Working with Stata

Inference on proportions

Nicola Orsini

Biostatistics Team
Department of Public Health Sciences
Karolinska Institutet

Outline

- Inference on one population proportion
- Principle of maximum likelihood
- Risk, odds, and their relationship
- Introducing logistic regression

Inference on one proportion

Suppose we randomly pick a sample of n observations y_i from a certain population.

y_i defines a random variable that follows a Bernoulli distribution (special case of the binomial distribution).

The observations are independent from one another.

For each individual we observed either 1 or 0 with probability p and $(1 - p)$.

We assume that there is underlying population proportion (p) that is the same for all the individuals.

We don't know p, we wish to quantify it.

The method of maximum likelihood provides an estimate of p that maximize the probability (likelihood) of obtaining the data included in the sample.

Likelihood function

The contribution to the likelihood function for each observation is

$$p^{y_i}(1-p)^{1-y_i}$$

the likelihood function is obtained as their product if they are independent

$$L(p|y_i) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

Mathematically and computationally easier to work with logarithms.

$$l(p|y_i) = \sum_{i=1}^{n} y_i \log(p) + (1 - y_i)\log(1 - p)$$

Given the data, we maximize the log-likelihood function with respect to p.

Maximization typically requires an iterative algorithm.

Miniature example

Suppose we collect the following sample of 5 observations

	y	
1.	1	p
2.	0	$1-p$
3.	1	p
4.	0	$1-p$
5.	0	$1-p$

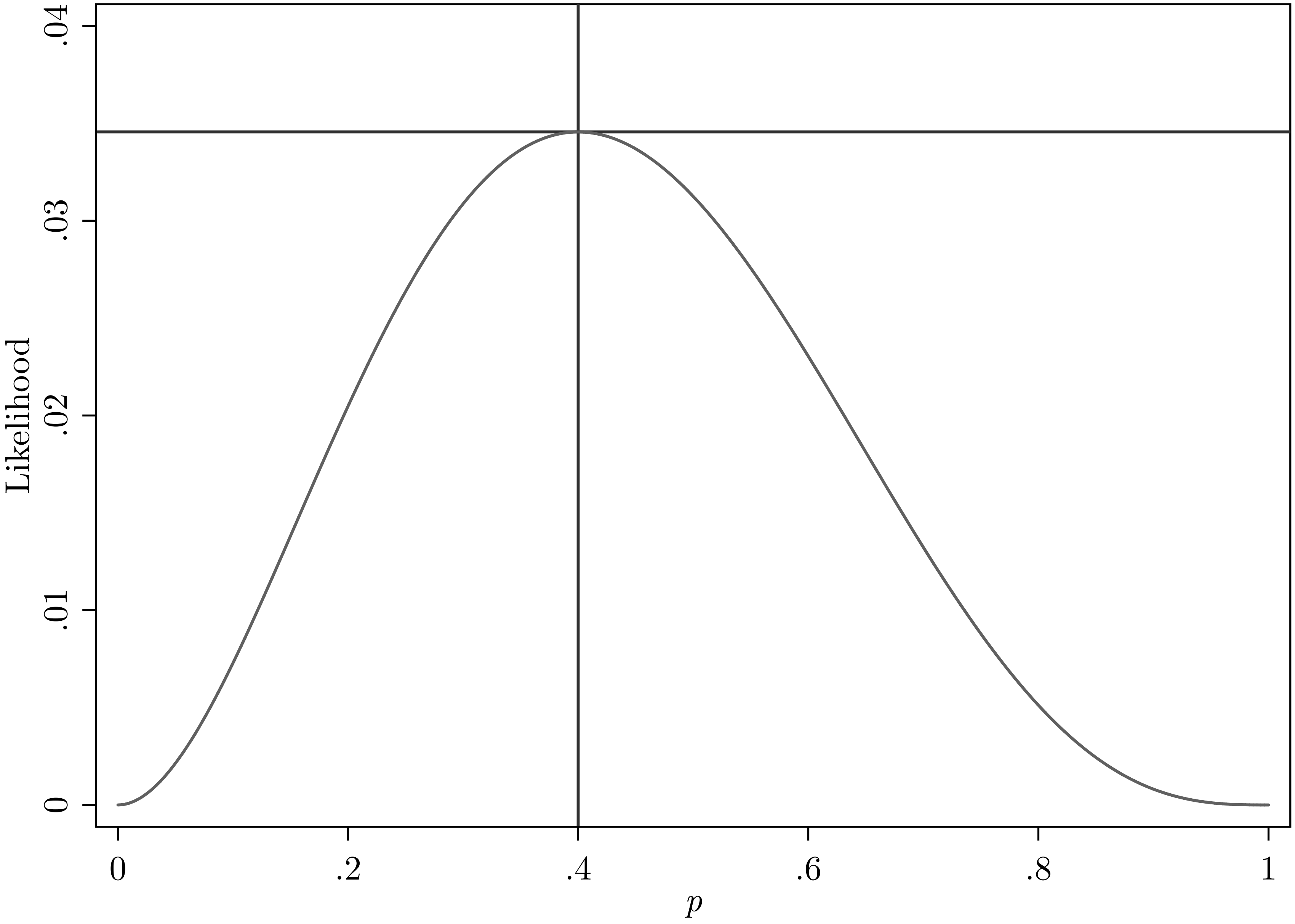
We need to find the value of p that maximizes the likelihood of the observed data.

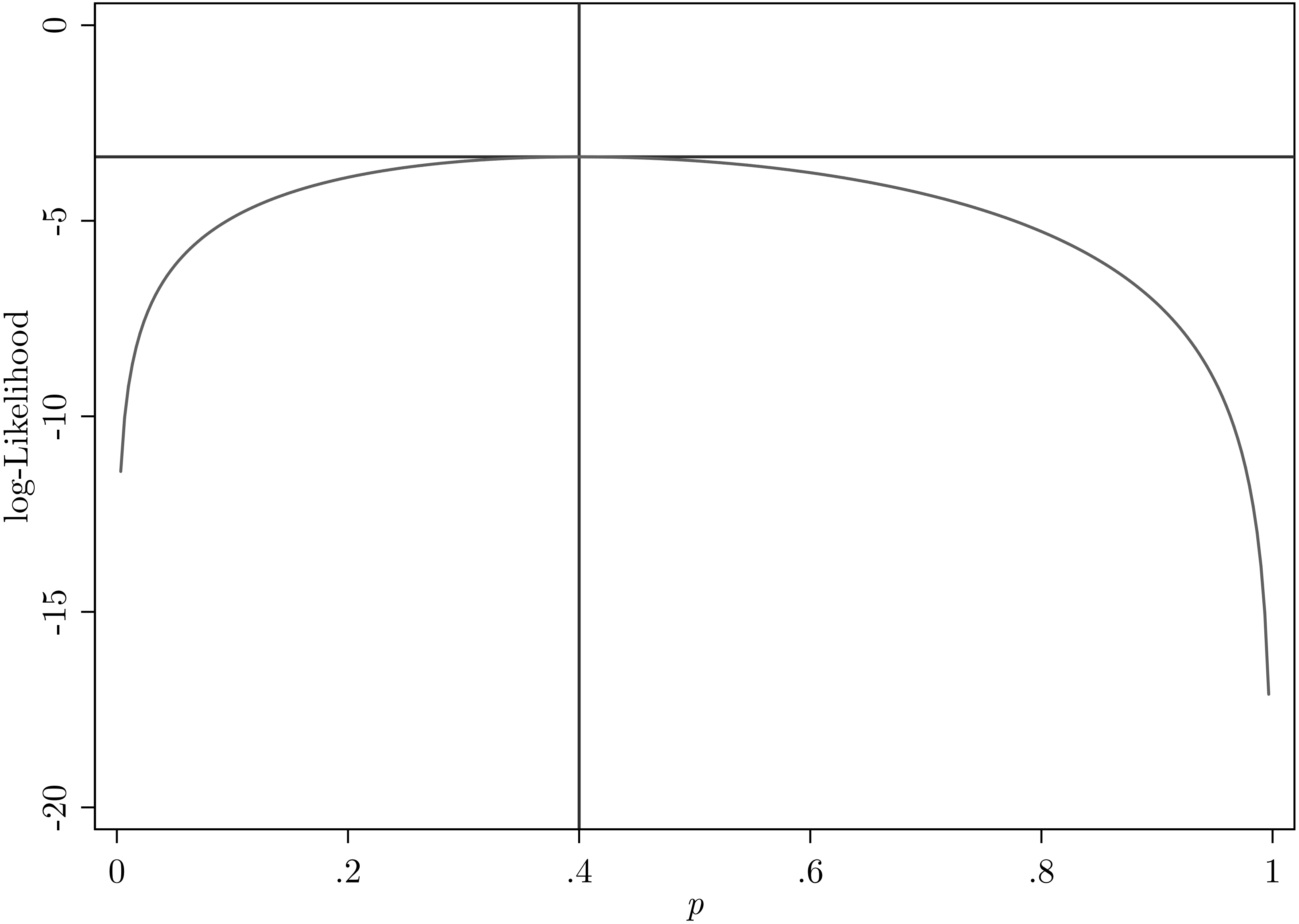
```
. mlexp (y*ln({p})+(1-y)*ln(1-{p}))
```

```
Iteration 0:   log likelihood = -3.4657359
Iteration 1:   log likelihood = -3.3650586
Iteration 2:   log likelihood = -3.3650583

Maximum likelihood estimation

Log likelihood = -3.3650583                     Number of obs     =          5

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          /p |         .4    .219089     1.83   0.068    -.0294066    .8294066
------------------------------------------------------------------------------
```

The maximum likelihood estimate of the population proportion *p* is 0.4. Actually it is the sample proportion, 2 out of 5.

```
. scalar p = .4

. di log(p)+log(1-p)+log(p)+log(1-p)+log(1-p)
-3.3650583
```


Likelihood
.04
.03
.02
.01
0
0
.2
.4
.6
.8
1
p


0
-5
-10
-15
-20
log-Likelihood
0
.2
.4
.6
.8
1
p

Large sample

Suppose that we draw 500 observations and 50 of them experienced the outcome of interest (y_i=1).

What is the estimated population proportion? And what are the limits of a 95% confidence interval?

The population proportion p is estimated to be 50/500=0.1 (10%).

A 95% confidence interval for population proportion p can be obtained as

$$p \pm 1.96 \times \sigma/\sqrt{n}$$

where $\sigma = \sqrt{p(1-p)} = \sqrt{0.1(1-0.1)} = 0.3$

$$\text{standard error} = 0.3/\sqrt{500} = 0.013$$

$$0.1 \pm 1.96 \times \frac{0.3}{\sqrt{500}} = 0.07 - 0.13$$

Simulations

1. Fix a sample size $n = 500$
2. Draw i.i.d. observations y_i (0/1) from a population where the proportion is $p = 0.1$
3. Estimate the sample proportion of the binary outcome y_i in the sample

Repeat Steps 1 to 3 a large number of times, for example $s = 1000$.

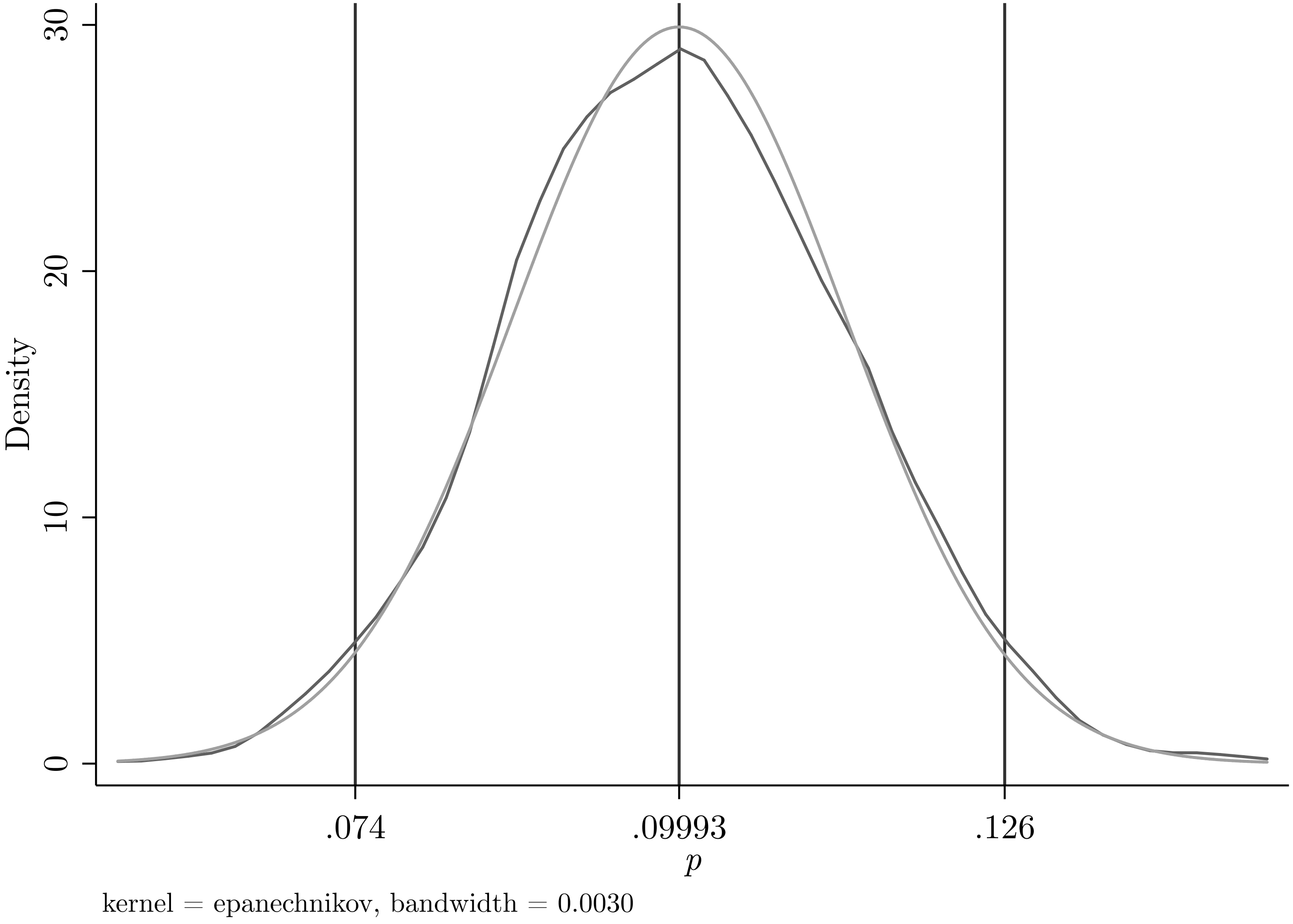
```
. su sample_p
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    sample_p |      1,000      .09993    .0133354       .058       .144
```

The mean of the all 1,000 sample proportions is 0.1, the true population proportion.

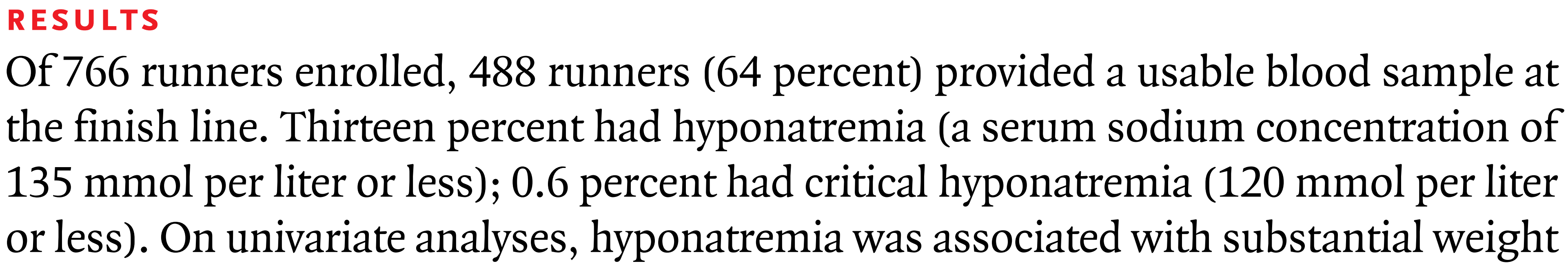
The standard deviation of the sample proportions is close to the standard error used in our hand calculations according to the theory

$$\sqrt{p(1-p)/n} = \sqrt{0.1(1-0.1)/500} = 0.013$$



Vertical lines are the 2.5 th percentile, Mean, and 97.5th percentile

Hyponatremia risk



RESULTS

Of 766 runners enrolled, 488 runners (64 percent) provided a usable blood sample at the finish line. Thirteen percent had hyponatremia (a serum sodium concentration of 135 mmol per liter or less); 0.6 percent had critical hyponatremia (120 mmol per liter or less). On univariate analyses, hyponatremia was associated with substantial weight

```
. count if na <= 135
   62

. di %2.0f 62/488*100      // Incidence of hyponatremia
13

. count if na <= 120       // Incidence of critical hyponatremia
    3

. di %2.1f 3/488*100
0.6
```

Table of counts

```
. tabulate nas135

      Serum |
     sodium |
concentrati |
  on <= 135 |
 mmol/liter |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        426       87.30       87.30
          1 |         62       12.70      100.00
------------+-----------------------------------
      Total |        488      100.00
```

About thirteen percent of the marathon runners had hyponatremia (a serum sodium concentration of 135 mmol per liter or less).

Confidence interval for the risk

The risk (*p*) corresponds to the proportion of 1′s (62/488).

```
. summarize nas135

    Variable |       Obs        Mean    Std. Dev.
-------------+-------------------------------------
      nas135 |       488    .1270492    .3333698
```

The 95% confidence interval for the population prevalence can be calculated as if the binary variable for hyponatremia was a numerical variable.

```
.127 +/- 1.96*.333/sqrt(488)
```

We are 95% confident that the population risk of hyponatremia is between 10% and 16%.

For proportions, not for continuous variables, knowing the population proportion implies knowing the population standard deviation, $[p(1-p)]^{1/2}$. Thus, we can use $(0.127 \times 0.873)^{1/2}$ instead of the sample standard deviation.

```
.127 +/- 1.96*sqrt(.127*.873/488)
```

Table of counts

```
. tabulate nas135

       Serum |
      sodium |
concentrati |
  on <= 135 |
 mmol/liter |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        426       87.30       87.30
          1 |         62       12.70      100.00
------------+-----------------------------------
      Total |        488      100.00
```

About thirteen percent of the marathon runners had hyponatremia (a serum sodium concentration of 135 mmol per liter or less).

(range, 114 to 158). Thirteen percent (62 of 488) had hyponatremia, including 22 percent of women (37 of 166) and 8 percent of men (25 of 322). Three runners (0.6 percent) had critical hyponatremia (serum sodium concentrations, 119, 118, and 114 mmol per liter).

```
. tabulate nas135 female, col

     Serum |
    sodium |
concentrat |
ion <= 135 |        Female
mmol/liter |        No        Yes |     Total
-----------+----------------------+----------
        No |       297        129 |       426
           |     92.24      77.71 |     87.30
-----------+----------------------+----------
       Yes |        25         37 |        62
           |      7.76      22.29 |     12.70
-----------+----------------------+----------
     Total |       322        166 |       488
           |    100.00     100.00 |    100.00
```

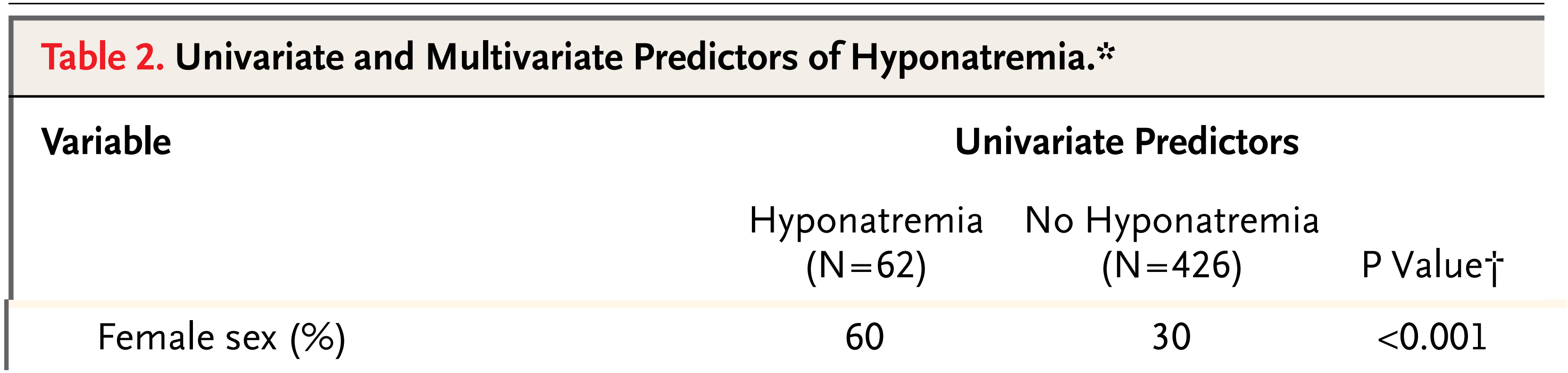


Table 2. Univariate and Multivariate Predictors of Hyponatremia.*

Variable	Univariate Predictors		
	Hyponatremia (N=62)	No Hyponatremia (N=426)	P Value†
Female sex (%)	60	30	<0.001

```
. tabulate nas135 female, row
     Serum |
    sodium |
concentrat |
ion <= 135 |        Female
mmol/liter |        No        Yes |     Total
-----------+----------------------+----------
        No |       297        129 |       426
           |     69.72      30.28 |    100.00
-----------+----------------------+----------
       Yes |        25         37 |        62
           |     40.32      59.68 |    100.00
-----------+----------------------+----------
     Total |       322        166 |       488
           |     65.98      34.02 |    100.00
```

The Chi-square test – 2 by 2 table

Let's compare the proportions of cases of hyponatremia in the populations of men and women.

We test the null hypothesis that the two proportions are identical.

The Pearson Chi-Square statistics is based on the comparison of observed (*o*) and expected (*e*) counts under the null hypothesis of independence.

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

```
. tabulate nas135 female, chi2 expected

     Serum |
    sodium |
concentrat |
ion <= 135 |        Female
mmol/liter |        No        Yes |     Total
-----------+----------------------+----------
        No |       297        129 |       426
           |     281.1      144.9 |     426.0
-----------+----------------------+----------
       Yes |        25         37 |        62
           |      40.9       21.1 |      62.0
-----------+----------------------+----------
     Total |       322        166 |       488
           |     322.0      166.0 |     488.0

          Pearson chi2(1) =  20.8365   Pr = 0.000
```

Expected counts are hypothetical counts that we would expect if no association were observed.

If overall 12.7% of the runners had hyponatremia I would expect approximately the same proportion of cases among males and females if no variation by gender were present.

You can calculate expected counts by hand like this:

322*0.127 = 41
166*0.127 = 21

The Pearson Chi2 statistic is 20.8 and the *p*-value is lower than 0.05.

We reject the null hypothesis that the proportions of hyponatremia in the populations of male and female are identical.

In particular, the risk of hyponatremia among females (22%) is approximately three times the risk of hyponatremia among males (8%).

Measures of disease occurrence

Risk = p = Cases / Total

Odds = p / (1-p) = Cases / Non-cases

p = Odds / (1+ Odds)

The probability (62/488) or risk of experiencing hyponatremia was 0.13. Incidence proportion of 13%. We expect 13 cases for every 100 marathon runners.

The odds (62/426) of experiencing hyponatremia was 0.15. Incidence odds of 15%. We expect 15 cases for every 100 non-cases.

Binary exposure

Are the odds of hyponatremia associated with gender?

The data can be summarized with a 2 by 2 table.

Various measure of exposure-disease association can be calculated with a binary outcome (risk difference, risk ratio, odds ratio).

This course will focus on the ratio of odds (Odds Ratio) which can be easily calculated using tables for epidemiologists (help epitab).

Table of counts

	$X=1$	$X=0$	Total
$Y=1$	a	b	n_1
$Y=0$	c	d	n_2

Odds Ratio

$$OR = \frac{a / c}{b / d}$$

The standard error (SE) of the **log** odds ratio ln(*OR*):

$$SE[\ln(OR)] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

```
. cs nas135 female, or woolf
```

```
                 | Female                 |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        37          25  |         62
        Noncases |       129         297  |        426
-----------------+------------------------+------------
           Total |       166         322  |        488
                 |                        |
            Risk |  .2228916    .0776398  |   .1270492
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .1452518       |    .0755191    .2149846
      Risk ratio |         2.870843       |    1.791408    4.600706
      Odds ratio |         3.407442       |    1.970056     5.89357
                 +-------------------------------------------------
                               chi2(1) =    20.84  Pr>chi2 = 0.0000
```

Interpretation of point estimates

The odds of hyponatremia is 0.29 (37/129) among female.

The odds of hyponatremia is 0.08 (25/297) among male.

The odds of hyponatremia among female is 3.4 times higher than males (0.29/0.08).

Interpretation of confidence interval estimates

The 95% CI calculated using the Woolf's method or Wald-test type is given by

$$\exp(\ \log(OR) \pm 1.96 \times \text{Standard error of } \log(OR)\ )$$

where the standard error of the log odds ratio is the square root of the sum of the inverse of the cell counts.

```
. di exp( log(3.4074)-1.96*sqrt(1/25 + 1/37 + 1/297 + 1/129) )
1.9700115

. di exp( log(3.4074)+1.96*sqrt(1/25 + 1/37 + 1/297 + 1/129) )
5.8935569
```

In repeated samples, 95% of confidence intervals include the true population value.

We are 95% confident that the odds ratio relating gender (female compared to male) to hyponatremia is between 1.97 and 5.89.

The size of confidence interval depends on sample size.

Your confidence interval may or may not include the true population value!!

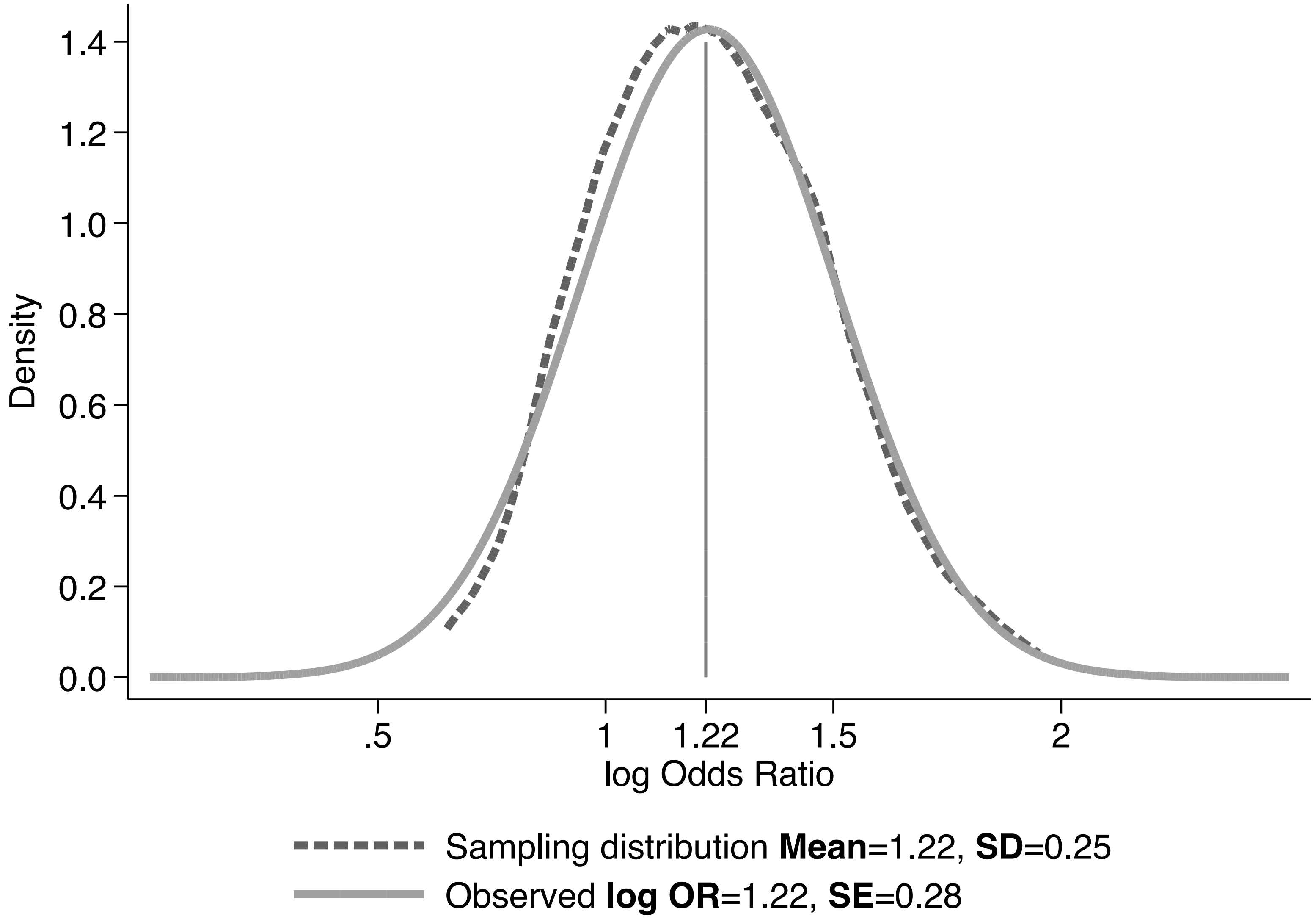
About 5% of your 95% confidence intervals do not, but you don’t know which ones!!

Sampling distribution of the log odds ratio

Let us simulate a population with characteristics similar to the hyponatremia study.

Incidence of hyponatremia = 13%
Sample size = 488
Percent of women = 32%
OR of hyponatremia women vs men = 3.4 = exp(1.22)

We sample from this population 300 times. Every time we save the estimate of the log Odds Ratio.


Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
.5
1
1.22
1.5
2
log Odds Ratio
Sampling distribution **Mean**=1.22, **SD**=0.25
Observed **log OR**=1.22, **SE**=0.28

The standard error of the log Odds Ratio estimated in the hyponatremia study

```
. di sqrt(1/25 + 1/37 + 1/297 + 1/129)
.28
```

is an estimate of the variability of the log Odds Ratios (SD=0.25 in our simulation) that one could sample from the same population.

Logistic function

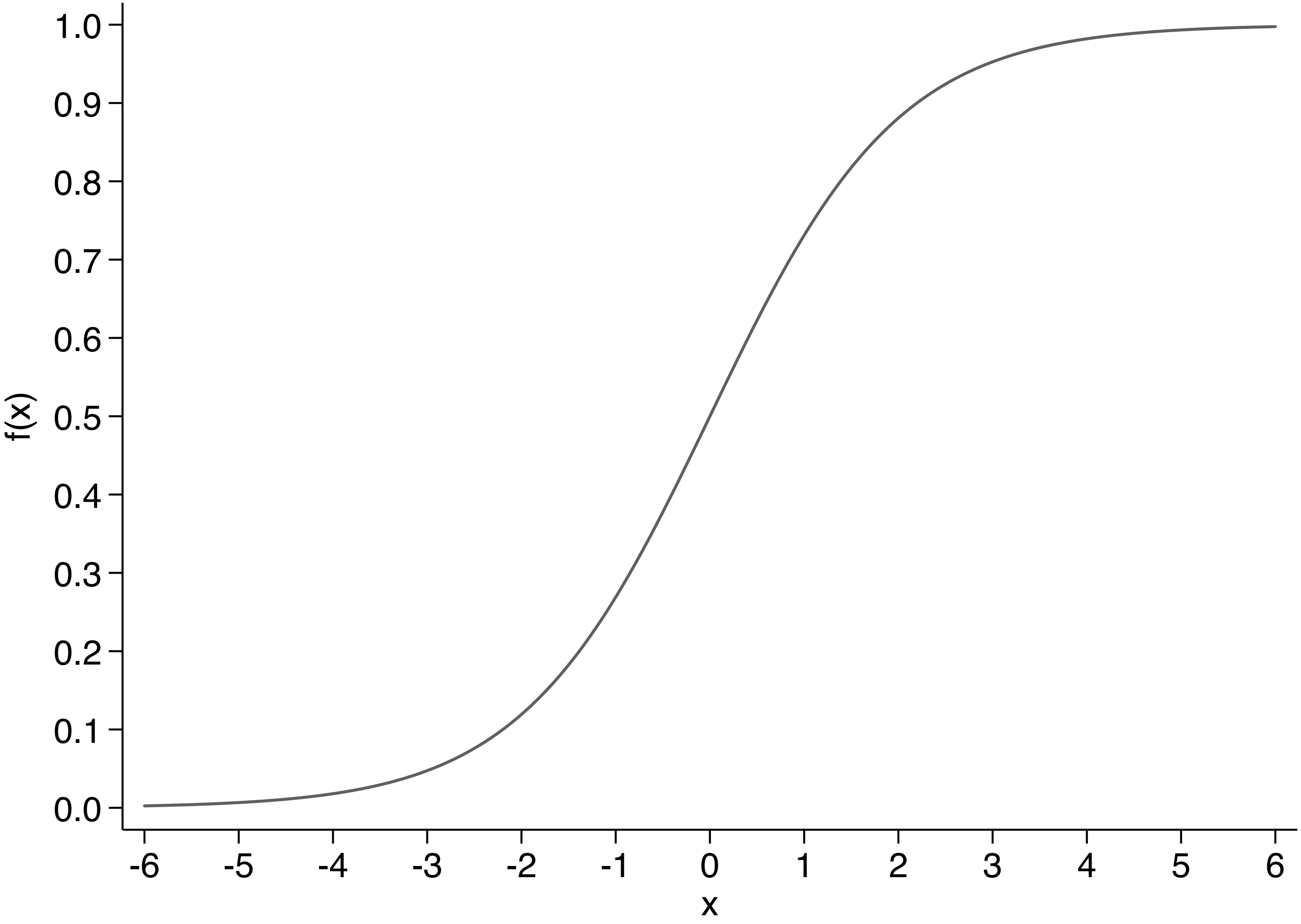
The **logistic function** describes the mathematical form on which the **logistic model** is based.

This function, called f(x), is given by 1 over 1 plus *e* to the minus *x*.

$$f(x) = \frac{e^{x}}{1+e^{x}} = \frac{1}{1+e^{-x}}$$

We can plot the values of this function as *x* varies from minus infinity to plus infinity.

```
twoway ///
  (function exp(x)/(1+exp(x)) , range(-6 6)) , ///
 plotregion(style(none)) scheme(s1mono) ///
  ylabel(0(.1)1, angl(horiz) format(%2.1fc)) ///
xlabel(-6(1)6) ytitle("f(x)")
```


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
f(x)
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
x

The logistic function *f*(*x*) **ranges between 0 and 1** and it is probably the main reason the logistic model is so popular.

The model is designed to describe a probability, which is always some number between 0 and 1.

In epidemiologic terms, such a probability gives the **risk** of an individual getting a disease.

Logistic regression model

It is mathematical model to make inference on the probability of a binary outcome given a set of covariates.

It can be used for any type of exposure: binary, continuous, or categorical covariate values.

It allows adjustment for confounding, assessment of effect modification (interaction).

Estimation method: Maximum likelihood (yields point estimates, standard error estimates, confidence intervals, and *p*-values)

No predictors

$$odds = \frac{P(y=1)}{1-P(y=1)} = \frac{p}{1-p}$$

$$logit(P(y=1)) = \log(odds|x) = \beta_0$$

$$odds = \exp(\beta_0)$$

$$P(y=1) = p = \frac{odds}{1+odds} = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$$

```
. logit nas135

Logistic regression                               Number of obs   =        488
Log likelihood = -185.80042                       Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
      nas135 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -1.927305   .1359281   -14.18   0.000    -2.193719   -1.660891
------------------------------------------------------------------------------
```

_b[_cons] = -1.927 is an estimate of the overall **log odds** of hyponatremia β_0

exp(-1.927) = 0.15 is an estimate of the overall **odds** hyponatremia

$\frac{\exp(-1.927)}{1+\exp\,(-1.927)}$=0.13 is an estimate of the overall **risk** of hyponatremia

Mathematical functions

```
. scalar p = 62/488

. scalar odds = 62/426

. display logit(p)
-1.927

. display invlogit(log(odds))
.127
```

The **`invlogit()`** function can be very useful to go from log odds to risk of the outcome.

Binary predictor

Logistic regression model with one binary (1/0) predictor *x*

$$P(y = 1|x) = \frac{\text{odds}}{1 + \text{odds}} = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$\text{logit}\big(P(y = 1|x)\big) = \log(\text{odds}|x) = \beta_0 + \beta_1 x$$

β_0 is the population log odds of the outcome when *x* = 0.

β_1 is the difference in the population log odds of the outcome (or log odds ratio) comparing *x*=1 vs. *x*=0.

$\exp(\beta_0)$ is the population odds of the outcome when *x*=0.

$\exp(\beta_0 + \beta_1 x)$ is the population odds of the outcome when *x*=1.

$\exp(\beta_1)$ is the population ratios of odds (OR) comparing *x*=1 vs. *x*=0.

There are different Stata commands available to estimate odds ratios because the binomial distribution of the outcome is a special case of a larger family of regression models known as generalized linear models (help glm).

We will use interchangeably logit or logistic and they are equivalent.

$$\exp(\beta_0 + \beta_1 x)$$

```
. logit nas135 female

Logistic regression                               Number of obs   =        488
                                                  LR chi2(1)      =      19.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -175.96547                       Pseudo R2       =     0.0529

------------------------------------------------------------------------------
      nas135 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.225962   .2795448     4.39   0.000      .678064     1.77386
       _cons |  -2.474856   .2082468   -11.88   0.000    -2.883013     -2.0667
------------------------------------------------------------------------------
```

The intercept **_cons**, that is an estimate of parameter β_0, is the log of the odds of hyponatremia among male (female = 0).

The coefficient of the variable **female**, that is an estimate of β_1, is the log of the odds ratio of hyponatremia for female (1) vs male (0).

Interpretation

```
. di exp(-2.474856)
.08417511
```

The odds of hyponatremia are 0.08 (8 cases for every 100 non-cases) among men.

```
. di exp(1.2259)
3.4072312
```

The odds of hyponatremia among women are 3.4 times higher than men. In other words, you have to multiply by 3.4 the odds of hyponatremia among male (0.08) to get the one among female (0.29).

```
. logit nas135 female , or

Logistic regression                               Number of obs   =        488
                                                  LR chi2(1)      =      19.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -175.96547                       Pseudo R2       =     0.0529

------------------------------------------------------------------------------
      nas135 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   3.407442   .9525368     4.39   0.000     1.970056     5.89357
       _cons |   .0841751   .0175292   -11.88   0.000     .0559658     .126603
------------------------------------------------------------------------------
```

The odds of hyponatremia among female was significantly higher than males (OR=3.41).

We are 95% confident that the odds ratio relating gender (being female compared to male) to hyponatremia is between 1.97 and 5.89.

```
. lincom _cons, eform

 ( 1)  _cons = 0

------------------------------------------------------------------------------
      nas135 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0841751   .0175292   -11.88   0.000     .0559659    .1266029
------------------------------------------------------------------------------
. lincom _cons + female, eform

 ( 1)  female + _cons = 0

------------------------------------------------------------------------------
      nas135 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .2868217   .0534894    -6.70   0.000     .1990084    .4133831
------------------------------------------------------------------------------
. lincom female, eform

 ( 1)  female = 0

------------------------------------------------------------------------------
      nas135 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   3.407442   .9525327     4.39   0.000      1.97006    5.893556
------------------------------------------------------------------------------
```

Predicted probabilities

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

What is the estimated probability of hyponatremia among male female? Remember that p = Odds / (1+ Odds).

p among male = **`exp(-2.4748)/(1+ exp(-2.4748))`**= 0.08

p among female =
`exp(-2.4748+1.2259)/(1+exp(-2.4748+1.2259))`= 0.22

The command **predict** will generate a new variable containing the predicted probabilities (or other statistics according to the specified options) for each individual.

```
. predict pr_nas135
(option pr assumed; Pr(nas135))

. list nas135  pr_nas135 in 1/5

     +-------------------+
     | nas135   pr_n~135 |
     |-------------------|
  1. |      0    .0776398 |
  2. |      1    .2228916 |
  3. |      0    .0776398 |
  4. |      0    .0776398 |
  5. |      1    .2228916 |
     +-------------------+
. table female, c(mean pr_nas135 mean nas135) f(%3.2f)

-------------------------------------------
   Female | mean(pr_n~135)     mean(nas135)
----------+--------------------------------
       No |           0.08             0.08
      Yes |           0.22             0.22
-------------------------------------------
```

The predicted probabilities are the same as the observed probabilities.

Remember that the mean of binary (0/1) variable is a probability.

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.

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

Continuous predictor

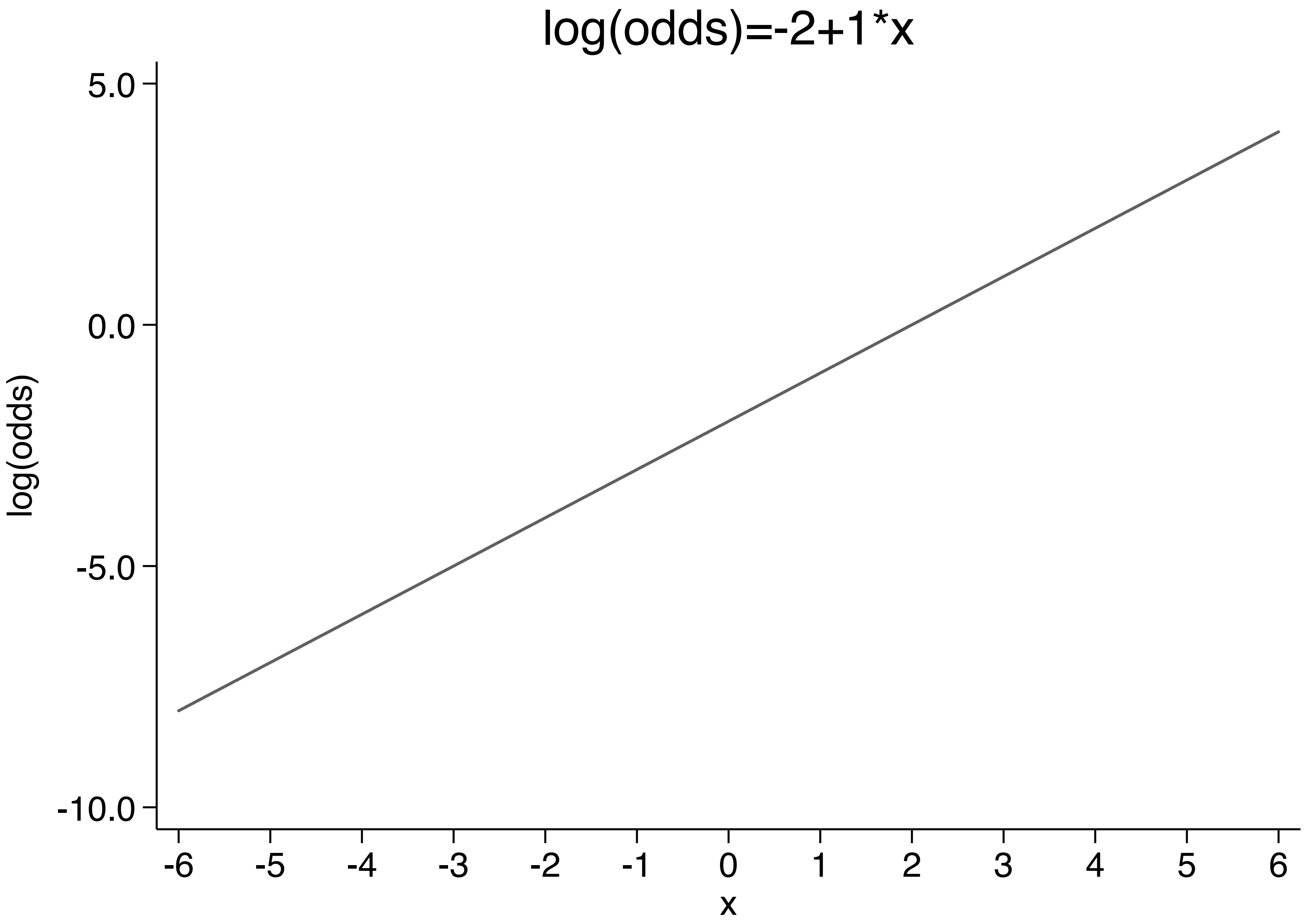
What is the change in the odds of the outcome for every 1 unit increase of the quantitative predictor *x* ?

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

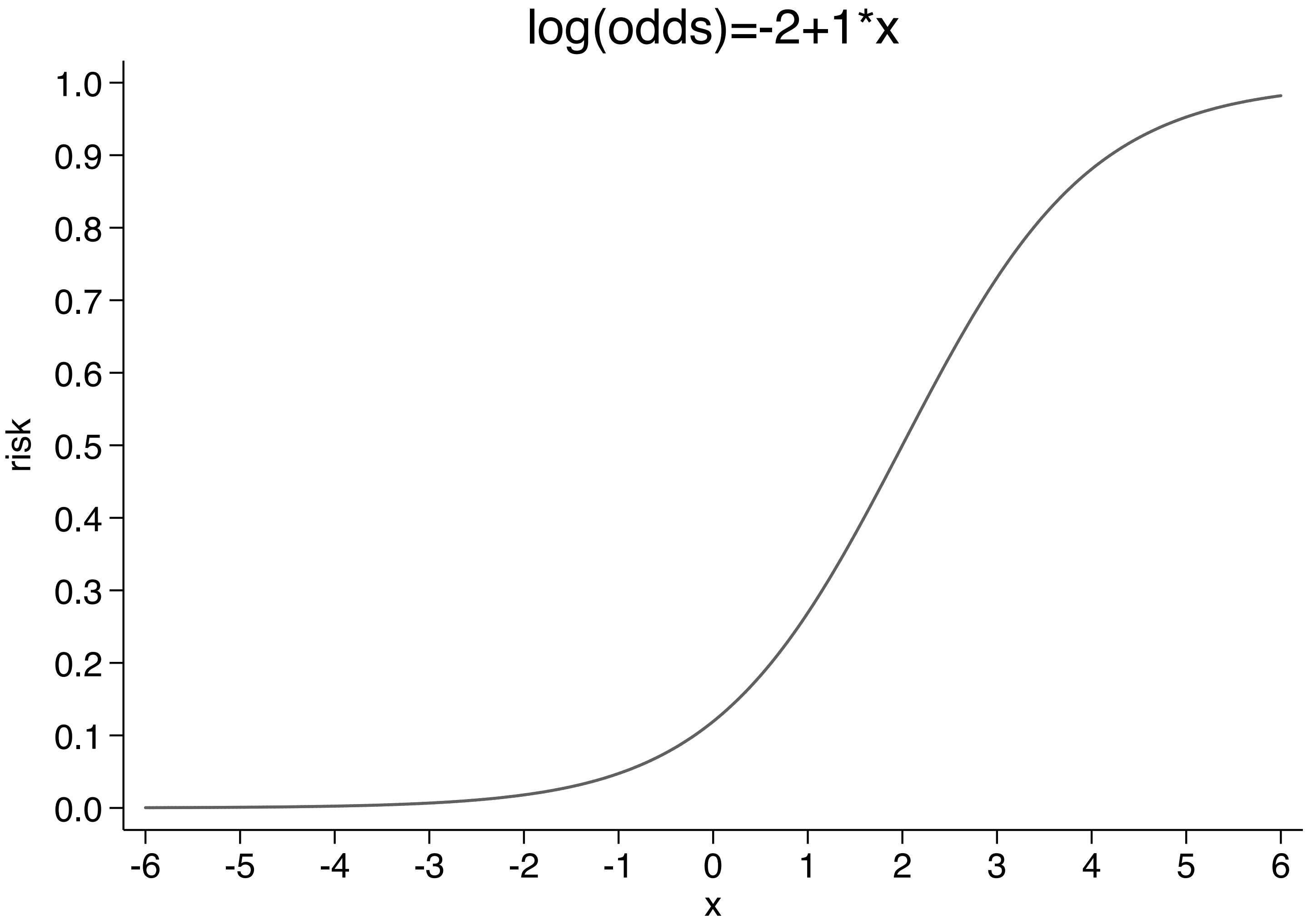
β_0 is the log odds of the outcome when $x = 0$.

β_1 is the change in the log odds of the outcome (or log odds ratio) for every 1 unit increase of the continuous predictor.

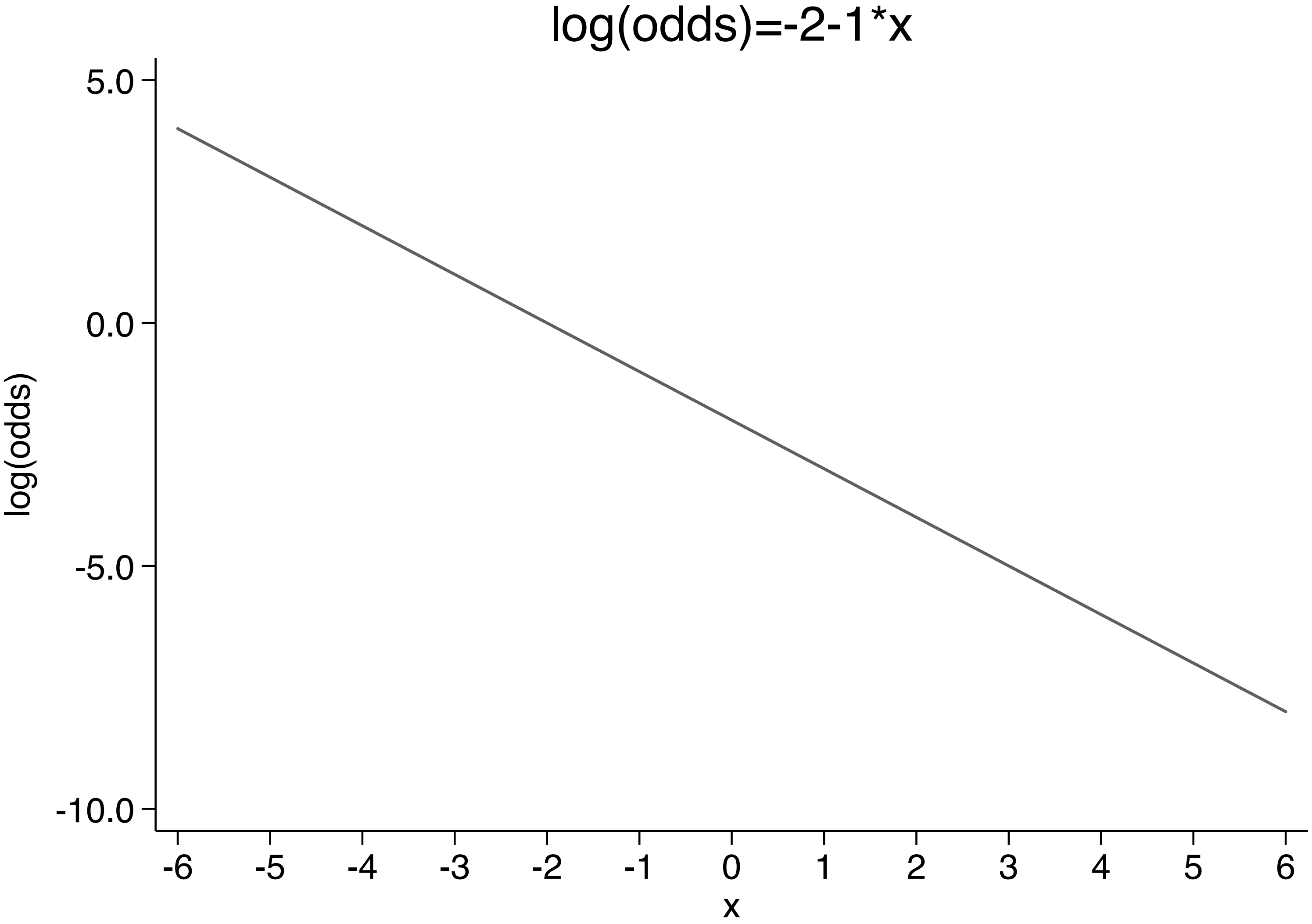
If β_1 is positive the log odds increases (linearly) with x



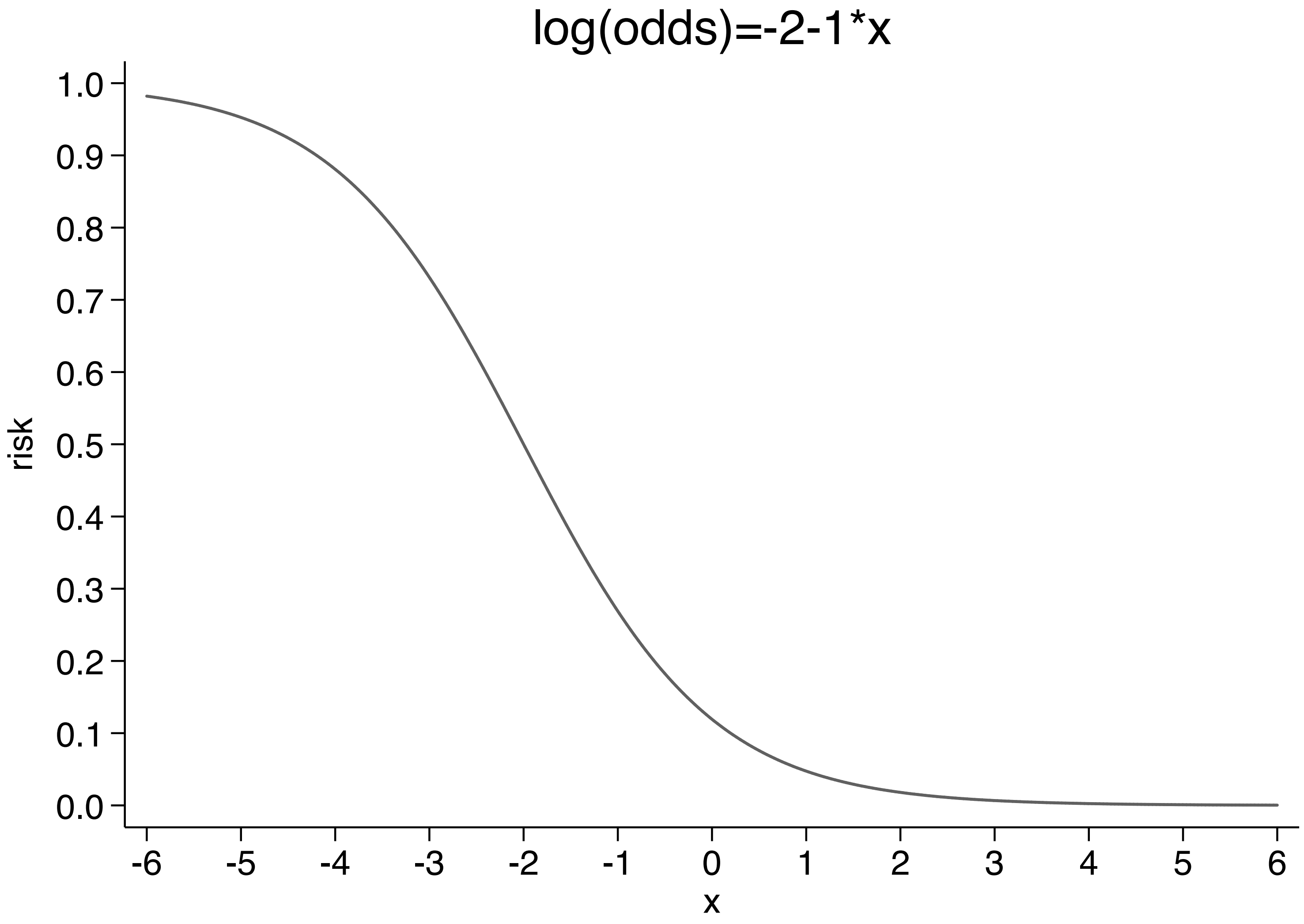
If β_1 is positive the risk increases with x



If β_1 is negative the log odds decreases (linearly) with x



If β_1 is negative the risk decreases with x



Let's investigate the association between weight change (either increase or decrease) during the marathon and the risk of hyponatremia.

Let's assume a linear relation between weight change and the log odds of hyponatremia as represented in Figure 2 of the NEJM paper.

Remember that the linearity assumption is strong and it may not hold.

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

```
. logit nas135 wtdiff, or

Logistic regression                               Number of obs   =        455
                                                  LR chi2(1)      =      54.40
                                                  Prob > chi2     =     0.0000
Log likelihood =  -144.4733                       Pseudo R2       =     0.1584

------------------------------------------------------------------------------
      nas135 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      wtdiff |   2.071862   .2285847     6.60   0.000     1.668973    2.572008
       _cons |   .1518379   .0240941   -11.88   0.000     .1112523    .2072295
------------------------------------------------------------------------------
```

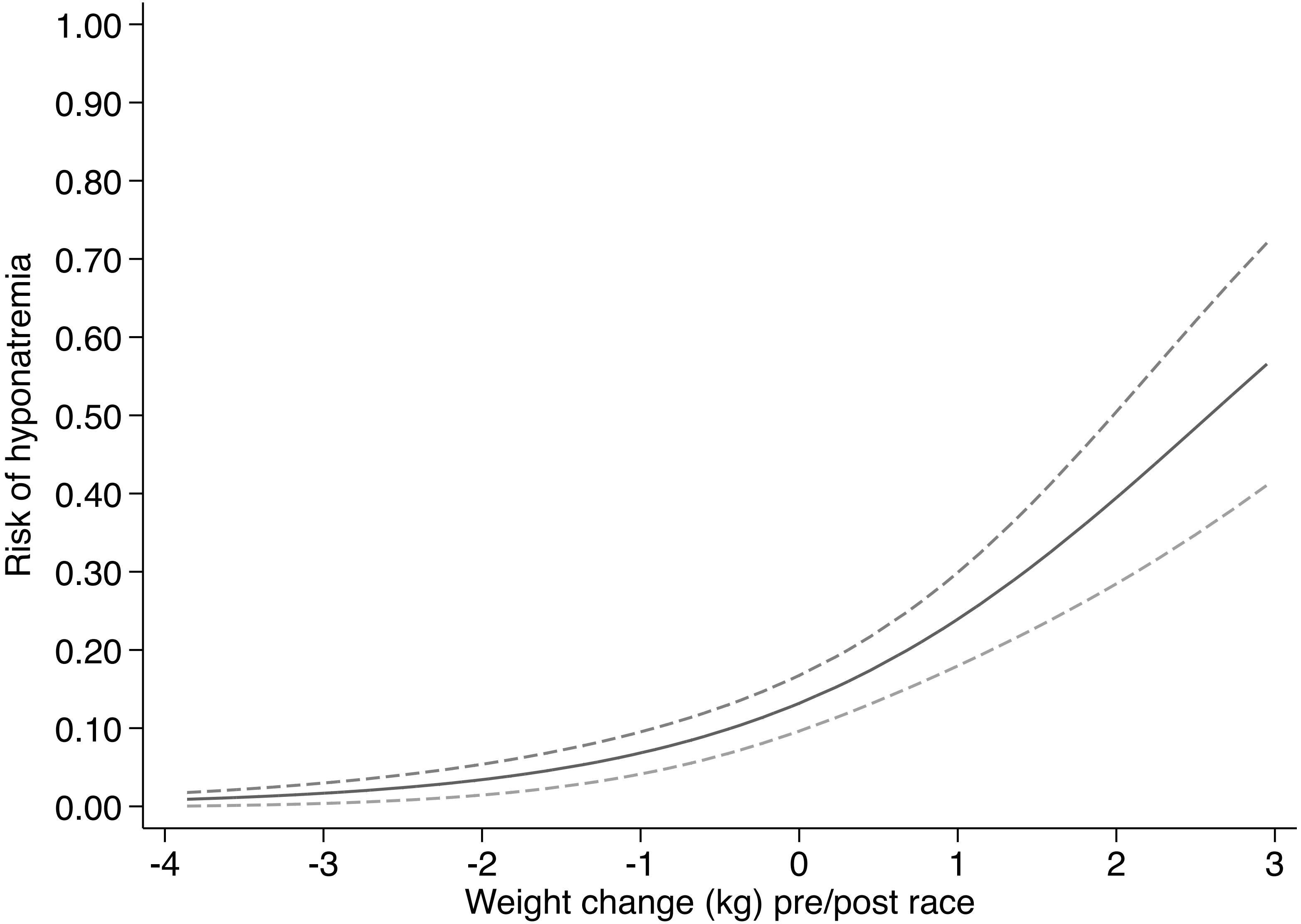
The odds of hyponatremia doubles (OR=2, 95% CI = 1.7-2.6) for every one kilogram increase in weight.

0.15 is the odds of hyponatremia for those runners who did not change weight (wtdiff=0).

$$P(\text{nas135} = 1|\text{wtdiff}) = \frac{\exp(-1.9 + 0.73\ \text{wtdiff})}{1 + \exp(-1.9 + 0.73\ \text{wtdiff})}$$

Graphical presentation of the predicted risk as function of weight change

```
predictnl prob = invlogit(_b[_cons]+_b[wtdiff]*wtdiff), ///
ci(lprob uprob)


tw (line prob lprob uprob wtdiff, sort lp(l - -)) ///
if inrange(wtdiff, -4, 3) , ///
ytitle("Risk of hyponatremia") ///
ylabel(0(.1)1, angle(horiz) format(%3.2fc)) ///
legend(off) ///
plotregion(style(none)) ///
xlabel(-4(1)3)
```


Risk of hyponatremia
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-4
-3
-2
-1
0
1
2
3
Weight change (kg) pre/post race

Constancy of the odds ratio

```
wtdiff        risk        odds          rr          or
    -4    .0081728    .0082402           .           .
    -3    .0167859    .0170725     2.05387    2.071862
    -2    .0341635    .0353719    2.035244    2.071862
    -1    .0682817    .0732857    1.998674    2.071862
     0    .1318223    .1518379    1.930567    2.071862
     1     .239305    .3145873     1.81536    2.071862
     2     .394593    .6517814    1.648913    2.071862
     3    .5745407    1.350401    1.456034    2.071862
```

Every 1 kg increase in weight change is associated with a two-fold increase odds of hyponatremia.

The log odds of the outcome for any two values of x are

$$\log(\text{odds}|x = x_1) = \beta_0 + \beta_1 x_1$$

$$\log(\text{odds}|x = x_2) = \beta_0 + \beta_1 x_2$$

The quantity

$$\log(\text{odds}|x = x_1) - \log(\text{odds}|x = x_2) = \beta_1(x_1 - x_2)$$

is the difference between two log odds of the outcome associated with a x_1 - x_2 unit increase of the covariate x

The log odds ratio is given by

$$\log\left(\frac{\text{odds}|x = x_1}{\text{odds}|x = x_2}\right) = \beta_1(x_1 - x_2)$$

Taking the exponential of both sides we get the odds ratio of the outcome

$$\text{OR} = \frac{\text{odds}|x = x_1}{\text{odds}|x = x_2} = \exp(\beta_1(x_1 - x_2))$$

comparing the sub-population having x_1 vs x_2 of the quantitative covariate x.

Confidence intervals

$$\log(\mathrm{OR}) = \beta_1(x_1 - x_2)$$

$$\mathrm{Var}(\beta_1(x_1 - x_2)) = \mathrm{Var}(\beta_1)(x_1 - x_2)^2$$

$$\mathrm{SE}(\mathrm{OR}) = \sqrt{\mathrm{Var}(\beta_1)(x_1 - x_2)^2}$$

95% CI = $\beta_1(x_1 - x_2) \pm 1.96\sqrt{\mathrm{Var}(\beta_1)(x_1 - x_2)^2}$

95% CI for the log Odds Ratio = $\log(\mathrm{OR}) \pm 1.96\ \mathrm{SE}(\log(\mathrm{OR}))$

95% CI for the Odds Ratio = $\exp\left(\log(\mathrm{OR}) \pm 1.96\ \mathrm{SE}(\log(\mathrm{OR}))\right)$

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

```
. lincom _b[wtdiff]*(2--1)

 ( 1)  3*[nas135]wtdiff = 0

------------------------------------------------------------------------------
      nas135 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   8.893702   2.943677     6.60   0.000     4.648878    17.01441
------------------------------------------------------------------------------
```

The odds of hyponatremia among those who increased 2 kg was 9 times the odds for those runners who lost 1kg.

Question 2. How do you plot of all the odd ratios of hyponatremia comparing the observed values of weight change using no change in weight as reference group (Figure 2)?

An easy way to get the predicted log odds ratios and the 95% confidence limits is using the post-estimation command **predictnl** (help predictnl).

```
predictnl logorl = _b[wtdiff]*(wtdiff-0), ci(lol hil)
```

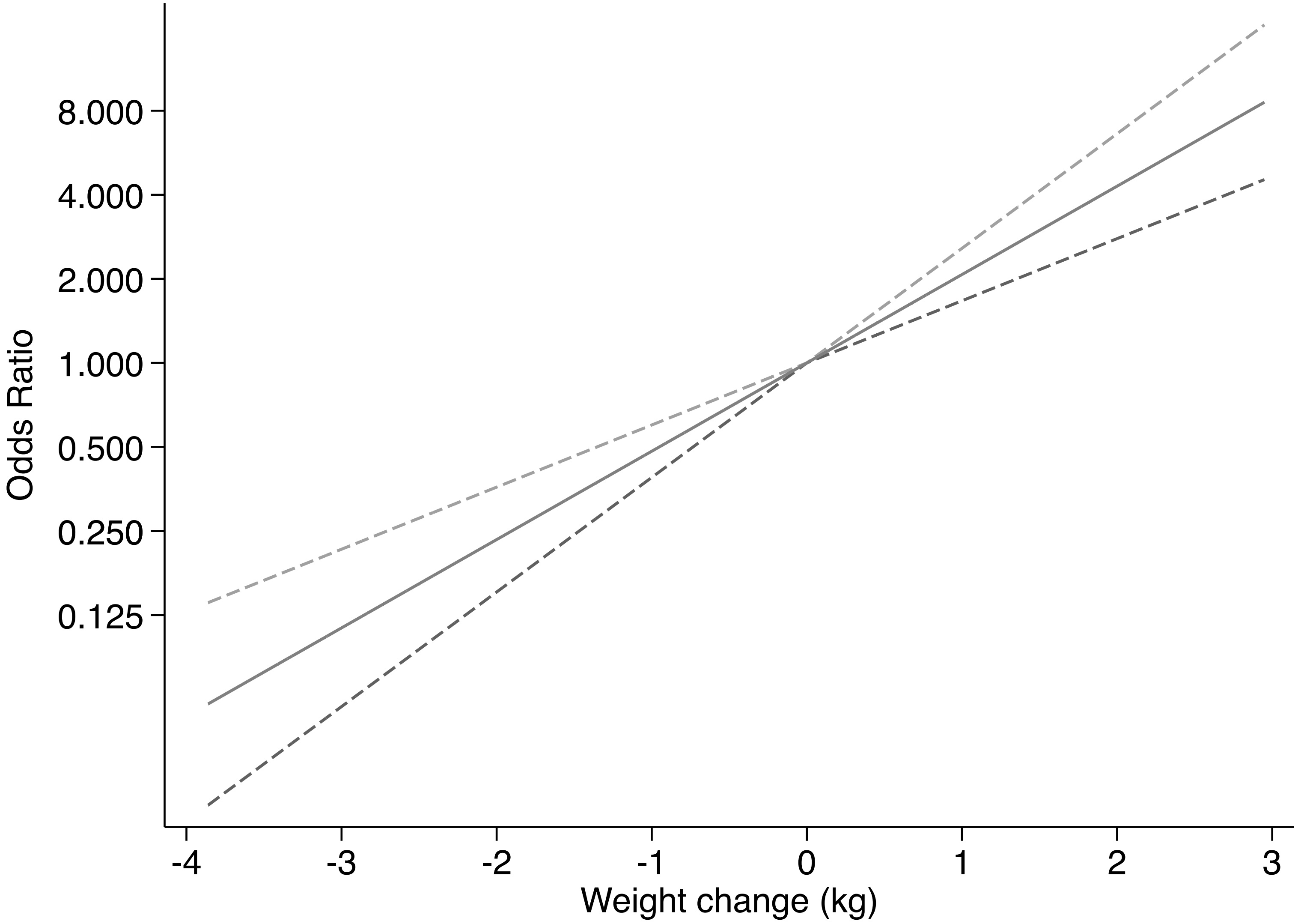
`_b[wtdiff]` is the way to refer to the coefficient of **`wtdiff`**

Once you have point estimate and confidence limits you can plot this information in many ways.

If you want to plot the odds ratios instead of the log odds ratios you can exponentiate the three variables of interest and re-run the graph.

The odds ratios are usually plotted on the **log scale**. In some journals, like *American Journal of Epidemiology* is compulsory. This is actually what the authors of the *NEJM* paper did in Figure 2.

```
gen orl = exp(logorl)
gen lbl = exp(lol)
gen ubl = exp(hil)

tw (line lbl ubl orl wtdiff, sort lp(- - l) ) ///
if inrange(wtdiff, -4,3) , yscale(log) ///
xlabel(-4(1)3) scheme(s1mono) legend(off) ///
ylabel(.125 0.25 0.25 0.5 1 2 4 8 ///
, angle(horiz) format(%4.3fc))  ///
ytitle("Odds Ratio") ///
xtitle("Weight change (kg)") ///
plotregion(style(none))
```


Odds Ratio
8.000
4.000
2.000
1.000
0.500
0.250
0.125
-4
-3
-2
-1
0
1
2
3
Weight change (kg)

Dichotomized predictor

Dichotomization of a continuous covariate like weight change would provide just one odds ratio comparing two groups.

```
. codebook gainweight

                  type:  numeric (float)
                 label:  gw

                 range:  [0,1]                   units:  1
         unique values:  2                   missing .:  33/488

            tabulation:  Freq.   Numeric  Label
                           320         0  Post<=Pre
                           135         1  Post>Pre
                            33
```

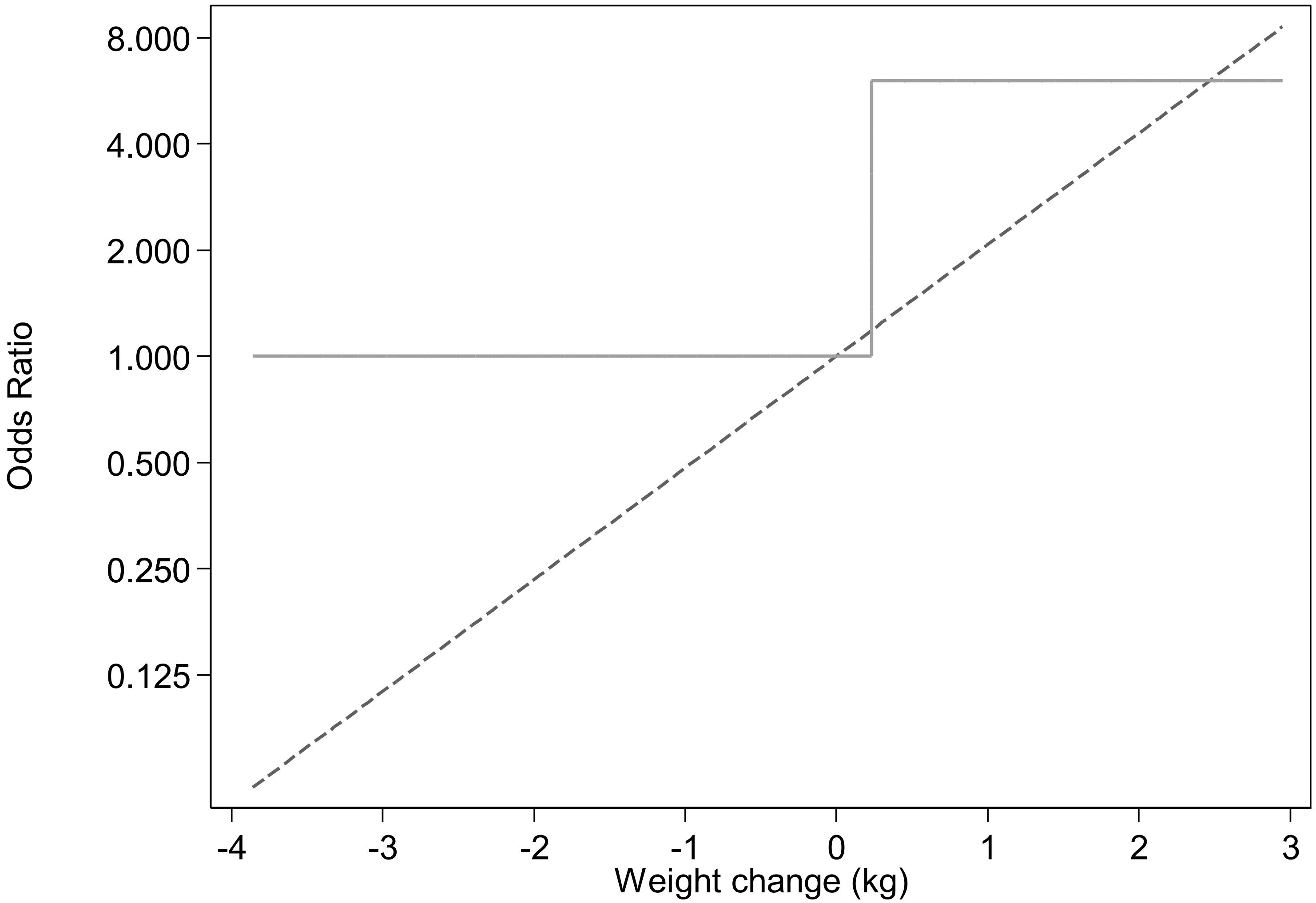
$$\log(\text{odds}|\text{gainweight}) = \beta_0 + \beta_1 \text{gainweight}$$

```
.  logistic nas135 gainweight

Logistic regression                                Number of obs   =        465
                                                   LR chi2(1)      =      38.08
                                                   Prob > chi2     =     0.0000
Log likelihood = -157.85449                        Pseudo R2       =     0.1076

------------------------------------------------------------------------------
      nas135 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gainweight |     6.0415   1.884663     5.77   0.000     3.278009    11.13472
------------------------------------------------------------------------------
```

Compared to those runners who lost weight or did not change weight, weight gain was associated with a 6 fold increase odds of hyponatremia.

Let's compare the linear-response and categorical models used for modeling weight change.

```
predictnl logor_hat2 = _b[gainweight]*gainweight

gen or2 = exp(logor_hat2)

tw (line or or2 wtdiff, sort c(l J) lp(- l) ) ///
if inrange(wtdiff, -4,3) , ///
xlabel(-4(1)3) scheme(s1mono) legend(off) ///
ylabel(.125 0.25 0.25 0.5 1 2 4 8 ///
, angle(horiz) format(%4.3fc))  ///
ytitle("Odds Ratio") ///
xtitle("Weight change (kg)") yscale(log)// Dichotomization
reg na gainweight
predict fit2
```


Odds Ratio
8.000
4.000
2.000
1.000
0.500
0.250
0.125
-4
-3
-2
-1
0
1
2
3
Weight change (kg)

Categorical predictor

A popular strategy among epidemiologists is to categorize the continuous covariate in 3 to 5 categories.

It is commonly used to present the data and findings in a tabular form and to avoid the assumption of linearity

Let's consider a categorized version of weight change (wtdiffc) as predictor of the hyponatremia risk.

A table of data (cases/non-cases) with odds or odds ratios can be obtained using the tabodds command.

```
.   tabodds nas135 wtdiffc
```

```
--------------------------------------------------------------------------
   wtdiffc  |     cases     controls       odds      [95% Conf. Interval]
------------+-------------------------------------------------------------
  3.0 to ~9 |         5            2    2.50000       0.48504  12.88565
  2.0 to ~9 |         7            9    0.77778       0.28966   2.08843
  1.0 to ~9 |        11           28    0.39286       0.19559   0.78909
  0.0 to ~9 |        18           78    0.23077       0.13823   0.38526
  -1.0 to~1 |         9          100    0.09000       0.04550   0.17802
  -2.0 to~1 |         6           93    0.06452       0.02826   0.14730
  -5.0 to~1 |         1           83    0.01205       0.00168   0.08654
--------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(6)  =    63.23
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    54.06
                                  Pr>chi2  =   0.0000
```

There is a strong association between weight change and risk of hyponatremia.

We present odds ratios of hyponatremia using the largest category (runners who lost up to 1 kg) as reference group.

```
. tabodds nas135 wtdiffc, base(5) or

---------------------------------------------------------------------------
     wtdiffc |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
  3.0 to ~9 |   27.777778      24.52       0.0000      3.718322 207.514311
  2.0 to ~9 |    8.641975      15.62       0.0001      2.380986  31.366730
  1.0 to ~9 |    4.365079       9.71       0.0018      1.585271  12.019342
  0.0 to ~9 |    2.564103       4.89       0.0270      1.078861   6.094042
  -1.0 to~1 |    1.000000          .            .             .          .
  -2.0 to~1 |    0.716846       0.37       0.5418      0.244796   2.099165
  -5.0 to~1 |    0.133869       4.80       0.0285      0.016094   1.113538
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(6)  =    63.23
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    54.06
                                  Pr>chi2  =   0.0000
```

In the context of logistic regression and categorical variables one need to be familiar with the code used to classify subjects in order to correctly interpret the regression coefficients.

```
. codebook wtdiffc

  range:  [1,7]                                    units:  1
  unique values:  7                            missing .:  38/488

           tabulation:  Freq.   Numeric  Label
                            7         1  3.0 to 4.9
                           16         2  2.0 to 2.9
                           39         3  1.0 to 1.9
                           96         4  0.0 to 0.9
                          109         5  -1.0 to -0.1
                           99         6  -2.0 to -1.1
                           84         7  -5.0 to -2.1
                           38         .
```

Indicator variables

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

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

You can generate indicator variables in many ways. The prefix command, however, **xi** makes it easy to generate indicator variables as well as all interactions terms.

By default, Stata uses the lowest value of the categorical variable as referent.

Suppose one want to define weight change between -0.1 to -1 kg as reference group rather than the default lowest value.

```
. char wtdiffc[omit] 5
. xi: logistic nas135 i.wtdiffc

Logistic regression                               Number of obs   =        450
                                                  LR chi2(6)      =      54.39
                                                  Prob > chi2     =     0.0000
Log likelihood = -143.80497                       Pseudo R2       =     0.1590

------------------------------------------------------------------------------
      nas135 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Iwtdiffc_1 |   27.77778   25.17088     3.67   0.000     4.703051    164.0648
 _Iwtdiffc_2 |   8.641975   5.292667     3.52   0.000     2.601989    28.70255
 _Iwtdiffc_3 |   4.365079   2.172628     2.96   0.003       1.6456    11.57871
 _Iwtdiffc_4 |   2.564103   1.116157     2.16   0.031     1.092462    6.018171
 _Iwtdiffc_6 |   .7168459   .3916698    -0.61   0.542      .245667    2.091726
 _Iwtdiffc_7 |   .1338688   .1425033    -1.89   0.059     .0166179    1.078408
       _cons |        .09   .0313209    -6.92   0.000     .0455004    .1780202
------------------------------------------------------------------------------
```

The intercept, exp(_cons), is the odds of hyponatremia among those runners who lost up to 1 kg (reference group), 9 cases every 100 non-cases.

Compared to runners who lost up to 1 kg, the odds of hyponatremia among those who gained up to 1 kg increased by 2.56 times.

Compared to runners who lost up to 1 kg, the odds of hyponatremia among those who lost between 1 and 2 kg was (1-.716) 28% lower.

And so on so forth.

Summary

- In large samples, we have done inference on one population proportion and saw the correspondence with an empty logistic regression model.
- We compared hand calculations based on a 2 by 2 table and the results of a logistic regression model with a binary predictor. Extension to more than 2 levels.
- We interpreted and presented the linearity assumption when modelling quantitative predictors.