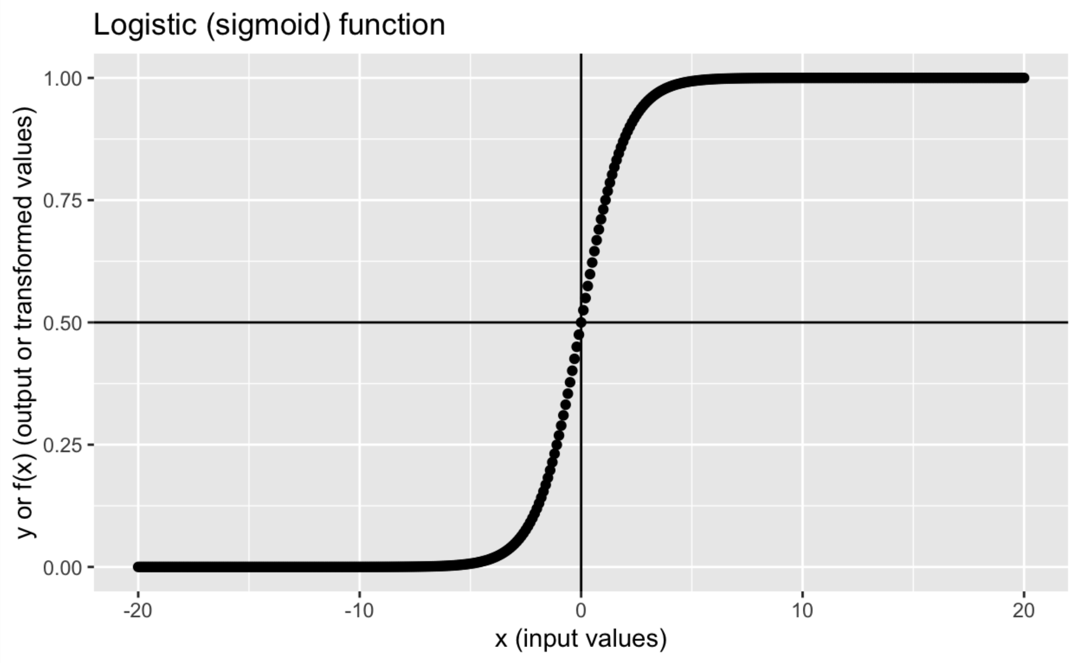
Logarithm of odds

- Asymmetry can be corrected with natural logarithm
- Natural log of 9 is 2.217: $\ln(.9/.1) = 2.217$.
- The natural log of 1/9 is -2.217: $\ln(.1/.9) = -2.217$
- Log odds of being male is exactly opposite to the log odds of being female
- Logarithm of odds is called the Logit

$$\ln\left(\frac{P}{1-P}\right)$$

Logit properties

- Logit = 0 when odds = 1
- When odds < 1 logit is negative
- When odds > 1 logit is positive



Linear property of the logit

- Logit is a linear function of X
- Equation can be rearranged with *P* as the dependent variable
- Logit allows S-shaped curve to be replaced with linear function for binary dependent variable

$$\mathrm{Logit(p)} = \ln\left(\frac{P}{1-P}\right) = a + bX$$

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

Multiple logistic regression

- Independent variables (X_1, X_2, X_3, ... X_n) can be continuous, categorical, or binary
- Dependent variable must be binary.
- A continuous variable can be used if divided into 2 categories.

$$\ln\left(\frac{P}{1-P}\right) = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_nX_n$$

Predictors of lead toxicity

Variable	Odds Ratio	95% CI	Coefficient (b)	SE	P value
Age	0.96	0.75–1.21	−0.045	0.12	0.71
Eye Pencil	6.00	1.5–24	1.800	0.70	0.01
HAZ	1.08	0.77–1.5	0.079	0.16	0.64
Sex	1.13	0.45–2.8	0.120	0.47	0.79
Constant			−0.39	1.0	0.83

Interpreting coefficients

- If b is the logistic regression coefficient for the variable Age, then e^b is the odds ratio corresponding to a one unit change in age.
- For binary variables e^b is the odds ratio of the characteristic assigned value of 1 compared with that assigned value of 0.

Model building

- Refers to deciding on the variables that provide the best prediction of outcome
- Methods include forward selection and backward elimination
- Put all of the variables you would like to explore in the model.
- Limit the number of variables in model to the number of subjects in your sample divided by 10.

Model building

- Variables that you would like to control for as confounding variables in a specific analysis should be retained in the model.
- This allows you to obtain adjusted odds ratios, because they are adjusted for specific confounders.
- Remove the variable with the least significant *p* value.
- Continue to remove variables until all variables in the model are significant (*backwards elimination*).
- Note the changes in Final –2LogLikelihood

–2 Log likelihood

- The statistic –2LogLikelihood is a “badness-of-fit” indicator.
- A large number means poor fit of the model to the data.
- The difference between values of –2LogLikelihoo of two models is known as the likelihood ratio.
- If the –2LogLikelihood does not change much upon removing a variable, that variable adds little to the model.
- Generally, if the difference in –2LogLikelihood between 2 models that differ by one variable is less than 3.84 (the χ^2 value corresponding to p = 0.05), then the difference in the 2 models is not significant.

Collinearity

- If the independent variables are highly correlated with each other, they are said to be collinear (e.g. height and weight).
- The regression coefficients may become inflated, so the observed value may be far from the true value.
- Choose one of the variables to include in the predictive model – typically the one with the greatest predictive value.

Dummy variables

- Categorical variables with more than 2 values must be recoded with dummy variables.
- One cannot simply code them as 1, 2, and 3, because they will be treated as numeric values.
- It will not make sense that the second and third categories are equal to 2 and 3 times the value of the first category.
- A categorical variable with k values must be coded in $k-1$ dichotomous dummy variables that each have two values: 0 value indicating no or least exposure. The "0" value is the reference value.

Dummy variables: example

- You would like to include location of residence as a variable in your model
- Your study has included subjects from 3 different locations: urban, semi-urban, and rural.
- Dichotomous dummy variables for Location (semi-urban & urban) will need to replace the original coding.

	Dummy Variables	
Location	Semi-urban	Urban
Rural	0	0
Semi-urban	1	0
Urban	0	1

Dummy variables: example

- Two dichotomous dummy variables are enough to locate the three initial values of the Location variable.
- In the logistic regression model, only the variables Urban and Semi-urban will appear, and the risk computed for each will be in reference to Rural.

	Dummy Variables	
Location	Semi-urban	Urban
Rural	0	0
Semi-urban	1	0
Urban	0	1

Interaction

- Interaction means that the odds ratio for a variable varies with the value of another variable.
- For example, if the outcome is renal failure, the effect of hypertension differs greatly between blacks and whites.
- Renal failure = race + hypertension does not tell the whole story, and another term, called an **interaction** term is needed
- Renal failure = race + hypertension + race*hypertension

Interaction

- Interaction must be addressed early in forming a model, because the model must contain all single variables as interaction terms to be "hierarchically well structured."
- All pertinent interaction terms be evaluated for significance before eliminating any individual variables.