Linear and Logistic Regression

Zuber D. Mulla

From Hennekens CH, Buring JE. *Epidemiology in Medicine*. Boston: Little, Brown and Company, 1987.

- **Stratification as a technique to control for confounding has several advantages**

- **Easy, straightforward**
- **Allows for the evaluation of interaction**

Problem with Stratified Analysis

- Inability to control simultaneously for a moderate or large number of potential confounders

- For example, assume you have a binary exposure (sex) and a binary outcome (MI vs. no MI) and you have a small sample size and two possible confounders: Diabetic (2 strata) and Age (4 strata) → 8 strata

Epidemiology in Medicine, p. 315

"Multivariate analysis allows for the efficient estimation of measures of association while controlling for a number of confounding factors simultaneously, even in situations where stratification would fail because of insufficient numbers."

X variable

- Independent variable
- Predictor variable
- Covariate
- Explanatory variable
- Exposure variable

Y variable

- Dependent variable
- Outcome variable
- Response variable

Regression

- Why? Quantify and/or predict
- Linear, Logistic, Cox, Poisson, etc.
- Simple: One independent var.
- Multiple: > 1 independent vars.

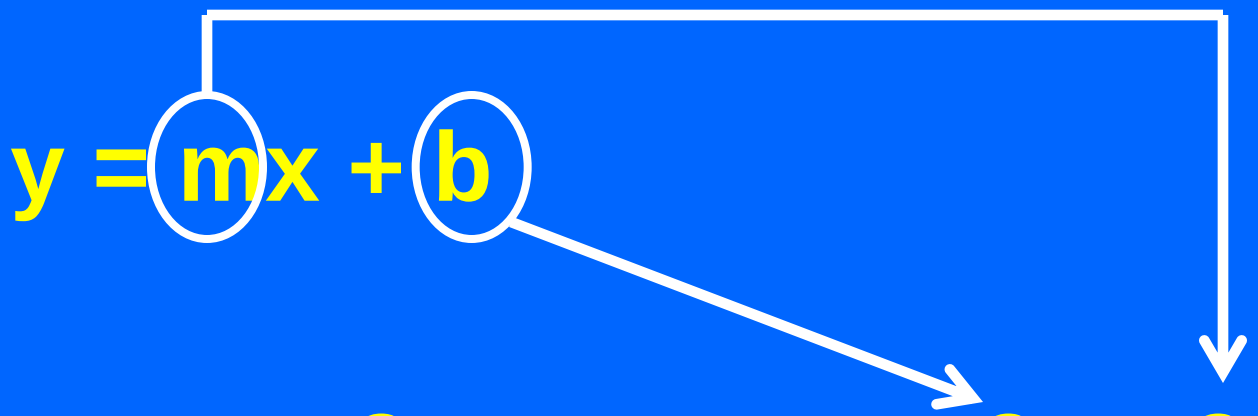
Linear Regression

y is a continuous variable

$y = mx + b$

$y = \alpha + \beta x$ or $y = \beta_0 + \beta_1 x_1$

Linear Regression

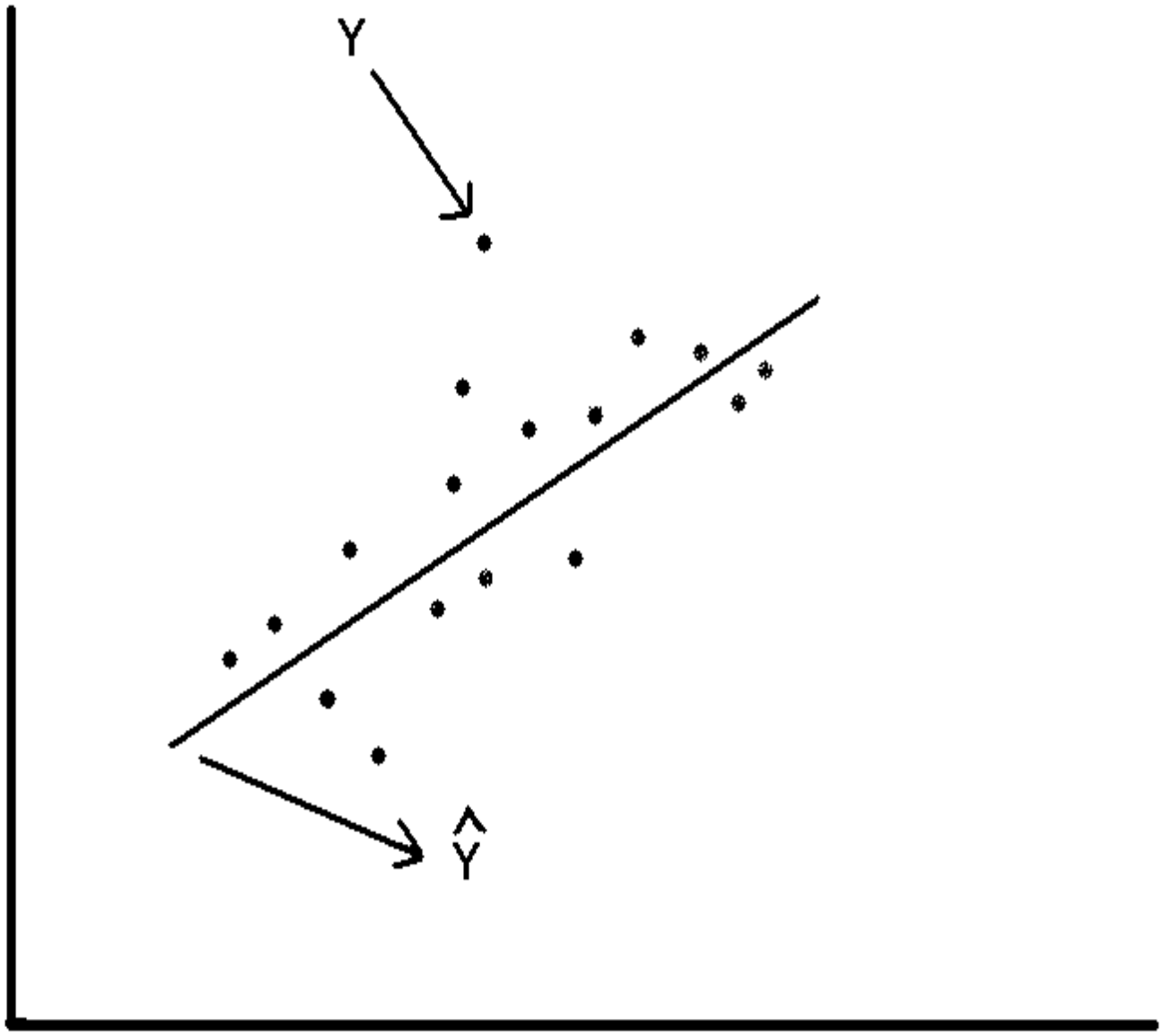


$y = \alpha + \beta x$ or $y = \beta_0 + \beta_1 x_1$

Beta, beta hat

$$\hat{y} = ß_0 + ßx$$

The hat over the *y* signifies that it is a value of *y* calculated from the linear equation rather than an observed value of the dependent variable.


Y
$\hat{Y}$

Assumptions needed for inference in linear regression analysis

Remember L.I.N.E.:

- **Linearity: the means of the subpopulations of *Y* all lie on the same straight line**
- **Independent**
- **Normality (outcome is bell-shaped)**
- **Equal variances**

Adapted from *Biostatistics: A Foundation for Analysis in the Health Sciences Fifth Edition* by Daniel, p. 368

Slope in red, *Y* intercept in green

$y = mx + b$

$y = \alpha + \beta x$

$y = \beta_0 + \beta_1 x_1$

Interpreting the Parameter Estimates from a Linear Regression Model

$$\hat{y} = ß_0 + ßx$$

$$\hat{y} = 2 + 5x$$

The slope (beta) is 5. The slope is the amount of change in *y* for each one unit increase in *x*.

Interpreting the Parameter Estimates from a Linear Regression Model

$$\hat{y} = 2 + 5x$$

Assume *x* is age in years and *y* is systolic blood pressure (SBP) measured in millimeters of mercury. For each one-year increase in age, there is a 5 mmHg increase in SBP.

Simple vs. Multiple Linear Regression

SIMPLE → $y = \beta_0 + \beta_1 x_1$

MULTIPLE → $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

More Review of Linear Regression

- What's the most common method for estimating parameters in linear regression?

- Least squares

Least Squares Method

- We choose values of β_0 and β_1 which minimize the sum of the squared deviations of the observed values of Y from the predicted values based upon the model

Another Well-known Algorithm for Estimating Parameters

- **Weighted-least-squares:**
  Find the $ß_0$ and $ß_1$ that minimize the sum of the squared differences between the predicted and actual risks weighted by the inverse of the variances

From *Applied Logistic Regression Second Edition* by Hosmer & Lemeshow

- Under the usual assumptions for linear regression, the method of least squares yields estimators with a number of desirable statistical properties

From Hosmer & Lemeshow (2nd edition)

- **Unfortunately, when this method is applied to a model with a dichotomous outcome the estimators no longer have these same properties**

Logistic Regression

- Binary outcome variable (usually)
- Powerful!
- Very common model in the epidemiologic and biomedical literature

Logistic Regression

- These regression coefficients, unlike those of linear regression, can be directly converted to odds ratios

Logistic Regression

- **Dependent variable:**
  Typically binary (1 or 0)
- **Ordinal logistic regression**
- **Multinomial logistic regression**

Logistic Regression with A Binary Outcome

- Very common in the public health and medical literature

- Independent variable:
  Binary (1 or 0) or
  Continuous or
  Categorical with 3 or more levels

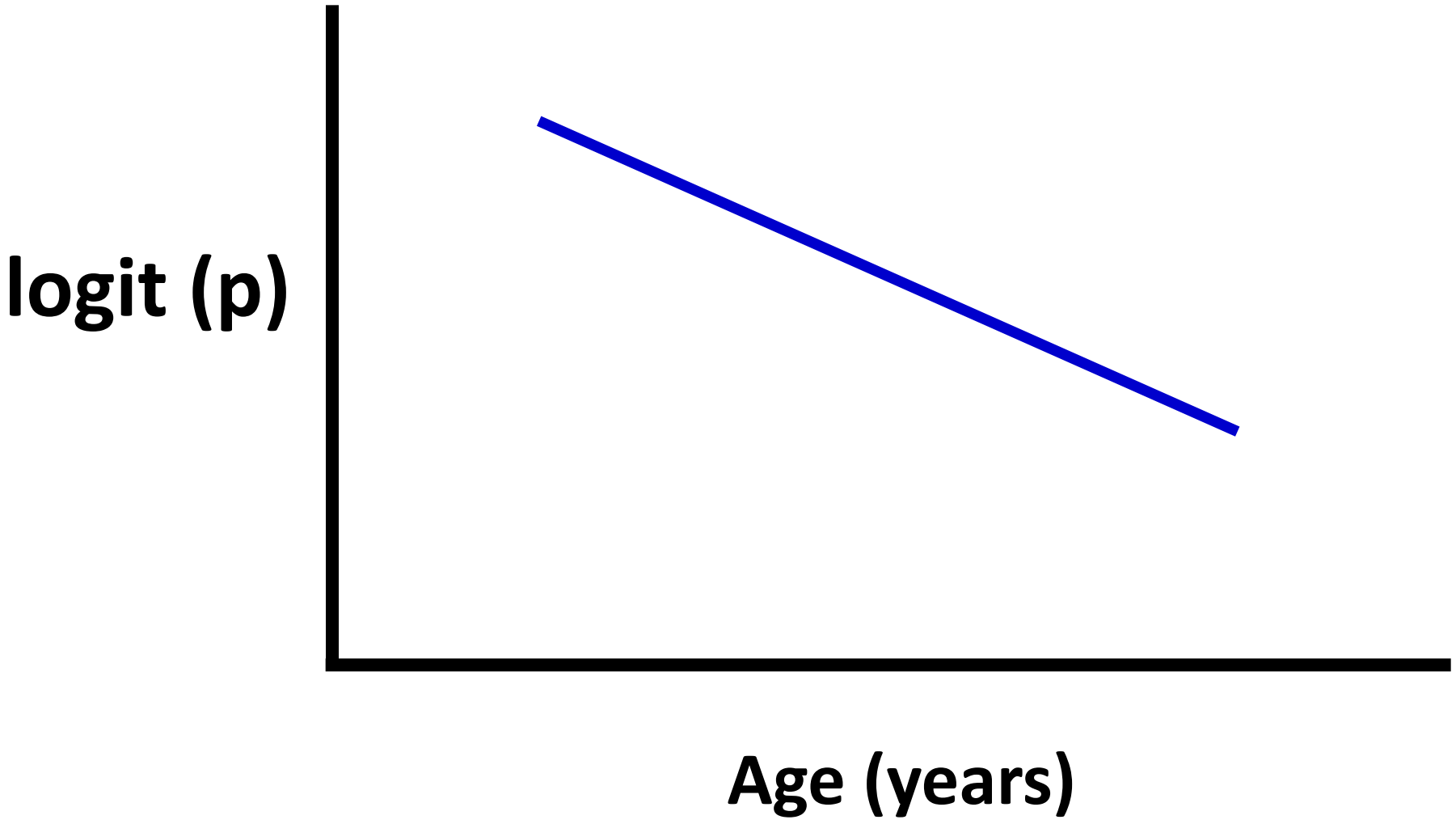
What is a logit (pronounced low-jit)?

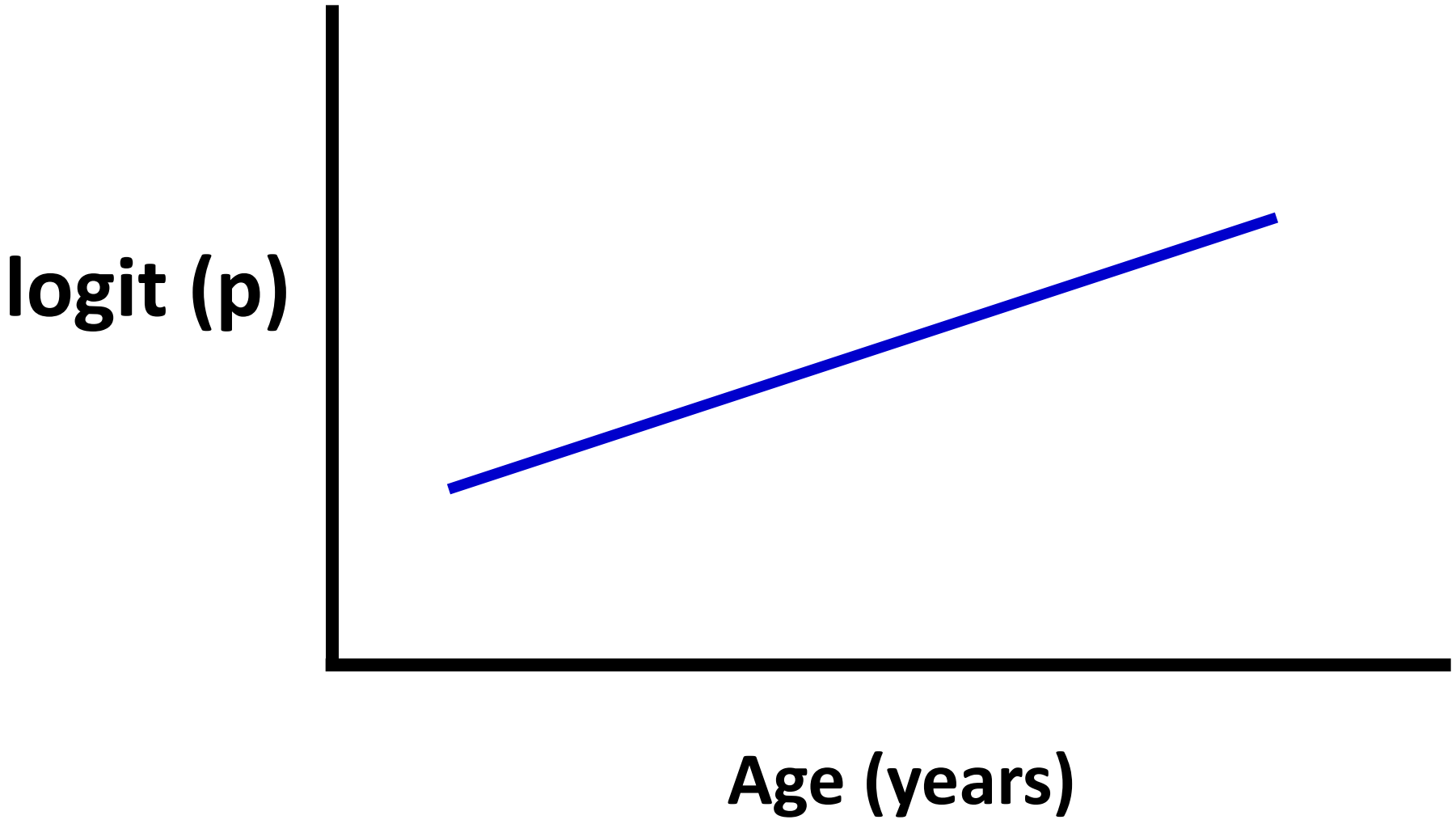
- It is the natural logarithm of the odds of the outcome
- Natural logarithm is abbreviated “ln”

Assumption in Logistic Regression Analysis

- If you have a continuous *X* variable, the logit increases or decreases in a linear fashion as *X* increases

- In other words, the natural log of the odds of the outcome varies in a linear fashion with *X*


logit (p)
Age (years)


logit (p)
Age (years)

Sample Model

Y = Heart attack (1 = yes, 0 = no)

X1 = Age (continuous variable)

X2 = Race (binary: Black or Other)

X3 = Systolic BP (binary: Hi or Low)

Fake Data for Three Subjects

Pt ID	*Y*	*X1*	*X2*	*X3*
A	**1**	**40**	**0**	**1**
B	**1**	**29**	**0**	**0**
C	**0**	**85**	**1**	**1**

Maximum Likelihood (Hosmer and Lemeshow)

- **The general method of estimation that leads to the least squares function under the linear regression model is called *maximum likelihood*.**

Maximum Likelihood (Hosmer and Lemeshow)

- We will use this method to estimate logistic regression parameters

- This method yields values for unknown parameters which maximize the probability of obtaining the observed set of data.

Steps in the Method of Maximum Likelihood (ML)

- **1. Construct a function. This is the *likelihood function*.**

This function expresses the probability of the observed data as a function of the unknown parameters.

Steps in the Method of ML

- **2. Maximize this function.**

The *maximum likelihood estimators* (MLE) of the parameters are chosen to be those values that maximize this function.

Steps in the Method of ML

- Thus, the resulting estimators are those which agree most closely with the observed data.

Definitions

n = Number of independent observations

i = Index of summation

y_i = Observed value of y

$\pi(x_i)$ = Predicted value of y

Log Likelihood

$$\sum_{i=1}^{n} \left(y_i \ln [\pi(x_i)] + (1-y_i) \ln [1-\pi(x_i)] \right)$$

Review of MLE

- **Find the β_0 and β_1 that maximizes the probability of getting the pattern of outcome events given the values of the independent variables.**

Iterative Weighted Least Squares

- It's a hybrid of MLE and Weighted Least Squares

- This is how PROC LOGISTIC calculates your parameter estimates: $ß_0$ and $ß_1$ and $ß_2$ etc.

Simple Logistic Regression

$$\ln\left[\frac{P}{1 - P}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

Probability of the outcome

$$\ln\left[\frac{P}{1-P}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

Odds of the outcome

$$\ln\left[\frac{P}{1-P}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

Log odds of the outcome (logit)

$$\ln\left[\frac{P}{1-P}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

Logistic Regression

- **These regression coefficients, unlike those of linear regression, can be directly converted to odds ratios (OR):**

$$\text{OR} = \exp\left(\hat{\beta}_1\right)$$

ORs from Logistic Regression

$$\ln(OR) = \hat{\beta}_1$$

$$\log_e(OR) = \hat{\beta}_1$$

$$e^{\hat{\beta}_1} = OR$$

$$OR = e^{\hat{\beta}_1}$$

$$OR = \exp(\hat{\beta}_1)$$

OR for what?

- **For one level of exposure relative to the next lower level**

Exposure odds ratio from a case-control study

	Case	Control
Exposed	A	B
Unexposed	C	D

$$OR = \frac{\frac{A}{C}}{\frac{B}{D}}$$

Exposure odds ratio from a case-control study

	Case	Control
Exposed	A	B
Unexposed	C	D

$$OR = \frac{\frac{A}{C}}{\frac{B}{D}}$$

Exposure odds ratio from a case-control study

	Case	Control
Exposed	A	B
Unexposed	C	D

$$OR = \frac{\frac{A}{C}}{\frac{B}{D}} = \frac{A}{C} \times \frac{D}{B} = \frac{AD}{BC}$$

Incidence odds ratio from a cohort study

	Got ill	Didn't get ill
Exposed	A	B
Unexposed	C	D

$$OR = \frac{\frac{A}{B}}{\frac{C}{D}} = \frac{A}{B} \times \frac{D}{C} = \frac{AD}{BC}$$

Multiple Logistic Regression

$$\ln\left[\frac{P}{1 - P}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Multivariate

- **More than one independent variable in the model**
- **Each regression parameter is adjusted for the remaining parameters in the model**
- **Some books say “multivariable” rather than “multivariate”**

Example using the Oswego Outbreak Investigation

- **"Oswego-An Outbreak of Gastrointestinal Illness Following a Church Supper," Case Studies in Applied Epidemiology No. 401-303, Centers for Disease Control and Prevention**

- **Outcome: Gastrointestinal illness**
- **Culprit (exposure variable): vanilla ice cream**
- **Possible confounder: age of the respondent**

The Oswego Example (CDC case study)

$$\ln\left[\frac{P}{1 - P}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Vanilla ice cream:
1= Ate it, 0 = Didn't eat it

Age:
1: ≤ 5 years, 0: > 5 years

$$\ln\left[\frac{P}{1-P}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Vanilla ice cream:
1: Ate it, 0: Didn't eat it

95% Confidence Interval for the Population (True) OR for Vanilla Ice Cream

$$= \exp(\hat{\beta}_1 \pm 1.96[S.E.])$$

$S.E.$ **in the above equation is the standard error of** $\hat{\beta}_1$

Interaction

- **In a linear regression model, the data analyst will sometimes multiply two terms and enter this product term in the model to look for statistical interaction**

Interaction in Linear Regression

- **See page 201and the last paragraph on page 209 of the text *Epidemiology An Introduction 2nd Edition* by Kenneth J. Rothman: a linear regression model is based on additivity of effects**

- **Adding a product term to a linear regression model is a way of evaluating if there is a departure from additivity**

Product term: x_1 multiplied by x_2

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 (x_1 * x_2)$$

Table 11-1 from *Epidemiology An Introduction 2nd Edition* by Rothman, p. 201

Smoke exposure	Not exposed to asbestos	Exposed to asbestos
Non-smokers	1	5
Smokers	10	50

Separate effect of smoking is 10 cases per 100,000.

Separate effect of asbestos is 5 cases per 100,000.

Separate effects of smoking & asbestos are not additive when both are present: $10 + 5 \neq 50$

However…

- **Adding a product term to a logistic regression model, "…amounts to the evaluation of departures from a multiplicative model rather than departures from additivity. Therefore, statistical evaluation of interaction using these models [models that use a logarithmic transformation] will not yield an appropriate assessment of biologic interaction." p. 209, *Introduction to Epidemiology 2nd Edition* by K. Rothman**

Product term

$$\ln\left[\frac{P}{1-P}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 (x_1 * x_2)$$

Table 11-1 from *Epidemiology An Introduction 2nd Edition* by Rothman, p. 201

Smoke exposure	Not exposed to asbestos	Exposed to asbestos
Non-smokers	1	5
Smokers	10	50

Risk ratio of smoking alone = 10/1 = 10

Risk ratio of asbestos alone = 5/1 = 5

Multiply the risk ratios

10 x 5 = 50

Table 11-1 from *Epidemiology An Introduction 2nd Edition* by Rothman, p. 201

Smoke exposure	Not exposed to asbestos	Exposed to asbestos
Non-smokers	1	5
Smokers	10	50

Risk in those exposed to both factors = 50 cases per 100,000

Risk in those exposed to neither factor = 1 case per 100,000

RR=50/1=50

Interaction

- **"If a statistical model is based on the multiplication of relative effects, as is the case for many popular statistical models used in epidemiologic applications (logistic regression is one example), the data in Table 11-1 would indicate no statistical interaction, because the relative effects of smoking and asbestos are multiplicative." p. 209, *Introduction to Epidemiology 2nd Edition* by K. Rothman**