Logistic Regression

Rosanna Irene Comoretto

Department of Public Health and Pediatrics,
University of Torino

Simple Logistic Regression

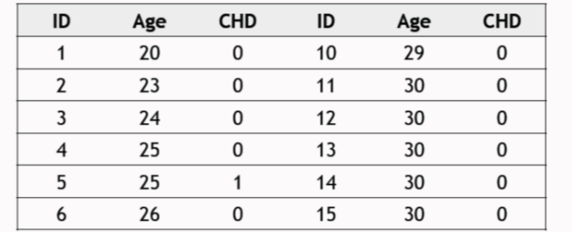
Outline

- The case for Logistic Regression
- How logistic regression relates a function of the probability (proportion) of a binary outcome through a linear relationship
- Interpret the resulting intercept and slope from a logistic regression model

The case for Logistic Regression

Evidence status (CHD): 100 subjects selected from a hospital population and screened for evidence CHD:

- average age 45 years, range 20 to 64
- 43% showed evidence of CHD



ID	Age	CHD	ID	Age	CHD
1	20	0	10	29	0
2	23	0	11	30	0
3	24	0	12	30	0
4	25	0	13	30	0
5	25	1	14	30	0
6	26	0	15	30	0

- **Aim**: to determine whether age is a risk factor for CHD and estimate the magnitude of this outcome exposure

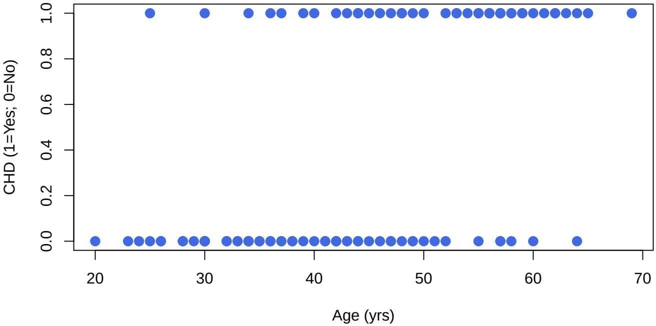
Outcome (response/dependent variable)

- Presence/absence of CHD evidence from screening result
  - $Y = 1$ if there is CHD evidence
  - $Y = 0$ if there is NOT CHD evidence

Y only takes on two values: - 1 (yes/presence) - 0 (no/absence)

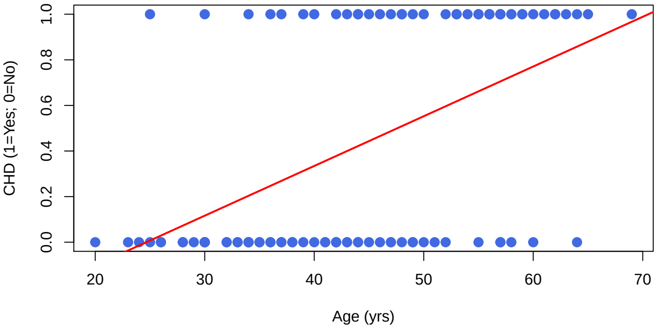
Example: CHD and age

- Could we use linear regression?



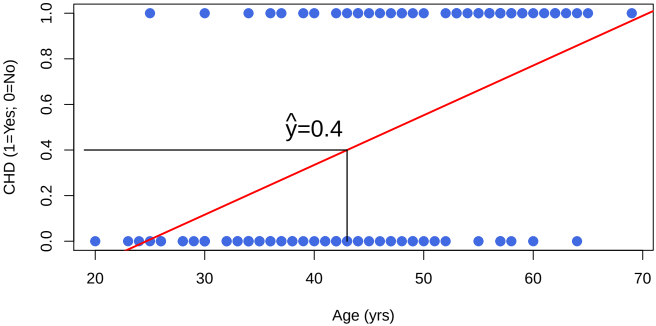
Example: CHD and age

- Could we use linear regression?



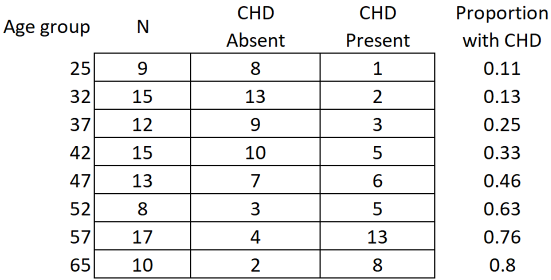
Example: CHD and age

- Could we use linear regression?



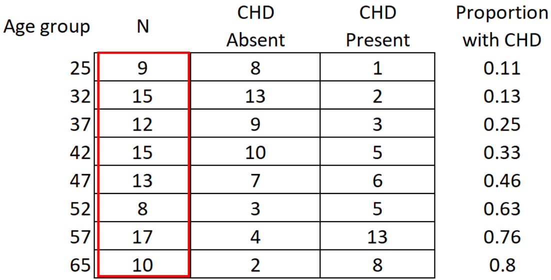
Example: CHD and age

- What about creating age intervals?



Age group	N	CHD Absent	CHD Present	Proportion with CHD
25	9	8	1	0.11
32	15	13	2	0.13
37	12	9	3	0.25
42	15	10	5	0.33
47	13	7	6	0.46
52	8	3	5	0.63
57	17	4	13	0.76
65	10	2	8	0.8

Example: CHD and age

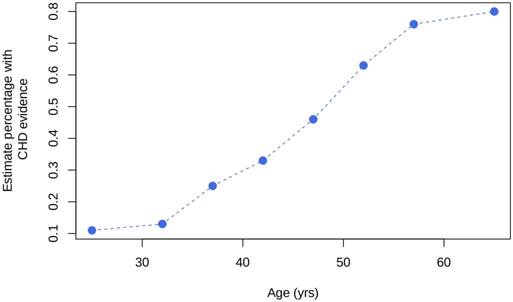


Age group	N	CHD Absent	CHD Present	Proportion with CHD
25	9	8	1	0.11
32	15	13	2	0.13
37	12	9	3	0.25
42	15	10	5	0.33
47	13	7	6	0.46
52	8	3	5	0.63
57	17	4	13	0.76
65	10	2	8	0.8

- Beware, each of the age intervals contain very few observations

Example: CHD and age

- It seems to be some structure/pattern here (percentage with CHD tends to increase with age)



Logistic Regression

- Wouldn't it be nice to model this relationship without having to categorize age and compute proportions?
- Logistic regression allows for such a curve relating (equation) the proportion with outcome to age

Objective of Logistic Regression

- Estimating a magnitude of outcome/exposure relationships
  - To evaluate the association of a **binary outcome** with a set of predictors
- Prediction
  - Develop a model to determine the probability/likelihood that an individual with *Xs* risk factors has the condition ($Y = 1$)

Different type of regression models

- **Linear Regression model**
  - Outcome variable Y is continuous
- **Proportional hazard (Cox) Regression model**
  - Outcome variable is time-to-event
- **Logistic Regression model**
  - Outcome variable Y is binary (dichotomous)
- **What does my outcome look like?** is the only (data type) question you need ask when choosing a regression method
  - Either regression method allows for many Xs (independent variables)
  - Xs can be either continuous or discrete

Logistic Regression

- Linear regression: outcome variable Y is continuous

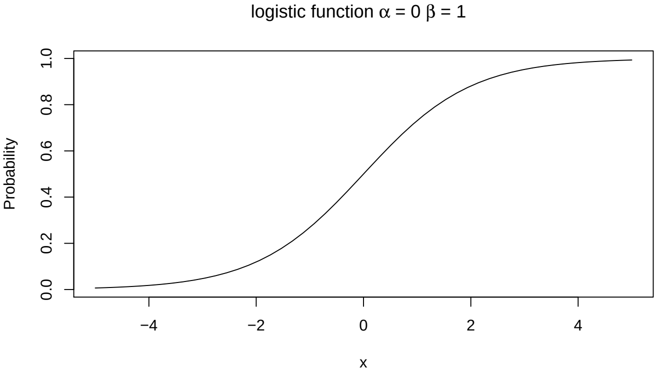
$$\mu_{Y|X} = \alpha + \beta x$$

- Logistic Regression: Outcome variable Y is binary (dichotomous)

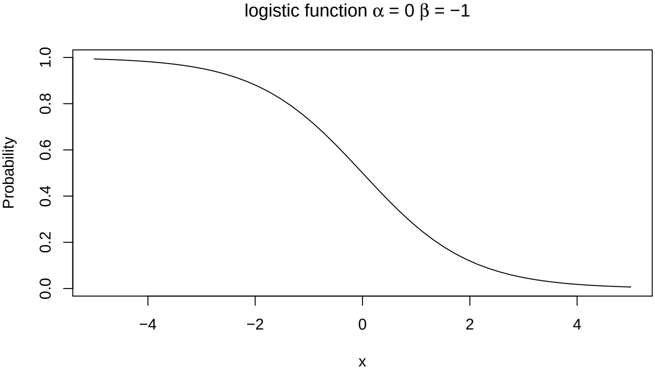
$$\mu_{Y|X} = P(Y = 1|x) = p_{Y|X}$$

1. $p_{Y|X} = \alpha + \beta x$
2. $p_{Y|X} = e^{\alpha + \beta x}$
3. $p_{Y|X} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$

Logistic function



Logistic function



Logistic Regression model

$$p_{Y|X} = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

becomes

$$\ln \frac{p}{1-p} = \alpha + \beta x$$

sometimes written as following:

$$\log \frac{p}{1-p} = \alpha + \beta x$$

- where ln (or log) is the logarithm to base e, or natural logarithms (e is the natural constant 2.718)
- The regression models the log odds of a binary outcome as a function of the predictor X
  - X can be binary, nominal, categorical or continuous

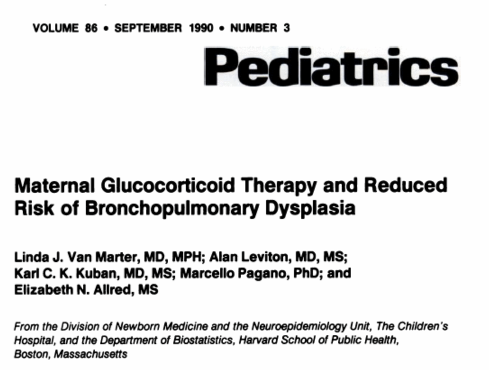
The odds

- The *odds* of an event is defined as following:

$$odds = \frac{p}{1-p}$$

- p is the probability (proportion) of $Y = 1$

Yet another example



VOLUME 86 • SEPTEMBER 1990 • NUMBER 3

Pediatrics

Maternal Glucocorticoid Therapy and Reduced Risk of Bronchopulmonary Dysplasia

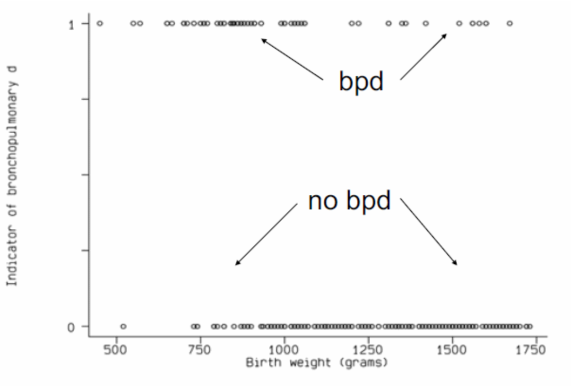
Linda J. Van Marter, MD, MPH; Alan Leviton, MD, MS;
Karl C. K. Kuban, MD, MS; Marcello Pagano, PhD; and
Elizabeth N. Allred, MS

From the Division of Newborn Medicine and the Neuroepidemiology Unit, The Children's Hospital, and the Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts

- Bronchopulmonary dysplasia (BPD) gets measured on 223 premature infants at about age 29 days after birth
  - 76 infants with BPD
  - 147 infants without BPD

Yet another example

- The study was carried out on infants born weighing less than 1750 grams.
- Each child was categorized as 0 (no bpd), or 1 (bpd)
- Results are plotted above, as a function of birth weight



Yet another example

Birthweight	BPD	N	Prop	Odds
0-950	49	68	0.721	2.58
951-1350	18	80	0.225	0.29
1351-1750	9	75	0.120	0.14
Total	76	223	0.341	0.52

The logistic regression model

- For BPD-weight dataset:

$$\ln \frac{p}{1-p} = \alpha + \beta x$$

- p: probability of BPD evidence (proportion of newborns with BPD)
- x: weight
- α and β are called *regression coefficients*

The logistic regression model

- Recall, the higher the odds of an event, the larger the probability of an event
- A predictor x that is positively associated with the odds will also be positively associated with the probability of the event (the estimated coefficient β will be positive)
- A predictor x that is negatively associated with the odds will also be negatively associated with the probability of the event (the estimated coefficient β will be negative)

The logistic regression model for BPD dataset

- Results from logistic regression of log odds of BPD evidence on birthweight:

$$\ln \frac{p}{1-p} = 4.03 - 0.0042 \times X$$

- p is the estimated probability of evidence (i.e. the estimated proportions of newborns with BPD evidence) among newbors of a given birthweight

BPD and birthweight

- The estimated coefficient (β_1) of birthweight X is negative
  - a negative association between birthweight and $\log \mathrm{odds}$ of BPD
  - a negative association between birthweight and BPD evidence
- How can we actually interpret the value 0.0042?

BPD and birthweight

- Consider two groups of newborns who differ in birthweight by 100 gr
  - **group 1:** birthweight = k gr
  - **group 2:** birthweight = k+100 gr
- The resulting equation estimating the $\log \mathrm{odds}$ of BPD in each birthweight group is:

$$\begin{aligned} \ln(\mathrm{odds\ of\ BPD; X = k + 100}) &= \alpha + \beta(k + 100) \\ \ln(\mathrm{odds\ of\ BPD; X = k}) &= \alpha + \beta k \end{aligned}$$

- Thus

$$100\beta = \ln(\mathrm{odds\,of\,BPD; X = k{+}100}) - \ln(\mathrm{odds\,of\,BPD; X = k})$$

BPD and birthweight

$$100\beta = \ln(\text{odds of BPD}; X = k + 100) - \ln(\text{odds of BPD}; X = k)$$

- From the properties of logarithms:

$$100\beta = \ln\left(\frac{\text{odds of BPD}; X = k + 100}{\text{odds of BPD}; X = k}\right) = \ln(\text{OR})$$

- β, the estimated slope of X is the natural log of an estimated odds ratio
- To get the estimated odds ratio, exponentiate β:

$$\text{OR} = e^{\beta}$$

BPD and birthweight

- In our example $\beta = -0.0042$ and $100\beta = \log(\mathrm{OR})$

BPD and birthweight

- In our example $\beta = -0.0042$ and $100\beta = \log(\mathrm{OR})$
- Here, $OR = e^{100\beta} = e^{-0.42} \approx 0.96$

BPD and birthweight

- In our example $\beta = -0.0042$ and $100\beta = \log(\mathrm{OR})$
- Here, $OR = e^{100\beta} = e^{-0.42} \approx 0.96$
- The estimated odds ratio of BPD evidence for 100 gr birthweight difference is 0.96

BPD and birthweight

- In our example $\beta = -0.0042$ and $100\beta = \log(\mathrm{OR})$
- Here, $OR = e^{100\beta} = e^{-0.42} \approx 0.96$
- The estimated odds ratio of BPD evidence for 100 gr birthweight difference is 0.96
- If we were to compare two groups of newborns who differ by 100 gr at birth, the estimated odds ratio for BPD evidence is 0.96

BPD and birthweight

- In our example $\beta = -0.0042$ and $100\beta = \log(\mathrm{OR})$
- Here, $OR = e^{100\beta} = e^{-0.42} \approx 0.96$
- The estimated odds ratio of BPD evidence for 100 gr birthweight difference is 0.96
- If we were to compare two groups of newborns who differ by 100 gr at birth, the estimated odds ratio for BPD evidence is 0.96
- 500 gr to 600 gr

BPD and birthweight

- In our example $\beta = -0.0042$ and $100\beta = \log(\mathrm{OR})$
- Here, $OR = e^{100\beta} = e^{-0.42} \approx 0.96$
- The estimated odds ratio of BPD evidence for 100 gr birthweight difference is 0.96
- If we were to compare two groups of newborns who differ by 100 gr at birth, the estimated odds ratio for BPD evidence is 0.96
- 500 gr to 600 gr
- 950 gr to 1050 gr

BPD and birthweight

- In our example $\beta = -0.0042$ and $100\beta = \log(\mathrm{OR})$
- Here, $OR = e^{100\beta} = e^{-0.42} \approx 0.96$
- The estimated odds ratio of BPD evidence for 100 gr birthweight difference is 0.96
- If we were to compare two groups of newborns who differ by 100 gr at birth, the estimated odds ratio for BPD evidence is 0.96
- 500 gr to 600 gr
- 950 gr to 1050 gr
- 1300 gr to 1400 gr

BPD and birthweight

- In our example $\beta = -0.0042$ and $100\beta = \log(\mathrm{OR})$
- Here, $OR = e^{100\beta} = e^{-0.42} \approx 0.96$
- The estimated odds ratio of BPD evidence for 100 gr birthweight difference is 0.96
- If we were to compare two groups of newborns who differ by 100 gr at birth, the estimated odds ratio for BPD evidence is 0.96
- 500 gr to 600 gr
- 950 gr to 1050 gr
- 1300 gr to 1400 gr
- This is valid for birthweight comparisons within our original range of data, 450-1730 gr

General interpretation: slope in Logistic regression

- β is the estimated change in the log odds of the outcome for a one unit increase in X
  - change in the $\log \mathrm{odds}$ of BPD for 100gr increase in birthweight
- It estimates the $\log \mathrm{odds\ ratio}$ for comparing two groups of observations
  - one group with x n-units higher than the other
- This estimated slope can be exponentiated to get the corresponding estimated odds ratio

What about the Intercept

- The resulting equation

$$\ln \frac{p}{1-p} = 4.03 - 0.0042 \times X$$

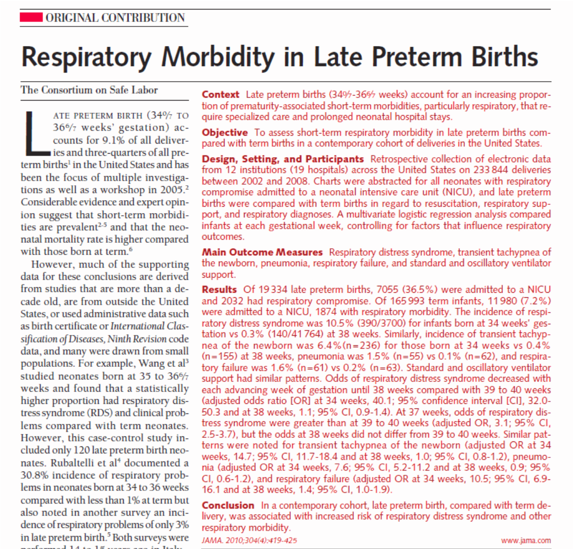
- Here, the intercept estimate α is in just a *place holder*
  - it is the estimated $\ln \mathrm{odds}$ of BPD evidence for newborns of birthweight 0
- The intercept is mathematically necessary to specify the entire equation and use the entire equation to estimate the $\ln \mathrm{odds}$ of the outcome for any group given X

Coefficients estimate in Logistic Regression

- The estimated regression coefficients are not the true population parameter regression coefficients
  - We will need to estimate a range of plausible values which takes into account error associated with an imperfect sample
  - We will need to test for a statistical significant association in the population
- We will need tools for doing inference

Example 2: Respiratory Failure and gestational age

- *Respiratory Morbidity in Late Preterm Births: The Consortium on Safe Labor, JAMA, 2010;304(4):419-25*



ORIGINAL CONTRIBUTION

Respiratory Morbidity in Late Preterm Births

The Consortium on Safe Labor

Context Late preterm births ($34\frac{0}{7}$-$36\frac{6}{7}$ weeks) account for an increasing proportion of prematurity-associated short-term morbidities, particularly respiratory, that require specialized care and prolonged neonatal hospital stays.

Objective To assess short-term respiratory morbidity in late preterm births compared with term births in a contemporary cohort of deliveries in the United States.

Design, Setting, and Participants Retrospective collection of electronic data from 12 institutions (19 hospitals) across the United States on 233 844 deliveries between 2002 and 2008. Charts were abstracted for all neonates with respiratory compromise admitted to a neonatal intensive care unit (NICU), and late preterm births were compared with term births in regard to resuscitation, respiratory support, and respiratory diagnoses. A multivariate logistic regression analysis compared infants at each gestational week, controlling for factors that influence respiratory outcomes.

Main Outcome Measures Respiratory distress syndrome, transient tachypnea of the newborn, pneumonia, respiratory failure, and standard and oscillatory ventilator support.

Results Of 19 334 late preterm births, 7055 (36.5%) were admitted to a NICU and 2032 had respiratory compromise. Of 165 993 term infants, 11 980 (7.2%) were admitted to a NICU, 1874 with respiratory morbidity. The incidence of respiratory distress syndrome was 10.5% (390/3700) for infants born at 34 weeks' gestation vs 0.3% (140/41 764) at 38 weeks. Similarly, incidence of transient tachypnea of the newborn was 6.4% (n=236) for those born at 34 weeks vs 0.4% (n=155) at 38 weeks, pneumonia was 1.5% (n=55) vs 0.1% (n=62), and respiratory failure was 1.6% (n=61) vs 0.2% (n=63). Standard and oscillatory ventilator support had similar patterns. Odds of respiratory distress syndrome decreased with each advancing week of gestation until 38 weeks compared with 39 to 40 weeks (adjusted odds ratio [OR] at 34 weeks, 40.1; 95% confidence interval [CI], 32.0-50.3 and at 38 weeks, 1.1; 95% CI, 0.9-1.4). At 37 weeks, odds of respiratory distress syndrome were greater than at 39 to 40 weeks (adjusted OR, 3.1; 95% CI, 2.5-3.7), but the odds at 38 weeks did not differ from 39 to 40 weeks. Similar patterns were noted for transient tachypnea of the newborn (adjusted OR at 34 weeks, 14.7; 95% CI, 11.7-18.4 and at 38 weeks, 1.0; 95% CI, 0.8-1.2), pneumonia (adjusted OR at 34 weeks, 7.6; 95% CI, 5.2-11.2 and at 38 weeks, 0.9; 95% CI, 0.6-1.2), and respiratory failure (adjusted OR at 34 weeks, 10.5; 95% CI, 6.9-16.1 and at 38 weeks, 1.4; 95% CI, 1.0-1.9).

Conclusion In a contemporary cohort, late preterm birth, compared with term delivery, was associated with increased risk of respiratory distress syndrome and other respiratory morbidity.

LATE PRETERM BIRTH ($34\frac{0}{7}$ TO $36\frac{6}{7}$ weeks' gestation) accounts for 9.1% of all deliveries and three-quarters of all preterm births[1] in the United States and has been the focus of multiple investigations as well as a workshop in 2005.[2] Considerable evidence and expert opinion suggest that short-term morbidities are prevalent[2-5] and that the neonatal mortality rate is higher compared with those born at term.[6]

However, much of the supporting data for these conclusions are derived from studies that are more than a decade old, are from outside the United States, or used administrative data such as birth certificate or *International Classification of Diseases, Ninth Revision* code data, and many were drawn from small populations. For example, Wang et al[3] studied neonates born at 35 to $36\frac{6}{7}$ weeks and found that a statistically higher proportion had respiratory distress syndrome (RDS) and clinical problems compared with term neonates. However, this case-control study included only 120 late preterm birth neonates. Rubaltelli et al[4] documented a 30.8% incidence of respiratory problems in neonates born at 34 to 36 weeks compared with less than 1% at term but also noted in another survey an incidence of respiratory problems of only 3% in late preterm birth.[5] Both surveys were performed 14 to 15 years ago in Italy

Example 2: Respiratory Failure and gestational age

- *Respiratory Morbidity in Late Preterm Births: The Consortium on Safe Labor, JAMA, 2010;304(4):419-25*

GestationalAge	Prop	Total
34 weeks	0.02	3700
35 weeks	0.03	5477
36 weeks	0.05	10157
37-40 weeks	0.90	165993

Example 2: Respiratory Failure and gestational age

- Gestational age categories are ordinal
  - authors didn't want to assume linearity of $\ln \text{odds}$ of respiratory failure and gestational age
- There are four categories:
  - make one category the reference and make binary X's indicators for the others
  - authors used *37-40 weeks* as the reference category

$$
\begin{aligned}
X_1 = 1 &\quad \text{if gestational age} = 34 \text{ weeks}\\
X_2 = 1 &\quad \text{if gestational age} = 35 \text{ weeks}\\
X_3 = 1 &\quad \text{if gestational age} = 36 \text{ weeks}
\end{aligned}
$$

Example 2: Respiratory Failure and gestational age

- *Respiratory Morbidity in Late Preterm Births: The Consortium on Safe Labor, JAMA, 2010;304(4):419-25*

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Example 2: Respiratory Failure and gestational age

- *Respiratory Morbidity in Late Preterm Births: The Consortium on Safe Labor, JAMA, 2010;304(4):419-25*

$$\begin{aligned}
\alpha &= \ln(\text{odds of respiratory failure, gest age} = 37-40) \\
\beta_1 &= \ln\left(\frac{\text{odds of respiratory failure, gest age=34}}{\text{odds of respiratory failure, gest age=37}-40}\right) \\
\beta_2 &= \ln\left(\frac{\text{odds of respiratory failure, gest age=35}}{\text{odds of respiratory failure, gest age=37}-40}\right) \\
\beta_3 &= \ln\left(\frac{\text{odds of respiratory failure, gest age=36}}{\text{odds of respiratory failure, gest age=37}-40}\right)
\end{aligned}$$

Example 2: Respiratory Failure and gestational age

- *Respiratory Morbidity in Late Preterm Births: The Consortium on Safe Labor, JAMA, 2010;304(4):419-25*

$$\ln\left(\frac{p}{1-p}\right) = -5.5 + 3.4X_1 + 2.8X_2 + 2.0X_3$$

- $\hat{\beta}_1 = 3.4 \rightarrow e^{3.4} = 30$
- $\hat{\beta}_2 = 2.8 \rightarrow e^{2.8} = 16.4$
- $\hat{\beta}_3 = 2.0 \rightarrow e^{2.0} = 7.4$

Example 2: Respiratory Failure and gestational age

- *Respiratory Morbidity in Late Preterm Births: The Consortium on Safe Labor, JAMA, 2010;304(4):419-25*

$$\ln\left(\frac{p}{1-p}\right) = -5.5 + 3.4X_1 + 2.8X_2 + 2.0X_3$$

- $\hat{\beta}_1 = 3.4 \rightarrow e^{3.4} = 30$
- $\hat{\beta}_2 = 2.8 \rightarrow e^{2.8} = 16.4$
- $\hat{\beta}_3 = 2.0 \rightarrow e^{2.0} = 7.4$
- $\ln\left(\frac{p}{1-p}\right) = -5.5 \rightarrow \mathrm{odds} = \mathrm{e}^{-5.5} = 0.004$

Example 3: Risk of obesity and HDL

- Data from 2009-10 NHANES
- Sample of over 6,400 residents US, 16-80 years old
- HDL levels
  - mean: 52.4 mg/dl
  - sd = 16
  - range = 11-144
  - 15% of the sample is obese (by BMI)

Example 3: Risk of obesity and HDL

- **Question:** does a line reasonably describe the general shape of the relationship between obesity and HDL?
- The line we estimate is in the form

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X$$

- p is the probability of being obese (proportion of individuals who are obese) for a given value of HDL cholesterol X

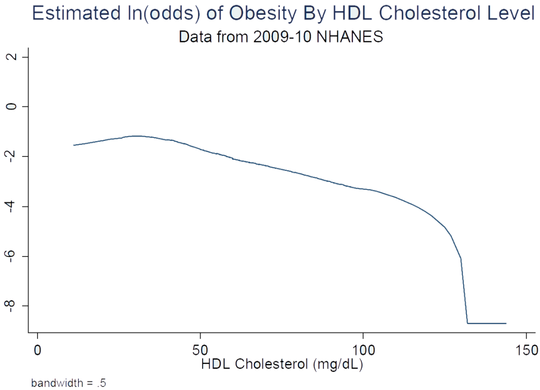
Example 3: Risk of obesity and HDL

- This formulation makes a strong assumption about the nature of the relationship between the ln(odds) of obesity and HDL cholesterol

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X$$

- How to investigate this assumption?

Example 3: Risk of obesity and HDL



Example 3: Risk of obesity and HDL

- Equation of the regression line relating ln(odds) of obesity to HDL

$$\ln\left(\frac{p}{1-p}\right) = -0.05 - 0.033X$$

Example 3: Risk of obesity and HDL

- The $OR = e^{-0.033} \approx 0.97$
- The OR of being obese for two groups of persons who differ by one mg/dL in HDL levels is 0.97, higher to lower HDL
  - higher HDL subjects have 3% lower odds (risk) of beig obese when compared to lower (by 1 mg/dL) HDL subjects
- The estimate is for any two groups who differ by 1 mg/dL in HDL in the population from which the sample was drawn
  - 60 mg/dL to 59 mg/dL
  - 44 mg/dL to 43 mg/dL
  - ...

Example 3: Risk of obesity and HDL

- What is the OR of being obese for persons with HDL of 100 mg/dL compared to persons with 80 mg/dL
- Using properties of logarithms

$$\ln\left(\frac{\mathrm{odds\ of\ obesity, X = 100}}{\mathrm{odds\ of\ obesity, X = 80}}\right) = \ln(\mathrm{OR}) = 20\beta$$

$$OR = e^{20\beta} = e^{20\times(-0.033)} \approx 0.51$$

- **Beware:**

$$OR = e^{20\beta} = (e^{\beta})^{20} = (e^{-0.033})^{20} = 0.97^{20} \approx 0.51$$

- Why?

Summary

- Logistic regression is a method for relating a binary outcome to a predictor X via a linear equation
  - the predictor can be binary, categorical or continuous
- The resulting linear equation relates the $\ln(\text{odds})$ of the binary outcome to the predictor X
- Slopes from logistic regression have $\ln(\text{odds})$ interpretation and can be eponentiated to estimate **odds ratios**
- The intercept estimates the $\ln(\text{odds})$ for the groups with $X = 0$

More Examples of Simple Logistic Regression

CHD and age

Variable	Estimated Coefficient	Standard Error
Age (yrs)	0.135	0.036
Constant	-6.54	1.73

CHD and Age

$$\ln \frac{p}{1-p} = -6.54 + 0.135 \times X$$

- p is the estimated probability of evidence (i.e. the estimated proportions of individuals with CHD evidence) among persons of a given age

CHD and Age

- In our example $\beta = 0.135$
- Here, $OR = e^{\beta} = e^{0.135} \approx 1.14$
- The estimated odds ratio of CHD evidence for a one-year age difference is 1.14, older to younger
- If we were to compare two groups of people who differ by one year of age, the estimated odds ratio for CHE evidence is 1.14 (this is valid for age comparisons within our original range of data, 20-69 years)
    - 60 years old to 59 years old
    - 45 years old to 44 years old
    - 27 years old to 26 years old

Death in the ICU: patients with sepsis

- Sample of 106 patients admitted to the ICU at a large U.S. hospital (Pine. et al.)
- All patients in sample had sepsis (blood infection) at time of admission to ICU
  - information also on whether patient died while in ICU,
  - patient's age at admission (range 17-94 years)
  - whether patient was in shock at time of admission
- Using age as predictor X, let's use logistic regression to relate death to patient age

Death in the ICU: patients with sepsis

Variable	Estimated Coefficient	Standard Error
Age of Patients (yrs)	0.052	0.015
Constant	-4.38	0.98

- $\beta = 0.052$ is the estimated $\ln \text{odds ratio}$ of death in ICU for one year difference in age
  - $\beta = 0.052$ is the estimated $\ln \text{odds ratio}$ of death in ICU for two groups of patients who differ by one year in age, older to younger
- The corresponding odds ratio estimate is $\text{OR} = \text{e}^{\beta} = \text{e}^{0.052} \approx 1.05$
  - in this sample a one year difference in age is associated with a 5% higher odds of death, older to younger
  - the older patients have 1.05 times the odds of death compared to the younger patients

Death in the ICU: patients with sepsis

- We could also use logistic regression to estimate the association between death and whether the patient was in shock at the time of admission to ICU (9% of the sample was in shock)

Variable	Estimated Coefficient	Standard Error
Shock (1 = yes)	2.61	0.75
Constant	-1.77	0.29

- $\beta = 2.61$ is the estimated ln $\mathrm{odds\ ratio}$ of death in ICU for for those in shock compared to those not in shock
- The corresponding odds ratio estimate is $\mathrm{OR} = \mathrm{e}^{\beta} = \mathrm{e}^{2.61} \approx 13.75$

Incorporating Sampling Variability

Outline

- 95% Confidence intervals for the intercept and slope and 95% Confidence intervals for OR
- Estimate p-values for testing the null $H_0 : \beta = 0$ (and hence OR $= 1$)

Coefficients estimate in Logistic Regression

- The estimated regression coefficients are not the true population parameter regression coefficients
- We will need to estimate a range of plausible values which takes into account error associated with an imperfect sample
- We will need to test for a statistical significant association in the population

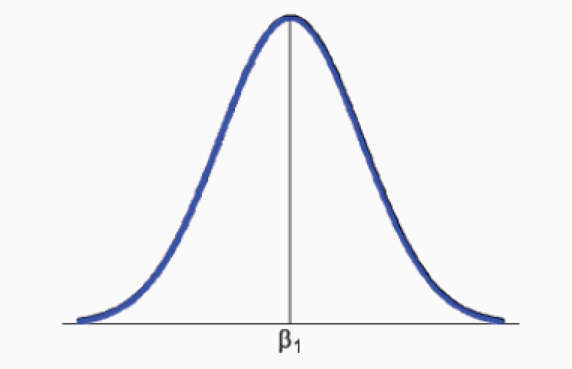
We will need tools for doing inference

Method of Estimation

- The method used to estimate the regression coefficients in logistic regression is called the **method of maximum likelihood**
- the resulting estimates of the slope and intercept are the values that make the observed data most likely among all choices of values for α and β
- This method is computationally intensive and is of course best handled by computers
- Along with estimates of α and β this method yields estimates of the **standard errors**
  - *standard errors* can be used to create confidence intervals and do hypothesis tests

Sampling Behavior of Logistic Regression Coefficients

- Random sampling behavior of estimated regression coefficients is normal for *large samples* and centered at true population value



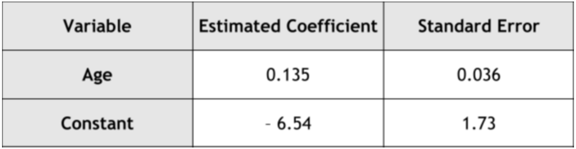
- we can use standard statistical reasoning to derive 95% CI or get *p-values*

Sampling Behavior of Logistic Regression Coefficients

- **Beware** the coefficients from logistic regression are on the $\ln \mathrm{odds}$ scale
- The sampling distribution of *odds* and *odds ratios* is not necessarily normal, but the sampling distribution of the ln of such quantities is
- We will create confidence intervals on the coefficient scale and will need to exponentiate the results to get corresponding CIs on the odds (ratio) scale
- Hypothesis testing and p-value will also be obtained on the coefficient scale

Example CHD and Age

- Recall the results from logistic regression of log odds of CHD evidence on age:



Variable	Estimated Coefficient	Standard Error
Age	0.135	0.036
Constant	- 6.54	1.73

- $\hat{\beta} = 0.135$ is the estimated $\ln \mathrm{odds\ ratio}$ of CHD evidence for two groups who differ by one year in age
  - the corresponding *odds ratio* is: $OR = e^{\hat{\beta}} = e^{0.135} \approx 1.14$

Example CHD and Age

- How to get 95%CI of β, the population value of ln $\mathrm{odds\ ratio}$?
- Same old approach: $\hat{\beta} \pm 1.96 \times \mathrm{SE}(\hat{\beta})$
  - for this example: $0.135 \pm 1.96 \times 0.036 = (0.06, 0.21)$
  - Notice, the 95% CI does not include 0, which would indicate no relationship between CHD and age on the ln $\mathrm{odds\ ratio}$
- To get the corresponding 95% CI for the odds ratio relating CHD to age, exponentiate the endpoints of the 95%CI
  - for this example: $(e^{0.06}, e^{0.21}) = (1.06; 1.23)$
  - Notice, the 95% CI does not include 1, which would indicate no relationship between CHD and age on the odds ratio scale

Example CHD and Age

- pvalue for testing

$$\begin{array}{ll} H_0 : \beta = 0 & \quad H_0 : e^{\beta} = 1 \ (\mathrm{OR} = 1) \\ H_1 : \beta \neq 0 & \quad H_1 : e^{\beta} \neq 1 \ (\mathrm{OR} \neq 1) \end{array}$$

- Assume null true and compute the standardized distance of $\hat{\beta}$ from 0

$$z = \frac{\hat{\beta} - 0}{\mathrm{SE}(\widehat{\beta})} = \frac{\hat{\beta}}{\mathrm{SE}(\widehat{\beta})} = \frac{0.135}{0.036} \approx 3.75$$

- **p-value** is the probability of being 3.75 or more standard errors away from 0 on a normal curve: $p < .001$ (very low)

Example CHD and Age

- How about confidence intervals for the odds ratio when the comparison is on two groups who differ by more than one unit of X?
- What does the CHD/age results estimate as the odds ratio of CHD evidence for 60 year olds compared to 50 olds? What is a 95% CI for this odds ratio?
- The estimated odds ratio is found by taking $e^{10\hat{\beta}} = e^{10\times 0.135} = e^{1.35} \approx 3.9$
  - it is the same as taking $OR^{10} = 1.14^{10}$
- Properties of 95% CI similar on a coefficient scale: 95%CI for $10\hat{\beta}$:

$$10\hat{\beta} \pm 1.96 \times \mathrm{SE}(10\widehat{\beta}) \Longrightarrow 10 \times [\widehat{\beta} \pm 1.96 \times \mathrm{SE}(10\widehat{\beta})]$$

95% Confidence Intervals

- On odds ratio scale, 95% CI for $e^{10\hat{\beta}}$ will be given by

  $$e^{10\times[\hat{\beta}\pm 1.96\times \mathrm{SE}(10\widehat{\beta})]} = \left(e^{10\times[\hat{\beta}-1.96\times \mathrm{SE}(10\widehat{\beta})]}; e^{10\times[\hat{\beta}+1.96\times \mathrm{SE}(10\widehat{\beta})]}\right)$$

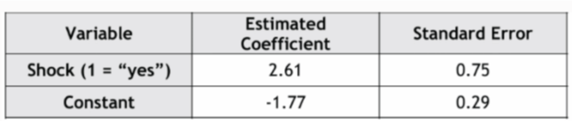
  which can be written down as:

$$\left([e^{\hat{\beta}-1.96\times \mathrm{SE}(10\widehat{\beta})}]^{10}; [e^{\hat{\beta}+1.96\times \mathrm{SE}(10\widehat{\beta})}]^{10}\right)$$

which is just $(L^{10}; U^{10})$ where L and U are the lower and upper endpoints respectively for the 95% CU for e^{β}

Death in the ICU: Patients with Sepsis

- Recall the results from logistic regression of log odds of death on shock status at the time of ICU admission:



Variable	Estimated Coefficient	Standard Error
Shock (1 = “yes”)	2.61	0.75
Constant	-1.77	0.29

- $OR = e^{\hat{\beta}} = e^{2.61} \approx 13.75$

Death in the ICU: Patients with Sepsis

- How to get 95%CI of β, the population value of $\ln \mathrm{odds\ ratio}$?
- Same old approach: $\hat{\beta} \pm 1.96 \times \mathrm{SE}(\hat{\beta})$
  - for this example: $2.61 \pm 1.96 \times 0.75 = (1.11, 4.11)$
  - Notice, the 95% CI does not include 0, which would indicate no relationship between CHD and age on the $\ln \mathrm{odds\ ratio}$

- To get the corresponding 95% CI for the odds ratio relating CHD to age, exponentiate the endpoints of the 95%CI
  - for this example: $(e^{1.11}, e^{4.11}) = (3.0; 61.0)$
  - Notice, the 95% CI does not include 1, which would indicate no relationship between CHD and age on the odds ratio scale

Death in the ICU: Patients with Sepsis

- pvalue for testing

$$\begin{array}{ll} H_0 : \beta = 0 & H_0 : e^{\beta} = 1 \ (\mathrm{OR} = 1) \\ H_1 : \beta \neq 0 & H_1 : e^{\beta} \neq 1 \ (\mathrm{OR} \neq 1) \end{array}$$

- Assume null true and compute the standardized distance of $\hat{\beta}$ from 0

$$z = \frac{\hat{\beta} - 0}{\mathrm{SE}(\widehat{\beta})} = \frac{\hat{\beta}}{\mathrm{SE}(\widehat{\beta})} = \frac{2.61}{0.75} \approx 3.5$$

- p-value is the probability of being 3.5 or more standard errors away from 0 on a normal curve: $p < .001$ (very low)

Estimating Risk and Functions of Risk

Study Design and Allowable Estimates

- Because the associations given in logistic regression are estimated odds ratios, this method can be used to analyze results from all types of study designs, including randomized studies, observational studies and case-control studies
- In randomized, or observational non case-control studies, we are not limited to odds ratios as measures of association
  - we can also estimate probability (proportions, risk), risk differences, and relative risks
  - can we get such association measures from logistic regression as well?

Study Design and allowable estimates

- Recall the generic equation for simple logistic regression:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X$$

- For any single value of x the equation estimates a $\ln \mathrm{odds}$ for a single group, the group with value x
- If we can get a $\ln \mathrm{odds}$, we can get an odds
  - it turns out if we can estimate the odds for any single group given x, we can estimate the probability of the outcome as well

Example: REspiratory failure and gestational age

- *Respiratory Morbidity in Late Preterm Births: The Consortium on Safe Labor, JAMA, 2010;304(4):419-25*

$$\ln\left(\frac{p}{1-p}\right) = -5.5 + 3.4X_1 + 2.8X_2 + 2.0X_3$$

- $\hat{\beta}_1 = 3.4 \to e^{3.4} = 30$
- $\hat{\beta}_2 = 2.8 \to e^{2.8} = 16.4$
- $\hat{\beta}_3 = 2.0 \to e^{2.0} = 7.4$
- $\ln\left(\frac{p}{1-p}\right) = -5.5 \to \mathrm{odds} = \mathrm{e}^{-5.5} = 0.004$

Example: REspiratory failure and gestational age

- To compute estimate risk of respiratory failure for reference group (37-40 weeks)

$$\ln\left(\frac{p}{1-p}\right) = -5.5 \rightarrow \text{odds} = \frac{p}{1-p} = e^{-5.5} = 0.004$$

$$p = \frac{\text{odds}}{1+\text{odds}} = \frac{0.004}{1.004} \approx 0.004\ (0.4\%)$$

Example: CHD and age

- Recall the resulting equation from our example relating CHD evidence to age

$$\ln\left(\frac{p}{1-p}\right) = -6.54 + 0.135 \times \mathrm{Age}$$

- p is the estimated probability of CHD (i.e., the estimated proportions of individual who had CHD) amongst those of a given age
- What does the above estimate for 57 year old individuals?

$$\ln\left(\frac{p}{1-p}\right) = -6.54 + 0.135 \times 57 = 1.16$$

Example: CHD and age

- This is the estimated $\ln \mathrm{odds}$ of CHD evidence of 57year old individuals in the sample

$$\ln\left(\frac{p}{1-p}\right) = -6.54 + 0.135 \times 57 = 1.16$$

- To get the corresponding odds, exponentiate
  - the *odds* of CHD for 57 year old individuals is $e^{1.16} \approx 3.19$
- Notice: $\mathrm{odds} = \frac{\widehat{\mathrm{p}}}{1-\widehat{\mathrm{p}}} \Longrightarrow \widehat{\mathrm{p}} = \frac{\mathrm{odds}}{1+\mathrm{odds}}$
- The above result translated into an estimated probability of

$$\hat{p} = \frac{\widehat{\mathrm{odds}}}{1+\widehat{\mathrm{odds}}} = \frac{3.19}{4.19} \approx 0.76$$

Example: CHD and age

- An estimated 76% of 57 year old individuals had CHD in the sample
- What about the estimated proportion of 55 year old individuals?

$$\ln\left(\frac{p}{1-p}; \mathrm{Age} = 55\right) = -6.54 + 0.135 \times 55 = 0.89$$

- The corresponding *odds* is $e^{0.89} \approx 2.44$
- The corresponding estimated probability is:

$$\hat{p} = \frac{\widehat{\mathrm{odds}}}{1 + \widehat{\mathrm{odds}}} = \frac{2.44}{3.44} \approx 0.71$$

Example: CHD and age

- An estimated 71% of 55 year old individuals had CHD in the sample
- An estimated 76% of 57 year old individuals had CHD in the sample
- What about the estimated relative risk of CHD for 57 year old invididuals to 55 year old ones?

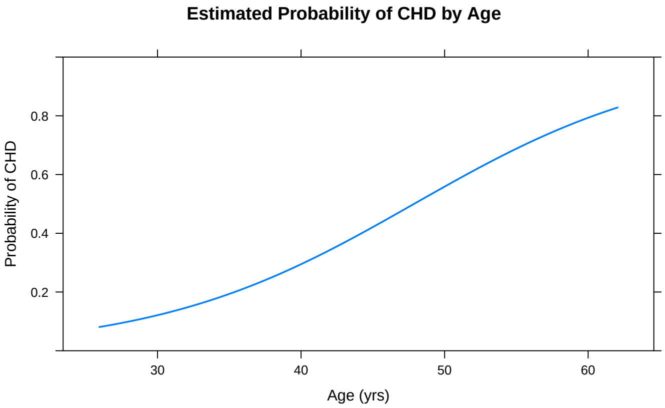
$$\widehat{\mathrm{RR}} = \frac{\widehat{\mathrm{p}}_{\mathrm{age}=57}}{\widehat{\mathrm{p}}_{\mathrm{age}=55}} = \frac{0.76}{0.71} \approx 1.07$$

- The estimated risk difference for the same age comparison

$$\hat{p}_{\mathrm{age}=57} - \hat{p}_{\mathrm{age}=55} = 0.76 - 0.71 = 0.05$$

Example: CHD and age

- Sometimes the graph of the estimated probability of the outcome across the values of a continuous variable in the sample is shown



Example: Death in the ICU and Shock

- Recall the resulting equation from our example on 106 ICU patients with severe sepsis relating death to shock

$$\ln\left(\frac{p}{1-p}\right) = -1.77 + 2.61 \times \mathrm{shock}$$

- p is the estimated probability of death (i.e., the estimated proportions of patients who die) amongst patients of a given shock status at admission
    - $\mathrm{shock} = 1$ if patient is in shock at the time of the admission

Example: Death in the ICU and Shock

- What does the equation estimate for patients in shock at the time of admission to the ICU?

$$\ln\left(\frac{p}{1-p}\right) = -1.77 + 2.61 \times \mathrm{shock}$$

$$\ln\left(\frac{p}{1-p}; \mathrm{shock} = 1\right) = -1.77 + 2.61 \times 1 = 0.84$$

- The corresponding *odds* is $e^{0.84} \approx 2.32$
- The corresponding estimated probability is:

$$\hat{p} = \frac{\widehat{\mathrm{odds}}}{1 + \widehat{\mathrm{odds}}} = \frac{2.32}{3.32} \approx 0.70$$

Example: Death in the ICU and Shock

- What does the equation estimate for patients NOT in shock at the time of admission to the ICU?

$$\ln\left(\frac{p}{1-p}\right) = -1.77 + 2.61 \times \mathrm{shock}$$

$$\ln\left(\frac{p}{1-p}; \mathrm{shock} = 1\right) = -1.77 + 2.61 \times 0 = -1.77$$

- The corresponding *odds* is $e^{-1.77} \approx 0.17$
- The corresponding estimated probability is:

$$\hat{p} = \frac{\widehat{\mathrm{odds}}}{1 + \widehat{\mathrm{odds}}} = \frac{0.17}{1.17} \approx 0.15$$

Example: Death in the ICU and Shock

- An estimated 70% of patients in shock died in this sample of ICU patients
- An estimated 15% of patients NOT in shock died
- What about the estimated relative risk of death for patients in shock compared to those NOT in shock?

$$\widehat{\mathrm{RR}} = \frac{\widehat{\mathrm{p}}_{\mathrm{shock=1}}}{\widehat{\mathrm{p}}_{\mathrm{shock=0}}} = \frac{0.70}{0.15} \approx 4.7$$

- The estimated risk difference for the same age comparison

$$\hat{p}_{\mathrm{shock=1}} - \hat{p}_{\mathrm{shock=0}} = 0.7 - 0.15 = 0.55$$