Logistic Regression

HIP 529 Applied Biostatistics II

DATE: March 28, 2022

PRESENTED BY: Emile Latour, MS, Associate Biostatistician, OHSU Knight Cancer Institute

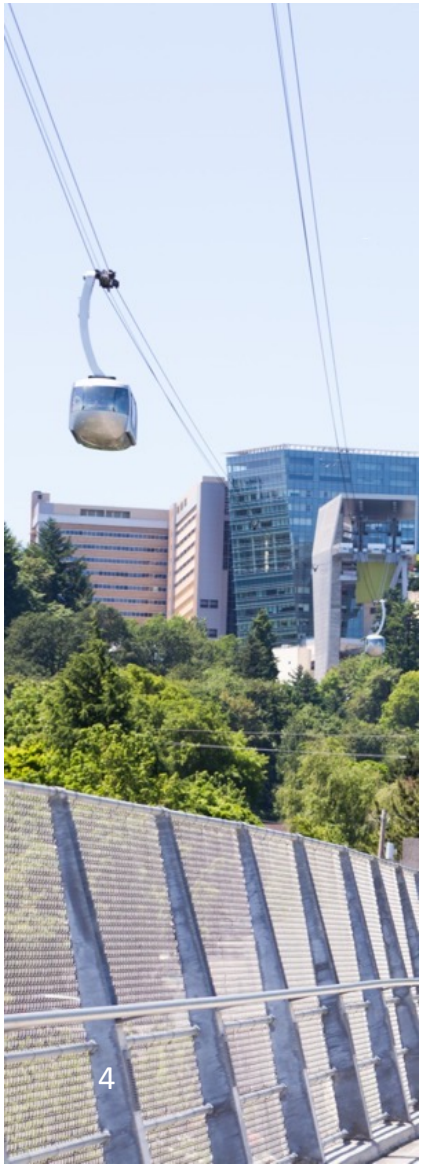
Slides are available

- https://bit.ly/hip-529-logistic-regression_2022-03-28

OHSU

About me

- Pension actuary for 16 years
- MS in Biostatistics from OHSU in 2017
- Associate Biostatistician with OHSU Knight Cancer Institute since 2016
- Long term support for OHSU Dermatology Department since 2018



Goals

- Exposure to statistics background on logistic regression models and what sets it apart
- Understand terms: log odds, odds ratio, and relative risk
- Univariable logistic regression models (association models):
  - Continuous predictor, categorical predictors

OHSU


Statistical modeling

Goal of statistical modeling

- To find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome variable and a set of independent variables.

Describe a relationship

- A statistical model is a mathematical relationship between an outcome (dependent variable) and one or more independent variables.

Describe a relationship

- A statistical model is a mathematical relationship between an outcome (dependent variable) and one or more independent variables.


Types of data /
variables

Types of data / variables

- Discrete (numerical) – quantities or counts, likely with units
- Continuous (numerical) -- quantities, likely with units
- Ordinal (categorical) – labels with order
- Nominal (categorical) – labels with no order

OHSU

Discrete (numerical) data

- Actual quantities; not numerical coding of categories
- Restricted to a set of isolated points
- Intermediate values are impossible
- Examples:
  - Occupants per household: 1, 2, 3, ….
  - Number of tumors: 0, 1, 2, 3, 4, …
  - Age to nearest whole year: …, 35, 36, 37, …

OHSU

Continuous (numerical) data

- Actual quantities
- Not restricted to a set of isolated points
- Intermediate values are(theoretically) possible
- Examples:
  - Change in body weight during pandemic: +19.25 lbs
  - Cholesterol level: 162 mg/dL
  - Age (fractional): 35.75 years

OHSU

Ordinal (categorical) data

- Categories or classes with meaningful order
- Examples:
  - Pain: None, Mild, Moderate, Severe
  - Likert: Disagree, Neutral, Agree
  - Smoking status: Never, Occasionally, Frequently
- Numerical codes can be assigned, but must show the ranking (order)

Nominal (categorical) data

- Unordered categories or classes
- Examples:
  - Biological sex: M = Male, F = Female
  - Blood type: A, B, O, AB
  - Smoking status: Yes/No (1/0)
- Data can be given numerical codes (e.g. 1=A, 2=B, 3=O, 4=AB)

Special case – Binary data

- Nominal data with only two levels
- Examples: Male/Female, Yes/No, Case/Control, Tx/No Tx
- Labels can be replaced with the codes 0 and 1
  - Sum of the numeric codes gives the total with code 1
  - The average of the codes gives the proportion (or percentage) assigned the code 1
  - Example: Sample of n = 10 people coded as smoker (1) or non-smoker (0):
    - {1, 0, 1, 0, 1, 1, 1, 1, 1, 0}
    - Total = 7
    - Average = Total / n = 7 / 10 = 0.70 = 70.0%

OHSU


Regression models

Regression analysis

- Mathematically describe the relationship between a dependent variable (outcome) and a set of one or more independent variables

OHSU

Choosing a model

- Most often, the type of data of your outcome will determine the type of model that is used

OHSU

Terms

- Dependent variable, aka
  - Outcome
  - Response
  - Typically denoted with Y
- Independent variable(s)
  - Predictors
  - Explanatory variables
  - Typically denoted with X_i

OHSU

Classes of models

- "General" Linear models:
  - Refers to conventional linear regression models
  - Continuous outcome variable
  - Continuous and categorical predictors
  - Also includes multiple linear regression, ANOVA, and ANCOVA

https://online.stat.psu.edu/stat504/lesson/6/6.1

Classes of models

- "Generalized" Linear models (GLM):
  - Broad class of models
  - Includes ordinary regression/ANOVA models for continuous outcomes
  - Also, includes models that can handle categorical outcomes

OHSU

3 components of a GLM

- Random component
- Systematic component
- Link function

https://online.stat.psu.edu/stat504/lesson/6/6.1

OHSU

Random component

- Specifies the response variable Y and a probability distribution for it
  - If Y is continuous, then assumes normal distribution of Y
  - If Y is binary, then assumes a binomial distribution of Y
  - If Y is a count, then assumes a Poisson or negative binomial distribution of Y

Binomial distribution

- Success/Failure
- Conditions:
  - The number of trials is fixed (n)
  - Each trial has one of two outcomes: success or failure
  - The probability of success (p) is the same for each trial
  - Trials are independent; the outcome of one trial does not affect the outcome of another

Binomial distribution

- Examples
  - Number of heads out of 20 coin flips
  - Number of times you hit red light on your commute to work
  - Number of passengers on the Titanic that survived
  - Number of patients with a disease in a sample of 1000 patients.

https://shiny.rit.albany.edu/stat/binomial/

Systematic component

- Specifies the explanatory variables (x_1, x_2,… x_k) which enter the model as a linear combination

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Here the x_i can be based on other explanatory variables:
  - For example, $x_3 = x_1 * x_2$ or $x_3 = {x_1}^2$

OHSU

Link function

- Connects the predictors in a model with the expected value of the response variable in a linear way
- Specifies the link between the random component and the systematic component

Link function

- Let the expected value of the response be

$$\mu = E(Y)$$

- The link function specifies $g(.)$ that relates μ to the linear combination of predictors

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

OHSU

Link function

- Identity link
  - The link for a continuous response variable
  - $g(\mu) = \mu$
- Log link
  - Often used for a discrete count response variable
  - $g(\mu) = log(\mu)$
- Logit link
  - Often used for a categorical response variable
  - $g(\mu) = logit(\mu) = log\left(\frac{\mu}{1-\mu}\right)$
  - A GLM that uses a logit link function is called a **logistic regression model**

OHSU

GLM for binary data

- Identity link is not appropriate when the dependent variable is categorical
- Assumptions that underlie ordinary linear regression (OLR) are violated:
  - Linear relationship between μ and x can be problematic
  - Normality assumption is violated
    - Serious when sample size is small; less serious when sample size is large. Still an issue either way
  - Homoscedasticity assumption is violated
    - Can lead to biased estimates of the standard error to an unknown degree


Example

CHD Study

- A study to evaluate the relationship between age and presence of coronary heart disease (CHD)

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression, 3rd ed., New York: Wiley

OHSU

CHD Study

- N = 100
- First 20 shown here

id	age	agegrp	chd
1	20	20-29	No
2	23	20-29	No
3	24	20-29	No
4	25	20-29	No
5	25	20-29	Yes
6	26	20-29	No
7	26	20-29	No
8	28	20-29	No
9	28	20-29	No
10	29	20-29	No
11	30	30-34	No
12	30	30-34	No
13	30	30-34	No
14	30	30-34	No
15	30	30-34	No
16	30	30-34	Yes
17	32	30-34	No
18	32	30-34	No
19	33	30-34	No
20	33	30-34	No

CHD Study

- Primary endpoint (dependent variable): CHD
  - CHD = “Yes”: patient has coronary heart disease
  - CHD = “No”: patient does not have coronary heart disease

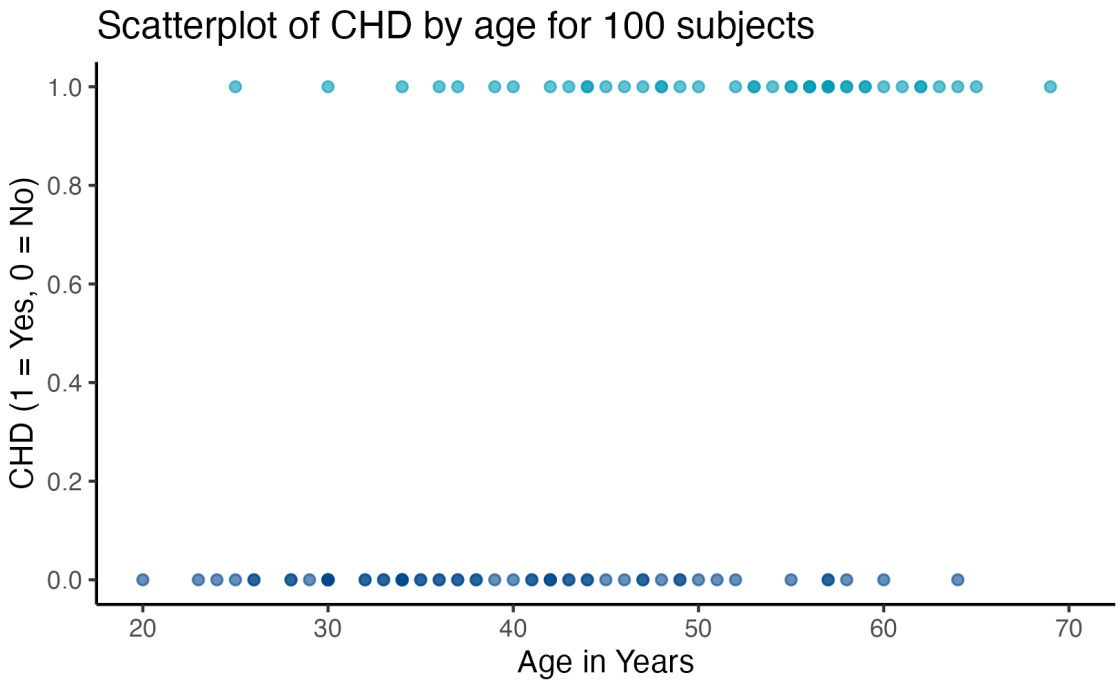
OHSU

CHD Study

- Primary endpoint (dependent variable): CHD
  - CHD = “Yes”: patient has coronary heart disease
  - CHD = “No”: patient does not have coronary heart disease

OHSU

CHD Study



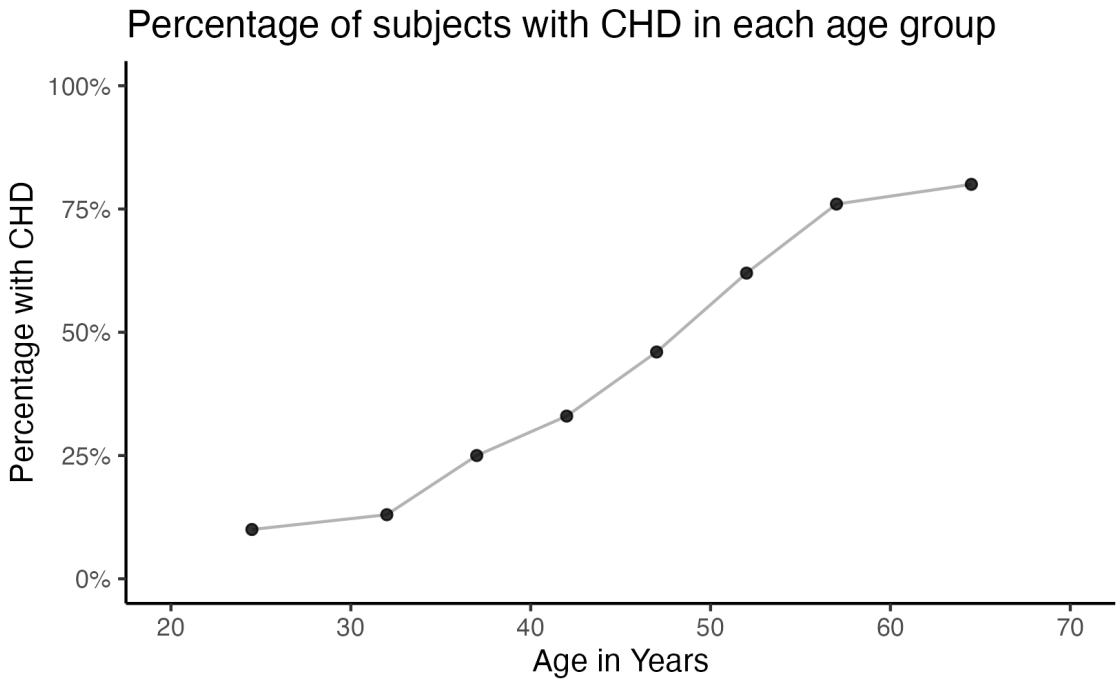
OHSU

CHD Study

Frequency table of CHD by age group				
		CHD		
Age group	n	Yes	No	Mean (proportion)
20-29	10	1	9	0.10
30-34	15	2	13	0.13
35-39	12	3	9	0.25
40-44	15	5	10	0.33
45-49	13	6	7	0.46
50-54	8	5	3	0.62
55-59	17	13	4	0.76
60-69	10	8	2	0.80
Total	100	43	57	0.43

OHSU

CHD Study



OHSU

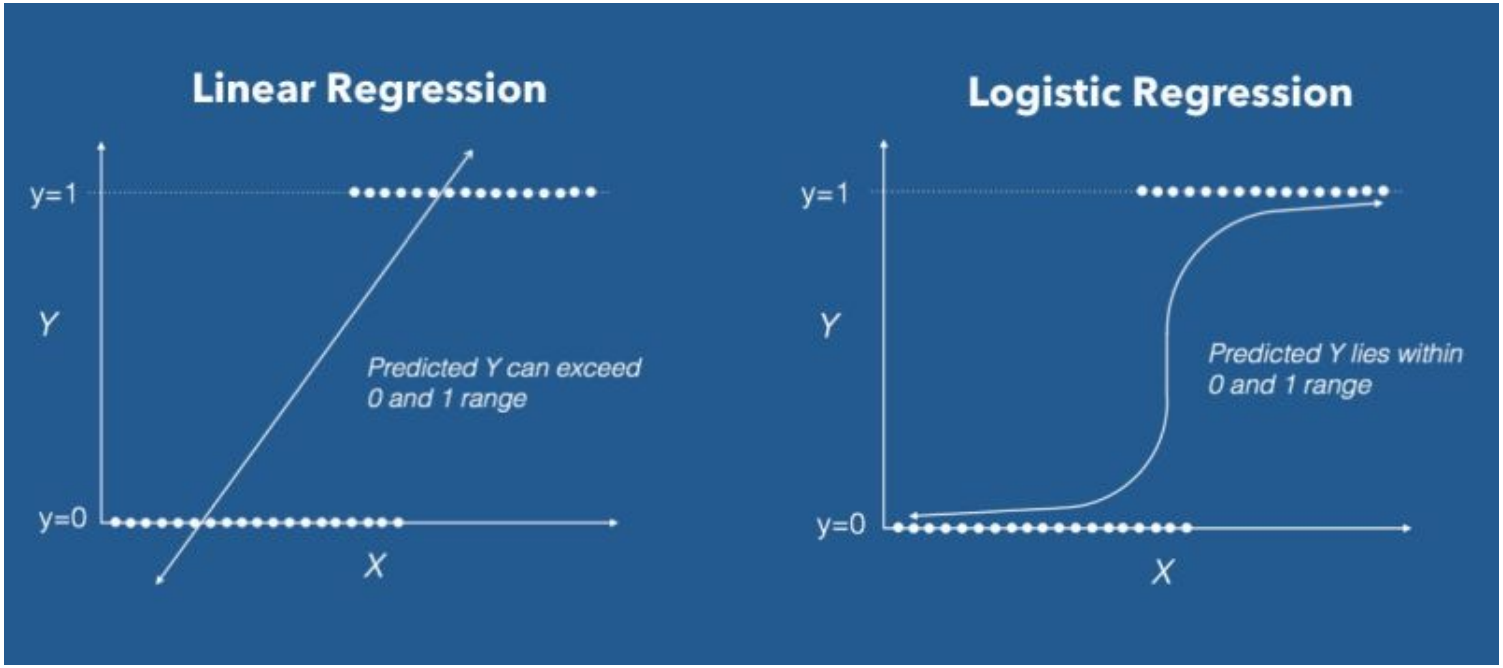
CHD Study

- Apparent trend: increasing proportion with CHD as age increases
- Relationship does not look linear in the two ends

OHSU


Making things linear

Another look at the issue



https://www.datacamp.com/community/tutorials/logistic-regression-R

OHSU

Issue

- We have binary data
- We want to be able to run something like a regression
- But where we model the probability of the outcome

Issue and solution

- Probabilities are limited between 0 and 1
- To use a linear model we need to transform the probabilities so that they range from $-\infty$ (infinity) to $+\infty$

OHSU

Odds

- Transform the probabilities
  - From $0\ to\ 1$
  - To $0\ to\ \infty$
- Can do this by converting probabilities to odds, expressed as a ratio

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

OHSU

Odds

- Probabilities and odds are two equivalent ways of expressing the same idea

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

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

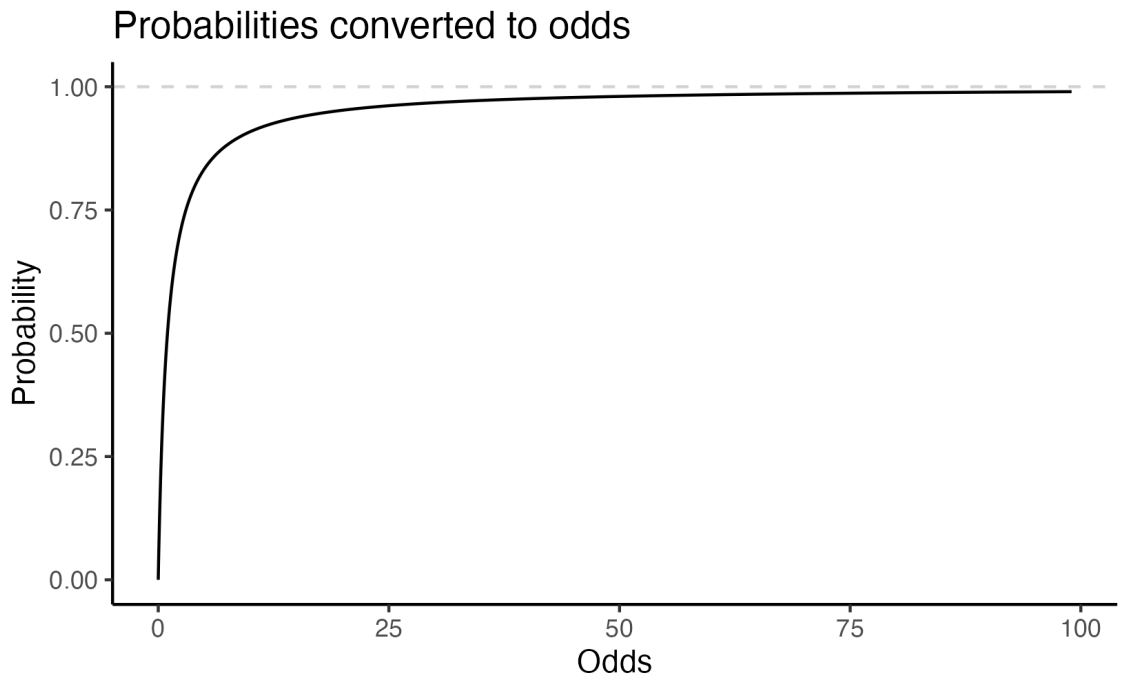
OHSU

Odds

- Say $p = 0.5$ then $odds = \frac{0.5}{1-0.5} = \frac{0.5}{0.5} = \frac{1}{1}$ or 1 $to 1$
- Say $p = 0.25$ then $odds = \frac{0.25}{1-0.25} = \frac{0.25}{0.75} = \frac{1}{3}$ or 1 $to 3$

OHSU

Odds


Probabilities converted to odds
1.00
0.75
0.50
0.25
0.00
Probability
0
25
50
75
100
Odds

Odds

- If we convert probability to odds, then the odds are always > zero
- For a linear model, we want our estimates to range from $-\infty$ to $+\infty$

OHSU

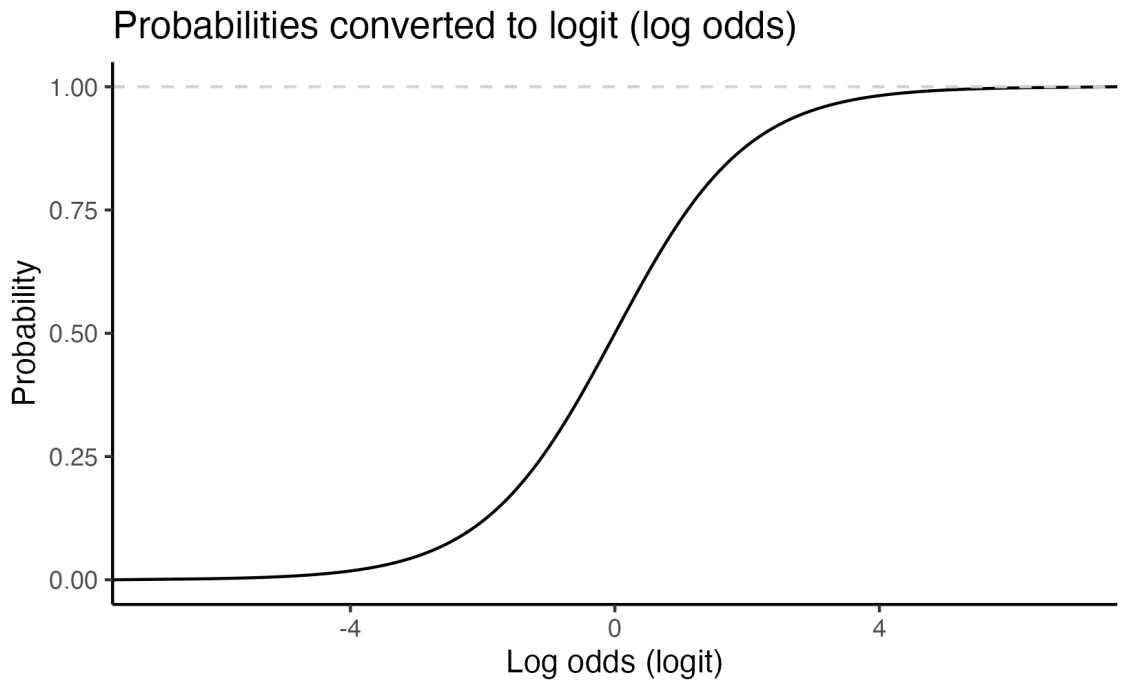
Log odds

- To solve this, we can take the logarithm of the odds to get our estimates on a scale $-\infty$ to $+\infty$
- Logarithm of the odds is called the **logit**

$$logit(p) = \log(odds) = log\left(\frac{p}{1-p}\right)$$

OHSU

Logit


Probabilities converted to logit (log odds)
1.00
0.75
0.50
0.25
0.00
Probability
-4
0
4
Log odds (logit)

OHSU

Logit

- By taking the log odds (logit) of the probability, our estimates on a scale $-\infty$ to $+\infty$

OHSU

Logistic regression model

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Random component follows a binomial distribution
- Systematic component
- Link function: logit link

OHSU

Logistic regression model

- Units are in log-odds
- In linear regression, a coefficient (β) is the change in the outcome for a unit change in the predictor.
- In logistic regression, a coefficient is the **change in log-odds of the outcome** being 1 for a unit change in the predictor.

OHSU

Logistic regression model

- Change in log-odds is the same as the ratio of odds.
- To get an odds ratio, need to transform the log-odds.
- Exponentiate the estimated model coefficient

$$Odds\ ratio = \exp(\beta) = e^{\beta}$$

https://benwhalley.github.io/just-enough-r/link-functions.html

OHSU


Simple logistic
regression

Hip fracture study

- A study to examine the relationship between hip fracture with age, sex, BMI, and bone mineral density (bone densitometries).

https://bookdown.org/tpinto_home/Regression-and-Classification/#datasets-used-in-the-examples

OHSU

Hip fracture study

- id: subject ID
- age: subject age in years
- sex: biological sex (Female/Male)
- fracture: hip fracture (Fracture/Nofracture)
- weight_kg: weight in kg
- height_cm: height in cm
- medication: prescribed medications
- waiting_time: time the subject waited for the densitometry (in minutes)
- bmd: bone mineral density measure in the hip
- bmi: calculated body mass index
- bmi_c: Body mass index category (Underweight, Healthy, Overweight, Obese)

https://bookdown.org/tpinto_home/Regression-and-Classification/#datasets-used-in-the-examples

OHSU

- N = 169; First 20 shown below

id	age	sex	fracture	weight_kg	height_cm	medication	waiting_time	bmd	bmi	bmi_c
00469	57.05	Female	No	64	155.5	Anticonvulsant	18	0.8793	26.46788	Overweight
08724	75.74	Female	No	78	162.0	No medication	56	0.7946	29.72108	Overweight
06736	70.78	Male	No	73	170.5	No medication	10	0.9067	25.11158	Overweight
24180	78.25	Female	No	60	148.0	No medication	14	0.7112	27.39226	Overweight
17072	54.19	Male	No	55	161.0	No medication	20	0.7909	21.21832	Healthy
03806	77.18	Male	No	65	168.0	No medication	7	0.7301	23.03005	Healthy
17106	56.18	Male	No	77	159.0	No medication	26	1.0096	30.45766	Obese
23834	49.92	Female	No	59	150.0	No medication	9	0.7310	26.22222	Overweight
02454	68.41	Male	No	64	167.0	Glucocorticoids	6	0.6893	22.94812	Healthy
02088	66.26	Male	No	72	159.5	No medication	10	0.9466	28.30161	Overweight
05364	45.87	Male	No	62	169.0	No medication	12	0.8015	21.70792	Healthy
08922	73.97	Female	No	68	164.0	No medication	5	0.5793	25.28257	Overweight
23890	60.56	Female	No	76	155.0	No medication	11	0.9760	31.63371	Obese
03047	64.21	Male	No	90	175.0	Glucocorticoids	28	0.9184	29.38776	Overweight
02179	53.40	Male	No	70	162.5	No medication	73	0.8020	26.50888	Overweight
03800	66.83	Male	No	76	171.0	No medication	13	0.8033	25.99090	Overweight
07528	57.93	Male	No	67	160.0	Glucocorticoids	5	0.7978	26.17187	Overweight
05288	40.23	Male	No	66	165.0	No medication	8	1.0390	24.24242	Healthy
00109	69.05	Female	No	72	154.0	No medication	11	0.7861	30.35925	Obese
08622	57.80	Female	No	50	152.0	No medication	21	0.8254	21.64127	Healthy

OHSU

Hip fracture Study

- Primary endpoint (dependent variable): fracture
  - fracture = “Yes”: patient had a hip fracture
  - fracture = “No”: patient did not have hip fracture

OHSU

Fracture and age

- Research question #1:
  - Is there a relationship between a hip fracture occurring and age?

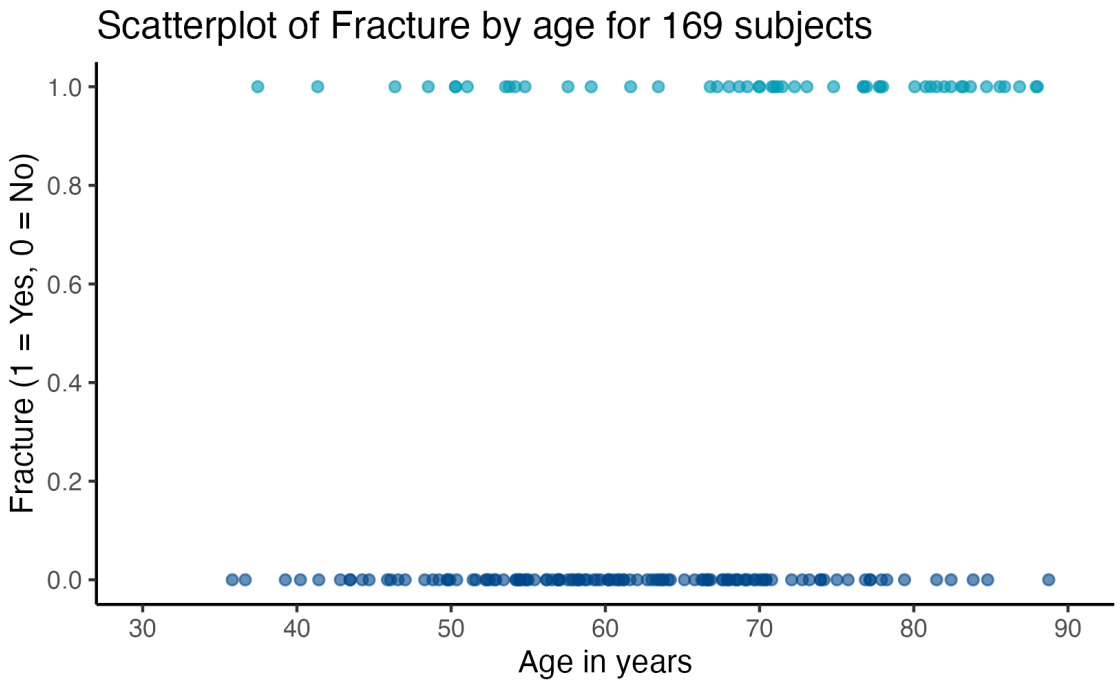
OHSU

Fracture and age

fracture	n	percent
Yes	50	29.6%
No	119	70.4%
Total	169	100.0%

fracture	variable	n	complete	missing	mean	sd	p0	p25	p50	p75	p100	range
All subjects	age	169	169	0	63.63	12.36	35.81	54.42	63.49	72.08	88.75	52.94
Yes	age	50	50	0	69.77	13.38	37.46	59.72	71.32	81.01	88.02	50.56
No	age	119	119	0	61.05	10.97	35.81	54.21	60.56	68.56	88.75	52.94

Fracture and age



Fracture and age

$$logit(p) = \beta_0 + \beta_1 Age$$

- p = probability of Fracture
- β_1 determines whether the curve ascends or descends
  - $\beta_1 > 0$, then p increases as Age increases
  - $\beta_1 < 0$, then p decreases as Age increases
  - $\beta_1 = 0$, then p is independent of Age

OHSU

Fracture and age – log-odds

term	estimate	lower_ci	upper_ci	p_value
(Intercept)	-5.0	-7.2	-3.0	< 0.001
age	0.1	0.0	0.1	< 0.001

- Interpretation: For every one unit increase in age, the log-odds of fracture occurring increases on average by 0.1 units (95% CI: 0.0 to 0.1 units increase in log-odds). There is a statistically significant association between hip fracture and age ($P < 0.001$).

OHSU

Fracture and age – log-odds

```
> res <- glm((fracture == "Yes") ~ age,
+     family = binomial(link = "logit"),
+     data = data)
> summary(res)

Call:
glm(formula = (fracture == "Yes") ~ age, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4444  -0.8553  -0.6216   1.0699   2.3290

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.01863    1.07672  -4.661 3.15e-06 ***
age          0.06341    0.01583   4.007 6.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 205.27  on 168  degrees of freedom
Residual deviance: 186.76  on 167  degrees of freedom
AIC: 190.76

Number of Fisher Scoring iterations: 4
```


OHSU

Fracture and age – log-odds

```
. logit fracture age

Iteration 0:   log likelihood = -102.63604
Iteration 1:   log likelihood = -93.602157
Iteration 2:   log likelihood = -93.378565
Iteration 3:   log likelihood =  -93.37797
Iteration 4:   log likelihood =  -93.37797

Logistic regression                             Number of obs     =        169
                                                LR chi2(1)        =      18.52
                                                Prob > chi2       =     0.0000
Log likelihood =  -93.37797                     Pseudo R2         =     0.0902

------------------------------------------------------------------------------
    fracture |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0634076   .0158255     4.01   0.000     .0323902     .094425
       _cons |  -5.018625    1.07672    -4.66   0.000    -7.128958   -2.908293
------------------------------------------------------------------------------
```


OHSU

Fracture and age – odds ratio

$$logit(p) = \beta_0 + \beta_1 Age$$

$$Odds\ ratio = OR = e^{\beta_1}$$

- $OR > 1$, then greater likelihood of outcome (hip fracture) as age increases
- $OR < 1$, then lesser likelihood of outcome as age increases
- $OR = 1$, then the same likelihood of outcome as age increases

OHSU

Fracture and age –odds ratio

term	estimate	lower_ci	upper_ci	p_value
(Intercept)	0.0	0.0	0.0	< 0.001
age	1.1	1.0	1.1	< 0.001

- Interpretation: For every one unit increase in age, the odds of fracture occurring increases 1.1 times (95% CI: 1.0 to 1.1 times). There is a statistically significant association between hip fracture and age ($P < 0.001$).

OHSU

Fracture and age –odds ratio

```
> res <- glm((fracture == "Yes") ~ age,
+     family = binomial(link = "logit"),
+     data = data)
> exp(cbind(OR = coef(res), confint(res)))
Waiting for profiling to be done...
                    OR        2.5 %    97.5 %
(Intercept) 0.00661361 0.0007122222 0.0494659
age         1.06546105 1.0341146816 1.1006313
```


OHSU

Fracture and age –odds ratio

```
. logit fracture age, or

Iteration 0:   log likelihood = -102.63604
Iteration 1:   log likelihood = -93.602157
Iteration 2:   log likelihood = -93.378565
Iteration 3:   log likelihood =  -93.37797
Iteration 4:   log likelihood =  -93.37797

Logistic regression                             Number of obs     =        169
                                                LR chi2(1)        =      18.52
                                                Prob > chi2       =     0.0000
Log likelihood =  -93.37797                     Pseudo R2         =     0.0902

------------------------------------------------------------------------------
    fracture | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.065461   .0168614     4.01   0.000      1.03292    1.099027
       _cons |   .0066136    .007121    -4.66   0.000     .0008016    .0545688
------------------------------------------------------------------------------
```


OHSU

Fracture and sex

- Research question #2:
  - Is there a relationship between a hip fracture occurring and biological sex?

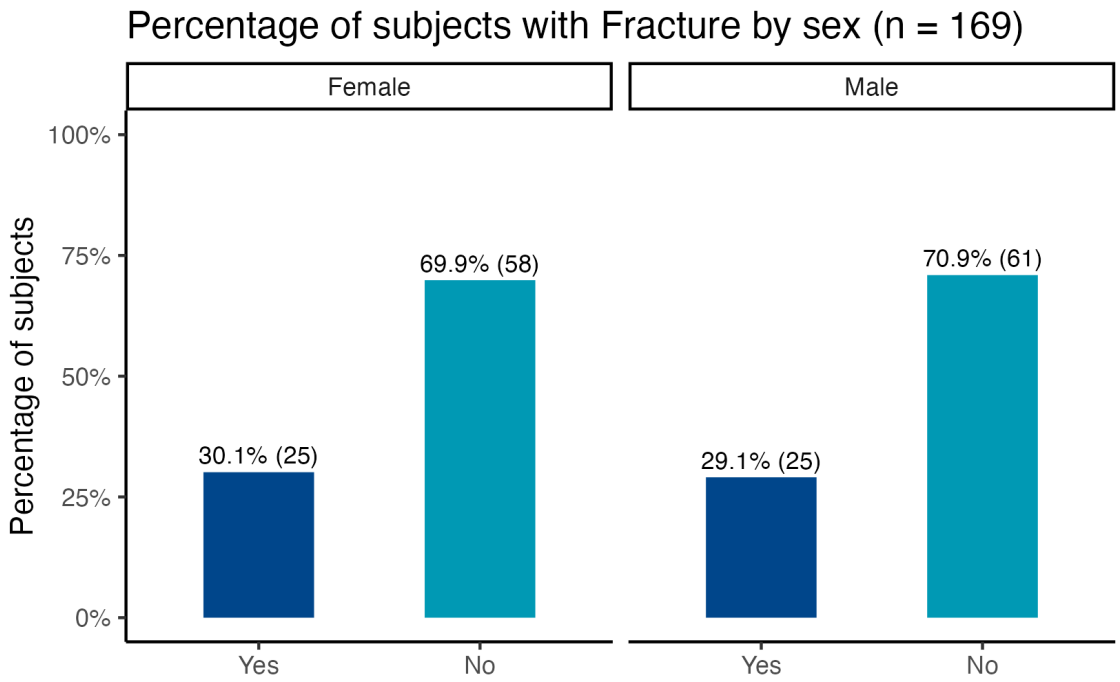
OHSU

Fracture and sex

fracture/sex	Male	Female	Total
Yes	25 (29.1%)	25 (30.1%)	50 (29.6%)
No	61 (70.9%)	58 (69.9%)	119 (70.4%)

OHSU

Fracture and sex



Fracture and sex

$$logit(p) = \beta_0 + \beta_1 Sex$$

- p = probability of Fracture
- β_1 determines whether the curve ascends or descends
  - $\beta_1 > 0$, then p higher in males than females
  - $\beta_1 < 0$, then p lower in males than females
  - $\beta_1 = 0$, then p the same in males than females

OHSU

Fracture and sex – odds ratio

$$logit(p) = \beta_0 + \beta_1 sex$$

$$Odds\ ratio = OR = e^{\beta_1}$$

- $OR > 1$, then greater likelihood of outcome (hip fracture) as in males than females
- $OR < 1$, then lesser likelihood of outcome in males than females
- $OR = 1$, then the same likelihood of outcome in males than females

OHSU

Fracture and sex

term	level	reference	estimate	lower_ci	upper_ci	p_value
(Intercept)	(Intercept)	Non-Baseline Category	0.4	0.3	0.7	< 0.001
sex	Female	Baseline Category				
sex	Male	Non-Baseline Category	1.0	0.5	1.8	0.881

- Interpretation: The odds of hip fracture among male subjects is 1.0 times (95% CI: 0.5 to 1.8 times) the odds of hip fracture among female subjects. There is no statistically significant association between hip fracture and sex ($P < 0.881$).

OHSU

Fracture and sex

```
> res <- glm((fracture == "Yes") ~ sex,
+            family = binomial(link = "logit"),
+            data = data)
> summary(res)

Call:
glm(formula = (fracture == "Yes") ~ sex, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8466  -0.8466  -0.8288   1.5492   1.5719

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.84157    0.23925  -3.517 0.000436 ***
sexMale     -0.05043    0.33710  -0.150 0.881078
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 205.27  on 168  degrees of freedom
Residual deviance: 205.25  on 167  degrees of freedom
AIC: 209.25

Number of Fisher Scoring iterations: 4

> ## odds ratios and 95% CI
> exp(cbind(OR = coef(res), confint(res)))
Waiting for profiling to be done...
                   OR     2.5 %    97.5 %
(Intercept) 0.4310345 0.2651600 0.6804247
sexMale     0.9508197 0.4898953 1.8448256
```


OHSU

Fracture and sex

```
. logit fracture i.sex, or

Iteration 0:   log likelihood = -102.63604
Iteration 1:   log likelihood = -102.62485
Iteration 2:   log likelihood = -102.62485

Logistic regression                             Number of obs     =        169
                                                LR chi2(1)        =       0.02
                                                Prob > chi2       =     0.8811
Log likelihood = -102.62485                     Pseudo R2         =     0.0001

------------------------------------------------------------------------------
    fracture | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |   .9508197   .3205191    -0.15   0.881     .4910889    1.840925
       _cons |   .4310345   .1031257    -3.52   0.000     .2696874    .6889115
------------------------------------------------------------------------------
```


OHSU

Reference levels

- For interpretation it's vital to know which level of a variable is the Reference Level.
- The level of a categorical variable considered the "baseline" or "usual" value
- Other levels are compared to the Reference Level

Reference levels

- In the Fracture and Sex example, *Female* is the reference level.
- The models give us an odds ratio for the *Male* level.
- Odds ratio is in terms of *Male* compared to *Female*.

Reference levels

- When in doubt, check it.
- Can do this “by hand” with a 2x2 table.

OHSU

Odds ratio – 2x2 table

fracture/sex	Male	Female	Total
Yes	25 (29.1%)	25 (30.1%)	50 (29.6%)
No	61 (70.9%)	58 (69.9%)	119 (70.4%)

		Predictor	
		Yes	No
Outcome	Yes	a	b
	No	c	d

OHSU

Odds ratio – 2x2 table

- a = 25
- b = 25
- c = 61
- d = 58

fracture/sex	Male	Female	Total
Yes	25 (29.1%)	25 (30.1%)	50 (29.6%)
No	61 (70.9%)	58 (69.9%)	119 (70.4%)

OHSU

Odds ratio – 2x2 table

- a = 25
- b = 25
- c = 61
- d = 58

fracture/sex	Male	Female	Total
Yes	25 (29.1%)	25 (30.1%)	50 (29.6%)
No	61 (70.9%)	58 (69.9%)	119 (70.4%)

- $Odds\ ratio = OR = (a/b)/(c/d) = ad/bc$
- $OR = (25/25)/(61/58) = (25 * 58)/(25 * 61) = 0.95$

OHSU

Odds ratio – 2x2 table

		Predictor: Yes	Predictor: No
Outcome	Yes	a	b
	No	c	d

- The odds of [outcome = “Yes”] among those [predictor = “Yes”] is [odds ratio] times the odds of those with [predictor = “No].

OHSU

Odds ratio – 2x2 table

sex/fracture	Yes	No	Total
Male	25 (50.0%)	61 (51.3%)	86 (50.9%)
Female	25 (50.0%)	58 (48.7%)	83 (49.1%)

		Outcome	
		Yes	No
Predictor	Yes	a	b
	No	c	d

OHSU

Odds ratio – 2x2 table

- a = 25
- b = 61
- c = 25
- d = 58

sex/fracture	Yes	No	Total
Male	25 (50.0%)	61 (51.3%)	86 (50.9%)
Female	25 (50.0%)	58 (48.7%)	83 (49.1%)

- $OR = (25/61)/(25/58) = (25 * 58)/(61 * 25) = 0.95$
- Same result as before

OHSU

Change the reference level – 2x2 table

- a = 25
- b = 25
- c = 58
- d = 61
- $Odds\ ratio = OR = (a/b)/(c/d) = ad/bc$

fracture/sex	Female	Male	Total
Yes	25 (30.1%)	25 (29.1%)	50 (29.6%)
No	58 (69.9%)	61 (70.9%)	119 (70.4%)

- $OR = (25/25)/(58/61) = (25 * 61)/(25 * 58) = 1.05$
- Different than before! Pay attention to the reference level.

OHSU

Change the reference level – LR Model

term	level	reference	estimate	lower_ci	upper_ci	p_value
(Intercept)	(Intercept)	Non-Baseline Category	0.4	0.3	0.6	< 0.001
sex	Male	Baseline Category				
sex	Female	Non-Baseline Category	1.1	0.5	2.0	0.881

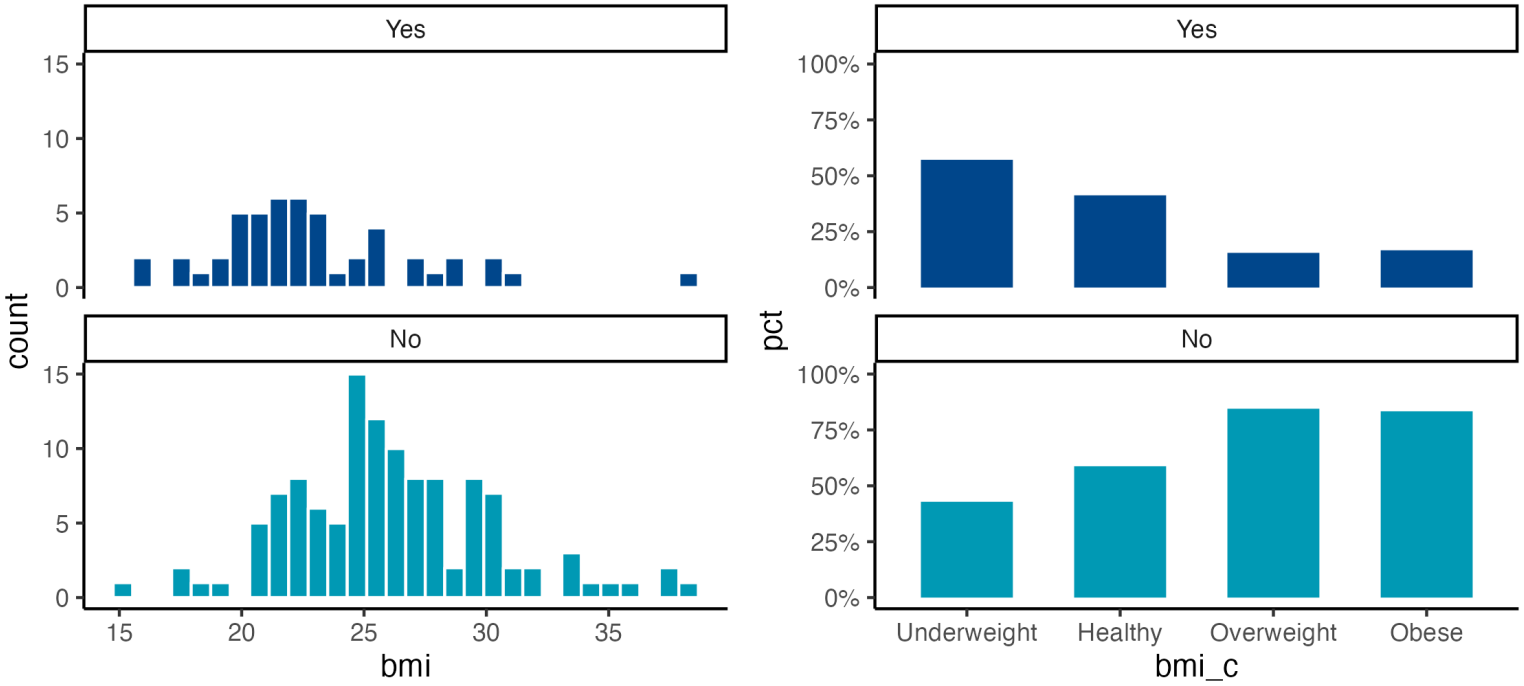
OHSU

Fracture and BMI

- Research question #2:
  - Is there a relationship between a hip fracture occurring and BMI?

OHSU

BMI – continuous or categorical?



OHSU

Fracture and BMI

- Categorical BMI
- “Healthy” is the reference level

OHSU

Fracture and BMI

term	level	reference	estimate	lower_ci	upper_ci	p_value
(Intercept)	(Intercept)	Non-Baseline Category	0.7	0.4	1.1	0.119
bmi_c	Underweight	Non-Baseline Category	1.9	0.4	10.2	0.421
bmi_c	Healthy	Baseline Category				
bmi_c	Overweight	Non-Baseline Category	0.3	0.1	0.6	0.002
bmi_c	Obese	Non-Baseline Category	0.3	0.1	0.8	0.034

Fracture and BMI

- Interpret relative to the Reference Level
  - Compared to Healthy subjects, Underweight subjects were 1.9 times (95% CI: 0.4 to 10.2 times) more likely to have hip fracture (P = 0.421).
  - The odds of hip fracture among Overweight subjects was 0.3 times (95% CI: 0.1 to 0.6 times) the odds of hip fracture among Healthy subjects. Overweight subjects were less likely to experience hip fracture (P = 0.002).

OHSU

Fracture and BMI

```
> data$bmi_c <- relevel(data$bmi_c, ref = "Healthy")
> res <- glm((fracture == "Yes") ~ bmi_c,
+            family = binomial(link = "logit"),
+            data = data)
> summary(res)

Call:
glm(formula = (fracture == "Yes") ~ bmi_c, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3018  -1.0314  -0.5807   1.3308   1.9304

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.3536     0.2271  -1.557  0.11944
bmi_cUnderweight  0.6413     0.7968   0.805  0.42090
bmi_cOverweight  -1.3410     0.4279  -3.134  0.00173 **
bmi_cObese       -1.2558     0.5929  -2.118  0.03418 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 205.27  on 168  degrees of freedom
Residual deviance: 189.69  on 165  degrees of freedom
AIC: 197.69

Number of Fisher Scoring iterations: 4

> ## odds ratios and 95% CI
> exp(cbind(OR = coef(res), confint(res)))
Waiting for profiling to be done...
                        OR      2.5 %     97.5 %
(Intercept)      0.7021277 0.44643887  1.0914867
bmi_cUnderweight 1.8989899 0.39391648 10.1755229
bmi_cOverweight  0.2615955 0.10772680  0.5853194
bmi_cObese       0.2848485 0.07743896  0.8363863
```


OHSU

Fracture and BMI

```
. logit fracture ib1.bmi_c2, or

Iteration 0:   log likelihood = -102.63604
Iteration 1:   log likelihood = -95.007409
Iteration 2:   log likelihood = -94.846001
Iteration 3:   log likelihood = -94.845749
Iteration 4:   log likelihood = -94.845749

Logistic regression                             Number of obs     =        169
                                                LR chi2(3)        =      15.58
                                                Prob > chi2       =     0.0014
Log likelihood = -94.845749                     Pseudo R2         =     0.0759

------------------------------------------------------------------------------
    fracture | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      bmi_c2 |
      Obese  |   .2848485   .1688985    -2.12   0.034     .0891048    .9105979
 Overweight  |   .2615955   .1119369    -3.13   0.002     .1130839    .6051457
Underweight  |    1.89899   1.513142     0.80   0.421     .3983576    9.052576
             |
       _cons |   .7021277   .1594613    -1.56   0.119     .4498818    1.095806
------------------------------------------------------------------------------
```


OHSU

Odds and ends

- Relative Risk
  - In case control studies and cohort studies where the outcome occurs in less than 10% of the exposed population, Odds Ratio provides a reasonable approximation to the Relative Risk

Odds ratios and risk ratios: what's the difference and why does it matter? (Viera)

Odds and ends

- Association vs. Prediction
  - Models we have seen here are focused on associations – relationships and interpretations
  - Prediction/classification – aims to predict. Topic for another time

Classification versus association models: Should the same methods apply? (Feng)

OHSU

Odds and ends

- Sample size (guideline)
  - Let p be the smallest proportion of negative or positive cases in the population
  - Let k be the number of independent variables to include
  - Then the minimum number of cases to include is

$$N = \frac{10k}{p}$$

- Example: 5 explanatory variables, proportion of positive cases in the population is 25%), then $N = 10 \times 5 / 0.25 = 200$
- If less than 100, increase to 100

[An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain (Park 2013)]

OHSU

Odds and ends

- Univariable – one explanatory variable
- Univariate – One response variable

- Multivariable – more than one explanatory variable
- Multivariate – more than one response variable

Multivariate or Multivariable Regression? (Hidalgo and Goodman)

OHSU

References

- Understanding logistic regression analysis (Sperandei)
- Using logistic regression in perinatal epidemiology: an introduction for clinical researchers. Part 1: basic concepts (Brand, Keirse)
- An introduction to logistic regression with an application to the analysis of language recovery following a stroke (Greenhouse et al)
- Explaining Odds Ratios (Szumilas)
- Penn State, STAT 504, Binary Logistic Regression

OHSU

Thank You