UCDAVIS
HEALTH

Logistic Regression

Susan Alber, Ph.D.

February 10, 2021

Learning goals

- Understand the basics of the logistic regression model
- Understand important differences between logistic regression and linear regression
- Be able to interpret results from logistic regression
  (focusing on interpretation of odds ratios)

If the only thing you learn from this lecture is how to interpret odds ratio then we have both succeeded.

Terminology for this lecture

In most public health research we have

One or more outcome variable(s) (indicators of disease are common outcomes)

One or more predictor variable(s) (factors that increase or decrease risk of disease are common predictors)

Research goal: determine if there is a significant relationship between the predictor variables and the outcome variable(s)

For this talk we will call any outcome variable “disease”

and any predictor variable “exposure” (exposure is anything that increases or decreases risk of disease)

Choice of statical methods

	Continuous (numerical) outcome (e.g. blood pressure)	**Binary outcome (e.g. disease)**
Categorical predictors (e.g. sex, race)	2 levels → T-test >2 levels →ANOVA	Chi-square test
Continuous predictors (e.g. age)	Linear regression	Logistic regression

ANOVA = analysis of variance

example

Outcome: coronary artery disease (CAD) (yes/no)

CAD = coronary artery disease

Predictors

Sex (categorical with 2 levels male/female)

Age (continuous 24-72)

Weight (categorical with 4 levels)

Body mass index (BMI)	weight
< 18.5	underweight
18.5 - 24.9	normal
25.0 - 29.9	overweight
> 30	obese

Outcome (CAD) is binary (disease / no disease)

and

One of the predictors is continuous (age)

Therefore we need to use logistic regression

Similarities between linear and logistic regression

- Based on a mathematical model of the dependence of a single outcome variable (e.g. disease) on one or more predictor (exposure) variables

  Predictors → outcome

- Predictor (exposure) variables can include any combination of continuous (e.g. age) and categorical (e.g. sex) predictors
- Model allows you to estimate exposure effects adjusted for confounders.
- Uses p-values to determine if predictors are significantly related to the outcome
- Uses confidence intervals for estimates of interest

key differences between linear and logistic regression

Linear regression: outcome is continuous (e.g. blood pressure)

Logistic regression: outcome is binary (e.g. disease / no disease)

Linear: dependence of the outcome on predictors quantified by

- Differences between means (for categorical predictors)
- Slopes (for continuous predictors)

Logistic regression: dependence of outcome on predictors quantified by odds ratios

Key challenge for understanding logistic regression is being able to interpret odds ratios (to be defined soon)

example: looking first at sex as a predictor of CAD

Sex	No Disease (CAD)	Disease (CAD)
male	162 (43%)	217 (57%)
female	136 (56%)	105 (44%)

Predictor is binary (male/female)
Outcome is binary (CAD / no CAD)

Therefore we use chi-square test

Chi-square p-value = 0.0009 → males have significantly larger risk of CAD

We do not need logistic regression for this because the predictor is not continuous (not a number)

example: looking first at sex as a predictor of CAD

Sex	No Disease (CAD)	Disease (CAD)
male	162 (43%)	217 (57%)
female	136 (56%)	105 (44%)

Three ways to quantify the (significant) sex effect

1. Risk difference
2. Relative risk
3. Odds ratios (what you would get from logistic regression)

Risk difference and relative risk

Sex	No Disease	disease
male	162 (42.7%)	217 (57.3%)
female	136 (56.4%)	105 (43.6%)

Risk difference = P(disease for male) – P(disease for females) = 0.573 – 0.436 = 0.137

Interpretation: males are about 14 percent more likely to have CAD

Relative risk (also called risk ratio) = $\frac{P(disease\ for\ males)}{P(disease\ for\ females)} = \frac{0.573}{0.436} = 1.31$

Interpretation: males are about 1.3 times as likely to have CAD

odds ratios

The plan

define odds

define odds ratios for sex and CAD

define odds ratios for weight (4 levels) and CAD

define odds ratios for age (continuous)

logistic regression with sex, weight, and age as predictors

What are the odds?

Suppose have game where the odds of winning are 2 to 1

on average for every 2 games won 1 game is lost

→ on average win 2 out of every 3 games

→ probability of winning is 2/3

→ probability of loosing is = 1 – (probability of winning) = 1 - 2/3 = 1/3

$$\text{odds} \ = \frac{P(win)}{P(lose)} = \frac{P(win)}{1-P(win)} = \frac{(2/3)}{(1/3)} = 2$$

More generally for any event

$$\text{odds} = \frac{P(\text{event})}{1 - \ P(\text{event})}$$

Odds ratios (OR)

odds ratio is a ratio of odds under two different conditions: for example exposed versus unexposed

$$\text{OR} = \frac{odds(disease\ for\ exposed)}{odds(disease\ for\ unexposed)}$$

Because odds are always positive OR > 0

OR = 1 → odds(disease for exposed) = odds(disease for unexposed)

→ Exposure does not affect risk of disease

OR > 1 → odds(disease for exposed) > odds(disease for unexposed)

→ exposed have higher risk of disease than unexposed

OR < 1 → odds(disease for exposed) < odds(disease for unexposed)

→ exposed have lower risk of disease than unexposed

Null values for hypothesis testing

Null hypothesis is P(disease given exposed) = P(disease given unexposed)

The null values are

Odds ratios (OR) $= \frac{odds(disease\ for\ exposed)}{odds(disease\ for\ unexposed)}$ = 1

Relative risk = $\frac{P(disease\ for\ exposed)}{P(disease\ for\ unexposed)} = 1$

Risk difference = P(disease given exposed) - P(disease given unexposed) = 0

Sex	CAD	n	Probability of disease	odds
male	217	379	217/379 = 0.573	0.573/(1-0.573) = 1.342
female	105	241	105/241 = 0.436	0.436/(1-0.436) = 0.773

OR $= \frac{odds(disease\ for\ male)}{odds(disease\ for\ female)}$ = 1.342/0.773 = 1.74

The odds of CAD for men is 1.74 times larger than for women

It is common practice to make the numerator the category we expect to have higher odds, but it is not necessary.

OR $= \frac{odds(disease\ for\ female)}{odds(disease\ for\ male)}$ = 0.773/1.342 = 0.574

To interpret the OR we need to know which is in the numerator.
1.74 is OR for male to female, 0.574 is odds ratio for female to male

95% confidence interval for the odds ratio

- OR = 1.7
- 95% confidence interval is (1.3, 2.4)

We say we are “95% confident” that the true odds ratios is between 1.3 and 2.4.

Why is this statement justified?

What does it actually mean?

95% confidence interval for the odds ratio

- 95% confidence interval is (1.3, 2.4)

- H_0: OR=1 (null hypothesis)

Two ways to test if null hypothesis is true at significance level (“alpha”) 0.05

1. p-value < 0.05 (0.0009 < 0.05 → significance)
2. 1 not in the confidence interval (1 is not in interval (1.3,2.4) → significance)

Questions?

RR and OR are close when the risk is small

Relative risk (RR) $= \frac{P(disease\ for\ exposed)}{P(disease\ for\ unexposed)}$

Odds ratios (OR) $= \frac{P(disease\ for\ exposed)/P(no\ disease\ for\ exposed)}{P(disease\ for\ unexposed)/P(no\ disease\ for\ unexposed)}$

If the disease is rare for <u>both</u> exposed and unexposed then
P(no disease for exposed) ~ 1
P(no disease for unexposed) ~ 1
→ RR ~ OR

Looking only at age <38 (to get data with small risks)

Sex	No Disease	disease
male	16 (84.21%)	3 (15.79 %)
female	10 (83.33%)	2 (16.67%)

Small risk in both sexes

Relative risk (RR) = 1.056

Odds ratio (OR) = 1.067

RR and OR very close because the risk for *both* males and females is small.

Looking only at age < 40

Sex	No Disease	disease
male	22 (88.00%)	3 (12.00%)
female	14 (77.78%)	4 (22.22%)

Relative risk (RR) = 1.85

Odds ratio (OR) = 2.095

Small risk for females, but risk is not small for men

RR and OR are NOT very close because the risk is NOT small for *both* male and female

Now that we understand how to interpret odds ratios for 2 groups we need to extend to

1. Categorical predictors with >2 groups
2. Continuous predictors

odds ratios for more than 2 categories

weight	No Disease	disease
underweight	7 (63.64)	4 (36.36)
normal	69 (51.49)	65 (48.51)
overweight	97 (41.99)	134 (58.01)
obese	125 (51.23)	119 (48.77)

Categorical predictor (4 levels) + binary outcome (disease / no disease)
chi-square test is appropriate

p-value=0.11

Since this p-value is not significant (0.11>0.05) we would normally not calculate any effect measures (such as risk difference, relative risk or odds ratios). Will do it here to learn about odds ratios.

odds ratios for more than 2 categories

weight	No Disease	disease
underweight	7 (63.64)	4 (36.36)
normal	69 (51.49)	65 (48.51)
overweight	97 (41.99)	134 (58.01)
obese	125 (51.23)	119 (48.77)

6 different OR can be calculated (corresponding to 6 different pairwise comparisons).
Generally a good idea to limit how many OR we calculate by making choices for which comparisons we want to focus on.

A "reference group" is a group that we choose to be the reference so that all odds ratios will be a comparison to the reference group.

Suppose choose normal as the reference group. Then we would compare underweight, overweight, and obese to normal

odds ratios for more than 2 categories

weight	Odds ratio (OR)	Interpretation
underweight	0.607	Less risk than normal weight
overweight	1.466	More risk than normal weight
obese	1.011	Approximately equal risk to normal weight
normal	1.000	

Sometimes, but not always papers will include this to indicate normal is the reference group: which means the OR for normal is odds(normal)/odds(normal) which of course is 1

Questions?

Relationship between age and CAD

Age (in years) is linear so now we need to use logistic regression.

From the logistic regression model we get

Odds ratio = 1.073, p-value < 0.0001, 95% confidence interval (1.054,1.093)

interpretation

Older age is a significant risk for CAD

For every one year increase in age the odds is 1.073 times larger

Logistic regression with all 3 predictors

Logistic regression allows us to look at all three predictors (sex, weight, and age) simultaneously.

Looking at relationships between each predictor and CAD separately is a good first step before proceeding to the full logistic regression model. It is important to understand these relationships first before looking at the full model.

Logistic regression results with all 3 predictors

	Odds ratio	P-value	95% confidence interval
sex (male vs female)	1.829	0.0007	(1.29,2.59)
Age (in years)	1.074	<0.0001	(1.015,1.095)
weight (obese vs normal)	1.225	0.4968	(0.780,1.952)
weight (overweight vs normal)	1.513	0.0903	(0.959,2.386)
weight (underweight vs normal)	0.695	0.4008	(0.177,2.724)

Can look at either the p-values or check if 1 is in the confidence interval to determine significance

conclusions

Male associated with increase risk of CAD
Risk of CAD increases with age with the odds increasing by 1.074 times for each one unit increase in age
Weight (with this categorization) not significantly associated risk of CAD
Same conclusions as we had with individual comparisons

A peak under the hood of the logistic regression model

One predictor: age → call X and one outcome → call this Y

Logistic regression model $\log\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right] = \mathrm{b}_0 + \mathrm{b}_1\mathrm{X}$

Compare to linear regression model $Y = b_0 + b_1X + e$

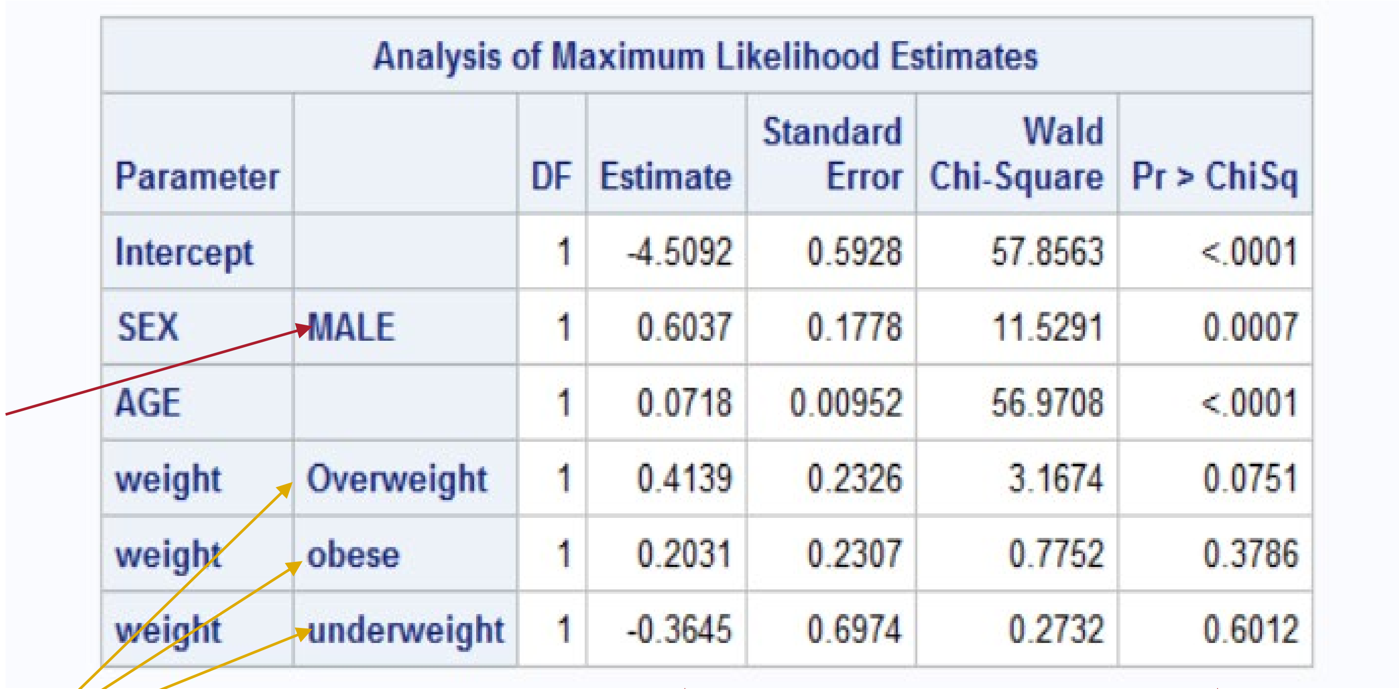
Term inside the square brackets is the odds conditional on the value of X

Entire term on left side of equals sign is a *log odds*

b_1 is a *log odds ratio* → odds ratio is $\mathrm{OR} = \exp(b_1)$

Note: here log is the natural log (with base e) which is some fields is written as ln(x)

(default) output from logistic regression in SAS



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.5092	0.5928	57.8563	<.0001
SEX	MALE	1	0.6037	0.1778	11.5291	0.0007
AGE		1	0.0718	0.00952	56.9708	<.0001
weight	Overweight	1	0.4139	0.2326	3.1674	0.0751
weight	obese	1	0.2031	0.2307	0.7752	0.3786
weight	underweight	1	-0.3645	0.6974	0.2732	0.6012

Absence of female tells you that is the reference group

Absence of normal tells you that is the reference group

log odds ratios

p-values

Converting from log odds ratios to odds ratios

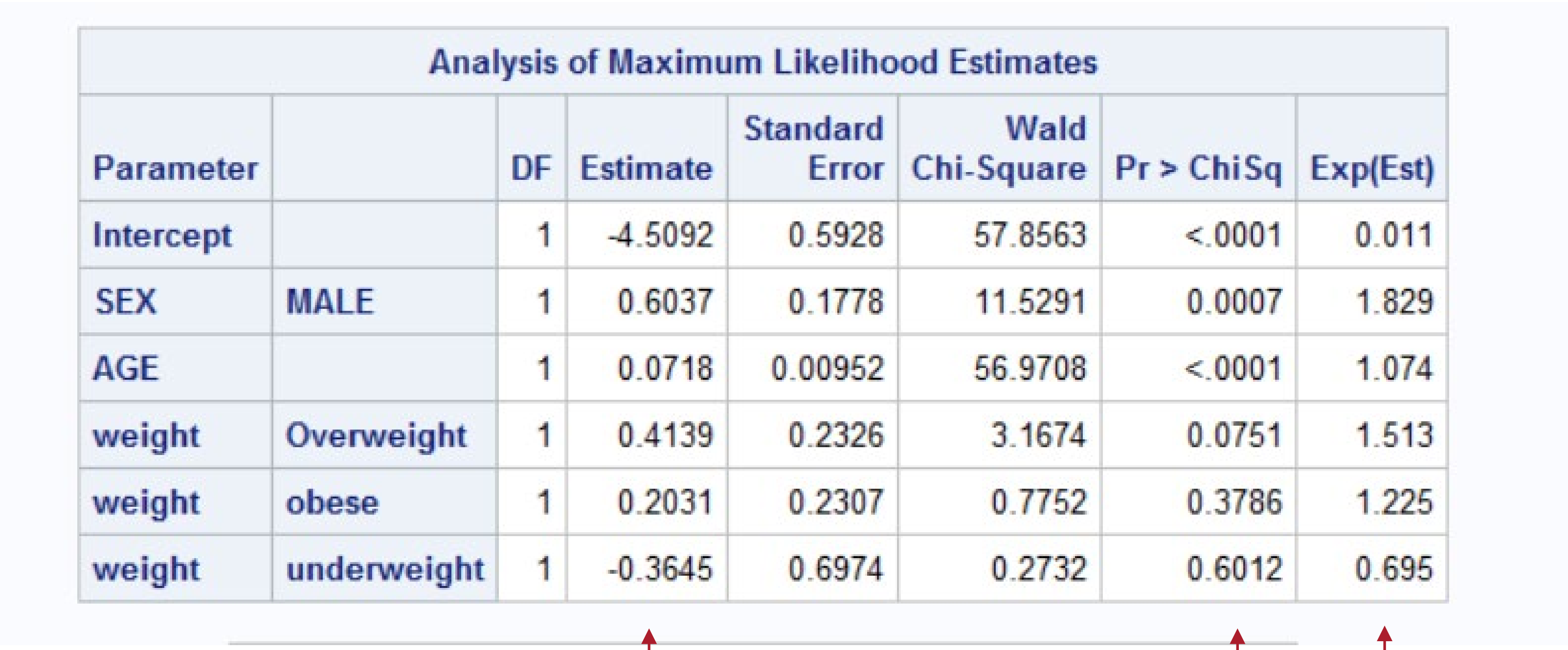
Odds ratio = exp(log odds ratio)

Example: to get the odds ratio for sex

Log odds ratio = 0.6037

Odds ratio = exp(0.6037) = 1.829

SAS output with both log odds ratios and odds ratios



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	-4.5092	0.5928	57.8563	<.0001	0.011
SEX	MALE	1	0.6037	0.1778	11.5291	0.0007	1.829
AGE		1	0.0718	0.00952	56.9708	<.0001	1.074
weight	Overweight	1	0.4139	0.2326	3.1674	0.0751	1.513
weight	obese	1	0.2031	0.2307	0.7752	0.3786	1.225
weight	underweight	1	-0.3645	0.6974	0.2732	0.6012	0.695

log odds ratios

odds ratios

p-values (same for log odds ratios and odds ratios)

Interpretation of odds ratio for age (from logistic)

Odds ratio = 1.074

Log odds ratio = log(1.074) = 0.072

For every 1 year increase in age the log odd increases by 0.072

→ For every 10 year increase in age the log odds increases by 0.072 x 10 = 0.72

Exponentiate to get back on the odds scale

exp(0.72) = 2.05

→ For every 10 year increase in age the odds doubles (i.e. twice the size)

How to tell if values are log odds ratios or odds ratios

- Some statistical software is nice enough to actually label outcome as being either odds ratios or log odds ratios, but otherwise
- If any of the estimates are negative then values are log odds ratios (minimum value for an odds ratio is 0)
- If the confidence intervals are symmetric around the estimate (i.e. distance between the estimate and the bounds are the same for lower and upper limit) then values are log odds ratios (confidence intervals for odds ratios are not symmetric around the estimate)

Questions?

Compare linear and logistic regression

linear regression

$$Y = b_0 + b_1X + e$$

Y is a linear function of X

We estimate the values of b_0 and b_1

can then use the estimates and the model to make prediction about the value of Y for a given value of X.

logistic: model dependence of Y on X through the log odds

logistic regression

$$\boxed{\log\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right]} = \mathrm{b}_0 + \mathrm{b}_1\mathrm{X}$$

Log odds

Two reasons why it is good to know this.

1. the log part is why b_1 is a <u>log</u> odds ratio not an odds ratio
2. model can be inverted (which means solved for P(Y=1|X)) and then used to estimate P(Y=1|X) for any X

Would usually use statistical software to calculate the inverse, so not necessary to know the function, but here it is

$$\log\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right] = \log\left[\frac{P(Y=1|X)}{1-\ P(Y=1|X)}\right] = \mathrm{b}_0 + \mathrm{b}_1\mathrm{X}$$

Invert: solve for P(Y=1|X)

$$P(Y=1|X) = \frac{\exp(b_0+b_1X)}{1+\exp(b_0+b_1X)} \qquad \text{(inverse function)}$$

Can now plug in any value of X to get P(Y=1|X)

The inverse function is more complicated when we have multiple predictors (as in our example).

But we can still use the logistic regression model to estimate the probability of disease for any combination of sex, weight, and age.

Example: If I wanted to compare probabilities of disease for overweight 30 year old men to women, using the logistic regression model.

Probability of disease for 30 year old overweight men is 0.21

Probability of disease for 30 year old overweight women is 0.13

(I used SAS to get these estimates from the model.)

Conditions needed for logistic regression

- Outcome is binary (can be extended to multinomial, but model is more complicated and a bit more difficult to interpret)
- Sample size needs to be large (larger than required for linear regression)

  necessary sample size is a function of

  1. number of predictors (more predictors requires larger sample size)
  2. probability values (close to 0 or 1 requires larger sample size)

Note: for linear regression sample size only needs to be large if the outcome is not normally distributed.

For logistic regression the outcome is binary, so not possible to be normally distributed.

what you need to know to interpret results from logistic regression

- Direction of the outcome (is the model for the probability of disease or for the probability of no disease).

When focus is on studying factors that increase the probability of disease we usually model the probability of disease (as we did in our example).

When focus is on studying factors that decrease the probability of disease then would usually model the probability of no disease.

However for statistical software there is usually a default choice which may or may not be the one you want.

(you get the same conclusions, but need to know to interpret results)

- For each categorical variable what is the reference group? (there are other ways to specify a model that do not use the reference group coding)
- Are the results odds ratios or log odds ratios?
- What are the scales for each continuous variable? For example, is age in years or some other unit

Thank you

Questions?