Review of Linear regression

Nicola Orsini

Motivating example

Dataset
marathon.dta

Reference
"Hyponatremia among Runners in the Boston Marathon",
New England Journal of Medicine, 2005, Volume 352:1550-1556.

Descriptive abstract
Hyponatremia has emerged as an important cause of race-related death and life-threatening illness among marathon runners. We studied a cohort of marathon runners to estimate the incidence of hyponatremia and to identify the principal risk factors.

Acknowledgement
Professor David Wypij, Harvard School of Public Health

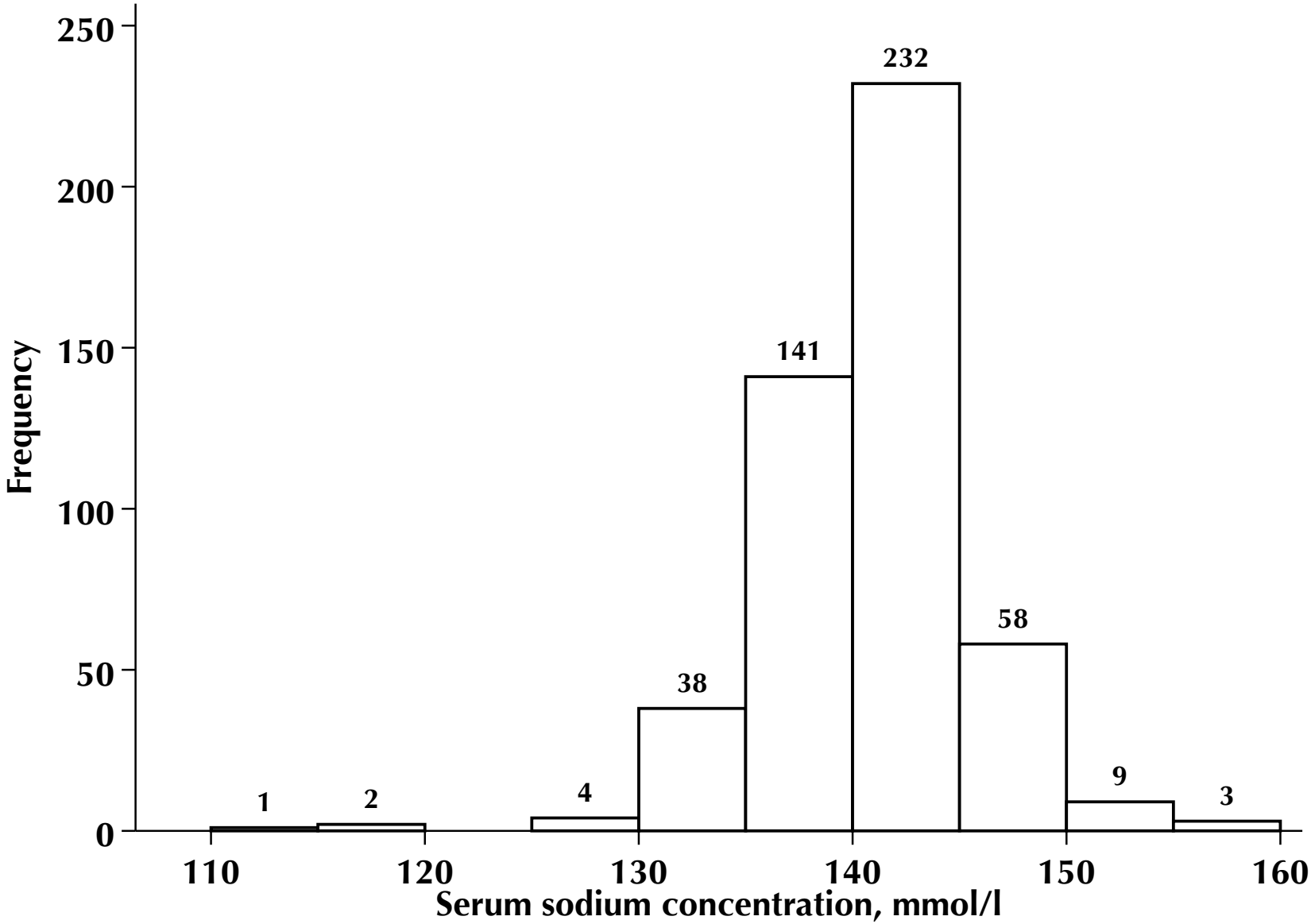
Univariable analysis

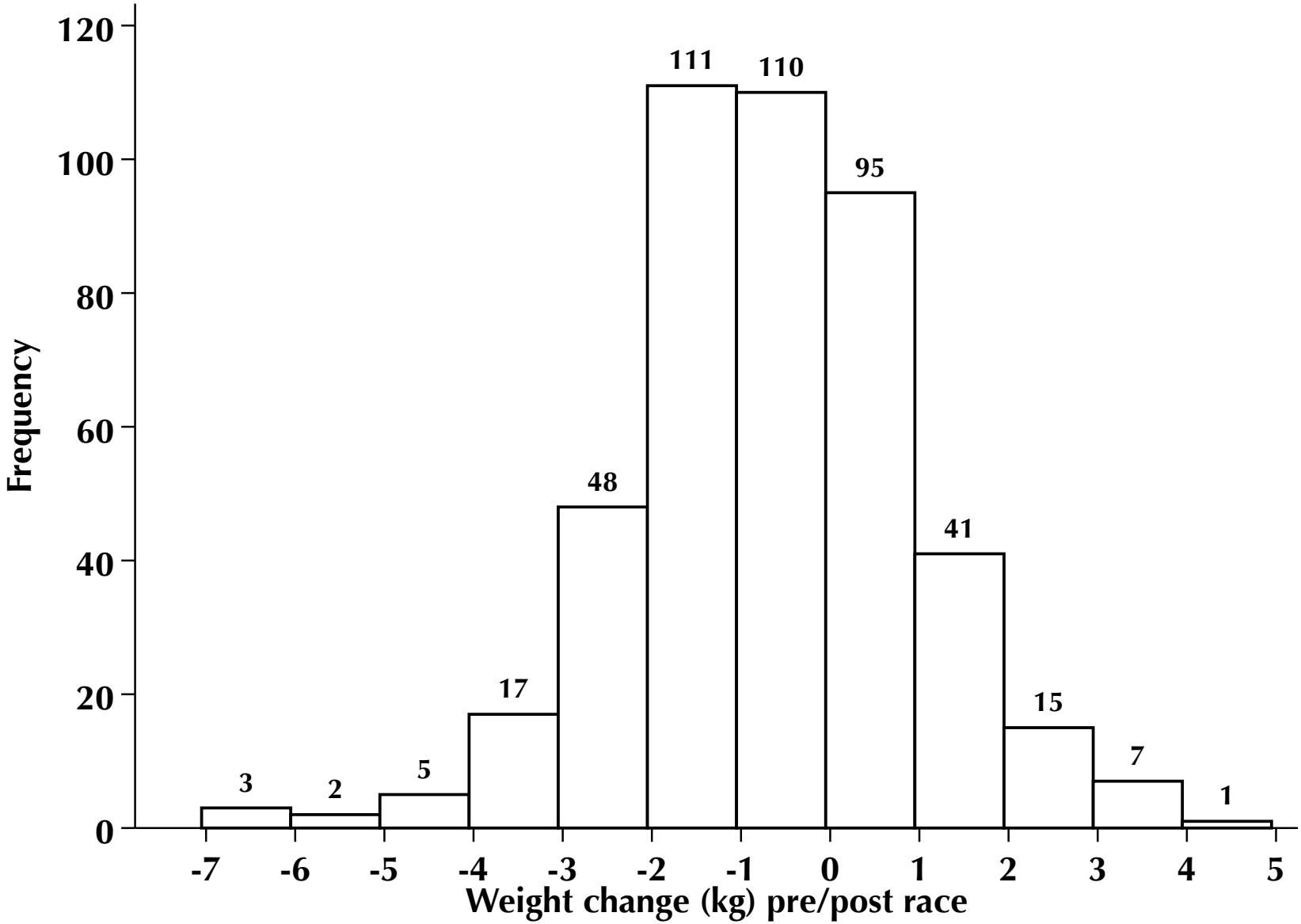
We want to investigate the relation between

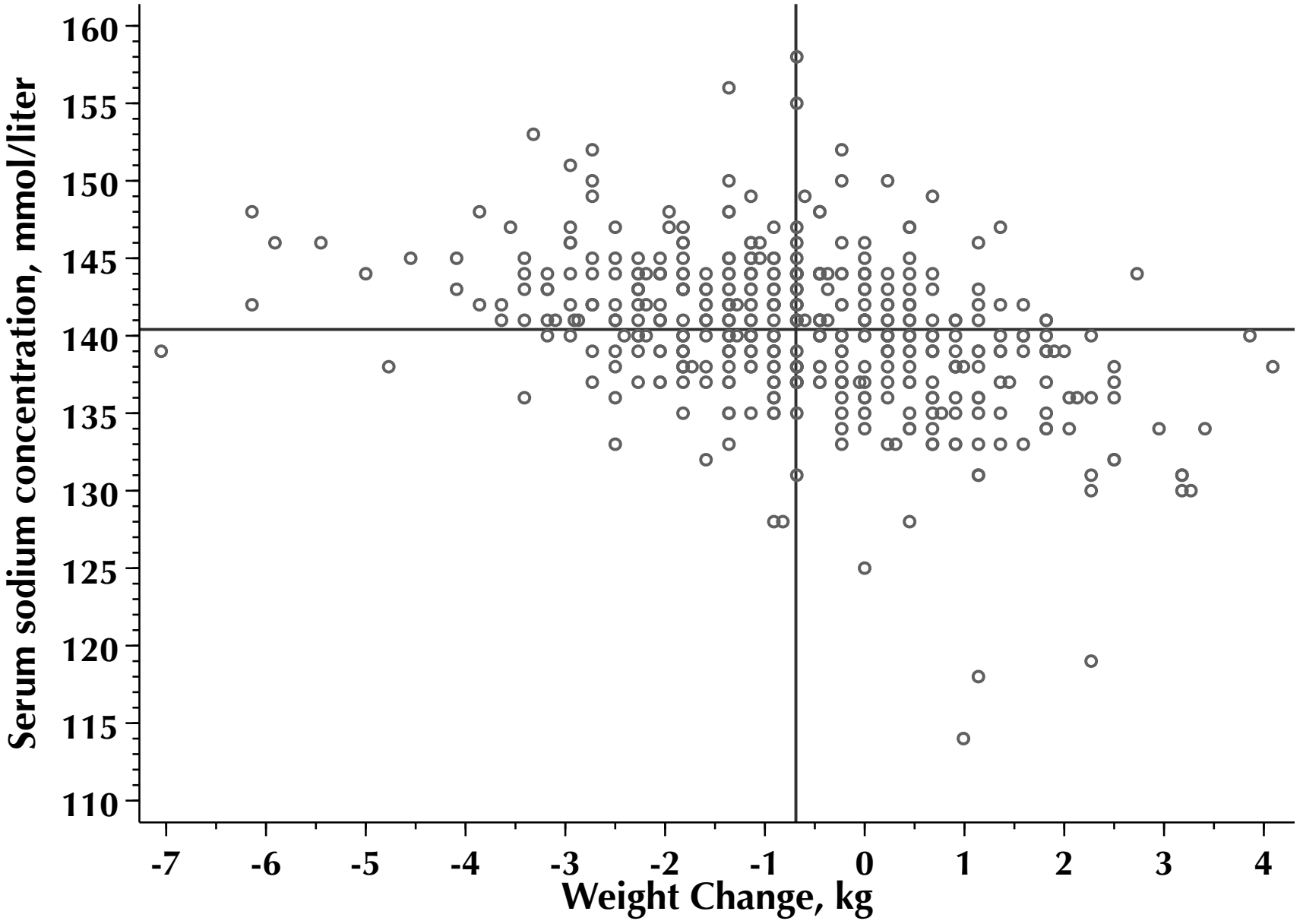
wtdiff = quantitative predictor (weight change, kg)

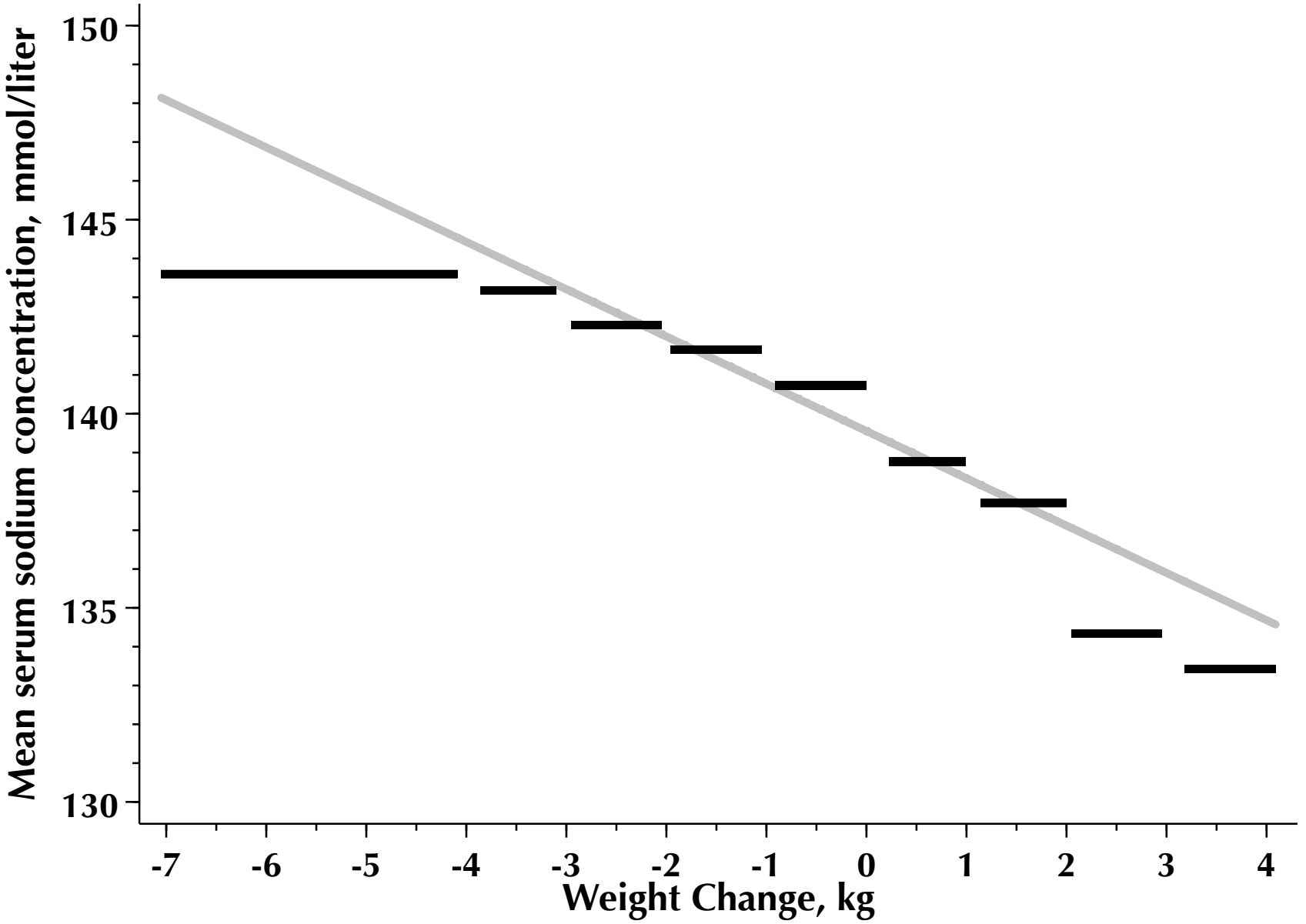
and

na = quantitative outcome (serum sodium concentration, mmol/liter)


250
200
150
100
50
0
Frequency
232
141
58
38
9
4
3
1
2
110
120
130
140
150
160
Serum sodium concentration, mmol/l


Frequency
120
100
80
60
40
20
0
3
2
5
17
48
111
110
95
41
15
7
1
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
Weight change (kg) pre/post race


Serum sodium concentration, mmol/liter
160
155
150
145
140
135
130
125
120
115
110
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
Weight Change, kg


Mean serum sodium concentration, mmol/liter
150
145
140
135
130
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
Weight Change, kg

Regression model for the mean

We assume a statistical model to make inference about the population mean outcome as linear function of (conditioning on) a quantitative covariate.

$$\text{Mean}(y|x) = \beta_0 + \beta_1 x$$

y represents individual values of independent outcomes

x represents individual values of a quantitative covariate

Basic assumptions of the model

A sample of *n* of independent observations

The response is equal to a "fixed" part that depends on the value of the predictor plus a random error

$$y = \beta_0 + \beta_1 x + \epsilon$$

The response, conditionally on the value of the predictor, is assumed to have a constant variance

$$\mathrm{Var}(y|x) = \sigma^2$$

The population mean outcome among individuals with a covariate x equal to 0 is given by

$$\text{Mean}(y|x = 0) = \beta_0$$

The difference in population mean outcome comparing individuals with a covariate value x_1 with individuals with a covariate value x_2 is given by

$$\text{Mean}(y|x = x_1) = \beta_0 + \beta_1 x_1$$

$$\text{Mean}(y|x = x_2) = \beta_0 + \beta_1 x_2$$

Given the specified model, one could explore variation in the population mean outcome.

The difference or contrast in population mean outcomes comparing individuals with a value of the covariate x_1 with individuals with a value of the covariate x_2 is given by

$$\text{Mean}(y|x = x_1) - \text{Mean}(y|x = x_2) = \beta_1(x_1 - x_2)$$

Every $(x_1 - x_2)$ unit increase in the predictor, is associated with a β_1 unit change in the mean response, regardless of where one begin the increase (x_2).

This is the linear-response assumption.

We specify a simple linear regression model for the mean sodium concentration with weight change as the only predictor.

$$\text{Mean}(\text{na}|\text{wtdiff}) = \beta_0 + \beta_1 \text{wtdiff}$$

Estimation procedures such as ordinary least-square or maximum likelihood provide estimates of unknown population parameters β_0 and β_1.

```
. regress na wtdiff

      Source |       SS       df       MS              Number of obs =     455
-------------+------------------------------           F(  1,   453) =   91.82
       Model |  1765.61492     1  1765.61492            Prob > F      =  0.0000
    Residual |  8710.96529   453   19.229504           R-squared     =  0.1685
-------------+------------------------------           Adj R-squared =  0.1667
       Total |  10476.5802   454  23.0761679           Root MSE      =  4.3851

------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      wtdiff |    -1.2181   .1271215    -9.58   0.000    -1.467921   -.9682791
       _cons |   139.5535   .2234862   624.44   0.000     139.1143    139.9927
------------------------------------------------------------------------------
```

The first variable name is the response followed by a list of covariates or predictors.

$$\text{Mean}(\text{na}|\text{wtdiff}) = 140 - 1.2\ \text{wtdiff}$$

The mean serum sodium concentration significantly decreases by 1.2 mmol per liter (95% CI = -1.5 to -1) for every 1 kg increase of weight change during race.

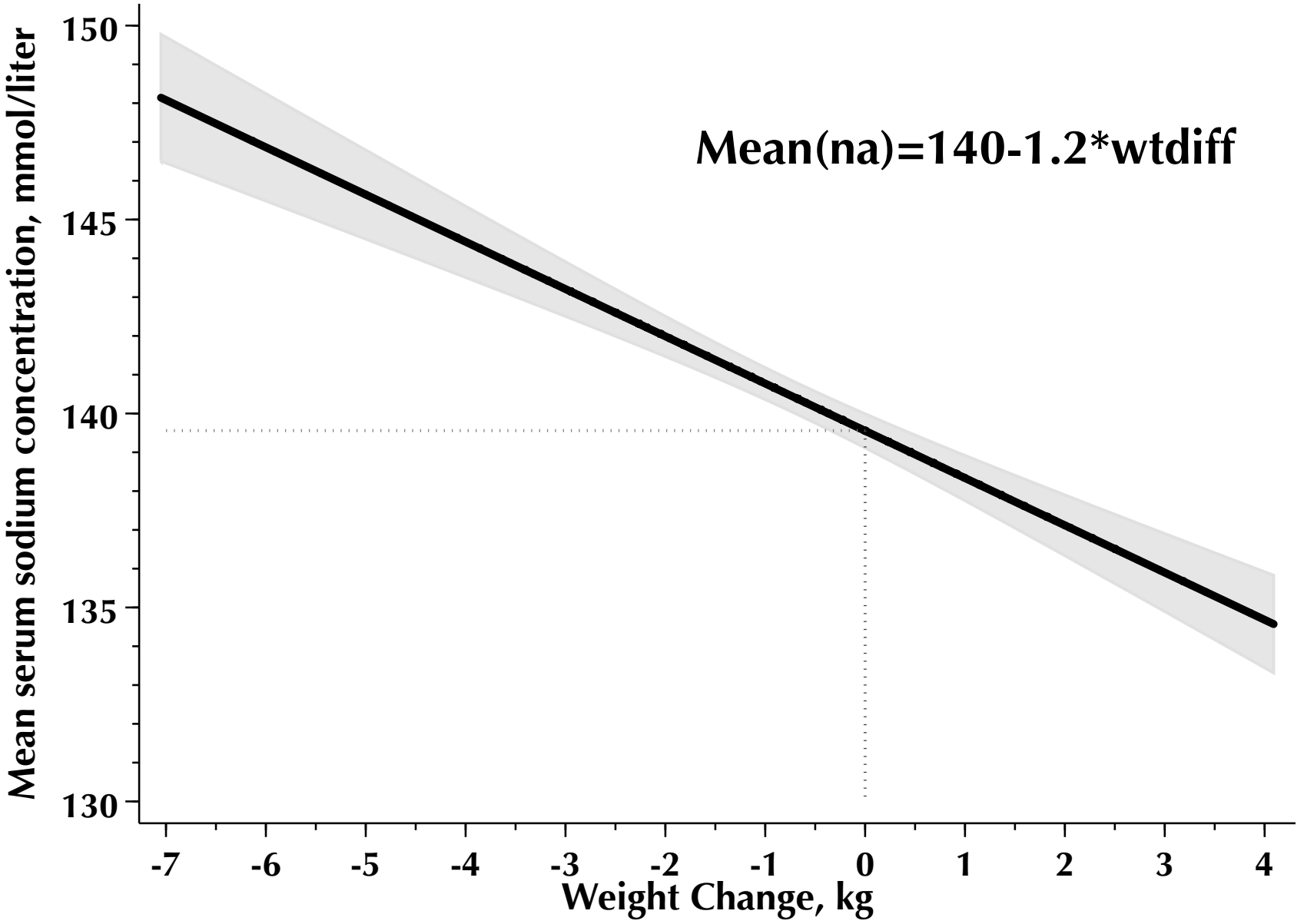
The intercept, _cons, is the estimated mean response when the predictor is set to zero. The population mean sodium concentration is 140 mmol/liter for those runners who did not change weight during the race.

1. What is the population mean serum sodium concentration among those runners who increased 3 kg during the marathon?

$$\text{Mean}(\text{na}|\text{wtdiff} = 3) = 140 - 1.2 \times 3$$

```
. lincom _b[cons] + _b[wtdiff]*3
```

na	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	135.8992	.5120958	265.38	0.000	134.8928	136.9056


Mean(na)=140-1.2*wtdiff
Mean serum sodium concentration, mmol/liter
150
145
140
135
130
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
Weight Change, kg

2. What is the change in the population mean serum sodium concentration associated with 2 kg increment?

$$\text{Mean}(\text{na}|\text{wtdiff} = x + 2) - \text{Mean}(\text{na}|\text{wtdiff} = x)$$
$$= -1.2 \times (x + 2 - x) = -1.2 \times 2$$

```
. lincom  _b[wtdiff]*2
```

na	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	-2.4362	.254243	-9.58	0.000	-2.935842	-1.936558

The mean serum sodium concentration decreases by 2.4 mmol per liter for every 2 kg increment in weight change.

2. What are the differences in the population mean serum sodium concentration comparing runners with any value of weight change ($x_1 = x$) relative to runners who did not change weight ($x_2 = 0$)?

x_1 represents any sub-population defined by x

x_2 represents the reference (or baseline) sub-population

$$\text{Mean}(\text{na}|\text{wtdiff} = x) - \text{Mean}(\text{na}|\text{wtdiff} = 0) =$$
$$= -1.2 \times (x - 0)$$

Tabulate mean differences

Weight change, kg				
x_1= -3	x_1= -1	x_2=0	x_1= 1	x_1= 2
$\beta_1(-3-0)$	$\beta_1(-1-0)$	Ref	$\beta_1(1-0)$	$\beta_1(2-0)$
3.7 (2.9 to 4.4)	1.2 (1.0 to 1.5)	0	-1.2 (-1.5 to -1.0)	-2.4 (-2.9 to -1.9)

In our example of weight change in predicting mean sodium concentration, we can estimate differences for any value x_1 relative to x_2 using the **lincom** postestimation command.

$\beta_1(-3-0)$

```
. lincom _b[wtdiff]*(-3-0), cformat(%2.1fc)
```

na	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	3.7	0.4	9.58	0.000	2.9	4.4

$\beta_1(2-0)$

```
. lincom _b[wtdiff]*(2-0), cformat(%2.1fc)
```

na	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	-2.4	0.3	-9.58	0.000	-2.9	-1.9

Plot mean differences

To present graphically the quantity

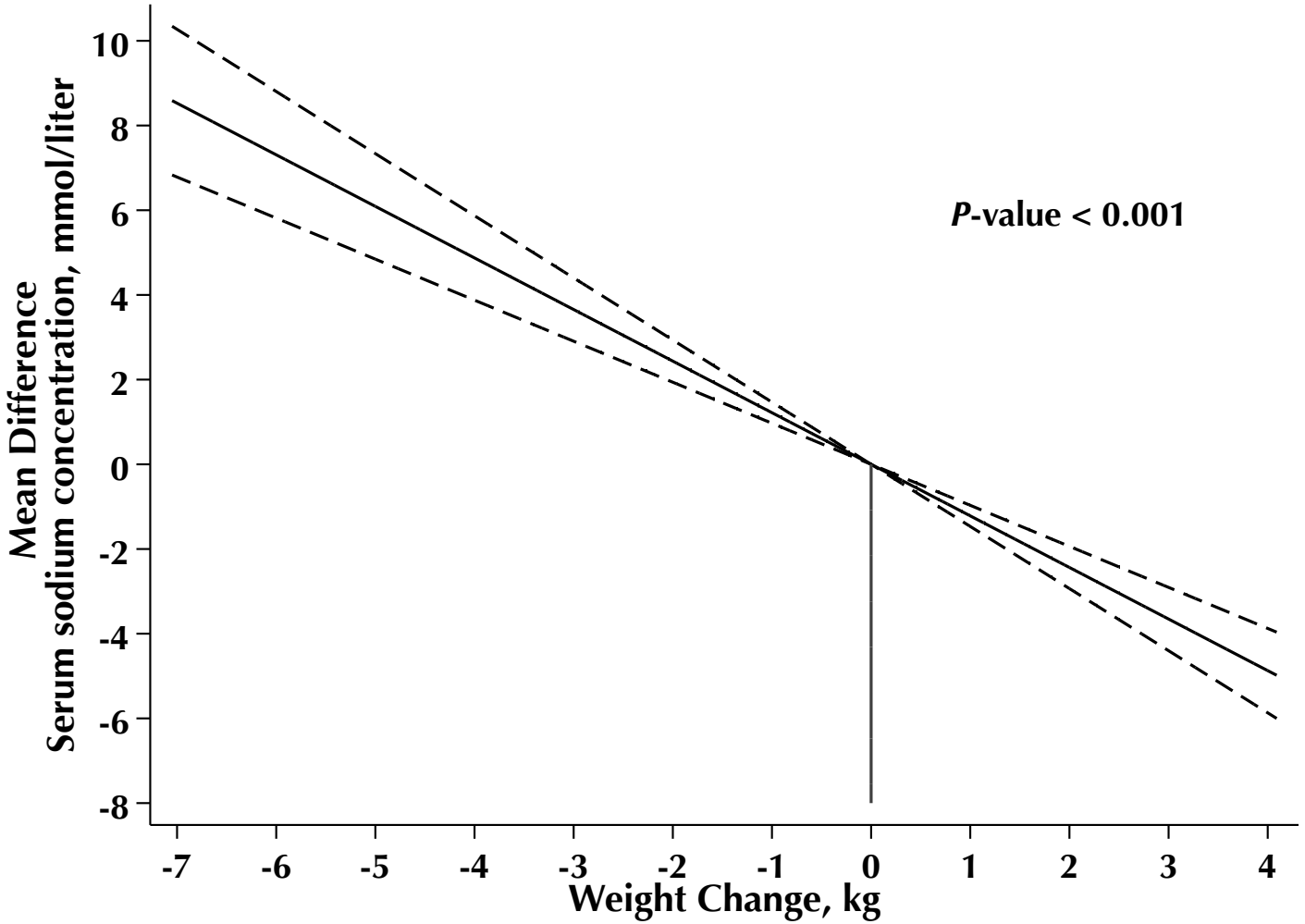
$$\beta_1(x_1 - x_2)$$

$$\beta_1(x - x_{ref})$$

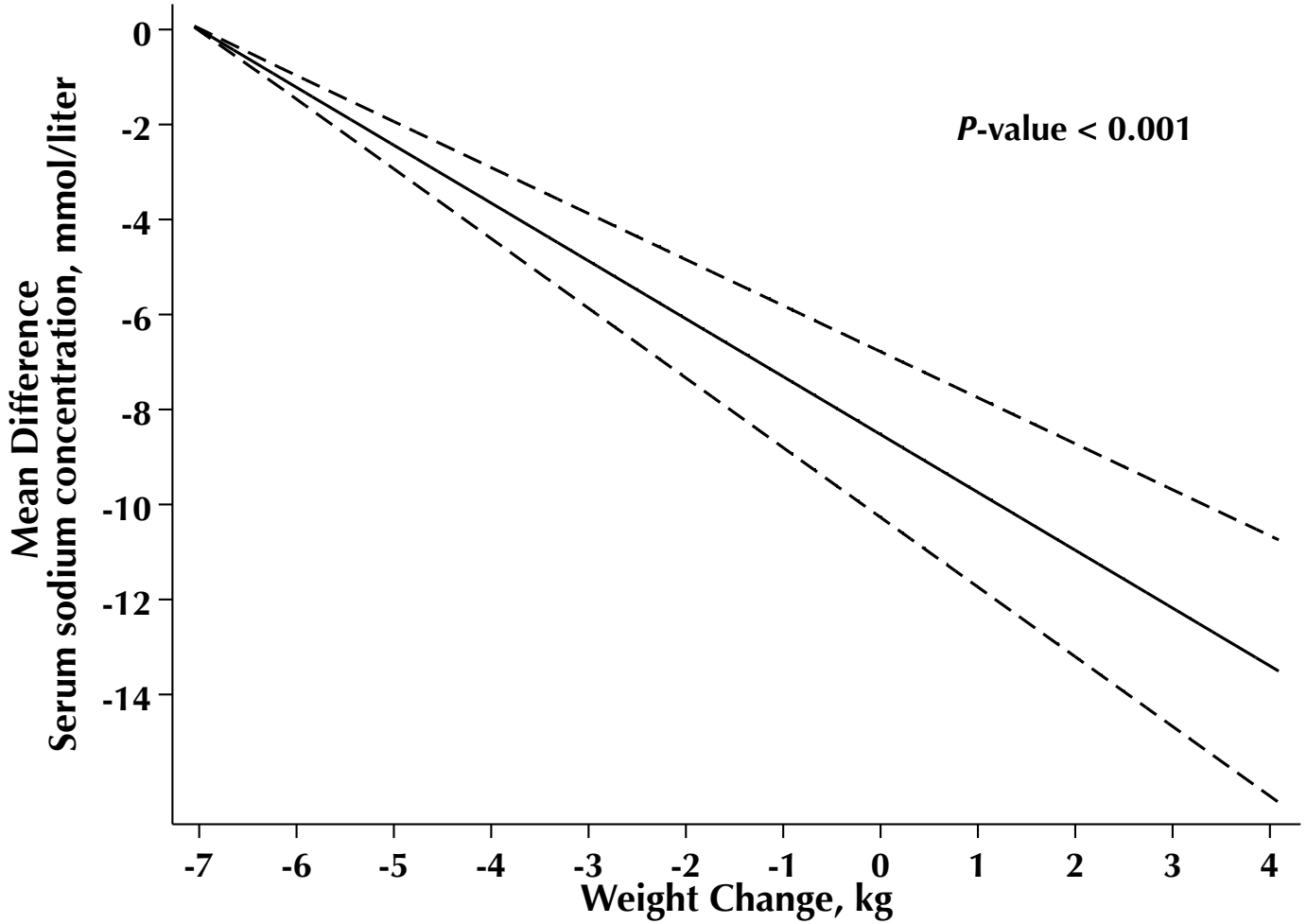
The post-estimation command **predictnl** is very useful to obtain the above quantity for any value of x with 95% confidence interval.

Any covariate value x_2 can be used as referent.

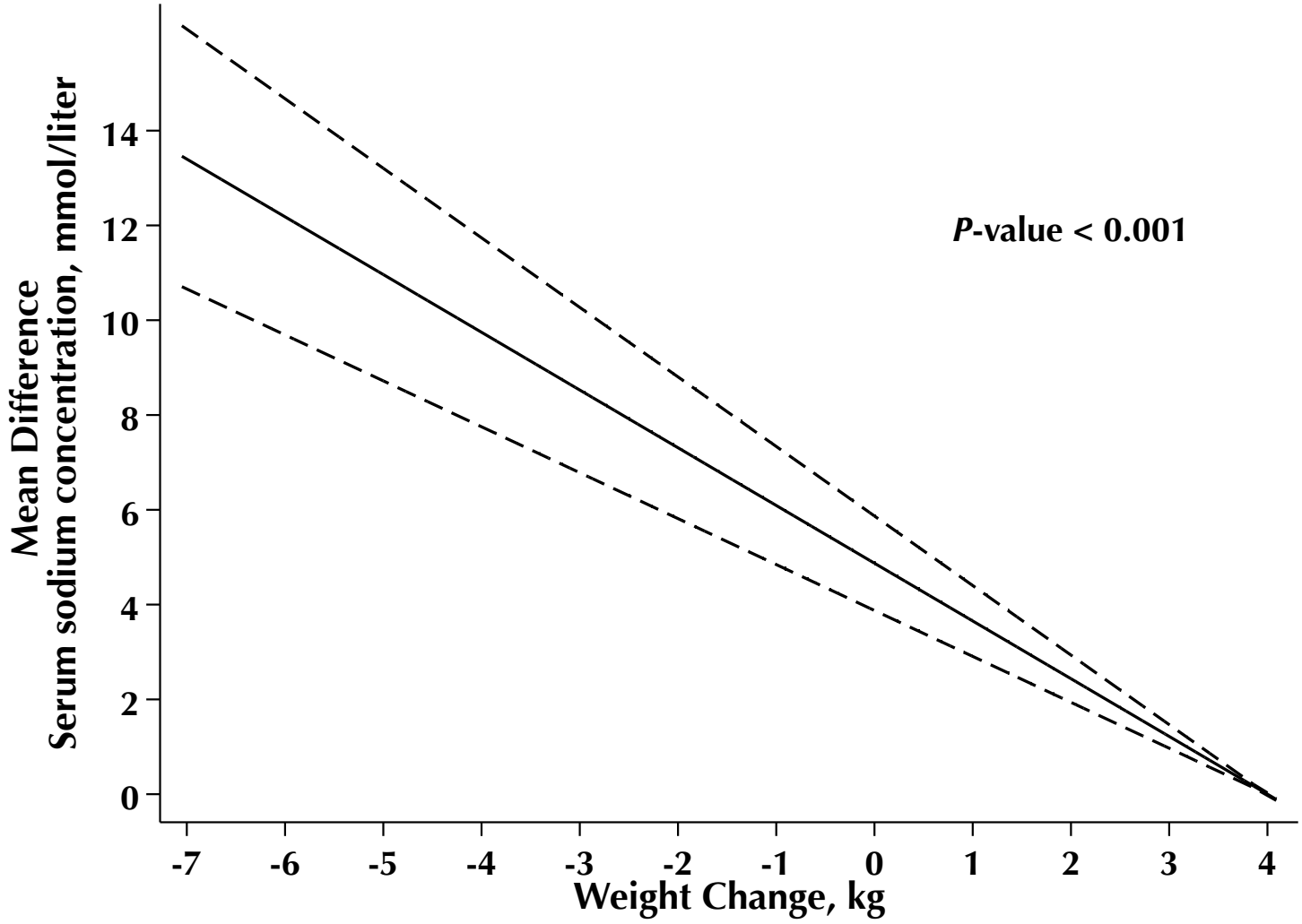
$$\mathrm{MD} = \beta_1(x - 0)$$



$$\mathrm{MD} = \beta_1(x - -7)$$



$$\mathrm{MD} = \beta_1(x - 4)$$


Mean Difference
Serum sodium concentration, mmol/liter
14
12
10
8
6
4
2
0
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
Weight Change, kg
P-value < 0.001

Confidence intervals for the mean outcome

$$\mathrm{Var}(\mathrm{Mean}(y|x)) = \mathrm{Var}(\eta) = \mathrm{Var}(\beta_0 + \beta_1 x)$$

$$\mathrm{Var}(\eta) = \mathrm{Var}(\beta_0) + \mathrm{Var}(\beta_1)x^2 + 2\mathrm{Cov}(\beta_0, \beta_1)x$$

$$\mathrm{SE}(\eta) = \sqrt{\mathrm{Var}(\eta)}$$

By the central limit theorem, we know that

$$\Pr\left[-1.96 < \frac{\eta}{\mathrm{SE}(\eta)} < 1.96\right] \simeq 95\%$$

Rearranging the terms,

$$\Pr[\eta - 1.96\,\mathrm{SE}(\eta) < \eta < \eta + 1.96\mathrm{SE}(\eta)] \simeq 95\%$$

Note: Before the sample is selected we can say there is 95% probability that η is included; after the sample is selected we can only say that there is 95% *confidence* that η is included.

A 95% confidence interval using the Standard normal distribution is computed using the constant of 1.96.

Using probability functions

```
display invnormal(.025)
-1.959964

display invnormal(.975)
1.959964

display normal(1.959964)-normal(-1.959964)
.95
```

```
. mat list e(V)

symmetric e(V)[2,2]
             wtdiff      _cons
wtdiff   .01615988
 _cons   .01114286   .04994607
```

$$\mathrm{Var}(\beta_0) = .04994607$$
$$\mathrm{Var}(\beta_1) = .01615988$$
$$\mathrm{Cov}(\beta_0, \beta_1) = .0111428$$

$$\mathrm{Var}(\eta) = .04994607 + .01615988 \times x^2 + 2{\times}.0111428{\times}x$$

$$\mathrm{Var}(\mathrm{Mean}(\mathrm{na}|\mathrm{wtdiff} = 0)) = .04994607$$

$$\mathrm{SE}(\mathrm{Mean}(\mathrm{na}|\mathrm{wtdiff} = 0)) = \sqrt{.04994607} = .22348618$$

95% CI for the mean serum sodium concentration among those who did not change weight is given by

$$\text{Mean}(\text{na}|\text{wtdiff} = 0) = 140$$

Lower Limit = 140 - 1.96*.22348618 = 139 mmol/liter

Upper Limit = 140 + 1.96* .22348618 = 140 mmol/liter

```
. lincom _b[_cons]

------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   139.5535   .2234862   624.44   0.000     139.1143    139.9927
------------------------------------------------------------------------------
```

95% CI for the mean serum sodium concentration among those who increased 4 kg is given by

$$\text{Mean}(\text{na}|\text{wtdiff} = 4) = 140 - 1.2 * 4 = 135$$

$$\text{SE}(\eta) = \sqrt{.04994607 + .01615988 \times 4^2 + 2\times.0111428\times4} = .6305925$$

Lower Limit = 135 - 1.96*.6305925 = 133 mmol/liter
Upper Limit = 135 + 1.96* .6305925 = 136 mmol/liter

```
. lincom _b[_cons] + _b[wtdiff]*4
------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   134.6811   .6305925   213.58   0.000     133.4418    135.9203
------------------------------------------------------------------------------
```

Confidence intervals for the difference in mean outcomes

$$\text{MD} = \text{Mean}(y|x = x_1) - \text{Mean}(y|x = x_2) = \beta_1(x_1 - x_2)$$

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

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

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

$$95\%\ \text{CI} = \text{MD} \pm 1.96\ \text{SE}(\text{MD})$$

What is the 95% CI for the mean difference in sodium concentration comparing those who lost 3 kg ($x_1 = -3$) compared to those runners who did not change weight ($x_2 = 0$)?

$$\text{MD} = -1.2181 \times (-3 - 0) = 3.65$$

$$\text{Var}(\beta_1) = .01615988$$

$$\text{SE(MD)} = \sqrt{.01615988(3-0)^2} = .38136451$$

$$95\%\ \text{CI} = 3.65 \times (3 - 0) \pm 1.96\ .38136451$$

$$95\%\ \text{CI} = 2.9 \text{ to } 4.4$$

Notes on Confidence Intervals

The width of the 95% confidence interval for the mean outcome is smaller at the mean value of the quantitative predictor.

The width of the 95% CI for the mean outcome is increasing moving away from the mean value of the predictor.

The width of the 95% CI for the difference in mean outcome is zero when the two values of the quantitative predictor being compared are the same ($x_1 = x_2$). MD = 0 and SE(MD)=0.

The width of the 95% CI for the difference in mean outcome is zero is increasing with the distance between the two values of the predictor being compared ($x_1 - x_2$).

Dichotomous predictor

Consider now a binary or dichotomous predictor. For example, an indicator variable of whether a runner increased or lost weight during the marathon.

```
. 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
```

A linear regression with a single binary (0/1) predictor provides a comparison of the mean response across the two subpopulations defined by the predictor.

This is equivalent to a comparison of two means for independent populations (help ttest).

Let's assume a piecewise constant association between weight change and mean serum sodium concentration with a knot at zero.

```
. regress na gainweight

      Source |       SS       df       MS              Number of obs =     455
-------------+------------------------------           F(  1,   453) =   70.55
       Model |  1411.72652     1  1411.72652           Prob > F      =  0.0000
    Residual |   9064.8537   453  20.0107146           R-squared     =  0.1348
-------------+------------------------------           Adj R-squared =  0.1328
       Total |  10476.5802   454  23.0761679           Root MSE      =  4.4733

------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gainweight |  -3.856019   .4590871    -8.40   0.000    -4.758223   -2.953814
       _cons |   141.5375    .250067   566.00   0.000     141.0461    142.0289
------------------------------------------------------------------------------
```

$$\text{Mean}(\text{na}|\text{gainweight}) = 142 - 4\,\text{gainweight}$$

The intercept (_cons), 142 mmol/liter is the mean sodium concentration at the referent value of gainweight, that is, individuals who lost or did not change weight (**Post<=Pre**).

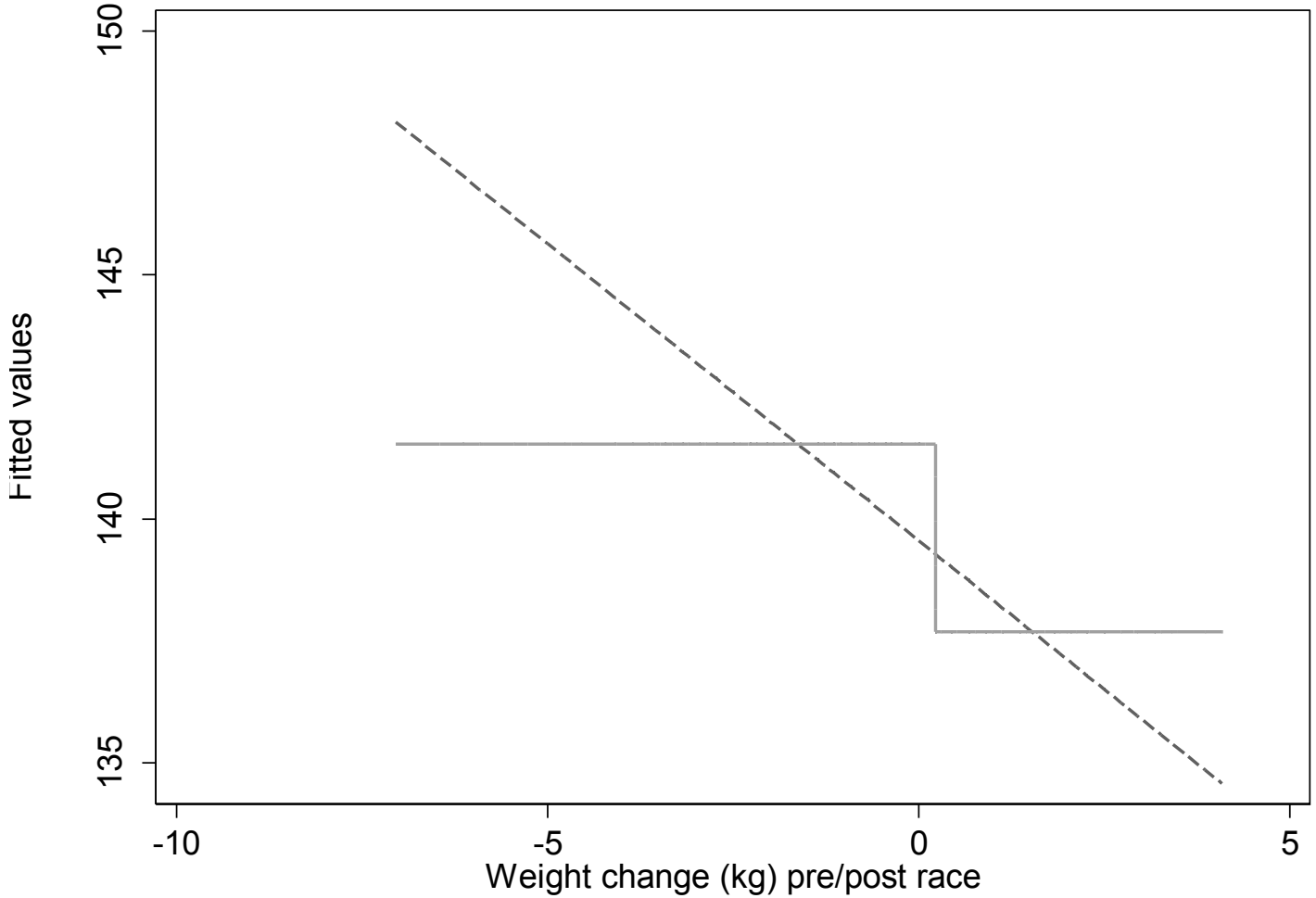
The mean sodium concentration among those who increased weight was 4 mmol/liter significantly lower (95% CI = -5, -3) compared to those who lost or did not change weight.

Both the linearity and the dichotomization of a continuous covariate make strong assumptions about the dose-response relationship. Let's compare the two approaches.

```
// Linear trend
reg na wtdiff
predict fit1

// Dichotomization
reg na gainweight
predict fit2
```

```
tw (line fit1 fit2 wtdiff, sort c(l J) lp(- l) ) , ///
scheme(s1mono) legend(off)
```



More than 2 categories

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

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

Let's consider a categorized version of weight change as predictor of serum concentration.

```
. table wtdiffc , c(freq mean na sd na) f(%3.0f)
-------------------------------------------------
Categorization |
of weight      |
change         |      Freq.    mean(na)       sd(na)
---------------+---------------------------------
   3.0 to 4.9  |           7         133           4
   2.0 to 2.9  |          16         135           6
   1.0 to 1.9  |          39         138           5
   0.0 to 0.9  |          96         139           5
 -1.0 to -0.1  |         109         141           5
 -2.0 to -1.1  |          99         142           4
 -5.0 to -2.1  |          84         143           4
-------------------------------------------------
```

To correctly interpret the regression coefficients of indicator variables we need to know how the variable is coded (meaning of the numbers).

```
. 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         .
```

Categorical variables – prefix xi

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 referent group.

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 reference.

$$\text{Mean(na)} = \beta_0 + \beta_1\_\text{Iwtdiffc\_2} + \ldots + \beta_7\_\text{Iwtdiffc\_7}$$

```
. xi: regress na i.wtdiffc

i.wtdiffc         _Iwtdiffc_1-7       (naturally coded; _Iwtdiffc_1 omitted)

      Source |       SS       df       MS              Number of obs =     450
-------------+------------------------------           F(  6,   443) =   16.48
       Model |  1888.39075     6  314.731791            Prob > F      =  0.0000
    Residual |   8462.1337   443  19.1018819            R-squared     =  0.1824
-------------+------------------------------           Adj R-squared =  0.1714
       Total |  10350.5244   449   23.052393            Root MSE      =  4.3706

------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Iwtdiffc_2 |    1.196429   1.980583     0.60   0.546    -2.696077    5.088934
 _Iwtdiffc_3 |    4.238095   1.794055     2.36   0.019     .7121797    7.764011
 _Iwtdiffc_4 |    5.644345   1.711087     3.30   0.001     2.281489    9.007201
 _Iwtdiffc_5 |    7.442988   1.704138     4.37   0.000     4.093789    10.79219
 _Iwtdiffc_6 |    8.227994   1.709324     4.81   0.000     4.868603    11.58739
 _Iwtdiffc_7 |    9.083333   1.719373     5.28   0.000     5.704192    12.46247
       _cons |    133.4286    1.65192    80.77   0.000      130.182    136.6751
------------------------------------------------------------------------------
```

The intercept (_cons) is the mean sodium concentration at the referent value of all predictors, that is, individuals who gained 3 to 4.9 kg during race.

The coefficient of **_lwtdiffc_2** is the difference in the mean sodium concentration comparing runners who gained 2 to 2.9 kg vs the referent.

The coefficient of **_lwtdiffc_7** is the difference in the mean sodium concentration comparing runners who lost 2.1 to 5 kg vs the referent.

Suppose you want to define weight change between 0 to 0.9 kg as your referent group rather than the default lowest value.

```
. char wtdiffc[omit] 4

. xi: regress na i.wtdiffc
i.wtdiffc         _Iwtdiffc_1-7       (naturally coded; _Iwtdiffc_4 omitted)

      Source |       SS       df       MS              Number of obs =     450
-------------+------------------------------           F(  6,   443) =   16.48
       Model |  1888.39075     6  314.731791            Prob > F      =  0.0000
    Residual |   8462.1337   443  19.1018819           R-squared     =  0.1824
-------------+------------------------------           Adj R-squared =  0.1714
       Total |  10350.5244   449   23.052393           Root MSE      =  4.3706

------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Iwtdiffc_1 |  -5.644345   1.711087    -3.30   0.001    -9.007201   -2.281489
 _Iwtdiffc_2 |  -4.447917   1.180189    -3.77   0.000    -6.767381   -2.128452
 _Iwtdiffc_3 |   -1.40625   .8299216    -1.69   0.091    -3.037323    .2248226
 _Iwtdiffc_5 |   1.798643    .611739     2.94   0.003     .5963719    3.000914
 _Iwtdiffc_6 |   2.583649   .6260401     4.13   0.000     1.353271    3.814027
 _Iwtdiffc_7 |   3.438988   .6529788     5.27   0.000     2.155667    4.722309
       _cons |   139.0729   .4460694   311.77   0.000     138.1962    139.9496
------------------------------------------------------------------------------
```

The intercept (_cons) is the mean sodium concentration at the referent value of all predictors, that is, individuals who gained 0 to 0.9 kg during race.

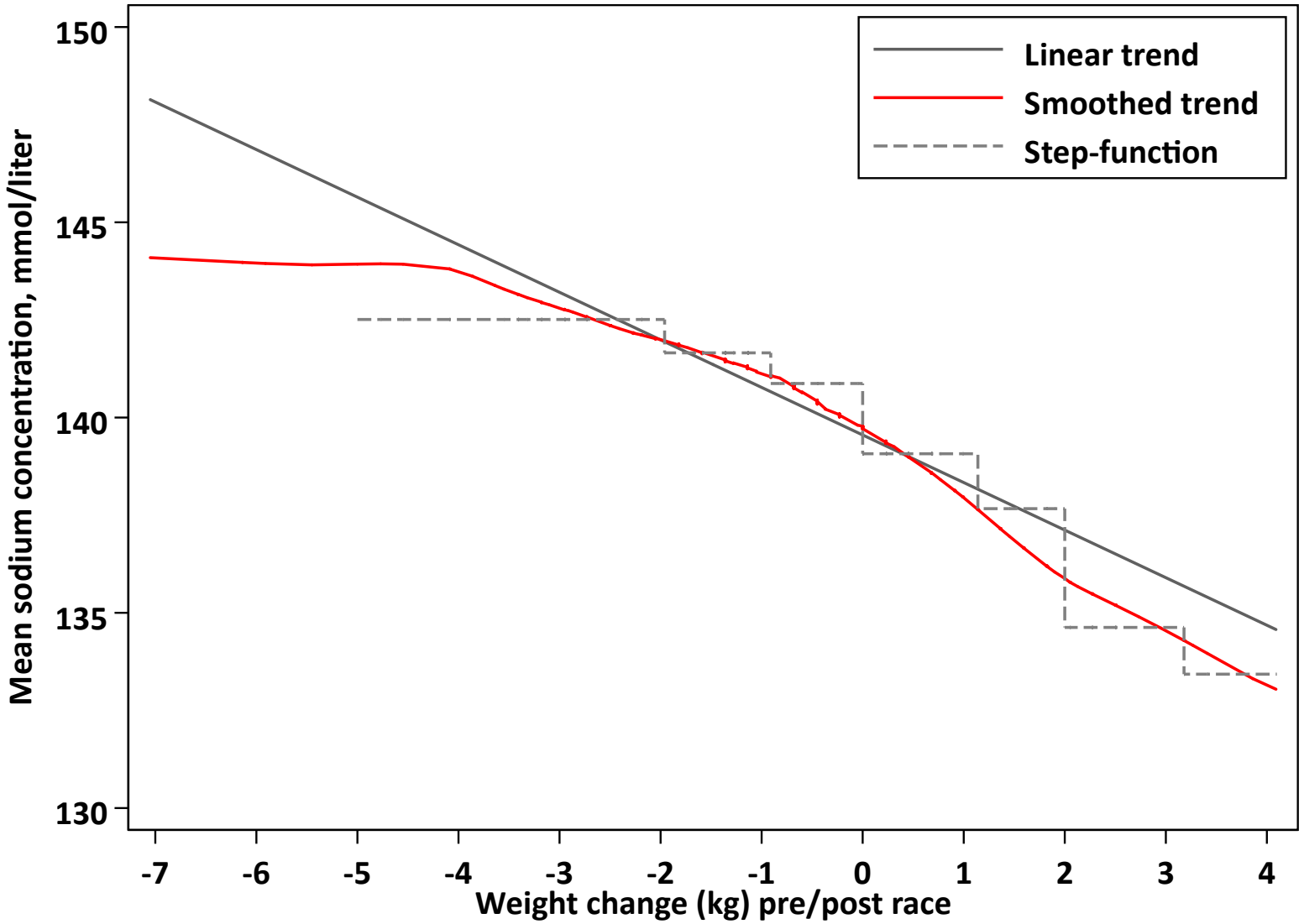
The coefficient of **_Iwtdiffc_2** is the difference in the mean sodium concentration comparing runners who gained 3 to 4.9 kg vs the referent.

And so on so forth.

The coefficient of **_Iwtdiffc_7** is the difference in the mean sodium concentration comparing runners who lost 2.1 to 5 kg vs the referent.

Comparing different approaches

```
tw   (lfit na wtdiff) ///
     (lowess na wtdiff, lc(red)) ///
     (line nahat2 wtdiff, c(J) lp(-) sort ) ///
     , legend(ring(0) pos(1) col(1) ///
      label(1 "Linear trend") ///
      label(2 "Smoothed trend") ///
      label(3 "Step-function") ) ///
      ytitle("Mean sodium concentration, mmol/liter")  ///
      xlabel(-7(1)4) ylabel(130(5)150, angle(horiz))
```


Linear trend
Smoothed trend
Step-function
Mean sodium concentration, mmol/liter
150
145
140
135
130
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
Weight change (kg) pre/post race

Lowess Regression

lowess regression (Locally Weighted Scatter plot Smoothing):

Fit a line through a scatter plot without any model assumption

Each observation (x_i, y_i) is fitted to a separate linear regression line based on adjacent observations

Each point in this range is weighted as a function of the distance from x_i

It provides a graph to easily detect strong departure from linearity.

Non-linear associations

A linear model can be used to model exposure-response relations that are not linear.

In our example, the flexible smoothed line for weight change suggests a possible non-linear relationship. The rate of change of sodium concentration among those who lost weight is not as steep as for those who increased weight during the race.

A way to detect strong departure from linearity is to fit a model that allows for non-linearity that includes the linear model as a special case.

A simple example is to fit a **regression model** in which is entered the exposure variable as it is and the exposure squared (to the power of 2), known as quadratic model.

Adding a quadratic transformation

The quadratic model for a quantitative exposure *x* is

$$Mean(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

The linear response model is nested in (special case of) the quadratic model.

A *p*-value for linearity is obtained by testing the coefficient β_2 equal to zero.

If the p-value is small (saying < 0.05), there is a departure from linearity that needs care and attention. Otherwise, the simpler linear model fits adequately the data.

We first generate a new variable containing weight change to the power of 2 (wtdiff squared).

```
. gen wtdiffsq = wtdiff^2
```

Then we fit the quadratic regression model

```
. regress na wtdiff wtdiffsq
```

```
      Source |       SS       df       MS              Number of obs =     455
-------------+------------------------------           F(  2,   452) =   51.04
       Model |  1930.17099     2  965.085495            Prob > F      =  0.0000
    Residual |  8546.40923   452   18.907985            R-squared     =  0.1842
-------------+------------------------------           Adj R-squared =  0.1806
       Total |  10476.5802   454  23.0761679            Root MSE      =  4.3483

------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      wtdiff |  -1.445263   .1477126    -9.78   0.000    -1.735552   -1.154974
    wtdiffsq |  -.1355339   .0459424    -2.95   0.003     -.225821   -.0452467
       _cons |   139.8157   .2387765   585.55   0.000     139.3465     140.285
------------------------------------------------------------------------------
```

Question 1. Is weight change overall predicting the mean sodium concentration?

We test simultaneously the two coefficients equal to zero

```
. testparm wtdiff wtdiffsq

 ( 1)  wtdiff = 0
 ( 2)  wtdiffsq = 0

       F(  2,   452) =   51.04
            Prob > F =    0.0000
```

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

Question 2. Is a quadratic model for weight change predicting the mean sodium concentration better compared to a simpler linear model?

We test the coefficient of the squared exposure equal to zero.

The test and its p-value is already in the output of regress command (p=0.003).

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

Question 3. What is the difference in the mean sodium concentration comparing those who increased 2 kg as compared to those who did not change weight?

To put it more generally, the predicted mean responses for any two values of *x* of a quadratic model are

$$Mean(y|x = x_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

$$Mean(y|x = x_2) = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2$$

The quantity

$$Mean(y|x = x_1) - Mean(y|x = x_2) = \beta_1(x_1 - x_2) + \beta_2(x_1^2 - x_2^2)$$

is the contrast between two predicted responses associated with a x_1 – x_2 unit change of the exposure x.

Compare to the linear response model, to quantify the change in the mean response is now more complicated because we need to involve two regression coefficients and two variables.

In health-related fields, the value of the covariate $x=x_2$ is called a reference value, and it is used to compute and interpret a set of comparisons of subpopulations defined by different covariate values.

You can easily estimate the above quantity with the postestimation commands **lincom** or **predictnl**.

The postestimation command **xblc** carries out these computations.

Orsini N., Greenland S. A procedure to tabulate and plot results after flexible modeling of a quantitative covariate. Stata Journal. 2011. 11, Number 1, pp. 1–29.

Example, using the post-estimation **lincom** command.

```
. lincom _b[wtdiff]*(2-0) + _b[wtdiffsq]*(4-0)

 ( 1)  2*wtdiff + 4*wtdiffsq = 0

------------------------------------------------------------------------------
          na |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -3.432662   .4214855    -8.14   0.000    -4.260976   -2.604347
------------------------------------------------------------------------------
```

Compare to those runners who did no change weight, those runners who increased 2 kg had a 3.4 mm/liter significantly lower mm/liter mean sodium concentration.

One can tabulate differences in mean responses for a list of specific values of the exposure.

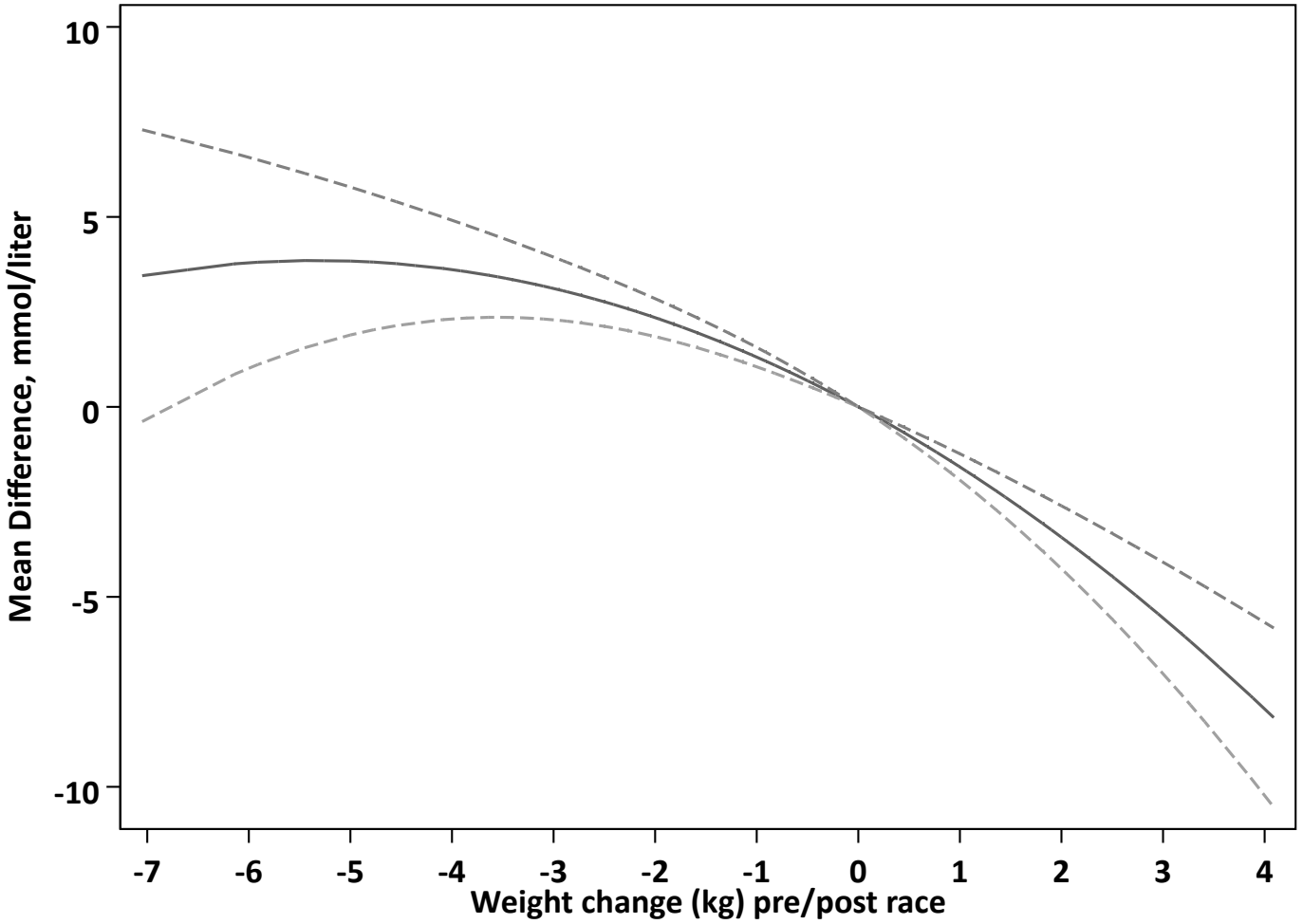
Question 4. How to plot the change in the mean response with 95% confidence intervals as function of the exposure using a specific exposure value as reference?

To create a plot we need to store the numbers we are interested in as variables. Once again, we can use the post-estimation command predictnl

```
predictnl diff = _b[wtdiff]*(wtdiff-0) +  ///
                 _b[wtdiffsq]*(wtdiffsq-0), ci(lb ub)
```

This gives us 3 new variables (diff, lb, and ub) in one line ready to be plotted with a standard twoway plot.

```
twoway (line diff lb ub wtdiff, sort lp(l - -)) , ///
   legend(off) scheme(s1mono) ytitle("Mean Difference")
```


10
5
0
-5
-10
Mean Difference, mmol/liter
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
Weight change (kg) pre/post race