Categorical Data Analysis - Simple Logistic Regression

Alessio Crippa

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Summary recap

For a binary outcome Y, we considered two *absolute* measures (of disease occurrence):

- p the probability ($E[Y]$)
- the odds ($E[Y]/(1 - E[Y])$)

If we want to relate the outcome Y with a binary predictor X we considered two *relative* measures (of association):

- RR the relative risk (p_1/p_0)
- the OR odds ratio ($\mathrm{odds}_1/\mathrm{odds}_0$)

What were the corresponding measures in linear regression?

The linearity assumption between the probability $P(Y|X)$ and X is most often not appropriate.

We assume a logistic function to describe an S-shaped relation

$\text{logistic}(x) = \frac{e^x}{1+e^x}$

The empty model to estimate the odds or probability

$P(Y) = \frac{exp(\beta_0)}{1+exp(\beta_0)}$

$\log(\text{odds}(Y)) = \beta_0$

Binary predictor

Binary predictor

We want to relate the probability of the outcome Y to a binary predictor X, i.e. we want to test if the probability of the outcome is different in the two groups defined by X ($X = 1$ vs $X = 0$)

$$\text{logit}\left(P[Y|X]\right) = \log\left(\mathrm{odds}(Y|X)\right) = \beta_0 + \beta_1 X$$

Or alternatively

$$P\left(Y|X\right) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

The model assumes that the **log odds** of the outcome *linearly* depends on X.

Or alternatively, that the probability of the outcome depends of X following a logistic function.

In general, all the modeling techniques apply to the linear model (log odds), while the results will be most often presented in terms of odds (or probability) and odds ratios (or risk ratio).

Interpretation

$$\log\left(\text{odds}(Y|X)\right) = \beta_0 + \beta_1 X$$

β_0 is the log odds of the outcome when $X = 0$.

What about β_1?

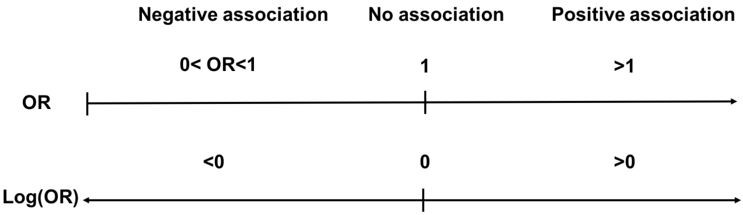
$$\beta_1 = \log\left(\text{odds}(Y|X=1)\right) - \log\left(\text{odds}(Y|X=0)\right) =$$
$$\log\left(\frac{\text{odds}(Y=1|X=1)}{\text{odds}(Y=1|X=0)}\right) = \log(OR)$$

β_1 is the the log odds ratio of the outcome comparing $X = 1$ vs. $X = 0$.

Interpretation of model coefficients is more informative on the exponential scale.

$\exp(\beta_0)$ is the odds of the outcome when $X = 0$.

$\exp(\beta_1)$ is the odds ratio of the outcome comparing $X = 1$ vs. $X = 0$.



Question: Is sex (`female`) a predictor of (the risk of) hyponatremia (`nas135`)?

Or, is the risk of hyponatremia different between men and women?

$$\log\left(\mathrm{odds}(\mathrm{nas135}|\mathrm{female})\right) = \beta_0 + \beta_1 \mathrm{female}$$

Estimation

We can use maximum likelihood to estimate the β coefficients.

$Y \sim \mathrm{Bernoulli}(p(x))$, $p(x) = P(Y|X = x)$.
The likelihood is defined as

$$\mathcal{L}\left(p|y, x\right) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

or alternatively the log-likelihood

$$\ell\left(p|y, x\right) = \sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))$$

$$p(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

so the log-likelihood becomes

$$\ell(\beta|y, x) = \\ \sum_{i=1}^{n} y_i \log\left(\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}\right) + (1 - y_i) \log\left(1 - \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}\right)$$

The $\hat{\beta}_0$ and $\hat{\beta}_1$ which maximizes $\ell(\beta|y, x)$ can be obtained using iterative algorithms (there is no close formula).

Properties of MLE

The MLE $\hat{B}$ has the following nice properties:

1. Consistency: the distributions of the estimators become more and more concentrated near the true value of the parameter being estimated.
2. Asymptotically normal: $\hat{B} \sim N(\beta, \text{Var}(\hat{B}))$
3. Asymptotic optimality: MLE has the smallest asymptotic variance and we say that the MLE is asymptotically efficient.
4. Invariance property: the maximum likelihood estimate of a function of the parameter being estimated ($\tau = g(\beta)$) is the function evaluated at the maximum likelihood estimate of the parameter ($\hat{\tau} = g(\hat{\beta})$)

Hypothesis testing

If sex is not associated with (risk of) hyponatremia, it means that $p_1 = p_0$, and thus $\mathrm{odds}_1 = \mathrm{odds}_0$. In this case, the *OR* will be 1.

The null hypothesis of no association between sex and (risk of) hyponatremia can be written as $H_0 : \beta_1 = 0$ (if $\beta_1 = \log(OR) = 0$, in means that $OR = 1$).

Based on large sample, the distribution of betas is approximately normal and Z test can be adopted.

$$Z = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

which, under the null hypothesis, follows a standard normal distribution.

Confidence intervals for the OR

95% confidence intervals are first defined on the log scale (i.e. for the $\log(OR)$)

$\hat{\beta}_1 \pm 1.96 \cdot \mathrm{SE}(\hat{\beta}_1)$

and are exponentiated to obtain the corresponding confidence intervals for the OR (invariance property)

$\exp\left(\hat{\beta}_1 \pm 1.96 \cdot \mathrm{SE}(\hat{\beta}_1)\right)$

```r
mod <- glm(nas135 ~ female, data = marathon, family = "binomial")
summary(mod)
```

```
Call:
glm(formula = nas135 ~ female, family = "binomial", data = marathon)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7102  -0.7102  -0.4020  -0.4020   2.2608

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)   -2.4749     0.2082 -11.884   < 2e-16
femalefemale   1.2260     0.2795   4.386 0.0000116

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 371.60  on 487  degrees of freedom
Residual deviance: 351.93  on 486  degrees of freedom
AIC: 355.93

Number of Fisher Scoring iterations: 5
```

Interpretation

Alternative ways:

- the log odds of hyponatremia for male runners is -2.47;
- the odds of hyponatremia is 0.08 (8 cases for every 100 non-cases) among men;
- the log odds ratio of hyponatremia comparing female vs male runners is 1.23;
- the odds ratio of hyponatremia comparing female vs male runners is 3.41;
- the odds of hyponatremia among women is 3.41 times the odds for men;
- you multiply by 3.41 the odds of hyponatremia among male (0.08) to get the one among female (-1.24).

The 95% CI for the $\log(OR)$ is calculated as
$1.23 \pm 1.96 \cdot 0.28 = (0.68, 1.77)$

The 95% CI for the OR is calculated as
$\exp(0.68, 1.77) = (1.97, 5.87)$

The odds of hyponatremia among women was significantly higher than men ($OR = 3.41$).
We are 95% confident that the odds ratio relating sex (being woman compared to man) to hyponatremia is between 1.97 and 5.87.

The multiplicative model

The logistic model is a linear model in terms of the log odds
$\log(\mathrm{odds}(Y|X)) = \beta_0 + \beta_1$

If we take the exponential, we can write the previous model as a multiplicative model in terms of the odds
$\exp\left(\log(\mathrm{odds}(Y|X))\right) = \mathrm{odds}(Y|X) = \exp\left(\beta_0 + \beta_1\right) = \exp(\beta_0)\exp(\beta_1) = \mathrm{odds}(Y|X=0) \cdot OR_{x=1\mathrm{vs}x=0}$

What is the odds of hyponatremia among men? $\exp(\hat{\beta}_0) = 0.08$

What is the odds of hyponatremia among women?
$\exp(\hat{\beta}_0) \cdot \exp(\hat{\beta}_1) = 0.08 \cdot 3.41 = 0.27$

Linear function of regression coefficients

Question: What is the odds of hyponatremia among women?

Coefficients:

```
    b0     b1
-2.475  1.226
```

Covariance matrix:

```
       b0     b1
b0  0.043 -0.043
b1 -0.043  0.078
```

Let's first calculate the 95% CI for the *log odds*

$$\widehat{\mathrm{est}} = \log(\mathrm{odds}(Y|X=1)) = \hat{\beta}_0 + \hat{\beta}_1 = -2.47 + (1.23) = -1.249$$

$$\mathrm{Var}\left(\widehat{\mathrm{est}}\right) = \mathrm{Var}(\hat{\beta}_0) + \mathrm{Var}(\hat{\beta}_1) + 2\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) =$$
$$0.043 + 0.078 + 2(-0.043) = 0.035$$

95% confidence interval for $\widehat{\mathrm{est}}$ is

$$\widehat{\mathrm{est}} \pm 1.96\sqrt{\mathrm{Var}\left(\widehat{\mathrm{est}}\right)} = (-1.616, -0.882)$$

Now we can exponentiate both the point estimate and the confidence interval:

$\exp\left(\hat{\beta}_0 + \hat{\beta}_1\right) = 0.29$

95% confidence interval for $\mathrm{odds}(Y|X = 1)$ is $\exp(-1.616, -0.882) = (0.199, 0.414)$

The odds of hyponatremia is 0.29 (29 cases for every 100 non-cases) among women (95% CI: $0.199, 0.414$).

Predicted probabilities

Once the β coefficients have been estimated it is possible to calculate the predicted probabilities of the outcome for any covariate values (covariate pattern).

What is the estimated probability of hyponatremia among male female?

$P(Y = 1|X = 0) = \text{invlogit}(\beta_0) = \frac{\exp(-2.47)}{1+\exp(-2.47)} = \frac{0.08}{1+0.08} = 0.078$

$P(Y = 1|X = 1) = \text{invlogit}(\beta_0 + \beta_1) = \frac{\exp(-2.47+1.23)}{1+\exp(-2.47+1.23)} = \frac{0.29}{1+0.29} = 0.224$

The 95% CI for the predicted probabilities can be obtained as the `invlogit` of the 95% CI for the corresponding *log odds*

95% CI for $\log(\mathrm{odds}(Y|X=0))$ (−> $\hat{\beta}_0$)
$$\hat{\beta}_0 \pm z_{0.975}\mathrm{SE}(\hat{\beta}_0) = -2.47 \pm 1.96\sqrt{0.04} = (-2.86, -2.08)$$

95% CI for $P(Y|X=0)$
$$\left(\frac{\exp(-2.86)}{1+\exp(-2.86)}, \frac{\exp(-2.08)}{1+\exp(-2.08)}\right) = (0.05, 0.11)$$

```
      b0    b1
b0  0.04 -0.04
b1 -0.04  0.08
```

95% CI for $\log(\text{odds}(Y|X=1))$ (−>$\hat{\beta}_0 + \hat{\beta}_1$)

$\text{Var}(\beta_0 + \beta_1) = \text{Var}(\hat{\beta}_0) + \text{Var}(\hat{\beta}_1) + 2\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) =$

$0.04 + 0.08 + 2(-0.04) = 0.04$

$(\hat{\beta}_0 + \hat{\beta}_1) \pm z_{0.975}\text{SE}(\hat{\beta}_0 + \hat{\beta}_1) = -1.24 \pm 1.96\sqrt{0.04} = (-1.63, -0.85)$

95% CI for $P(Y|X=1)$

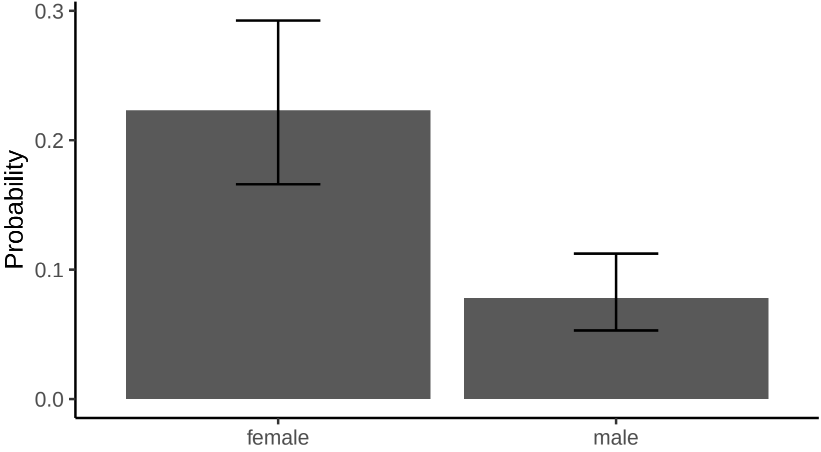
$\left(\frac{\exp(-1.63)}{1+\exp(-1.63)}, \frac{\exp(-0.85)}{1+\exp(-0.85)}\right) = (0.16, 0.3)$

Predicted response

X	odds	p	OR	RR
0	$\exp(\beta_0)$	$\text{invlogit}(\beta_0)$	ref	ref
1	$\exp(\beta_0+\beta_1)$	$\text{invlogit}(\beta_0+\beta_1)$	$\exp(\beta_1)$	p_1/p_0

X	odds	p	OR	RR
male	0.08	0.078	ref	ref
female	0.27	0.224	3.41	2.878

Graphical presentation of predicted probabilities



The logistic regression model estimated above with a binary covariate (female) is called saturated because the number of possible combination of covariate patterns (male, female) is equal to the number of parameters estimated.

cases	non-cases	female
25	297	male
37	129	female

The consequence is that the fitted values from the saturated model will exactly fit the observed data. No model is more complicated than that.

cases	non-cases	female	p obs	p pred
25	297	male	0.0776398	0.0776398
37	129	female	0.2228916	0.2228916

Continuous predictor

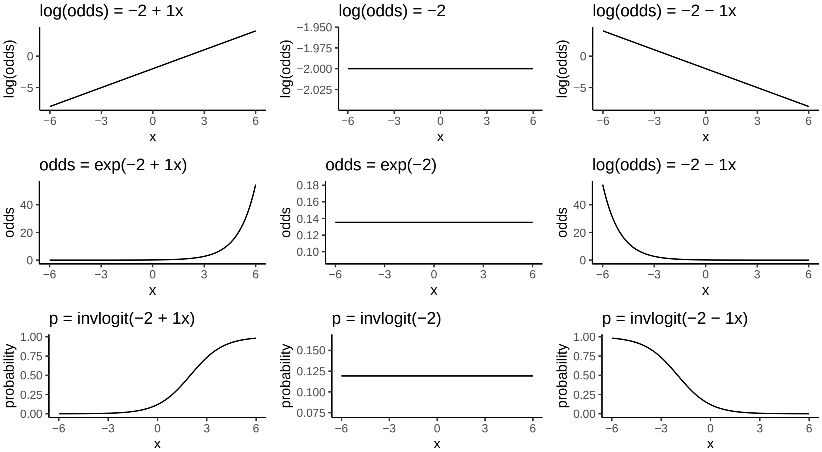
We want know to related the probability of the outcome Y to a continuous predictor X.

Question: is weight change (`wtdiff`) a predictor of the (risk of) hyponatremia (`nas135`)?

Or alternatively, does the risk oh hyponatremia vary as a function of weight change? if yes, how does it vary?

$$\log\left(\mathrm{odds}(\mathrm{nas135}|\mathrm{wtdiff})\right) = \beta_0 + \beta_1 \mathrm{wtdiff}$$

NB: we assume that the log odds of hyponatremia *linearly* varies as a function of weight change.


log(odds) = −2 + 1x
log(odds)
0
−5
−6 −3 0 3 6
x
log(odds) = −2
−1.950
−1.975
−2.000
−2.025
log(odds) = −2 − 1x
odds = exp(−2 + 1x)
odds
40
20
0
odds = exp(−2)
0.18
0.16
0.14
0.12
0.10
log(odds) = −2 − 1x
p = invlogit(−2 + 1x)
probability
1.00
0.75
0.50
0.25
0.00
p = invlogit(−2)
0.150
0.125
0.100
0.075
p = invlogit(−2 − 1x)

```r
mod <- glm(nas135 ~ wtdiff, data = marathon, family = "binomial")
summary(mod)
```

```

Call:
glm(formula = nas135 ~ wtdiff, family = "binomial", data = marathon)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6634  -0.5317  -0.3584  -0.2332   2.7314

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8850     0.1587 -11.880  < 2e-16
wtdiff        0.7284     0.1103   6.603 4.02e-11

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 343.35  on 454  degrees of freedom
Residual deviance: 288.94  on 453  degrees of freedom
  (33 observations deleted due to missingness)
AIC: 292.94

Number of Fisher Scoring iterations: 5
```

Interpretation

$\exp(\beta_0) = 0.15$ is the odds of hyponatremia for those runner who did not change weight.

$$\log\left(\text{odds}(Y|X = x+1)\right) - \log\left(\text{odds}(Y|X = x)\right) =$$
$$\log\left(\frac{\text{odds}(Y=1|X=x+1)}{\text{odds}(Y=1|X=x)}\right) = \log(OR) = \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x = \beta_1$$

Every one-unit increase in X, the log(OR) increases/decreases by β_1.

More in general, the log odds of the outcome for any two values of X (x_1 vs x_2) are

$\log(\mathrm{odds}(Y|X = x_1)) = \beta_0 + \beta_1 x_1$

$\log(\mathrm{odds}(Y|X = x_2)) = \beta_0 + \beta_1 x_2$

$\log(\mathrm{odds}(Y|X = x_1)) - \log(\mathrm{odds}(Y|X = x_2)) = \log(OR_{x1 \text{ vs } x_2}) = \beta_1(x_1 - x_2)$

is the log OR associated with a $(x_1 - x_2)$ change in X.

The odds ratio can be obtained exponentiating the log OR

$$OR = \exp(\beta_1(x_1 - x_2))$$

That is the OR comparing the sub-population having x_1 vs. x_2 of the quantitative covariate X.

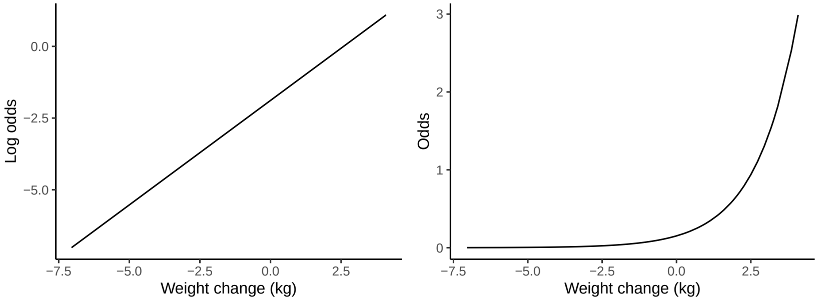
NB: The log(odds) is a linear function of the beta coefficients. In that sense, every one-unit change in X is associate with an increase in the log OR equal to β_1.

$$\log(\mathrm{odds}(Y|x+1)) = \log(\mathrm{odds}(Y|X)) + \beta_1$$

The odds can be obtained by exponentiating the log odds. On the odds scale, the odds for every unit change in X can be obtained by multiplying the baseline odds by $\exp(\beta_1)$

$$\mathrm{odds}(Y|x+1) = \mathrm{odds}(Y|x)\exp(\beta_1)$$

Change in the (log) odds



Every one kg increase in weight change is associated with a $\hat{\beta}_1 = 0.73$ increase in the **log** odds of hyponatremia.

Alternatively, the odds of hyponatremia doubles ($\exp(\hat{\beta}_1) = 2.08$) for every one kg increase in weight change.

If $\exp(\beta_1) > 1$, we usually interpret as $(\exp(\beta_1) - 1)\%$ increase in odds for every one-unit increase in x.
If $0 < \exp(\beta_1) < 1$, we usually interpret as $(1 - \exp(\beta_1))\%$ reduction in odds for every one-unit increase in x.

Constancy of the odds ratio

wtdiff	p	odds	rr	or
-4	0.00817	0.00824	NA	NA
-3	0.01679	0.01707	2.05379	2.07178
-2	0.03416	0.03537	2.03517	2.07178
-1	0.06828	0.07328	1.99860	2.07178
0	0.13181	0.15182	1.93051	2.07178
1	0.23928	0.31455	1.81532	2.07178
2	0.39455	0.65167	1.64891	2.07178
3	0.57449	1.35012	1.45605	2.07178

Confidence intervals for OR

$$\widehat{\text{est}} = \log(OR) = \beta_1(x_1 - x_2)$$

$$\text{Var}(\widehat{\text{est}}) = \text{Var}(\beta_1(x_1 - x_2)) = (x_1 - x_2)^2\text{Var}(\beta_1)$$

95% CI for **log** odds ratio

$\widehat{\text{est}} \pm z_{0.975}\sqrt{\text{Var}(\widehat{\text{est}})}$

95% CI for odds ratio (invariance property)

$\exp\left(\widehat{\text{est}} \pm z_{0.975}\sqrt{\text{Var}(\widehat{\text{est}})}\right)$

Question: What is the odds ratio of hyponatremia comparing those who increased 2 kg as compared to those who lost 1 kg?

Coefficients:

```
    b0      b1
-1.885   0.728
```

Covariance matrix:

```
        b0     b1
b0   0.025 -0.005
b1  -0.005  0.012
```

$$\widehat{\text{est}} = \hat{\beta}_1(x_1 - x_2) = 0.73(3) = 2.19$$

$$\text{Var}(\widehat{\text{est}}) = \text{Var}(3\hat{\beta}_1) = 9 \cdot \text{Var}(\hat{\beta}_1) = 9 \cdot 0.012 = 0.11$$

$$\text{SE}(\widehat{\text{est}}) = \sqrt{\text{Var}(\widehat{\text{est}})} = \sqrt(0.11) = 0.329$$

95% CI for **log** OR

$$2.19 \pm z_{0.975} \times 0.329 = (1.55, 2.83)$$

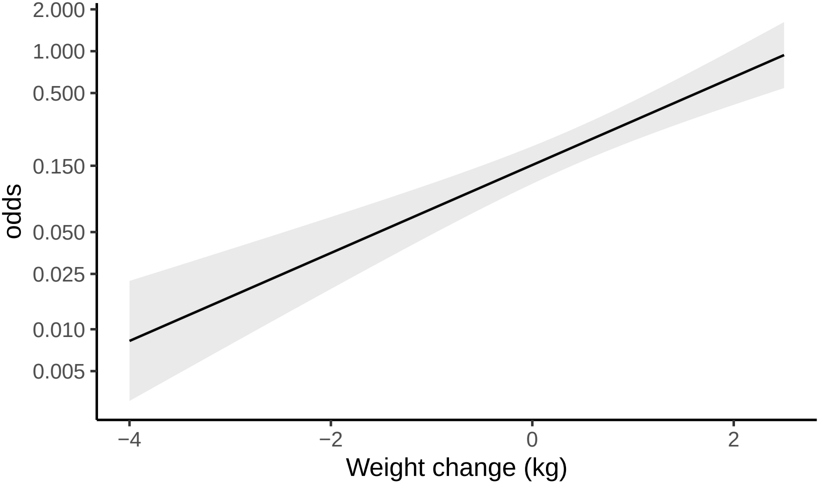
$$OR = \exp(3\hat{\beta}_1) = 8.94$$

95% CI for OR

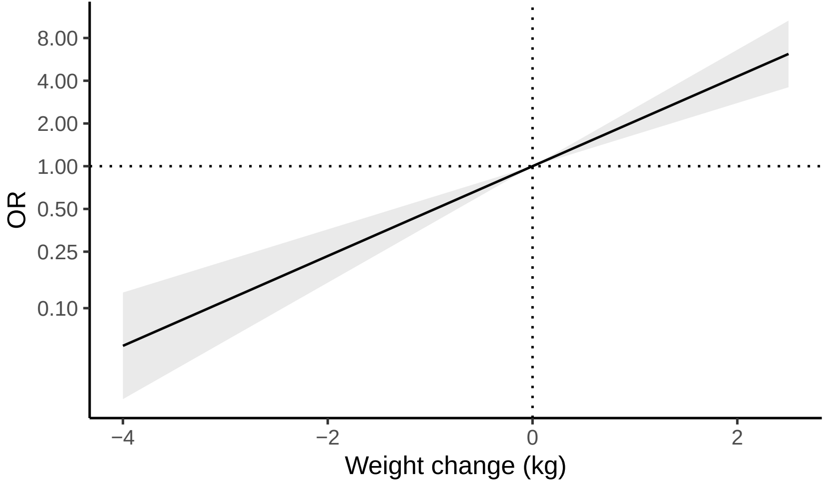
$$\exp(1.55, 2.83) = (4.69, 17.03)$$

The odds of hyponatremia among those who increased 2 kg was 9 (95% CI 4.69, 17.03) times the odds for those runners who lost 1kg.

Graphical presentation (odds)



Graphical presentation (odds ratio)



Question: What is the risk of hyponatremia for those runners who gained 1 kg during the race?

$\widehat{\text{est}} = \log\left(\text{odds}(\text{nas135}|\text{wtdiff} = 1)\right) = \hat{\beta}_0 + \hat{\beta}_1 =$
$-1.885 + 0.728 = -1.157$

$\text{Var}\left(\widehat{\text{est}}\right) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) + 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) =$
$0.0122 + 0.0122 + 2 \cdot (-0.0047) = 0.015$

Covariance matrix:

```
        b0     b1
b0   0.025 -0.005
b1  -0.005  0.012
```

95% CI for **log** odds

$-1.157 \pm 1.96 \times \sqrt{0.015} = (-1.397, -0.917)$

$P(\mathrm{nas135} = 1|\mathrm{wtdiff} = 1) = \frac{exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1)} = \frac{\exp(-1.157)}{1+\exp(-1.157)} =$ 0.2392

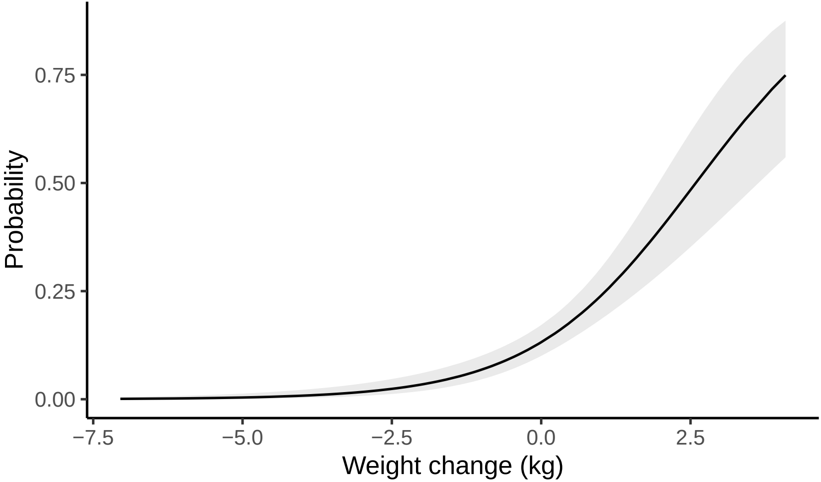
95% CI for $P(Y|X = 1)$ is the *invlogit* of the 95% CI for $\mathrm{logit}(P(Y|X = 1))$

$$\left(\frac{\exp(-1.397)}{1+\exp(-1.397)}, \frac{\exp(-0.917)}{1+\exp(-0.917)}\right) = (0.198, 0.286)$$

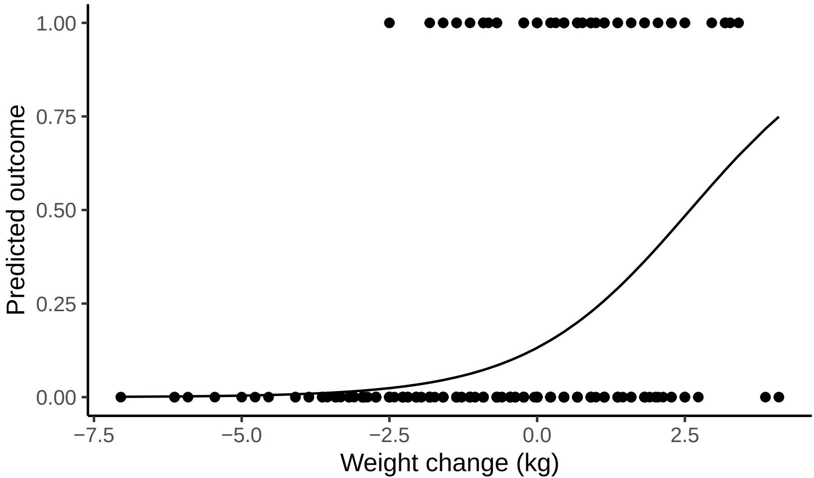
Predicted response

wtdiff	p	odds	rr	or
-4	0.008	0.008	0.062	0.054
-3	0.017	0.017	0.127	0.112
-2	0.034	0.035	0.259	0.233
-1	0.068	0.073	0.518	0.483
0	0.132	0.152	1.000	1.000
1	0.239	0.315	1.815	2.072
2	0.395	0.652	2.993	4.292
3	0.574	1.350	4.358	8.893

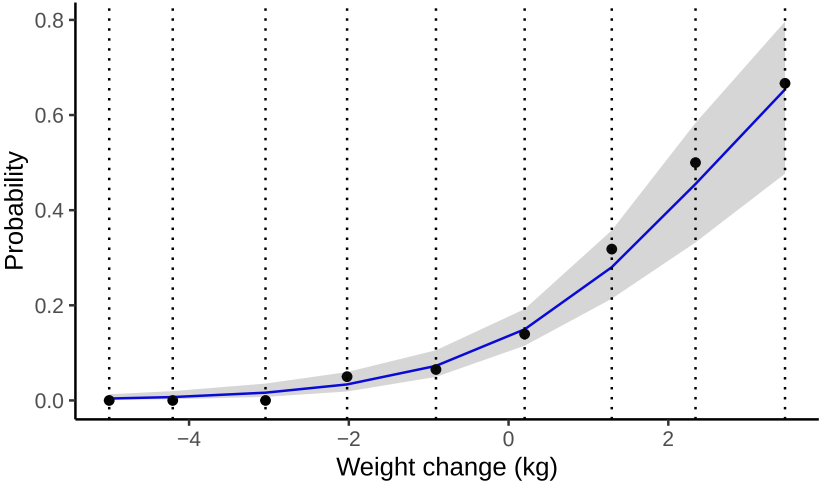
Graphical presentation (probability)



It would help to compare the predicted probabilities with the observed ones. But we only observe 0/1



We can divide the continuous X in k (e.g. 10) categories and calculate the proportion of the outcome.



Categorical predictor

The logistic regression model for a binary predictor X can be generalized to a categorical predictor with k levels (as in linear regression).

Creating a categorical variable from a continuous predictor can be a strategy to avoid the linearity assumption between the log odds of Y and X (but making another assumption).

Question: are categories of weight change (`wtdiffc`) associated with risk of hyponatremia (`nas135`)?

A descriptive table of data (cases/non-cases) can be easily presented.

nas135	3.0 to 4.9	2.0 to 2.9	1.0 to 1.9	0.0 to 0.9	-1.0 to -0.1	-2.0 to -1.1	-5.0 to -2.1
	wtdiffc						
na > 135	2	9	28	78	100	93	83
	28.6	56.2	71.8	81.2	91.7	93.9	98.8
na <= 135	5	7	11	18	9	6	1
	71.4	43.8	28.2	18.8	8.3	6.1	1.2

What test can be used to test the association between weight change and hyponatremia (without specifying a model)?

```
                  wtdiffc_nl
nas135             1 2    3    4    5    6    7
  na > 135         2 9    28   78   100  93   83
  na <= 135        5 7    11   18   9    6    1

  Odds ratio       1 0.33 0.16 0.1  0.04 0.03 0.01
  lower 95% CI       0.02 0.01 0.01 0    0    0
  upper 95% CI       2.79 1.19 0.64 0.27 0.21 0.08

Chi-squared = 63.37 , 6 d.f., P value = 0
```

There is evidence of association between weight change and risk of hyponatremia.

Indicator variables

Categorical variables with more than two levels are included in the regression model using indicator/dummy variables.

The indicator variable omitted from the model identifies the reference or baseline group.

$$\log\left(\mathrm{odds}(\mathrm{nas135}|\mathrm{wtdiff})\right) = \beta_0 + \beta_1 \mathrm{wtdiffc}_1 + \cdots + \beta_6 \mathrm{wtdiffc}_6$$

```r
mod_cat <- glm(nas135 ~ relevel(wtdiffc, 5), data = marathon, family = "binomia
summary(mod_cat)
```

```
Call:
glm(formula = nas135 ~ relevel(wtdiffc, 5), family = "binomial",
    data = marathon)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5829  -0.6444  -0.3536  -0.1548   2.9769

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -2.4079     0.3480  -6.919 4.54e-12
relevel(wtdiffc, 5)3.0 to 4.9     3.3242     0.9062   3.669 0.000244
relevel(wtdiffc, 5)2.0 to 2.9     2.1566     0.6124   3.521 0.000429
relevel(wtdiffc, 5)1.0 to 1.9     1.4736     0.4977   2.961 0.003069
relevel(wtdiffc, 5)0.0 to 0.9     0.9416     0.4353   2.163 0.030532
relevel(wtdiffc, 5)-2.0 to -1.1  -0.3329     0.5464  -0.609 0.542343
relevel(wtdiffc, 5)-5.0 to -2.1  -2.0109     1.0645  -1.889 0.058884

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 342.00  on 449  degrees of freedom
Residual deviance: 287.61  on 443  degrees of freedom
  (38 observations deleted due to missingness)
AIC: 301.61
```

Hypothesis testing

To test if there is an association between the categorical predictor and the risk of outcome, we need to test

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_6 = 0$

In other words, we need to compare the model with the dummy variables for `wtdiffc` with the nested empty model (with no dummy variables).

Two models are nested because are estimated on the same number of subjects but different number of parameters in the model.
It means that you can obtain the restricted model from the full model by putting constraints on the parameters you want to test.

Likelihood Ratio Test

The likelihood ratio test is a general procedure to compare the log-likelihoods (or deviances) of two nested models fitted on the same data.

$$\mathrm{LRT} = 2 \cdot |\log(L_R) - \log(L_U)|$$

where $\log(L_R)$ and $\log(L_U)$ are the log-likelihood of the restricted and unrestricted model, respectively.

Unrestricted model

$\log(\mathrm{odds}(Y|X_1, \ldots, X_6)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6$

Restricted model

$\log(\mathrm{odds}(Y|X_1, \ldots, X_6)) = \beta_0$

Under the null hypothesis (assuming the restricted model is true), the likelihood ratio test follows a χ^2 (chi-square) distribution with degrees of freedom equal to the difference in the number of parameters of the two models being compared.

The two log-likelihood are -143.805 and -171. The likelihood ratio test is given by

```
mod_0 <- glm(nas135 ~ 1, family = "binomial",
             data = subset(marathon, !is.na(wtdiffc)))
lrtest(mod_cat, mod_0)
```

```
Likelihood ratio test for MLE method
Chi-squared 6 d.f. =  54.39069 , P value =  6.15293e-10
```

The p-value associated with the LRT is very small ($p < 0.01$). There is evidence of association between categories of weight change and the risk of hyponatremia ($p < 0.01$).

To test which category of weight change is associated with a statistical significant higher risk, we look at the z test.

Interpretation

	exp(Est.)	2.5%	97.5%
(Intercept)	0.090	0.046	0.178
3.0 to 4.9	27.778	4.703	164.065
2.0 to 2.9	8.642	2.602	28.703
1.0 to 1.9	4.365	1.646	11.579
0.0 to 0.9	2.564	1.092	6.018
-2.0 to -1.1	0.717	0.246	2.092
-5.0 to -2.1	0.134	0.017	1.078

$\exp(\hat{\beta}_0) = 0.09$ is the odds of hyponatremia in the referent group, i.e. with a weight change between -1.0 to -0.1 kg.

$\exp(\hat{\beta}_3) = 4.365$ is the odds ratio of hyponatremia for those runner with a weight change between 1.0 to 1.9 kg compared to runner with a weight change between -1.0 to -0.1 kg.

Predicted probability

Question: What it the risk of hyponatremia for those runners who gain between 0 and 1 kg?

$\widehat{\text{est}} = \log(\text{odds}(Y|X_4 = 1)) = \hat{\beta}_0 + \hat{\beta}_4 = -2.408 + 0.942 = -1.466$

$P(Y|X_4 = 1) = \frac{\exp(-1.466)}{1+\exp(-1.466)} = 0.188$

Covariance matrix:

	b0	b1	b2	b3	b4	b5	b6
b0	0.121	-0.121	-0.121	-0.121	-0.121	-0.121	-0.121
b1	-0.121	0.821	0.121	0.121	0.121	0.121	0.121
b2	-0.121	0.121	0.375	0.121	0.121	0.121	0.121
b3	-0.121	0.121	0.121	0.248	0.121	0.121	0.121
b4	-0.121	0.121	0.121	0.121	0.189	0.121	0.121
b5	-0.121	0.121	0.121	0.121	0.121	0.299	0.121
b6	-0.121	0.121	0.121	0.121	0.121	0.121	1.133

95% CI for $\log(\mathrm{odds}(Y|X_4 = 1))$

$$-1.466 \pm 1.96 \times \sqrt{0.121 + 0.189 + 2 \cdot (-0.121))} = (-1.979, -0.954)$$

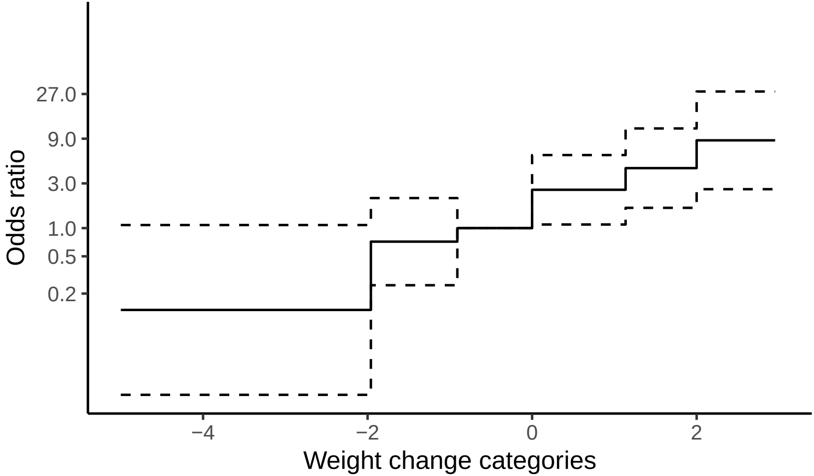
95% CI for $P(Y|X_4 = 1)$

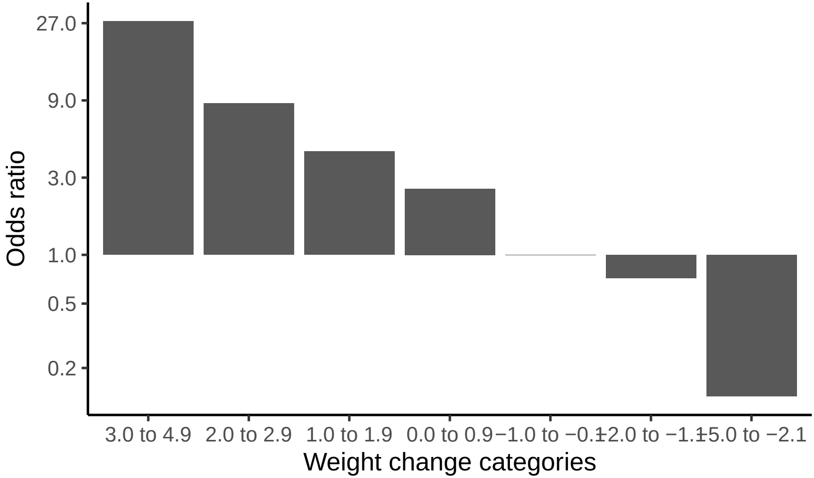
$$\text{invlogit}(-1.979, -0.954) = (0.121, 0.278)$$

Predicted response

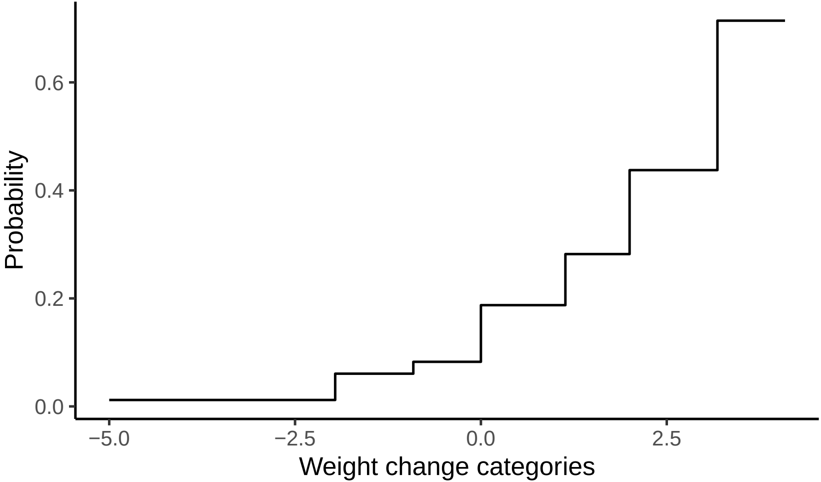
wtdiffc	p	odds	rr	or
3.0 to 4.9	1	2.500	8.651	27.778
2.0 to 2.9	0	0.778	5.299	8.642
1.0 to 1.9	0	0.393	3.416	4.365
0.0 to 0.9	0	0.231	2.271	2.564
-1.0 to -0.1	0	0.090	1.000	1.000
-2.0 to -1.1	0	0.065	0.734	0.717
-5.0 to -2.1	0	0.012	0.144	0.134

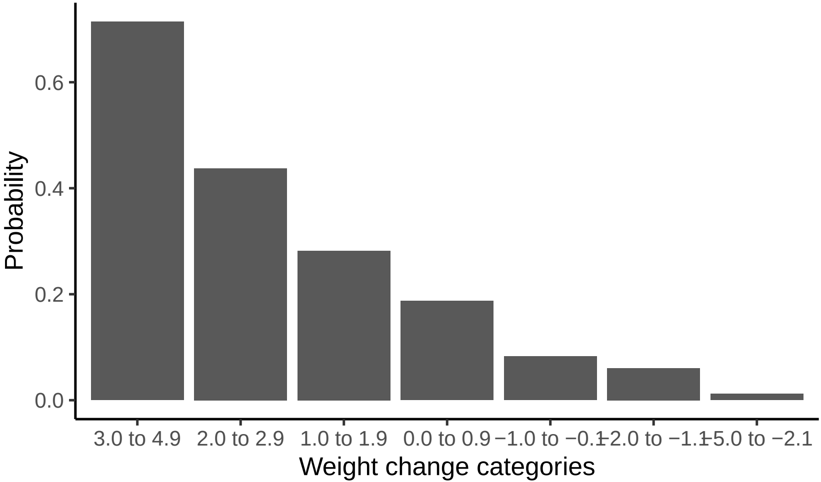
Graphical presentation (OR)




27.0
9.0
3.0
1.0
0.5
0.2
Odds ratio
3.0 to 4.9
2.0 to 2.9
1.0 to 1.9
0.0 to 0.9
−1.0 to −0.1
−2.0 to −1.1
−5.0 to −2.1
Weight change categories

Graphical presentation




Probability
0.0
0.2
0.4
0.6
3.0 to 4.9
2.0 to 2.9
1.0 to 1.9
0.0 to 0.9
−1.0 to −0.1
−2.0 to −1.1
−5.0 to −2.1
Weight change categories