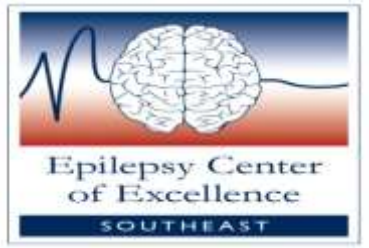
www.epilepsy.va.gov/Statistics

Statistics in Evidence Based Medicine (2014)

Lecture 3: Introduction to Logistic Regression

Rizwana Rehman, PhD
Regional Statistician
Southeast Epilepsy Center of Excellence
Durham VA Medical Center, Durham NC
Rizwana.Rehman@va.gov
(919)286-0411 ext: 5024



Course Outline

Understanding logistic regression in five lectures

Difference between relative risk and odds ratio ✓,
marginal and conditional odds ratios, ✓
terminology and interpretation of logistic regression

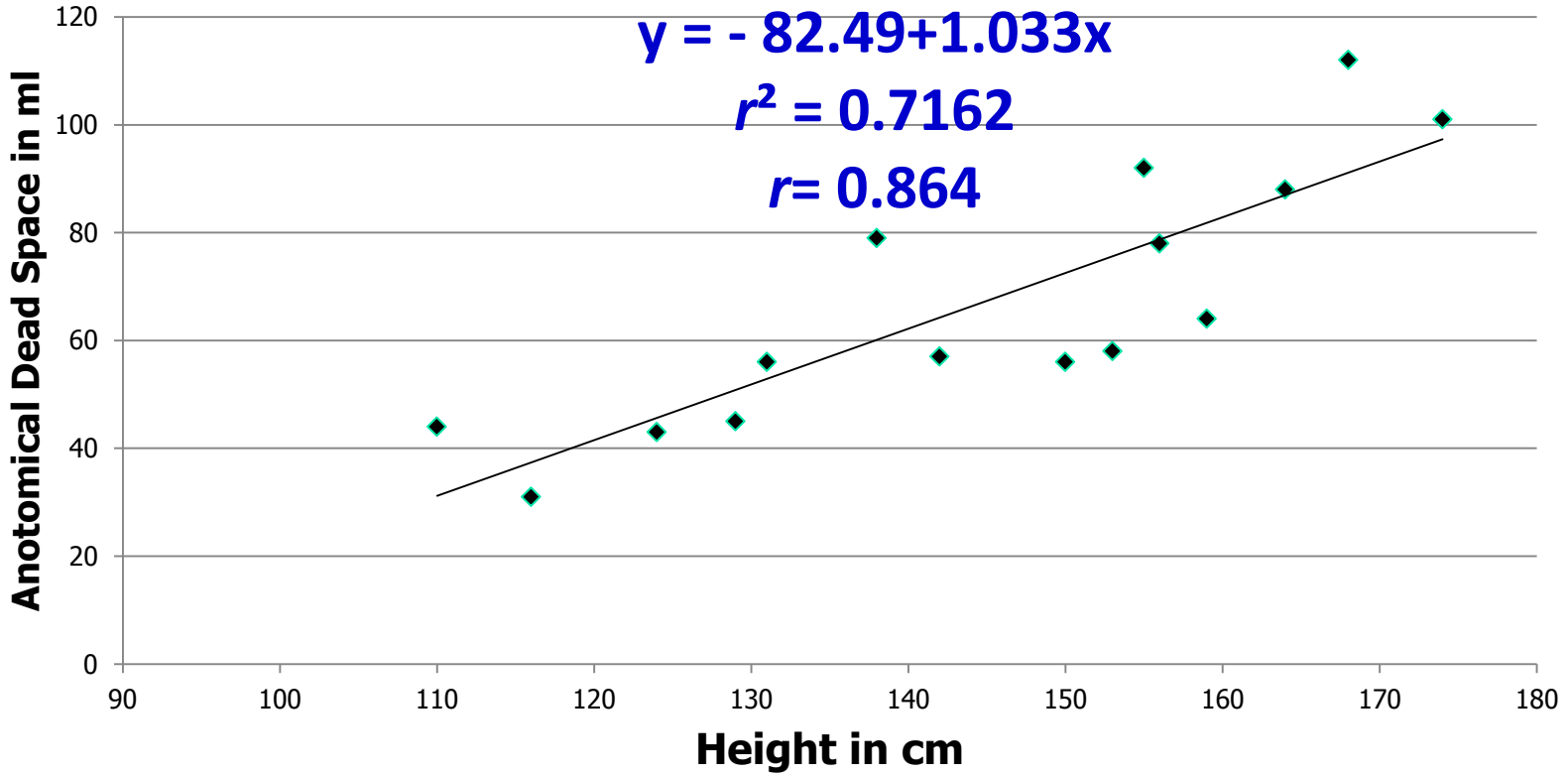
Suggested Book: Logistic Regression A Self-Learning Text by Kleinbaum & Klein

Third Edition Springer

Today's Lecture

- Comparison of linear and logistic regressions
- Logistic regression model
- Interpretation of coefficients
  - Examples
- Summary

Recap: Linear Regression for Continuous Outcome



Restriction: Assumptions

Binary Outcomes

A binary data takes only one of two values

Examples:

Alive or dead, Sick or Well, Exposed or Unexposed *etc*

We can find proportions for binary outcomes

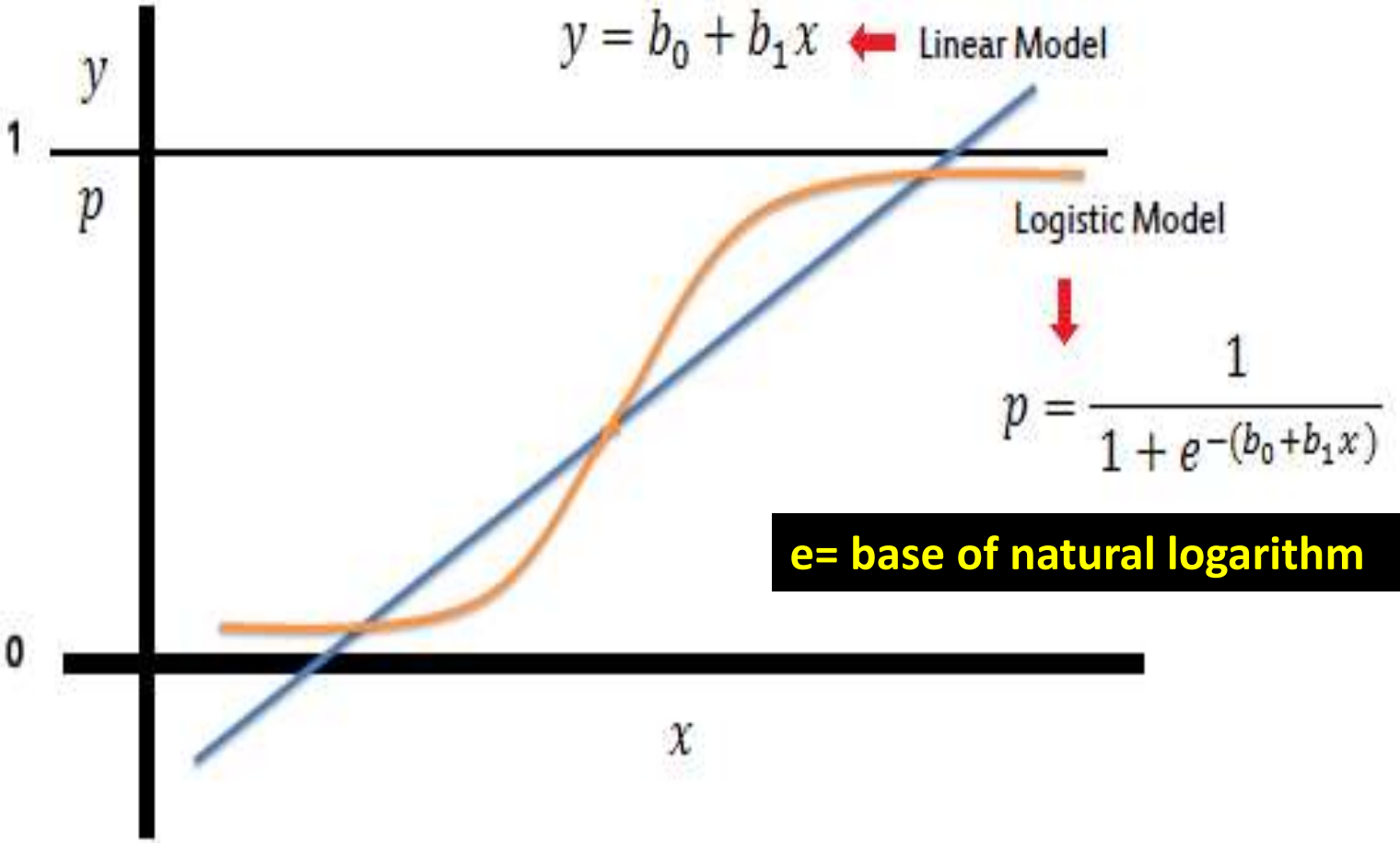
Why Linear Regression is not Suitable for a Binary Outcome?

- A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
- Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line (a violation of linear regression assumptions)

Properties of Logistic Regression

- Logistic regression is used for a dichotomous (binary) outcome variable
- Logistic regression produces a logistic curve whose values are between 0 and 1
- No assumptions required for normality of predictors or variance of predictors
- Can handle any number of categorical or continuous independent variables

Shape of a Logistic Curve



http://www.saedsayad.com/logistic_regression.htm

Relationship between Odds and Probability

- To calculate the odds (o) from Probability (p)

$$\text{Odds}=\frac{p}{1-p}$$

- To calculate the probability from Odds

$$\text{Probability}=\frac{o}{1+o}$$

Moving from Logistic Curve to Odds

- $p = \frac{1}{1+e^{-(b_0+b_1x)}}$

Estimated probability of response

- $\boxed{\frac{p}{1-p}} = \exp(b_0+b_1x) = e^{b_0+b_1x}$

Odds

- $\log_e(\frac{p}{1-p}) = b_0+b_1x$

Natural log of odds

Comparison between Linear Regression and Logistic Regression

- Linear Regression: Ordinary least square method is used to compute coefficients of the best fit line; Logistic Regression: Maximum likelihood estimation of model coefficients
- Linear Regression: r^2 is an indicator of the goodness of fit of model, Logistic Regression: a pseudo r^2 can be computed for model adequacy but not recommended

Interpretation of Intercept b_0

logit of estimated probability

$$\log_e\left(\frac{p}{1-p}\right)=b_0+b_1X$$

For a case control study ignore b_0

- For a prospective cohort study b_0 is the log of the background, or baseline, odds.
- By background odds we mean the odds that would result for a logistic model without X.

Interpretation of b_1 for a Dichotomous X

$$\log_e\left(\frac{p}{1-p}\right)=b_0+b_1X$$

	Case	Control	Total
Exposed	a	b	a+b
Not exposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

Odds of disease with risk factor present=$e^{b_0+b_1}$

Odds of disease with risk factor absent=e^{b_0}

$$\text{Odds Ratio} = \frac{e^{b_0+b_1}}{e^{b_0}} = e^{b_1}$$

$$\text{Log}_e(\text{Odds Ratio}) = b_1$$

Interpretation of b_1 for a Dichotomous X

$\log_e(\text{Odds Ratio}) = b_1$

The estimated regression coefficient b_1 is the natural log of the odds ratio associated with presence of risk.

$$\text{Odds Ratio} = e^{b_1}$$

Example: Smoking and Lung Cancer

Male Lung Cancer & Smoking (Doll and Hill 1950)

	Lung cancer (Case)	Control
Smokers	647	622
Non-smokers	2	27

$$\text{Odds Ratio} = \frac{647 \times 27}{2 \times 622} = 14.04$$

The odds of lung cancer in smokers were 14 times the odds of lung cancer in non-smokers

Logistic Regression for Association between Lung Cancer and Smoking

$$\log_e\left(\frac{p}{1-p}\right) = -2.6025 + 2.6419 \times \text{Smoking}$$

-2.6025 is the log odds of developing lung cancer

2.6419 is the increment to the log odds for smokers

Moving from non smokers to smokers increases the log odds (logit) of lung cancer by 2.6419

$\log_e$(odds ratio)=2.6419

Estimated odds ratio for smokers vs. non smokers= $e^{2.6419}$=14.04

Interpretation of Regression Coefficient for a Continuous Predictor

- $\log_e(\frac{p}{1-p})=b_0+b_1X$

b_1 represents the change in log odds that would result from a one unit change in the variable X

Example: Kyphosis and Age

Hastie and Tibshirani (1990)

Purpose: Determine age (in months) as a risk factor for Kyphosis

18 subjects with Kyphosis; 12, 15, 42, 52, 59, 73, 82, 91, 96, 105, 114, 120, 121, 128, 130, 139, 139, 157

22 subjects without Kyphosis; 1, 1, 2, 8, 11, 18, 22, 31, 37, 61, 72, 81, 97, 112, 118, 127, 131, 140, 151, 159, 177, 206

Logistic Regression for Kyphosis

$$\log_e\left(\frac{p}{1-p}\right) = -0.5727 + 0.00430 \times \text{Age}$$

$\log_e$(odds ratio)=0.0043

As the age increases by one month the expected change in log odds is 0.00430.

Estimated Odds Ratio= $e^{0.0043}$=1.004

Interpretation of Odds Ratio Like Relative Risk:

As the age increases by one month the odds of Kyphosis increase by .04%. Age is not associated with Kyphosis.

What if Age Changed by c Months?

Consider the risk associated with a c months increase

Odds Ratio = e^{cb_1}

Suppose c=6 months

Estimated Odds Ratio = $e^{6\times 0.0043}$= 1.026

The risk for Kyphosis increases by 2.6% for each 10 months increase in age

Recap: Confounding

- Means mixing
- Distortion (an error) of an association between two variables (exposure and outcome) due to a third factor
- May cause an overestimate of strength of relationship or vise versa
- May be responsible for all or partial relationship
- Can be controlled in study design and in analysis

Recap: Marginal and Conditional Odds Ratios

Given three variables A, B, C

- Marginal odds ratio between A and B is obtained by ignoring the variable C
- Conditional odds ratios are obtained by computing odds ratios between A and B at different levels of C

Recap: Confounding and Conditional Odds Ratios

- When marginal odds ratio is different from conditional odds ratios and conditional odds ratios are similar then confounding (a third factor participating in the association between a risk factor and disease) may be present.
- For a factor to qualify for a confounder the percent difference of marginal odds ratio from conditional odds ratio should be more than 10%.
- We can compute an adjusted odds ratio which takes confounding into consideration.

Adding a Confounder in the Logistic Regression

X_1 and X_2 are both dichotomous with 0, 1 coding

$$\log_e\left(\frac{P}{1-P}\right)=b_0+b_1X_1+b_2X_2$$

For a fixed level of X_2, when we increase X_1 by one unit, the log odds of outcome change by b_1.

For a fixed level of X_1, when we increase X_2 by one unit, the log odds of outcome change by b_2.

Example: Age as Confounding Factor

Age < 50, Age=C = 0

	D=Heart Disease=1	D=No Heart Disease=0
E=Inactive=1	10	90
E=Active=0	35	465

$OR_{ED/C=0} = 1.48$

Age > 50, Age C = 1

	Heart Disease=1	Heart Disease=0
E=Inactive=1	36	164
E=Active=0	25	175

$OR_{ED/C=1} = 1.54$

Common OR=1.52

Logistic Regression for Confounding

$$\log_e\left(\frac{p}{1-p}\right) = -2.592+0.415\times\text{Inactive}+0.6547\times\text{Age}$$

Odds ratio for physical inactivity adjusted for age $=e^{0.415}=1.52$ (Age adjusted odds ratio)

For a fixed age group the odds of CHD among inactive people were 1.52 times the odds of CHD among active people.

Odds ratio for age adjusted for physical inactivity $=e^{0.6547}=1.93$ (Physical activity status adjusted odds ratio)

Summary

- Logistic regression is used for a binary response variable
- Coefficients of logistic regression provide estimates of odds ratios. These odds ratios are adjusted when there are more than one predictors.
- For given values of X variables we can also compute probability of the response variable Y.

Uses of Logistic Regression in Study Designs

- **Logistic modeling is used for follow up, case control and cross sectional studies**
  - For a follow up study with rare disease assumption odds ratio will estimate relative risk
  - When rare disease assumption does not hold, alternative methods to compute adjusted relative risk from logistic modeling have been proposed
- Used to adjust the effects of confounders
- Used to study the simultaneous effect of a number of categorical variables on the outcome
- To predict a value of an outcome given inputs.

www.epilepsy.va.gov/Statistics

Questions/Comments

Rizwana.Rehman@va.gov

(919) 286-0411 ext: 5024

Thank you for being patient !