Categorical Data Analysis - Multivariable Logistic Regression

Alessio Crippa

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Summary recap

Multivariable logistic regression

Interaction analysis

Model building

GOF and non-linearity

Summary recap

Simple logistic regression: model the log odds of a binary outcome Y as a *linear* function of a predictor X (binary, continuous, or categorical)

1. Formulate a model (equation)

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

2. Estimate the model and interpret the coefficients

$\exp(\beta_0)$: odds of y when $X = 0$

$\exp(\beta_1)$: OR of y comparing $X = x + 1$ with $X = x$

Univariate test and likelihood ratio test

4. Predict quantities of interest ((log) odds, probabilities)

$$\log(\widehat{\text{odds}}(Y)) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\widehat{\text{odds}}(Y) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x)$$

$$\widehat{\text{P}}(Y) = \text{invlogit}(\log(\widehat{\text{odds}}(Y))) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

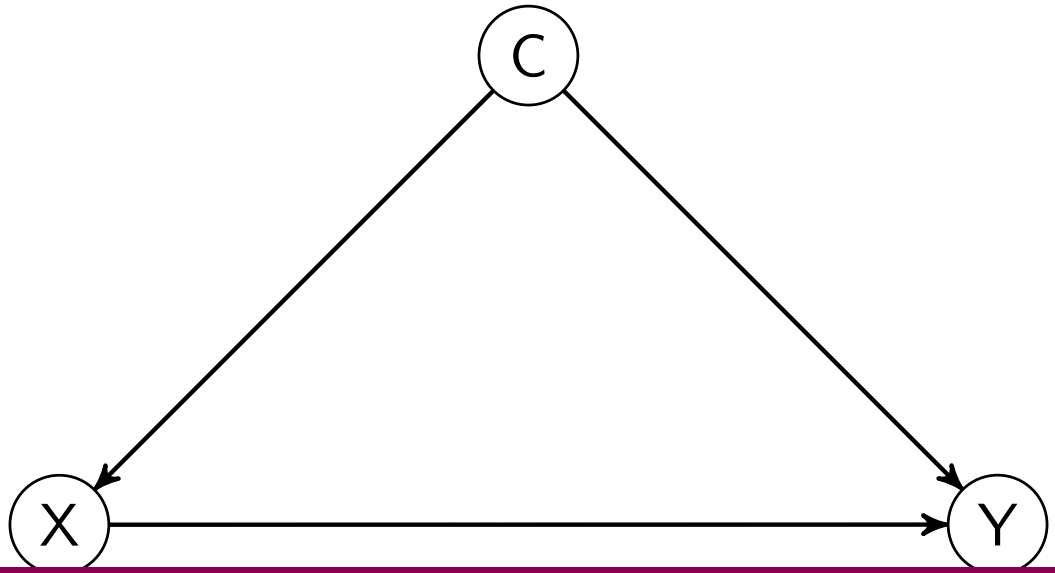
5. Present them in a graphical/tabular format

Multivariable logistic regression

Confounding

Multivariable models are often used to adjust/control the effect of predictors from the effect of potential confounders.

A confounding variable is one that is associated with both the outcome and predictor (or exposure) of interest but which is not, in itself, a focus of our investigation.



Question: Are slower runners at higher risk of hyponatremia?

$$\log(\text{odds}(\text{nas135})) = \beta_0 + \beta_1 \text{runtime}$$

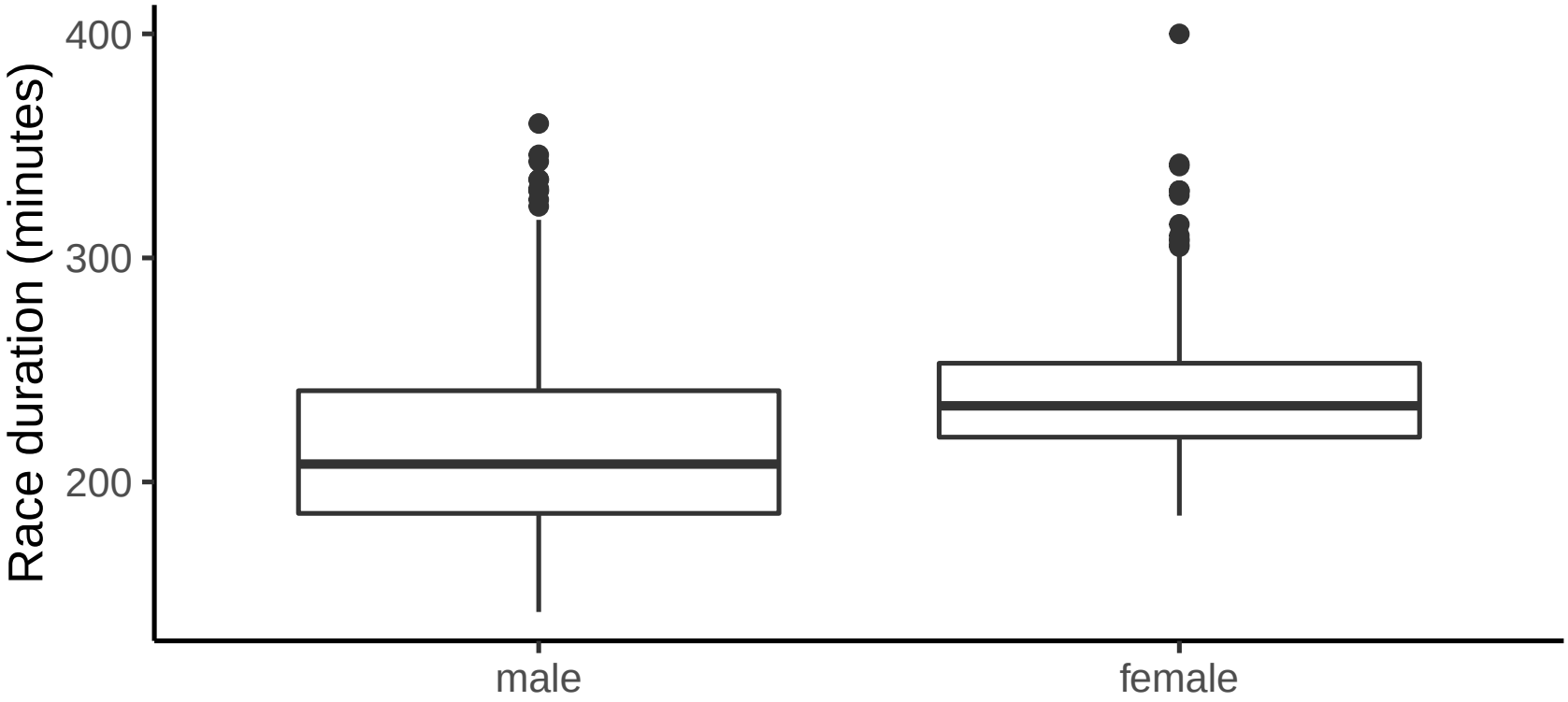
```
mod_r <- glm(nas135 ~ I(runtime/10), data = marathon, family = "binomial")
ci.exp(mod_r)
```

```
                exp(Est.)         2.5%     97.5%
(Intercept)    0.003725353 0.0008215826 0.0168921
I(runtime/10) 1.167680463 1.0990425509 1.2406050
```

Question: Is sex (`female`) a confounder of the association between race duration (`runtime`) and hyponatremia (`nas135`)?

1. What is the association between confounder and exposure?
2. What is the association between confounder and outcome?

Is sex associated with race duration?



	Estimate	StdErr	z	P	2.5%	97.5%
(Intercept)	217.71384	2.250627	96.734733	0.000000e+00	213.30269	222.12499
femalefemale	23.49371	3.898201	6.026809	1.672289e-09	15.85338	31.13404

Is sex associated with hyponatremia?

```
2 by 2 table analysis:
------------------------------------------------------
Outcome    : na <= 135
Comparing : female vs. male

       na <= 135 na > 135    P(na <= 135) 95% conf. interval
female        37      129          0.2229     0.166   0.2925
male          25      297          0.0776     0.053   0.1124

                                    95% conf. interval
            Relative Risk: 2.8708    1.7914   4.6007
          Sample Odds Ratio: 3.4074    1.9701   5.8936
Conditional MLE Odds Ratio: 3.3982    1.9037   6.1528
    Probability difference: 0.1453    0.0790   0.2186

             Exact P-value: 0
        Asymptotic P-value: 0
------------------------------------------------------
```

We observed that women on average run slower than men (22.5 minutes slower) and are more likely to get hyponatremia (22% vs 8%).

According to the previous definition, sex (`female`) could confound the association between run duration (`runtime`) and risk of hyponatremia (`nas135`).

How can we explore the confounding effect of sex?

Similarly to linear regression, we can extend the univariate setting regression to include the effect (and possibly the interaction) of k predictors.

The log odds of the outcome $\log\left(\text{odds}(Y = 1|X_1, \ldots, X_k)\right)$ depends on a linear combination of the k covariates or predictors.

$$\log\left(\text{odds}(Y = 1|X_1, \ldots, X_k)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

or alternatively

$$P(Y = 1|X_1, \ldots, X_k) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}$$

For any given combination of predictors (*covariance pattern*) we can estimate the probability of the outcome

Example: X continuous, C binary

To control/adjust the association between race duration and risk of hyponatremia for sex, we include it (`female`) in the multivariable logistic model

$$\log\left(\text{odds}(\text{nas135}|\text{runtime}, \text{sex})\right) = \beta_0 + \beta_1 \text{runtime} + \beta_2 \text{sex}$$

The effect of the two predictors is multiplicative on the odds

$$\text{odds}(\text{nas135}|\text{runtime}, \text{sex}) = \exp(\beta_0)\exp(\beta_1 \text{runtime})\exp(\beta_2 \text{sex})$$

In order to interpret β_0, we center the continuous variable `runtime` around the mean value (225). We divide the centered variable by 10 to interpret β_1 as change in the log odds for a 10 minutes increase.

```r
mod_mc <- glm(nas135 ~ I((runtime-225.5)/10) + female,
              data = marathon, family = "binomial")
summary(mod_mc)
```

```
Call:
glm(formula = nas135 ~ I((runtime - 225.5)/10) + female, family = "binomial",
    data = marathon)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2157  -0.5714  -0.3668  -0.2899   2.6427

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.51589    0.21913 -11.481  < 2e-16
I((runtime - 225.5)/10)  0.14214    0.03295   4.314 0.000016
femalefemale             0.96384    0.29105   3.312 0.000928

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 360.90  on 476  degrees of freedom
Residual deviance: 324.48  on 474  degrees of freedom
  (11 observations deleted due to missingness)
AIC: 330.48
```

Hypothesis testing

Similarly to multivariable linear regression, we first want to test if the fitted model is appropriate, i.e. at least one of the regression coefficient is different from zero.

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

We can either use a likelihood ratio test or a Wald test.

Wald-type test

A Wald-type test for the hypothesis $H_0 : \beta_1 = 0$ is equal to the ratio of the square of the maximum likelihood estimate of the parameter to an estimate of its variance.

$$W = \frac{(\hat{\beta}_1 - 0)^2}{\text{Var}(\hat{\beta}_1)}$$

Under the null hypothesis and assuming a large sample, this ratio follows a Chi-Square distribution with 1 degree of freedom.

NB The (univariate) Wald test are identical to the (square) of the Z test.

In a multivariable logistic models we may be interested in testing that more coefficients are simultaneously equal to zero

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

We then use the multivariate analogue of the Wald-test which, under the null hypothesis, follows a χ^2 distribution with 2 degrees of freedom.

```
Wald test:
----------

Chi-squared test:
X2 = 32.4, df = 2, P(> X2) = 9.2e-08
```

The p-value is small ($p < 0.001$), so we reject at a 95% confidence level that both the coefficients are simultaneously equal to zero.

The second (set of) hypothesis tests the significance of the individual predictors:

$$H_0 : \beta_i = 0 \qquad i = 1, \ldots, k$$

We can use the univariate z test presented in the main output. Both the predictors are significantly associated with the risk of hyponatremia.

Interpretation

$\exp(\hat{\beta}_0) = 0.08$ is the odds of hyponatremia for men (`female` $= 0$) who run the marathon in 225.5 min ($\sim$ 4:45 h) (`I((runtime - 225.5)/10) = 0`).

The effect of X_1 is the comparison between two log odds comparing $X_1 = x + 1$ with $X_1 = x$, fixing X_2 equal to x_2

$$\log(\text{odds}(Y|X_1 = x + 1, X_2 = x_2)) = \hat{\beta}_0 + \hat{\beta}_1(x + 1) + \hat{\beta}_2 x_2$$
$$\log(\text{odds}(Y|X_1 = x, X_2 = x_2)) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x_2$$

$$\log(\text{odds}|X_1 = x + 1, X_2 = x_2) - \log(\text{odds}|X_1 = x, X_2 = x_2) =$$
$$\log(OR_{x_1=x+1 \text{vs} x_1=x}) = \hat{\beta}_0 + \hat{\beta}_1(x+1) + \hat{\beta}_2 x_2 - (\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x_2) = \hat{\beta}_1$$

$\exp(\hat{\beta}_1) = 1.15$: every 10 minutes increase in race duration is associated with a 15% increase in the odds, holding `female` constant. Alternatively, the sex-adjusted odds ration comparing two groups of runners with a difference of 10 minutes in race duration is 1.15.

$\exp(\hat{\beta}_2) = 2.62$ is the odds ratio of hyponatremia comparing women to men, holding `runtime` constant.

Confidence intervals for adjusted OR

A 95% confidence interval for adjusted OR is constructed in a similar way as in the case of univariate logistic regression.

We first construct a confidence interval for the log(OR), i.e. β_1

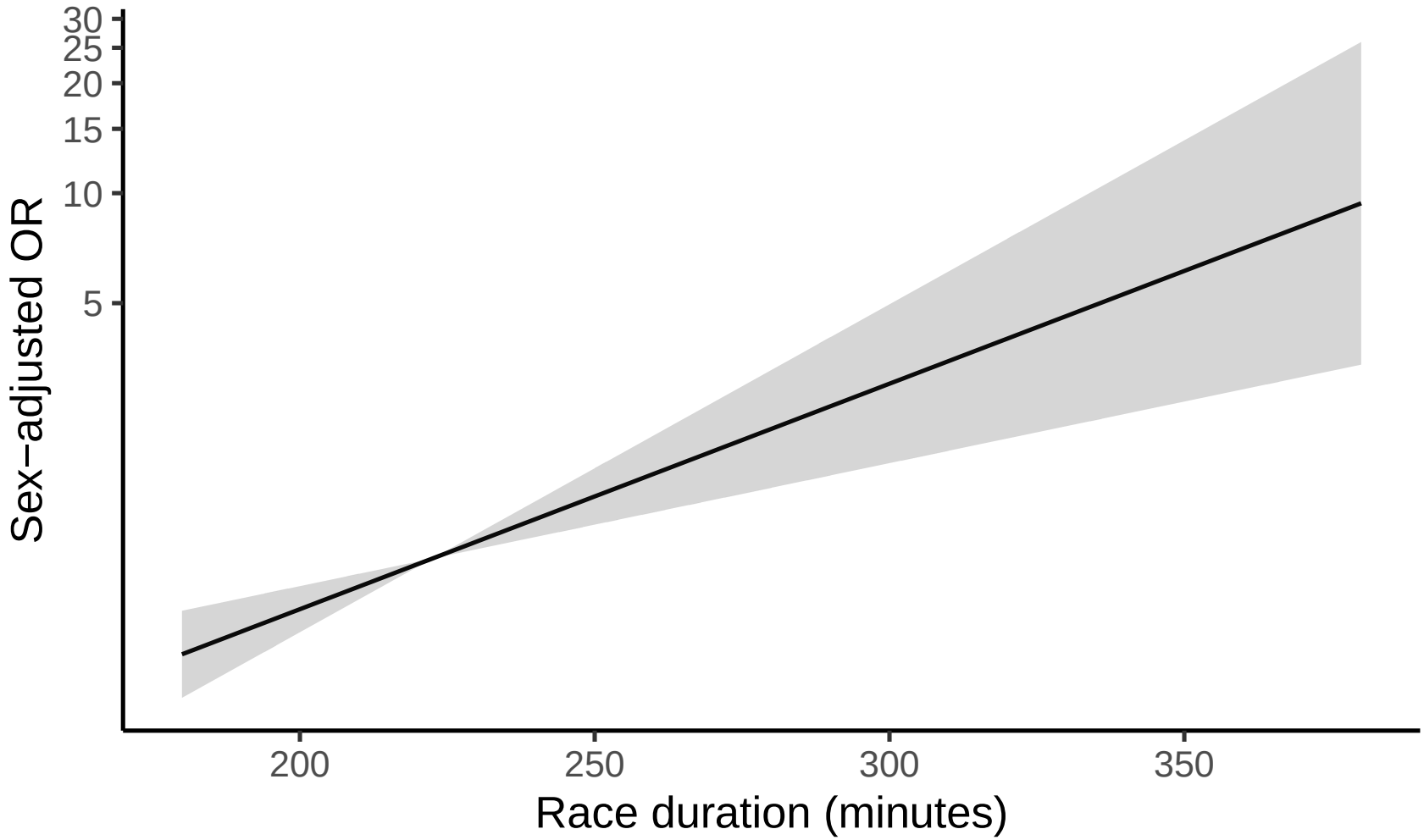
$\hat{\beta}_1 \pm 1.96\sqrt{\text{Var}(\hat{\beta}_1)}$

The corresponding confidence interval for the OR is

$\exp\left(\hat{\beta}_1 \pm 1.96\sqrt{\text{Var}(\hat{\beta}_1)}\right)$

We are 95% confident that the sex-adjusted odds ratio comparing two groups of runners with a difference of 10 minutes in race duration lies in the interval $\exp(0.14 \pm 1.96 \cdot 0.033) = (1.081, 1.23)$

Graphical presentation OR



Predicted (log) odds

To calculate the predicted odds, we need to specify the covariate patters (combination of predictor' values) we are interested in.

Question: what it the (log) odds of hyponatremia for women with a race duration equal to 4 h (240 minutes)?

$$\widehat{\text{est}} = \log(\text{odds}(Y|X_1 = 1, X_2 = 1)) = \hat{\beta}_0 + \frac{240-225.5}{10}\hat{\beta}_1 + \hat{\beta}_2 = -2.516 + 1.75 \cdot 0.142 + 0.964 = -1.346$$

Covariance matrix:

	b0	b1	b2
b0	0.048	-0.002	-0.044
b1	-0.002	0.001	-0.001
b2	-0.044	-0.001	0.085

$$\text{Var}\left(\widehat{\text{est}}\right) = \text{Var}(\hat{\beta}_0) + 1.75^2\text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) + 2 \cdot 1.75\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) +$$

$$2\text{Cov}(\hat{\beta}_0, \hat{\beta}_2) + 2 \cdot 1.75\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) =$$

$$0.048 + 1.75^2 \cdot 0.001 + 0.085 + 3.5 \cdot -0.002 + 2 \cdot -0.044 + 3.5 \cdot -0.001 = 0.038$$

95% CI for $\log(\text{odds}(Y|X_1 = 240, X_2 = 1))$:

$$-1.346 \pm 1.96 \cdot \sqrt{0.038} = (-1.736, -0.955)$$

Predicted probability

Question: what it the risk of hyponatremia for women with a race duration equal to 4 h?

We can use the `invlogit` function to calculate predicted probabilities from the predicted log odds (calculated before).

$$P(Y|X_1 = 240, X_2 = 1) = \frac{\exp(-1.346)}{1+\exp(-1.346)} = 0.207$$

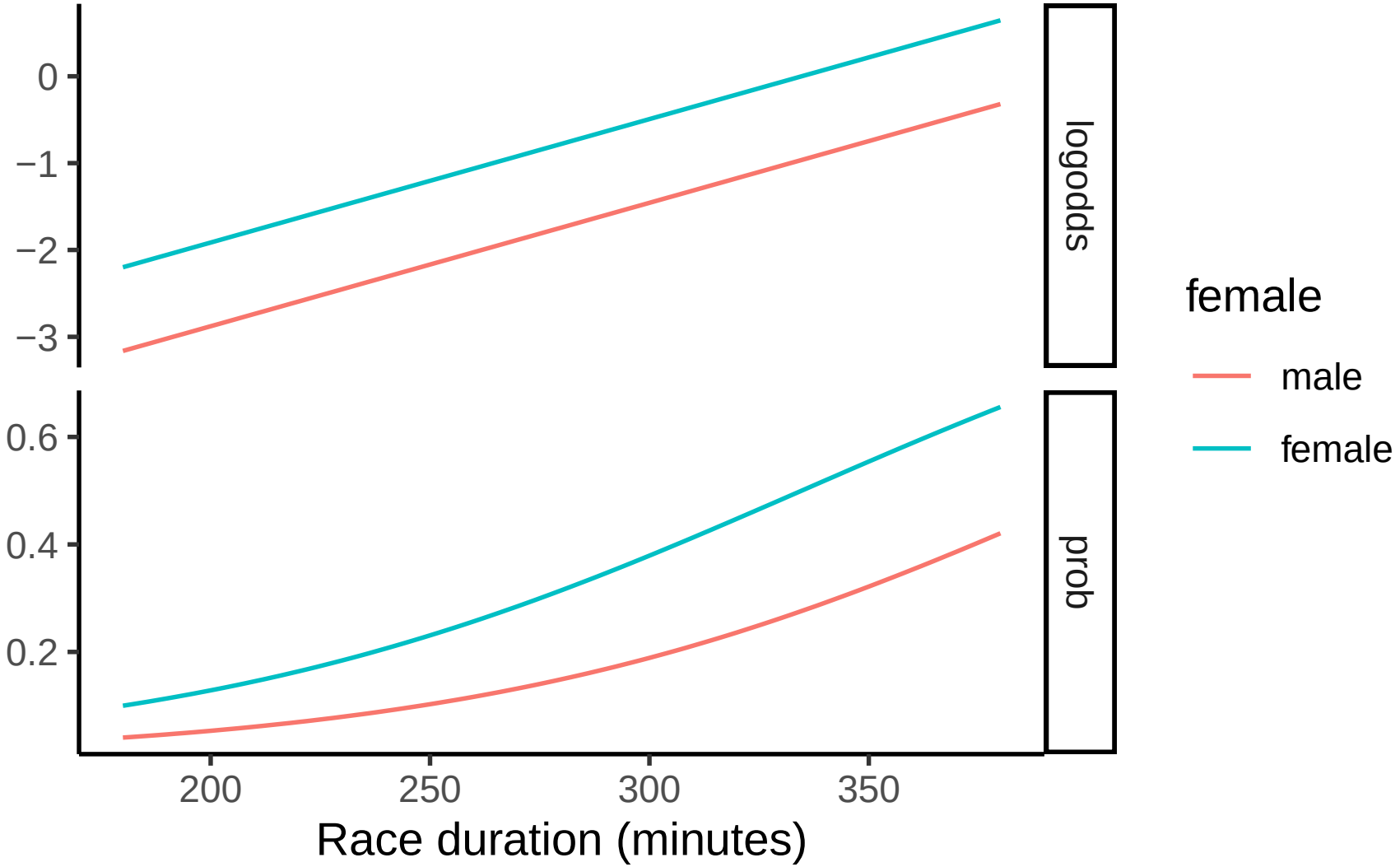
95% CI for $P(Y|X_1 = 240, X_2 = 1)$:

$$\text{invlogit}(-1.736, -0.955) = (0.15, 0.278)$$

Tabular presentation

runtime	female	prob	logodds
180	male	0.0405975	-3.1626049
210	male	0.0608710	-2.7361964
240	male	0.0903156	-2.3097879
300	male	0.1889311	-1.4569709
180	female	0.0998611	-2.1987685
210	female	0.1452491	-1.7723600
240	female	0.2065330	-1.3459515
300	female	0.3791555	-0.4931344

Graphical presentation odds and probabilities



Interaction analysis

Rationale

Question: Is the effect of the predictor X on the odds Y varying according to another predictor Z?

We can test the previous hypothesis by including an interaction term between X and Z in the logistic model

$$\log\left(\text{odds}(Y|X,Z)\right) = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

The β_3 is referred to as a *statistical interaction*, and it is additive on the log(odds), but multiplicative on the odds.

As in case of linear regression, the previous model can be re-written either as

$\log(odds(Y|X,Z)) = \beta_0 + (\beta_1 + \beta_3 Z)X + \beta_2 Z$

or $\log(odds(Y|X,Z)) = \beta_0 + \beta_1 X + (\beta_2 + \beta_3 X)Z$

so that the effect of X on the log odds of Y depends on Z via β_3, and viceversa.

Example: X continuous, Z binary

Question: Is the effect of race duraton (`runtime`) on the risk of hyponatremia (`nas135`) different among women and men (`female`)?

The logistic model to test for interarction is

$$\log\left(\text{odds}(\text{nas135}|\text{runtime}, \text{sex})\right) =$$
$$\beta_0 + \beta_1 \text{runtime} + \beta_2 \text{sex} + \beta_3 \text{runtime} \cdot \text{sex}$$

That is:

- for men: $\log\left(\text{odds}(\text{nas135}|\text{runtime}, \text{sex} = 0)\right) = \beta_0 + \beta_1 \text{runtime}$
- for women:

$$\log\left(\text{odds}(\text{nas135}|\text{runtime}, \text{sex} = 1)\right) = \beta_0 + (\beta_1 + \beta_3)\text{runtime} + \beta_2$$

```
Call:
glm(formula = nas135 ~ I((runtime - 225.5)/10) * female, family = "binomial",
    data = marathon)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1505  -0.5874  -0.3611  -0.2798   2.6786

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -2.53576    0.22705 -11.169  < 2e-16
I((runtime - 225.5)/10)              0.15392    0.04299   3.581 0.000343
femalefemale                         1.02285    0.32197   3.177 0.001489
I((runtime - 225.5)/10):femalefemale -0.02845    0.06657  -0.427 0.669123

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 360.9  on 476  degrees of freedom
Residual deviance: 324.3  on 473  degrees of freedom
  (11 observations deleted due to missingness)
AIC: 332.3

Number of Fisher Scoring iterations: 5
```

Hypothesis testing

To test if the interaction model fits the data, we test the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, i.e. that all the regression coefficients are simultaneously equal to zero. This can be done by eihter adoting a likelihood ratio test, or a multivariate Wald test

```
Likelihood ratio test for MLE method
Chi-squared 3 d.f. =  36.5978 , P value =  5.597194e-08
```

The p-value is small ($p < 0.01$) so we reject the null hypothesis at a 95% confidence level.

In an interaction model, the wald tests for the main effects (β_1 and β_2) are not longer interesting.

The main hypothesis is instead to test if there is evidence of interaction $H_0 : \beta_3 = 0$.

The p-value for the Wald test for the interaction term is greater than 0.05 ($p = 0.669$). At a 95% confidence level, we fail to reject the null hypothes, i.e. there is not (enough) evidence of (multiplicative) interaction.
If we can't reject the null hypothesis we can remove the interaction term from the model.

Interpretation

$\exp(\hat{\beta}_0) = 0.08$ is the odds of hyponatremia for men (`female` $= 0$) who run the marathon in 225.5 min (`I((runtime - 225.5)/10) = 0`).

$\exp(\hat{\beta}_1) = 1.17$ is the odds ration comparing two groups of **men** with a difference of 10 minutes in race duration is 1.17.

$\exp(\hat{\beta}_2) = 2.78$ is the odds ratio of hyponatremia comparing women who completed the marathon in 225.5 minutes vs men who completed the marathon in 225.5 minutes.

$\exp(\hat{\beta}_3) = 0.97$ is the additional (multiplicative) component in the combined effect of race duration and being female on the odds of the hyponatremia.

$$\text{odds}(Y|X = 225.5, Z = 1) = \exp(\hat{\beta}_0)\exp(\hat{\beta}_1)\exp(\hat{\beta}_2)\exp(\hat{\beta}_3)$$

Predicted response

$\log(odds)$	$Z = 0$	$Z = 1$
$X = 215.5$	$\hat{\beta}_0 - \hat{\beta}_1$	$\hat{\beta}_0 - \hat{\beta}_1 + \hat{\beta}_2 - \hat{\beta}_3$
$X = 225.5$	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\beta}_2$
$X = 235.5$	$\hat{\beta}_+\beta_1$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_2$

odds	$Z = 0$	$Z = 1$
$X = 215.5$	$\exp(\hat{\beta}_0 - \hat{\beta}_1) =$	$\exp(\hat{\beta}_0 - \hat{\beta}_1 + \hat{\beta}_2 - \hat{\beta}_3) =$
	$= \exp(\hat{\beta}_0)/\exp(\hat{\beta}_1)$	$= (\exp(\hat{\beta}_0)\exp(\hat{\beta}_2))/(\exp(\hat{\beta}_1)\exp(\hat{\beta}_3)$
$X = 225.5$	$\exp(\hat{\beta}_0)$	$\exp(\hat{\beta}_0 + \hat{\beta}_2) =$
		$= \exp(\hat{\beta}_0)\exp(\hat{\beta}_2)$
$X = 235.5$	$\exp(\hat{\beta}_0 + \hat{\beta}_1) =$	$\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3) =$
	$= \exp(\hat{\beta}_0)\exp(\hat{\beta}_1)$	$= \exp(\hat{\beta}_0)\exp(\hat{\beta}_1)\exp(\hat{\beta}_2)\exp(\hat{\beta}_3)$

OR	$Z = 0$	$Z = 1$
$X = 215.5$	$OR_{10} = \exp(-\hat{\beta}_1)$	$OR_{11} = \exp(-\hat{\beta}_1)\exp(\hat{\beta}_2)\exp(-\hat{\beta}_3)$
$X = 225.5$	1	$OR_{01} = \exp(\hat{\beta}_2)$
$X = 235.5$	$OR_{10} = \exp(\hat{\beta}_1)$	$OR_{11} = \exp(\hat{\beta}_1)\exp(\hat{\beta}_2)\exp(\hat{\beta}_3)$

p	$Z = 0$	$Z = 1$
$X = 215.5$	$p_{10} = \frac{\exp(\hat{\beta}_0 - \hat{\beta}_1)}{1+\exp(\hat{\beta}_0 - \hat{\beta}_1)}$	$p_{11} = \frac{\exp(\hat{\beta}_0 - \hat{\beta}_1 + \hat{\beta}_2 - \hat{\beta}_3)}{1+\exp(\hat{\beta}_0 - \hat{\beta}_1 + \hat{\beta}_2 - \hat{\beta}_3)}$
$X = 225.5$	$p_{00} = \frac{\exp(\hat{\beta}_0)}{1+\exp(\hat{\beta}_0)}$	$p_{01} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_2)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_2)}$
$X = 235.5$	$p_{10} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1)}$	$p_{11} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3)}$

$\log(odds)$	$X_2 = 0$	$X_2 = 1$
$X_1 = 215.5$	-2.69	-1.638
$X_1 = 225.5$	-2.536	-1.513
$X_1 = 235.5$	-2.382	-1.387

OR	$X_2 = 0$	$X_2 = 1$
$X_1 = 215.5$	0.861	2.456
$X_1 = 225.5$	1	2.785
$X_1 = 235.5$	1.165	3.165

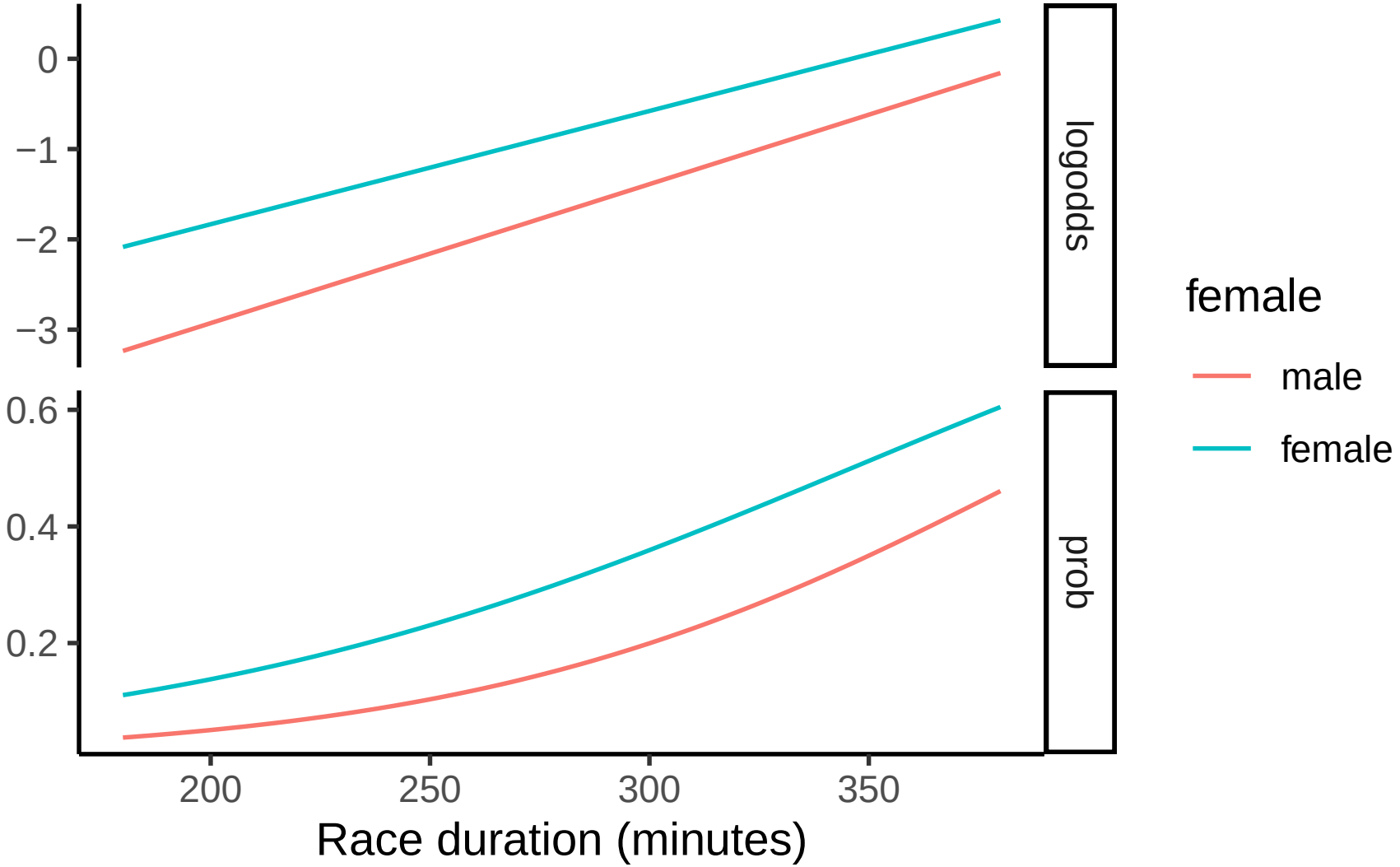
$odds$	$X_2 = 0$	$X_2 = 1$
$X_1 = 215.5$	0.068	0.194
$X_1 = 225.5$	0.079	0.22
$X_1 = 235.5$	0.092	0.25

p	$X_2 = 0$	$X_2 = 1$
$X_1 = 215.5$	0.064	0.163
$X_1 = 225.5$	0.073	0.18
$X_1 = 235.5$	0.085	0.2

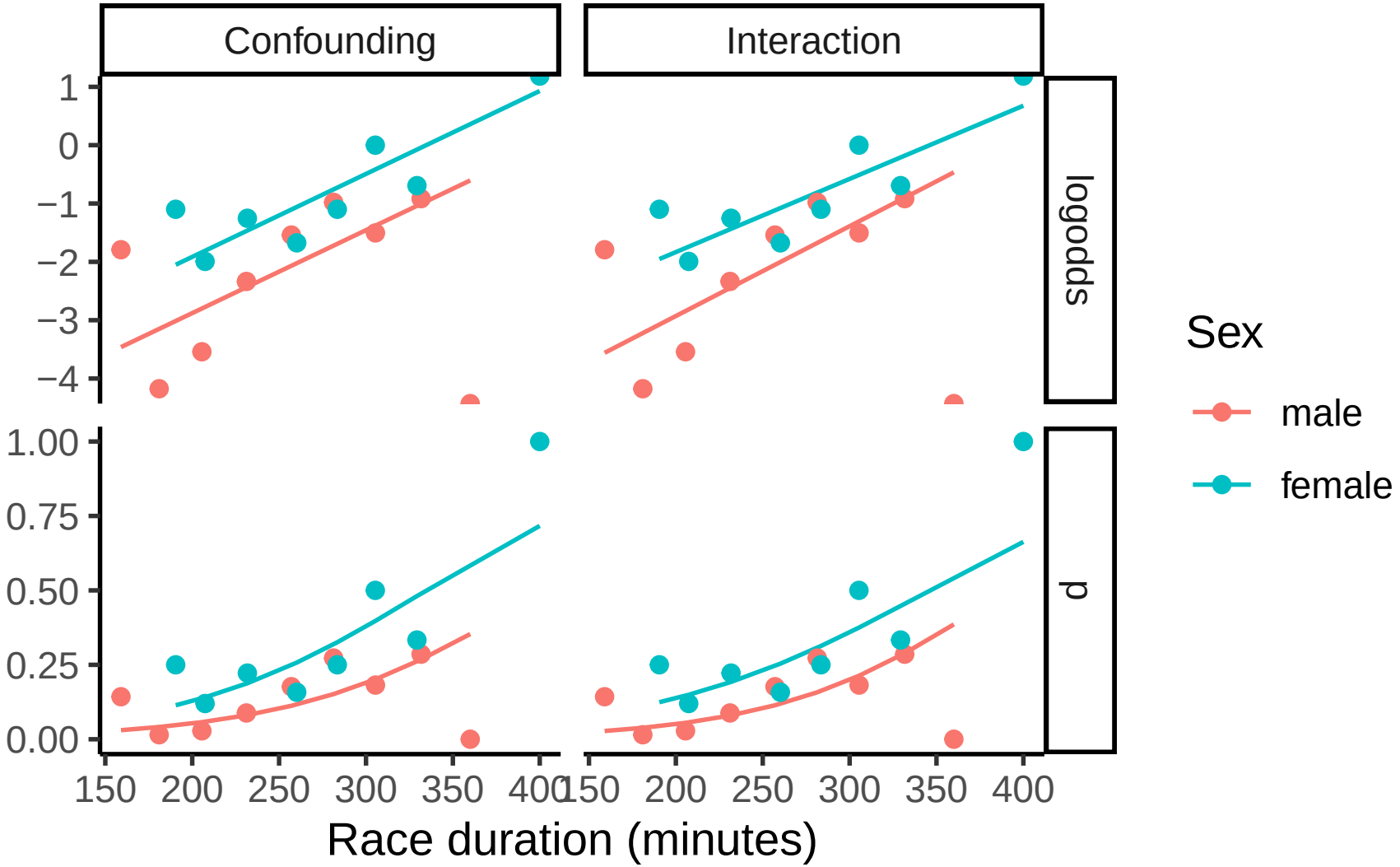
Tabular presentation

runtime	female	prob	logodds
180	male	0.0405975	-3.1626049
210	male	0.0608710	-2.7361964
240	male	0.0903156	-2.3097879
300	male	0.1889311	-1.4569709
180	female	0.0998611	-2.1987685
210	female	0.1452491	-1.7723600
240	female	0.2065330	-1.3459515
300	female	0.3791555	-0.4931344

Graphical presentation



Confounding vs Interaction



Measure of additive interaction

$$(p_{11} - p_{00}) - ((p_{10} - p_{00}) + (p_{01} - p_{00}))$$

We consider the contrast between the effects of both factors together versus the sum of each considered separately.

If the difference is not equal to zero we say that there is interaction on the difference scale.

Measure of multiplicative interaction

$$\frac{OR_{11}}{OR_{10}OR_{01}} = \beta_3$$

This quantity measures the extent to which, on the odds ratio scale, the effect of both exposures together exceeds the product of the effects of the two exposures considered separately.

Model building

Model building

- In most practical situations, we have a set of potential explanatory variables and we must decide which ones to include in the regression model and which ones to leave out. These decisions are often based on both statistical and non-statistical considerations.
- Ideally, we would have some prior knowledge as to which variables might be relevant and interesting to consider from a medical/epidemiologic perspective.

Possible strategies for model building

Some steps that may help:

- Begin with a careful univariable screening of each covariate. Comparison of continuous vs. categorical coding of covariates (e.g., continuous age vs. age categories).
- For categorical variables, you could consider recoding or collapsing some groups, if appropriate.
- Include variables known or thought to be important from subject matter considerations.

- Run bivariate association analyses (table and graphs)
- Be careful about covariates with missing values (using them leads to smaller sample sizes and possible bias).
- Use caution, logic, good sense, and biologic plausibility when using model building techniques.

Information criteria

How to compare the fit of different models?

The AIC is a popular measure for comparing maximum likelihood models.

AIC is defined as

$$\text{AIC} = -2\log(\text{likelihood}) + 2k$$

AIC is a measure that combine fit and complexity.

Fit is measured negatively by $-2\log(\text{likelihood})$; the larger the value, the worse the fit.
Complexity is measured positively, by $2k$.

Given two models fit on the same data, the model with the smaller value of the information criterion is considered to be better.

Example AIC

Which of the following models best fits the data?

1. *mod1*: `wtdiff`
2. *mod2*: *mod1* + `bmi` + `runtime`
3. *mod2*: *mod2* + `female` + `age` + `prevmara`
4. *mod4*: `height` + `prewt` + `age`

model	logLik	k	aic
mod2	-131.50	4	270.99
mod3	-130.47	7	274.95
mod1	-141.64	2	287.27
mod4	-164.30	4	336.61

GOF and non-linearity

Assessing the Goodness of Fit

A critical step in in logistic regression is how well the model (estimated probabilities) agrees with the observed data (observed frequencies): this is the goodness-of-fit of the model.

Summary measures of distance between observed and predicted outcomes (Hosmer-Lemeshow test, Pearson Chi-Square test)

Measures of discrimination between cases and non-cases (sensitivity, specificity, ROC curve, NRI)

Example

Question: Does a model with sex and bmi properly predict the risk of hyponatremia?

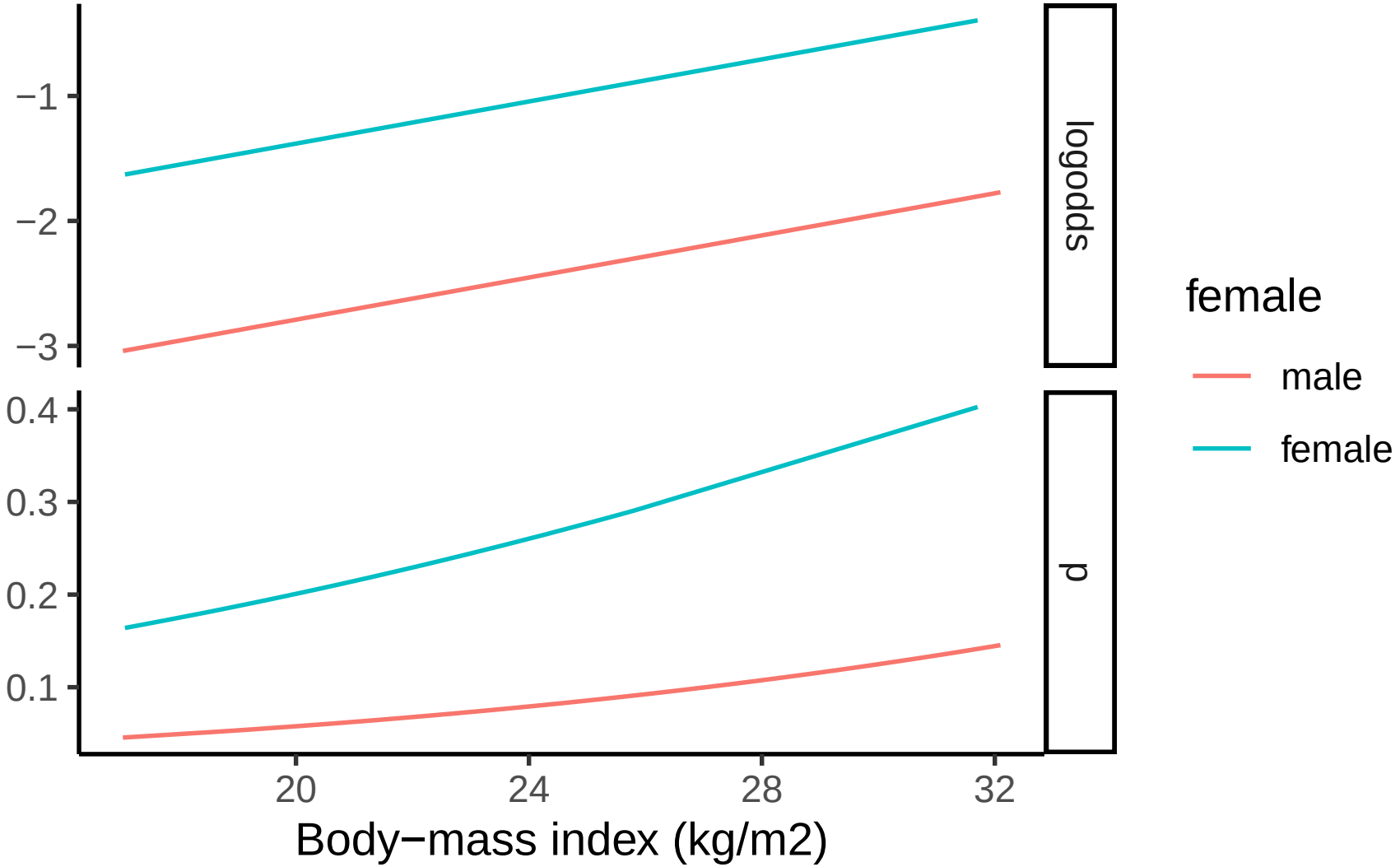
The logistic model is

$$\log\left(\text{odds}(\text{nas135}|\text{sex}, \text{bmi})\right) = \beta_0 + \beta_1(\text{sex}) + \beta_2(\text{bmi-23})$$

```
mod_b <- glm(nas135 ~ female + I(bmi - 23), data = marathon,
             family = "binomial")
ci.exp(mod_b)
```

```
                exp(Est.)       2.5%     97.5%
(Intercept)  0.07908786 0.05085438 0.1229961
femalefemale 4.09024715 2.16876951 7.7141078
I(bmi - 23)  1.08789422 0.97116197 1.2186575
```

Predicted response



Hosmer and Lemeshow test

Aim: Test whether or not the observed event proportions match the predicted probabilities

How: We order the predicted probabilities p and create deciles. So the first decile contains the smallest n/10 values of p, the second contains the next smallest n/10 values of p, and so on.
If the model holds then those who actually develop the outcome should have high values for p. Similarly, those who do not develop the outcome should have low values for p.

One can then compare observed frequencies of the outcome with the (mean) predicted probability across deciles.

A Chi-squared test test is used to test if the expected frequencies are statistically different from the predicted ones.

```
hoslem.test(mod_b$y, fitted(mod_b), g = 10)
```

```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  mod_b$y, fitted(mod_b)
X-squared = 20.496, df = 8, p-value = 0.008615
```

The p-value is small (0.009). Therefore, we reject the null hypothesis that the observed proportions are equal to the predicted probabilities (i.e. good fit of the specified multivariable logistic regression model).

Non-linear model

The previous model assumes that the (log) odds of hyponatremia linearly depends on BMI. This assumption may not be appropriate and thus results in a low fit.

A possible non-linear model can be obtained by including a quadratic term

$$\log\left(\text{odds}(\text{nas135}|\text{bmi}, \text{sex})\right) =$$
$$\beta_0 + \beta_1 \text{sex} + \beta_2(\text{bmi-23}) + \beta_3(\text{bmi-23})^2$$

```
mod_b2 <- update(mod_b, . ~ . + I((bmi-23)^2))
summary(mod_b2)
```

```
Call:
glm(formula = nas135 ~ female + I(bmi - 23) + I((bmi - 23)^2),
    family = "binomial", data = marathon)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4499  -0.6059  -0.3571  -0.3253   2.4404

Coefficients:
                Estimate Std. Error z value  Pr(>|z|)
(Intercept)     -2.92158    0.26184 -11.158   < 2e-16
femalefemale     1.29091    0.33906   3.807  0.000141
I(bmi - 23)     -0.02717    0.05758  -0.472  0.637038
I((bmi - 23)^2)  0.04473    0.01085   4.124 0.0000373

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 357.62  on 464  degrees of freedom
Residual deviance: 320.93  on 461  degrees of freedom
  (23 observations deleted due to missingness)
AIC: 328.93
```

Question 1. Is bmi predicting the odds of hyponatremia?

```
wald.test(vcov(mod_b2), coef(mod_b2), Terms = 2:3)
```

```
Wald test:
----------

Chi-squared test:
X2 = 21.0, df = 2, P(> X2) = 0.000028
```

```
modf <- glm(nas135 ~ female, data = filter(marathon, !is.na(bmi)),
            family = "binomial")
lrtest(mod_b2, modf)
```

```
Likelihood ratio test for MLE method
Chi-squared 2 d.f. =  18.45288 , P value =  0.00009840307
```

The p-value is small, so the answer is yes.

Question 2. Is a quadratic model for bmi predicting odds of hyponatremia better compared to a simpler linear-response model ($H_0 : \beta_3 = 0$)?

```
lrtest(mod_b2, mod_b)
```

```
Likelihood ratio test for MLE method
Chi-squared 1 d.f. =  16.38858 , P value =  0.00005159526
```

Or looking at the univariate p-value, there is evidence of departure from the linearity assumption for bmi and (log) odds of hyponatremia (p-value < 0.01).

Linear combination of regression coefficients

NB: the coefficients for BMI are not directly interpretable.

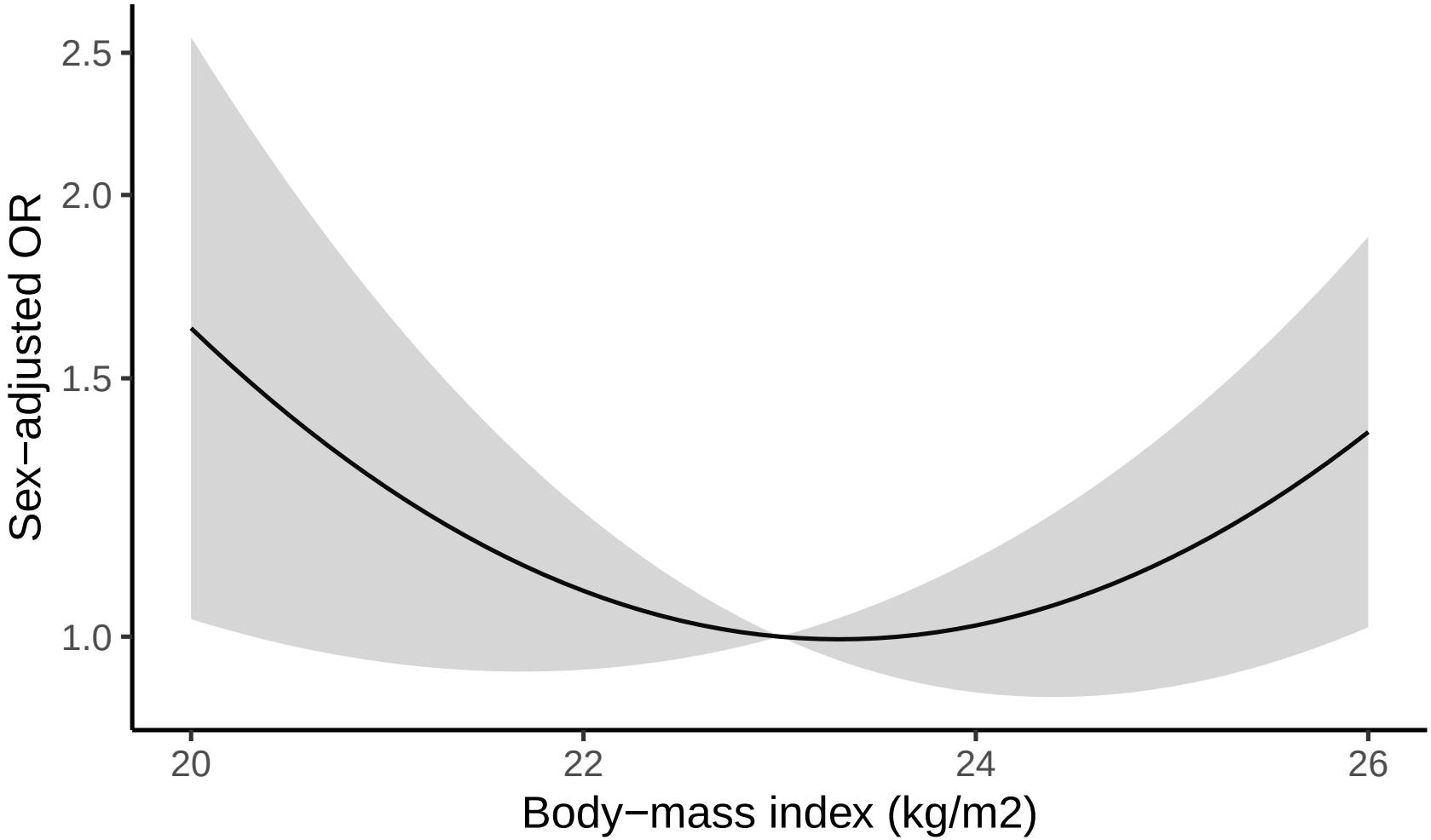
How to compare the log odds for two value of X: x2 compared to x1?

$$\log(\text{odds}(Y|C = c, X = x_1)) = \hat{\beta}_0 + \hat{\beta}_1 c + \hat{\beta}_2 x_1 + \hat{\beta}_3 x_1^2$$
$$\log(\text{odds}(Y|C = c, X = x_2)) = \hat{\beta}_0 + \hat{\beta}_1 c + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2$$

The adjusted OR is no longer constant but depends from the values to be compared

$$\log(\text{odds}(Y|C = c, X = x_2)) - \log(\text{odds}(Y|C = c, X = x_1)) = \beta_2(x_2 - x_1) + \beta_3(x_2^2 - x_1^2)$$

Graphical presentation OR



Predicted (log) odds and probabilities

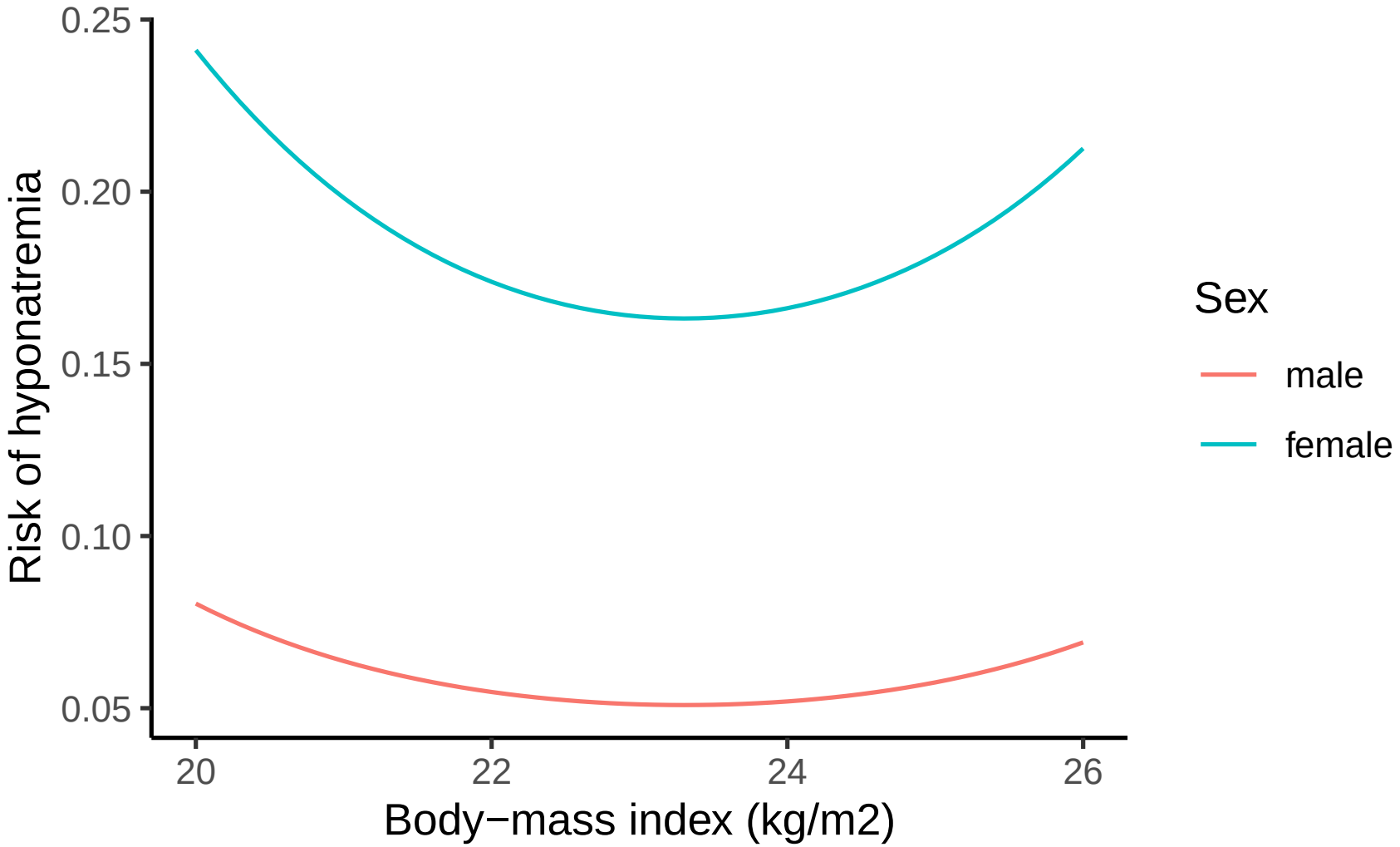
Question: what it the (log) odds of hyponatremia for men with a bmi of 25 kg/m^2?

$\widehat{\text{est}} = \log(\text{odds}(Y|X_1 = 0, X_2 = 25)) = \hat{\beta}_0 + (25 - 23)\hat{\beta}_2 + (25 - 23)^2\hat{\beta}_2 = -2.922 + 2 \cdot -0.027 + 4 \cdot 0.045 = -2.797$

Question: what it the risk of hyponatremia for men with a bmi of 25 kg/m^2?

$P(Y|X_1 = 0, X_2 = 25) = \text{invlogit}(\log(\text{odds}(Y|X_1 = 0, X_2 = 25))) = \text{invlogit}(-2.797) = 0.057$

Graphical presentation probabilities



```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  mod_b2$y, fitted(mod_b2)
X-squared = 6.2402, df = 8, p-value = 0.6203
```