Advanced Topics in Biostatistics 2020/21

Logistic regression

Tim Holland-Letz
t.holland-letz@dkfz-heidelberg.de
Div. Biostatistics – C060

October 21, 2020

DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT

Outline

- Motivation
- Simple logistic regression
- Multiple logistic regression
- Model-building strategies
- References

Motivation

Multiple linear regression:

- N independent explanatory variables $x_1, x_2, \ldots, x_N$
- a continuous outcome variable y, e.g. y ranging from $-\infty$ to $+\infty$

Model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_N x_N + \epsilon$$

with unknown coefficients $\beta_0, \beta_1, \ldots, \beta_N$ and an error term ϵ.

Example:

$$\text{weight} = \beta_0 + \beta_1 * \text{height} + \beta_2 * \text{age} + \epsilon,$$

with coefficients $\beta_0, \beta_1, \beta_2$ to be estimated.

Motivation

Now

- N independent explanatory variables $x_1, x_2, \ldots, x_N$
- a dichotomous (binary) outcome variable y , e.g. y is either 0 or 1.

Example: a study with 105 carcinoma patients where

- y is a radiodermatitis indicator, with $y = 1$ if a patient got radiodermatitis after radiotherapy and $y = 0$ otherwise
- x_1 is the form of the radiotherapy, with values in $\{A, B\}$
- x_2 is the degree of spread to regional lymph nodes, with values in $\{N0, N1, N2, N3\}$
- x_3 is the age of the patient; i.e. x_3 is a positive continuous variable

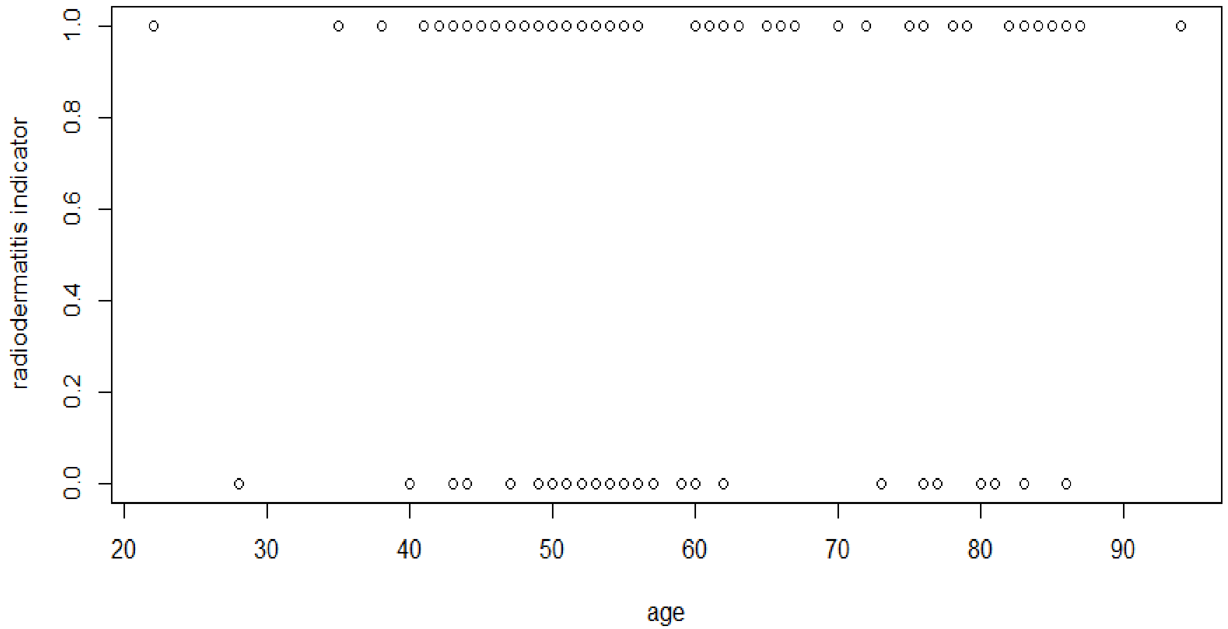
Motivation

Question: Can we use linear regression in the preceeding example?

If „yes“, then the following linear relation must hold

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

with some unknown coefficients β_0, β_1, β_2 and β_3. However, even graphically we see that



➡ we need to model a nonlinear relation between y and x_1, x_2, x_3

logistic regression

Simple logistic regression: definitions

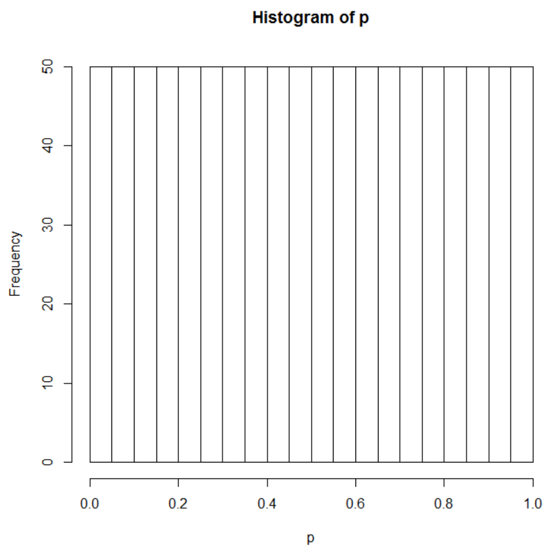
- Let y be (nonlinearly) related to $\beta_0 + \beta_1 x,$ where
  - y is a dichotomous outcome variable
  - x is an explanatory variable (also called *predictor*)
  - $\beta_0,\ \beta_1$ are some unknown coefficients
- Assumption: let $y \in \{0,1\}$
- Probability of an event occuring

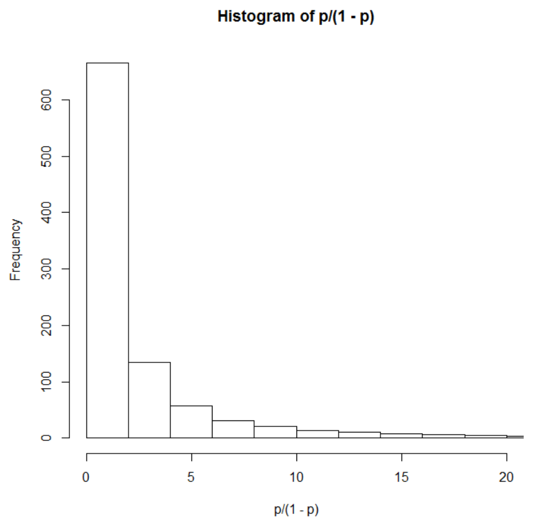
$$p(x) := P(y = 1|x)$$

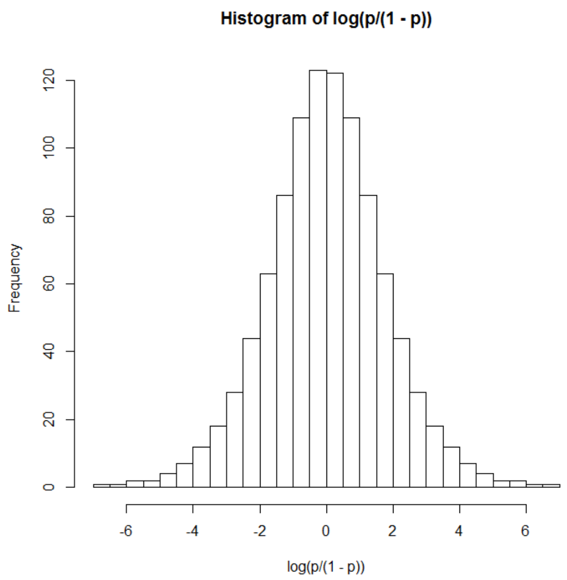
Thus $p(x)$ denotes the probability for y to be equal to 1 given a predictor value x.

Interpretation of $p(x)$ in the example above: $p(x)$ is the risk for the radiodermatitis to occur.

Motivation


Histogram of p
Frequency
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
p


Histogram of p/(1 - p)
Frequency
0
100
200
300
400
500
600
0
5
10
15
20
p/(1 - p)


Histogram of log(p/(1 - p))
Frequency
0
20
40
60
80
100
120
-6
-4
-2
0
2
4
6
log(p/(1 - p))

Simple logistic regression: definitions

- Logit transformation (logit function)

$$g(x) := \log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

- Note:

$g(x)$ may range from $-\infty$ to $+\infty$, depending on the range of x

linear regression can be applied

- Simple logistic regression model

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad (1)$$

where the right-hand side of (1) is called logistic function

Simple logistic regression: logistic curve

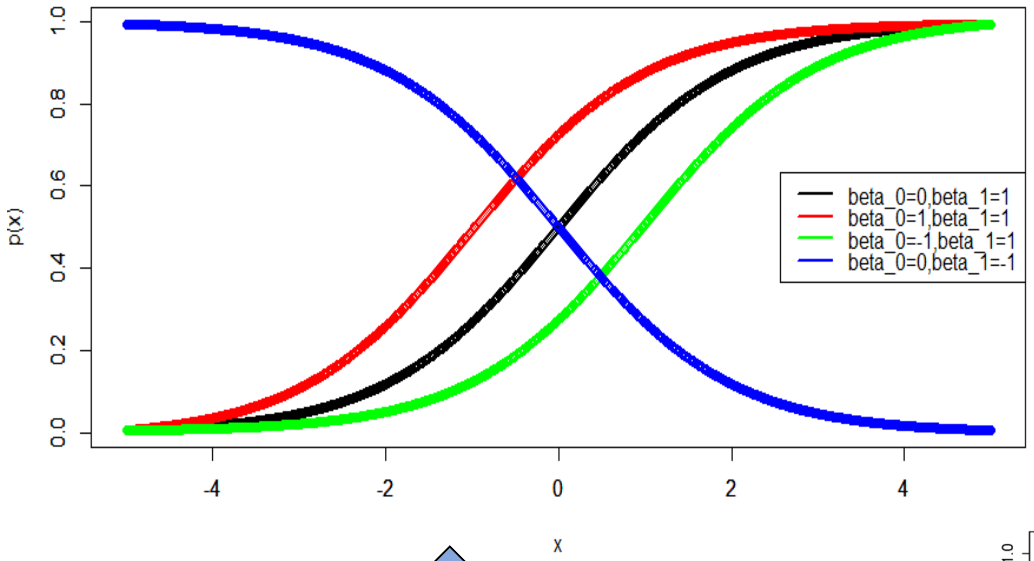


Figure 2: Percentage of patients with radiodermatitis in each age group

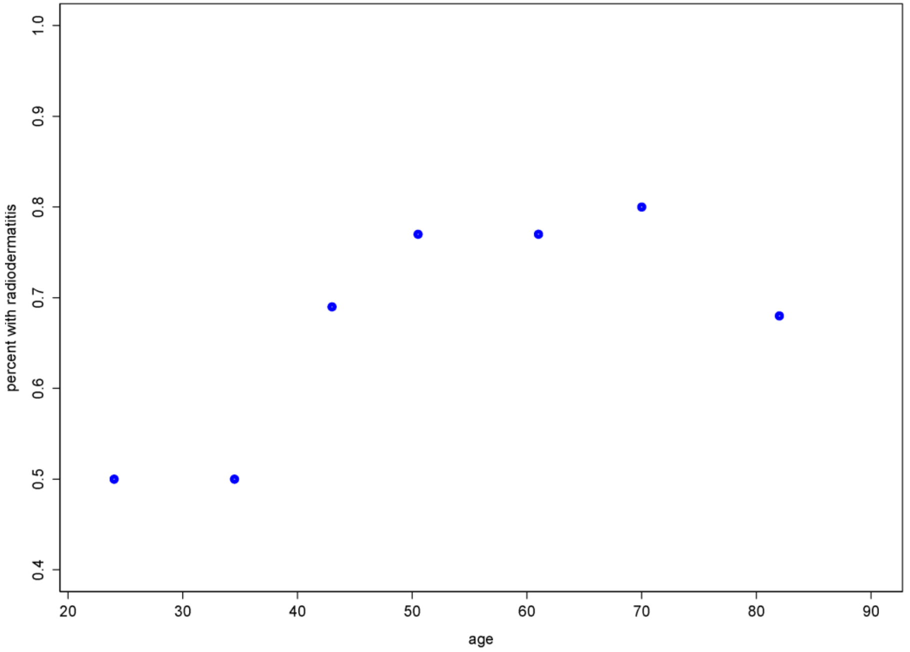


Figure 1: Logistic curve for different values of coefficients β_0 and β_1

Simple logistic regression: an example

- We consider a retrospective study with 105 carcinoma patients.
- 37 of those patients were treated with radiotherapy A and the remaining 68 patients with radiotherapy B.
- One of the possible complications from both therapies is radiodermatitis

Let us discuss the following three questions:

1. What impact has the applied radiotherapy on the occurence of radiodermatitis?
2. What impact has the degree of spread to regional lymph nodes on the occurence of radiodermatitis?
3. What impact has the age of the patients on the occurence of radiodermatitis?

Simple logistic regression: an example

- Let y denote the radiodermatitis occurence, with $y = 1$ if the patient did get radiodermatitis, and $y = 0$ otherwise
- Let x_1 denote the applied radiotherapy, with $x_1 \in \{A, B\}$
- Let x_2 denote the degree of spread to regional lymph nodes, with $x_2 \in \{N0, N1, N2, N3\}$
- Let x_3 denote the age of the patient, with $x_3 \in [22, 94]$.

We consider the following three univariable models:

1. y is (nonlinearly) related to x_1 (*binary predictor*)
2. y is (nonlinearly) related to x_2 (*polychotomous predictor*)
3. y is (nonlinearly) related to x_3 (*continuous predictor*)

Simple logistic regression: binary predictor

Model 1: *radiodermatitis ~ applied radiotherapy (A or B)*

An extract from the output in R:

```
Call:
glm(formula = as.factor(radioderm) ~ as.factor(radiotherapy),family = binomial)

..............

Coefficients:
                                  Estimate Std. Error z value  Pr(>|z|)
(Intercept)                         2.8622     0.7270   3.937  8.25e-05 ***
as.factor(radiotherapy) B          -2.3199     0.7693  -3.016   0.00256 **

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

...............
```

Question: How to interpret the coefficients in the output above?

Simple logistic regression: binary predictor

- Odds is the ratio of the probability of an event occurring to not occurring
- In our example, with the following 2x2 table

	applied radiotherapy		
radioderm	A	B	Total
no	2	25	27
yes	35	43	78
Total	37	68	105

we get

- the odds of radiodermatitis for patients who had radiotherapy A

$$odds(A) = \frac{P(radioderm="yes"|A)}{P(radioderm="no"|A)} = \frac{P(radioderm="yes"|A)}{1-P(radioderm="yes"|A)}$$

- the odds of radiodermatitis for patients who had radiotherapy B

$$odds(B) = \frac{P(radioderm="yes"|B)}{P(radioderm="no"|B)} = \frac{P(radioderm="yes"|B)}{1-P(radioderm="yes"|B)}$$

- Numerically: $\widehat{odds}(A) = \frac{35}{2} = 17.5$; $\widehat{odds}(B) = \frac{43}{25} = 1.72.$

Simple logistic regression: binary predictor

- Next, we introduce the odds ratio (OR) which in our example is defined as

$$OR(B,A) := \frac{odds(B)}{odds(A)}$$

- Numerically, we obtain the following estimate for the $OR(B,A)$:

$$\widehat{OR}(B,A) = \frac{1.72}{17.5} \cong 0.098 \qquad (2)$$

- Interpretation of $\widehat{OR}(B,A)$:

  In general, OR approximates how much higher/lower the odds for disease are for patients in the considered group than for patients in the reference group.

  Thus in our example the odds for radiodermatitis for patients with radiotherapy B are 10.17 times lower than for patients with radiotherapy A.

Simple logistic regression: binary predictor

Where are these Odds (i.e. the term $\frac{p}{1-p}$) in our logistic model?

- Logistic model: $\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x$

$$\Downarrow$$

- Odds: $\frac{p}{1-p} = e^{\beta_0 + \beta_1 x} = \frac{\mathrm{P}(Y=1)}{P(Y=0)}$

$$\Downarrow$$

- Odds Ratios: $\frac{odds(x+1)}{odds(x)} = \frac{\frac{p_{x+1}}{1-p_{x+1}}}{\frac{p_x}{1-p_x}} = \frac{e^{\beta_0+\beta_1 x_1+\beta_1}}{e^{\beta_0+\beta_1 x_1}} = e^{\beta_1}$

Simple logistic regression: binary predictor

- In our example $\widehat{\beta_1} = -2.3199$ and thus $e^{\widehat{\beta}_1} = e^{-2.3199} \cong 0.098 = \widehat{OR}(B, A)$ as given in (2).
- Further, the logit function

$$g(x) = \beta_0 + \beta_1 x$$

in our example is estimated by

$$\hat{g}(x) = \begin{cases} 2.8622 - 2.3199 = 0.5423, & \text{if a patient got radiotherapy } B \\ 2.8622, & \text{if a patient got radiotherapy } A \end{cases}$$

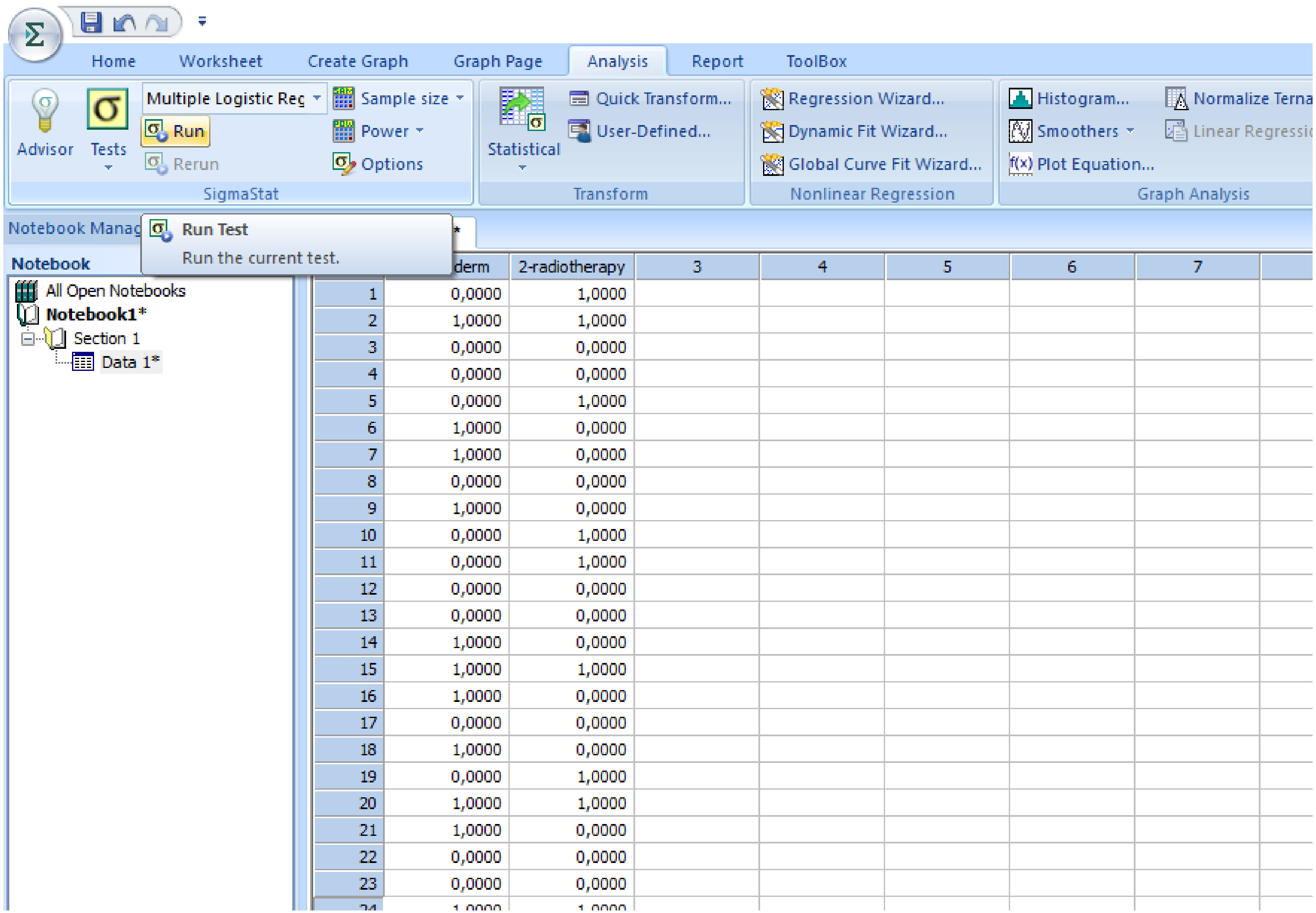
- A 95% confidence interval estimate for $\widehat{OR}\,(B, A)$ is given by:

$$e^{(\widehat{\beta_1} \pm z_{1-\alpha/2} \times \widehat{SE}(\widehat{\beta_1}))} = e^{(-2.3199 \pm 1.96 \times 0.7693)} = (0.022,\ 0.4439)$$

where $z_{1-\alpha/2}$ is the upper 97.5% point from the standard normal distribution and $\widehat{SE}(.)$ is an estimated standard error

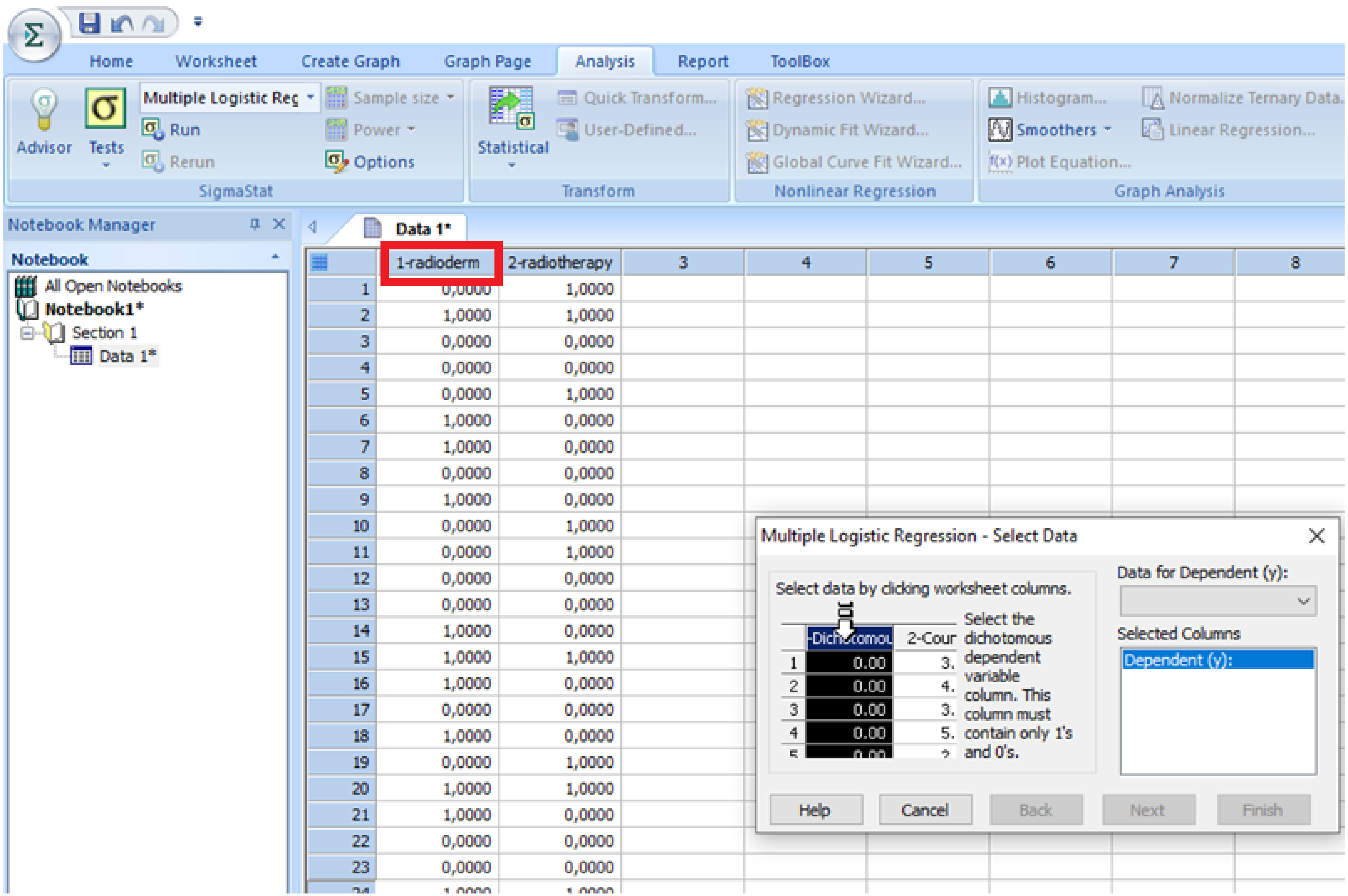
Simple logistic regression:
binary predictor, calculations in SigmaPlot

Select the multiple logistic regression function:



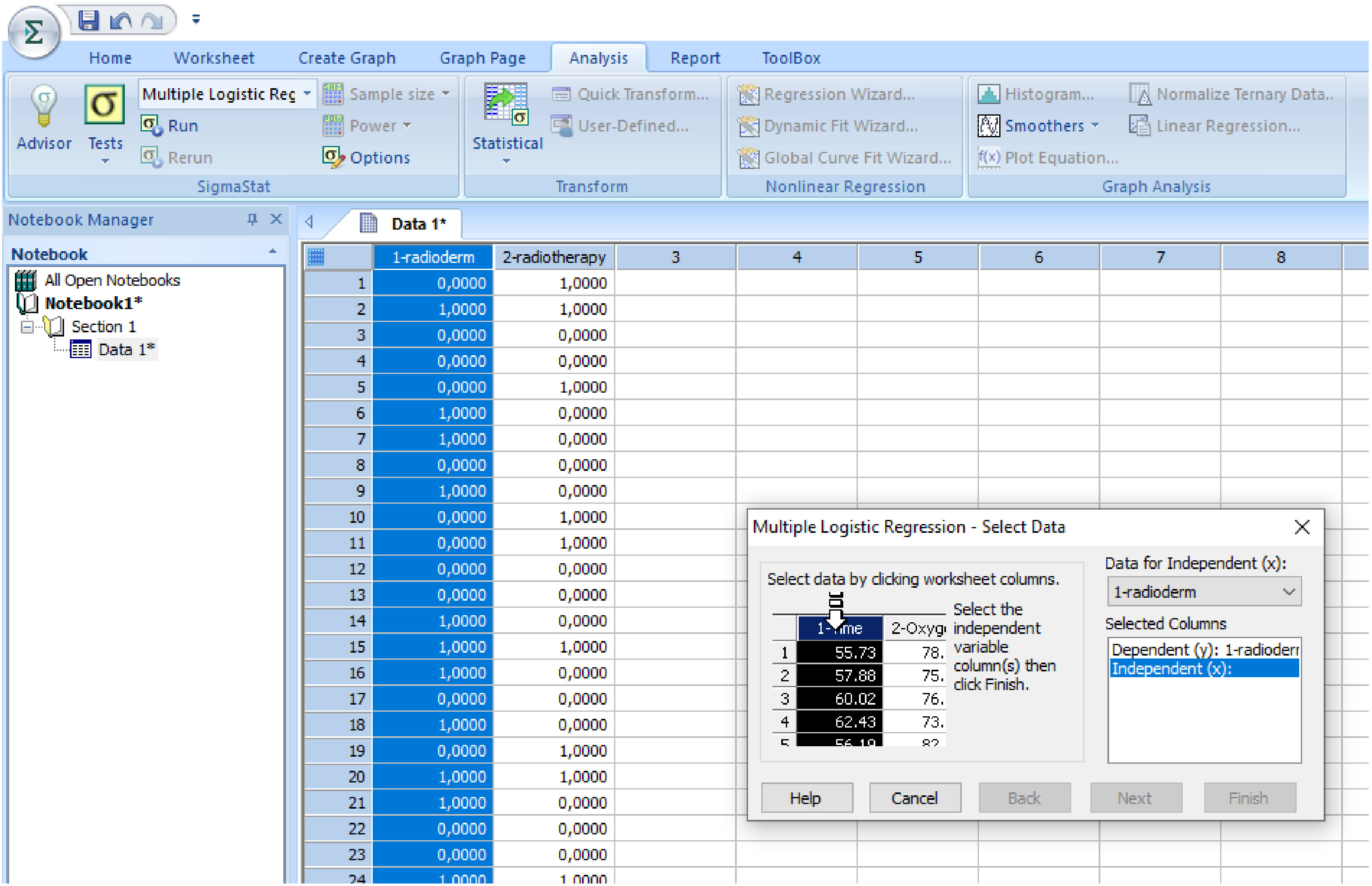
Simple logistic regression:
binary predictor, calculations in SigmaPlot

Select the outcome (dependent) variable:



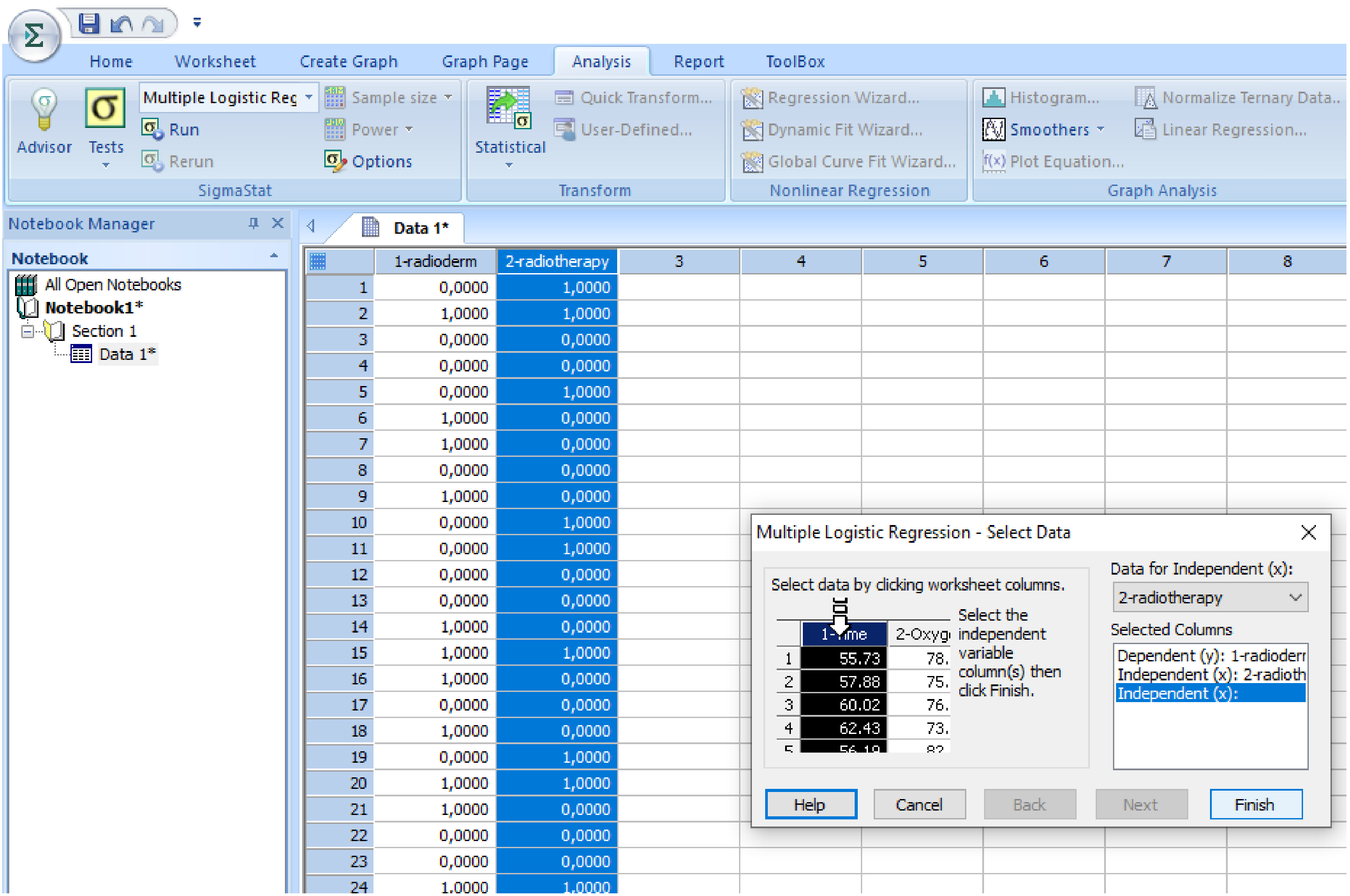
Simple logistic regression:
binary predictor, calculations in SigmaPlot

Select the predictor (independent) variable:



Simple logistic regression:
binary predictor, calculations in SigmaPlot

Click finish:



Simple logistic regression:
binary predictor, an extract from the output in SigmaPlot

Multiple Logistic Regression
Data source: Data 1 in data.xlsx

Logit P = 2,862 - (2,320 * radiotherapy)

N = 105
Estimation Criterion: Maximum likelihood
Dependent Variable: radioderm
Positive response (1): 1
Reference response (0): 0
Number of unique independent variable combinations: 2

……………………………………………………………….

Details of the Logistic Regression Equation

Ind. Variable	Coefficient	Standard Error	Wald Statistic	P value	VIF
Constant	2,862	0,727	15,499	<0,001	
radiotherapy	-2,320	0,769	9,094	0,003	1,000

Ind. Variable	Odds Ratio	5% Conf. Lower	95% Conf. Upper
Constant	17,500	4,209	72,759
radiotherapy	0,0983	0,0218	0,444

Simple logistic regression: polychotomous predictor

Model 2: *radiodermatitis* ~ *degree of spread to regional lymph nodes* (with levels *N0, N1, N2* or *N3*)

An extract from the output in R:

```
Call:
glm(formula = as.factor(radioderm) ~ as.factor(spread), family = binomial)

...........................

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.8899     0.2477   3.593 0.000327 ***
as.factor(spread)N1     0.3629     0.8392   0.432 0.665406
as.factor(spread)N2     1.1896     1.0892   1.092 0.274746
as.factor(spread)N3     1.0561     1.0974   0.962 0.335867

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

...........................
```

Question: How to interpret the coefficients in the output above?

Simple logistic regression: polychotomous predictor

We compute the odds estimates for all four categories of the degree of spread to regional lymph nodes:

	degree of spread				
radioderm	N0	N1	N2	N3	Total
no	23	2	1	1	27
yes	56	7	8	7	78
Total	79	9	9	8	105

$$\widehat{odds}(N0) = \frac{56}{23} \cong 2.43$$

$$\widehat{odds}(N2) = \frac{8}{1} = 8$$

$$\widehat{odds}(N1) = \frac{7}{2} = 3.5$$

$$\widehat{odds}(N3) = \frac{7}{1} = 7$$

Simple logistic regression: polychotomous predictor

Then we compute the corresponding odds ratios estimates:

$$\widehat{OR}(N0, N0) = \frac{\widehat{odds}(N0)}{\widehat{odds}(N0)} = 1$$

$$\widehat{OR}(N1, N0) = \frac{\widehat{odds}(N1)}{\widehat{odds}(N0)} = \frac{3.5}{2.43} \cong 1.44$$

$$\widehat{OR}(N2, N0) = \frac{\widehat{odds}(N2)}{\widehat{odds}(N0)} = \frac{8}{2.43} \cong 3.29$$

$$\widehat{OR}(N3, N0) = \frac{\widehat{odds}(N3)}{\widehat{odds}(N0)} = \frac{7}{2.43} \cong 2.88$$

Interpretation of $\widehat{OR}(N1, N0)$:

For the patients with *N1* degree of spread to regional lymph nodes the odds to get radiodermatitis are 1.44 times higher than for the patients with *N0* degree of spread to regional lymph nodes.

Simple logistic regression: polychotomous predictor

Using the R-output:

- $OR(N1, N0) = e^{\beta_1} = e^{0.3629} = 1.44$
- $OR(N2, N0) = e^{\beta_2} = e^{1.1896} = 3.29$
- $OR(N3, N0) = e^{\beta_3} = e^{1.0561} = 1.88$

Simple logistic regression: polychotomous predictor

- An estimate of the logit function

$$g(x) = \beta_0 + \beta_1 x$$

in our example is of the form

$$\hat{g}(x) = \begin{cases} 0.8899, & \text{for } N0 \text{ degree of spread} \\ 0.8899 + 0.3629 = 1.2528, & \text{for } N1 \text{ degree of spread} \\ 0.8899 + 1.1896 = 2.0795, & \text{for } N2 \text{ degree of spread} \\ 0.8899 + 1.0561 = 1.946, & \text{for } N3 \text{ degree of spread} \end{cases}$$

- A 95% confidence interval estimate for $OR(N1, N0)$ is given by

$$e^{(\widehat{\beta_1} \pm z_{1-\alpha/2} \times \widehat{SE}(\widehat{\beta_1}))} = e^{(0.3629 \pm 1.96 \times 0.8392)} = (0.278,\ 7.45)$$

where $z_{1-\alpha/2}$ is the upper 97.5% point from the standard normal distribution and $\widehat{SE}(.)$ is an estimated standard error of the corresponding parameter estimator. Recall that $\widehat{OR}(N1, N0) \cong 1.44$.

Simple logistic regression: continuous predictor

Model 3: *radiodermatitis ~ age*

An extract from the output in R:

```
Call:
glm(formula = as.factor(radioderm) ~ age, family = binomial)

......................

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.001238   0.936354   1.069    0.285
age         0.001031   0.015734   0.066    0.948

......................
```

Question: How to interpret the coefficients in the output above?

Simple logistic regression: continuous predictor

Assumption: The logit $g(x)$ is linear in the (continuous) predictor x.

In our example an estimate for the logit function takes the form

$$\widehat{g}\,(age) = 1.001 + 0.001 * age$$

Since $OR(c) = OR(x + c, x) = e^{c\beta_1}$, we obtain, e.g.

$$\widehat{OR}(20) = e^{20*\widehat{\beta_1}} = e^{20*0.001} \quad = 1.02$$

Interpretation of $\widehat{OR}(20)$:

For every increase of 20 years in age, the odds to get radiodermatitis increases by 2% (or 1,02 times).

Alternative approach: If reasonable, dichtomize a continuous predictor and consider the situation with a binary predictor.

Multiple logistic regression: motivation

- Till now:
  Logistic regression with a single explanatory variable in the fitted model.

- Issues:
  - possible associations between single variables
  - different distributions of a variable in different levels of the outcome

- Idea: Apply a multivariable analysis in order to (statistically) adjust the estimated effect of each single explanatory variable to possible associations with other explanatory variables considered in the model.

- Note: Each estimated coefficient gives an estimate of odds ratio *adjusted for possible effects of all other variables included in the model.*

Multiple logistic regression: definitions

- Let y be (nonlinearly) related to $\beta_0+\beta_1 x_1+\cdots+\beta_N x_N$ where
  - y is a dichotomous outcome variable
  - $x_1, x_2, \ldots, x_N$ are N independent explanatory variables
  - $\beta_0, \beta_1, \ldots, \beta_N$ are some unknown coefficients
- Assumption: let $y \in \{0,1\}$
- Analogously to the univariable case we introduce
  - The probability of an event occuring $p(\boldsymbol{x}) := P(y=1|\boldsymbol{x})$ with $\boldsymbol{x}=(x_1, x_2, \ldots, x_N)$
  - Multiple logistic regression model

$$p(\boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_N x_N}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_N x_N}}$$

  - Logit transformation

$$g(\boldsymbol{x}) := \log\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})} = \beta_0+\beta_1 x_1+\cdots+\beta_N x_N$$

Multiple logistic regression: an example

Model: *radiodermatitis ~ applied radiotherapy(A or B)+ age*

An extract from the output in R:

```
Call:
glm(formula = as.factor(radioderm) ~ as.factor(radiotherapy) + age, family = binomial)

.....................

Coefficients:
                          Estimate Std. Error z value  Pr(>|z|)
(Intercept)                3.50384    1.34817   2.599   0.00935 **
as.factor(radiotherapy)B -2.37966    0.77945  -3.053   0.00227 **
age                       -0.01031    0.01790  -0.576   0.56451

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

......................
```

Question: How to interpret the coefficients in the output above?

Multiple logistic regression: an example

- Note:
  - our primary interest is on the effect of applied radiotherapy
  - *age*-variable is usually called a covariate
  - outcome is assumed to be linear in both explanatory variables
- The estimated *age*-adjusted odds ratio in our example is given by

$$\widehat{OR}_{adj}(B, A) \cong e^{-2.38} \cong 0.093 \qquad (3)$$

- Comparing the value of $\widehat{OR}_{adj}(B, A)$ in (3) with the univariate (unadjusted) odds ratio estimate

$$\widehat{OR}_{unadj}(B, A) \cong 0.098$$

as given in (2), we state that in our example there is a little difference between the two considered groups due to differences in age.

Logistic regression: model-building strategies

- Issue: There are many independent variables we would like to include in the model
- Question: How to perform the variables selection that would lead to the „best“ model?
- A naive approach: Include all variables we have into the model.
- Possible disadvantages:
  - numerically instable model („overfit“)
  - model becomes even more dependent on the observed data
  - possibly time-consuming computations

Logistic regression: model-building strategies

- 1: Preselect variables for multivariable analysis depending on their clinical importance. Then use either

  - Forward variable selection
  - Backward variable elimination
  - Best subset regression

- 2: Choose which variables to include based on (for example)

  - p-values
  - likelihood ratio test

Logistic regression: Take home message

Logistic regression …

- is used for analysis of binary (0/1) endpoints
- allows construction of prognostic and predictive models
- expresses effects of predictors through Odds Ratios
- allows to consider several possible predictors at once
- can adjust effects of predictors for covariables

References

- D.W. Hosmer, S. Lemeshow. *Applied Logistic Regression.*- Second Edition. John Wiley &Sons, Inc., 2000.
- D.G. Kleinbaum, M. Klein*. Logistic Regression: A Self-Learning Text.* Second Edition. Springer, Berlin, 2002.
- F.C. Pampel. *Logistic Regression: A Primer.* Sage Publications, 2000.

NEXT LECTURE

Topic: „Dose-Response-Modeling“
When: October 28, 2020