Biostatistics

Logistic regression

Burkhardt Seifert & Alois Tschopp

Biostatistics Unit
University of Zurich

Logistic regression

Great importance for medical research

So far: "ordinary" regression

- explain an "outcome" variable y
  through explanatory variables $x_1, \ldots, x_k$
- quantitative outcome variable y (normally distributed)
- relation usually assumed to be linear

New with logistic regression: outcome **y** is binary

Examples

A. y = patients survive ($y = 0$) or die ($y = 1$)
   x_1 = therapy (x_1 = A, B; nominal)
   x_2 = age (in years; continuous)
   $x_3, \ldots$ = laboratory parameters.

B. case–control–study (epidemiology)
   y = case ($y = 1$) or control ($y = 0$)
   x_1 = exposed ($x_1 = 1$) or not ($x_1 = 0$)
   $x_2, \ldots$ = confounder.

Statistical analysis

with one independent variable x also:

- Mann–Whitney test (or unpaired t–test)
- Fisher's exact test (or χ^2–test)

Int. J. Cancer: **121,** 1764–1770 (2007)

Consistent expression of the stem cell renewal factor BMI-1 in primary and metastatic melanoma

Daniela Mihic-Probst[1]*, Ariana Kuster[1], Sandra Kilgus[1], Beata Bode-Lesniewska[1], Barbara Ingold-Heppner[1], Carly Leung[1], Martina Storz[1], Burkhardt Seifert[2], Silvia Marino[3], Peter Schraml[1], Reinhard Dummer[4] and Holger Moch[1]

[1]*Department of Pathology, Institute of Surgical Pathology, University Hospital Zurich, Zurich, Switzerland*
[2]*Department of Biostatistics, University of Zurich, Zurich, Switzerland*
[3]*Institute of Pathology, Barts and the London, Queen Mary School of Medicine and Dentistry, London, United Kingdom*
[4]*Department of Dermatology, University Hospital Zurich, Zurich, Switzerland*

Stem cell-like cells have recently been identified in melanoma cell lines, but their relevance for melanoma pathogenesis is controversial. To characterize the stem cell signature of melanoma, expression of stem cell markers BMI-1 and nestin was studied in 64 cutaneous melanomas, 165 melanoma metastases as well as 53 melanoma cell lines. Stem cell renewal factor BMI-1 is a transcriptional repressor of the Ink4a/Arf locus encoding $p16^{ink4a}$ and $p14^{Arf}$. Increased nuclear BMI-1 expression was detectable in 41 of 64 (64%) primary melanomas, 117 of 165 melanoma metastases (71%) and 15 of 53 (28%) melanoma cell lines. High nestin expression was observed in 14 of 56 primary melanomas (25%), 84 of 165 melanoma metastases (50%) and 21 of 53 melanoma cell lines (40%). There was a significant correlation between BMI-1 and nestin expression in cell lines ($p = 0.001$) and metastases ($p = 0.02$). These data indicate that cells in primary melanomas and their metastases may have stem cell properties. Cell lines obtained from melanoma metastases showed a significant higher BMI-1 expression compared to cell lines from primary melanoma ($p = 0.001$). Further, primary melanoma lacking lymphatic metastases at presentation (pN0, $n = 40$) was less frequently BMI-1 positive than melanomas presenting with lymphatic metastases (pN1; $n = 24$; 52% *versus* 83%; $p = 0.01$). Therefore, BMI-1 expression appears to induce a metastatic tendency. Because BMI-1 functions as a transcriptional repressor of the Ink4a/Arf locus, $p16^{ink4a}$ and $p14^{Arf}$ expression was also analyzed. A high BMI-1/low $p16^{ink4a}$ expression pattern was a significant predictor of metastasis by means of logistic regression analysis ($p = 0.005$). This suggests that BMI-1 mediated repression of $p16^{ink4a}$ may contribute to an increased aggressive behavior of stem cell-like melanoma cells.

Statistics

BMI-1, $p16^{ink4a}$, $p14^{Arf}$ and nestin expression in primary melanoma were compared between different patient groups using the Mann-Whitney test. Correlations between BMI-1, $p16^{ink4a}$, $p14^{Arf}$, nestin and Breslow tumor thickness were analyzed using Spearman's rank correlation. Differences in tumor-specific survival between groups were calculated by log rank test. A logistic regression was performed to evaluate the predictive power of BMI-1 and $p16^{ink4a}$ expression in primary malignant melanoma for lymph node metastasis. *p*-Values below 0.05 were considered as significant. SPSS 12.0.1 for windows (SPSS) was used for statistical analyses.

TABLE II – RELATIVE RISK OF LYMPH NODE METASTASIS ACCORDING TO BMI-1 AND $P16^{INK4A}$ EXPRESSION LEVELS IN PRIMARY MELANOMA

	n	Univariate OR	*p*-value	Multivariate OR	*p*-value
$p16^{ink4a}$ low *vs.* high[1]	35/29	3.0 (1.0–8.6)[2]	0.04	2.7 (0.89–8.1)	0.08
BMI-1 high *vs.* Low[1]	41/23	4.5 (1.3–15.6)	0.02	4.1 (1.2–14.6)	0.03
$p16^{ink4a}$ low/BMI-1 high *vs.* others[1]	22/42	3.2 (1.4–7.3)	0.005		

Odds ratio (OR)

Example: Identification of risk factors for lymph node metastases with prostate cancer (Brown, 1980)

$n = 52$ patients
y = nodal metastases (0 = none, 1 = metastases)
x = age, phosphatase, X-ray result, tumor size, tumor grade.

The first two x–variables are continuous, the rest binary.

Contingency table for the relation between nodal metastases and X-ray result

	X-ray result $x = 0$	X-ray result $x = 1$	
no nodal metastases ($y = 0$)	28	4	32
nodal metastases ($y = 1$)	9	11	20
	37	15	52

sensitivity $= 11/20 = 55\%$, specificity $= 28/32 = 87\%$
χ^2–test $p = 0.001$

Relative risk (RR) or **odds ratio** (OR)?

	$x = 0$	$x = 1$	
$y = 0$	28	4	32
$y = 1$	9	11	20
	37	15	52

"Risk" defined as

$$\mathrm{P}(y = 1|x) = p(x)\,,$$

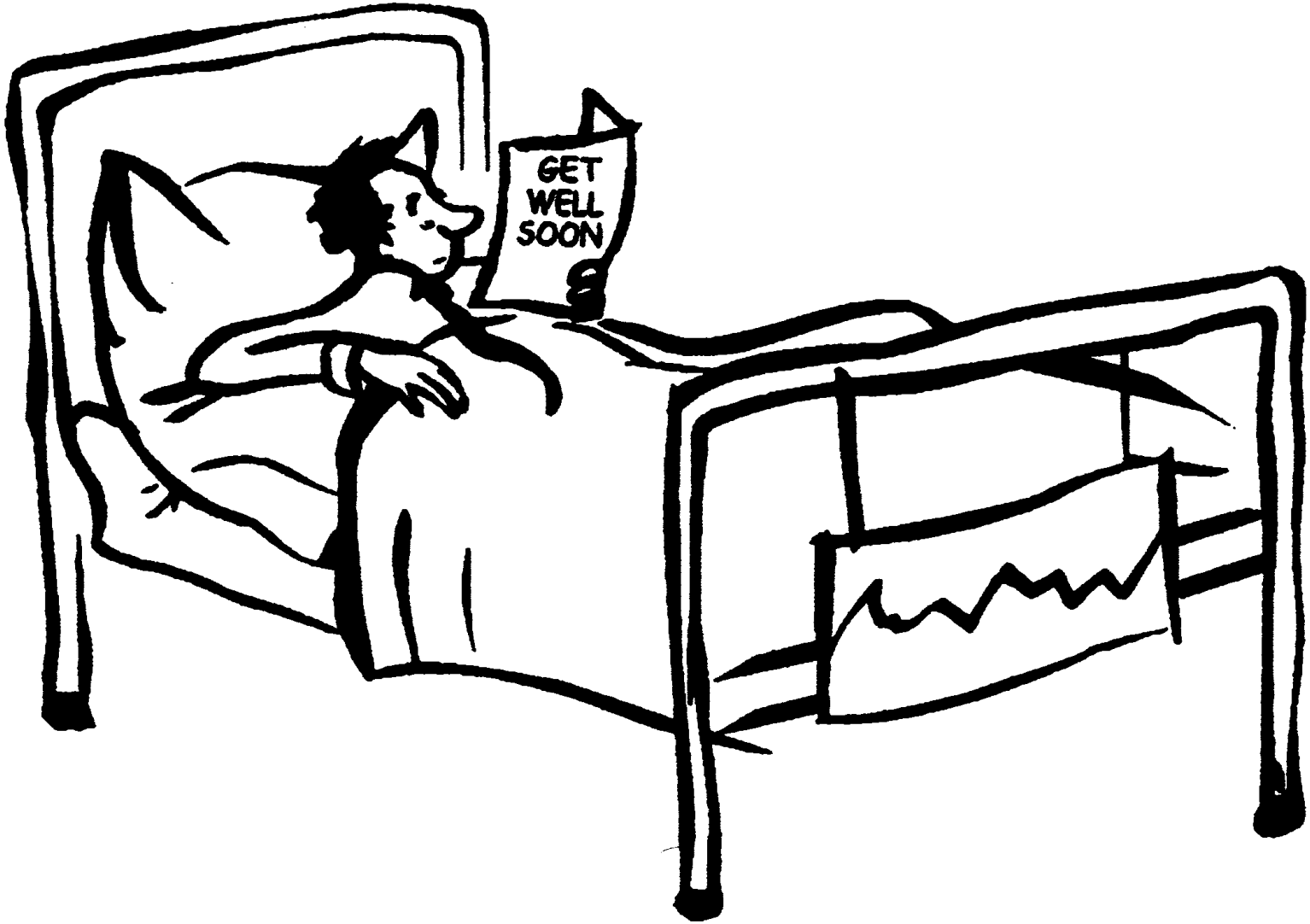
$\rightarrow p(0) = 9/37 = 24\%,\ p(1) = 11/15 = 73\%$

$$\mathrm{RR} = p(1)/p(0) = \frac{11 \times 37}{15 \times 9} = 3.0$$

RR only valid for representative sample

From betting we know "odds":

$$\frac{\mathrm{P}(y = 1|x)}{\mathrm{P}(y = 0|x)} = \frac{p(x)}{1 - p(x)}$$



Dear Mr. Goodman, CEO, the Members of the Board wish you a speedy recovery by a vote of 11 to 8.

Relative risk (RR) or **odds ratio** (OR)?

Im epidemiology the "odds ratio" is a measure for the relative risk:

	$x = 0$	$x = 1$
$y = 0$	28	4
$y = 1$	9	11

$$\mathrm{OR} = \frac{\mathrm{P}(y = 1|x = 1)}{1 - \mathrm{P}(y = 1|x = 1)} \Big/ \frac{\mathrm{P}(y = 1|x = 0)}{1 - \mathrm{P}(y = 1|x = 0)} = \frac{28 \times 11}{9 \times 4} = 8.6$$

OR is also valid for case–control studies

For rare diseases, OR and RR are nearly equal:

$$\mathrm{OR} = \frac{p(1)}{1 - p(1)} \Big/ \frac{p(0)}{1 - p(0)} \approx \frac{p(1)}{p(0)}$$

Modelling by means of logistic regression

What is fundamental for a (simple) regression?

Model: $$y_i = f(x_i, \beta) + \varepsilon_i \quad (i = 1, \ldots, n)$$

where: f = pre-specified function
e.g. linear $f(x_i, \beta_0, \beta_1) = \beta_0 + \beta_1 x_i$

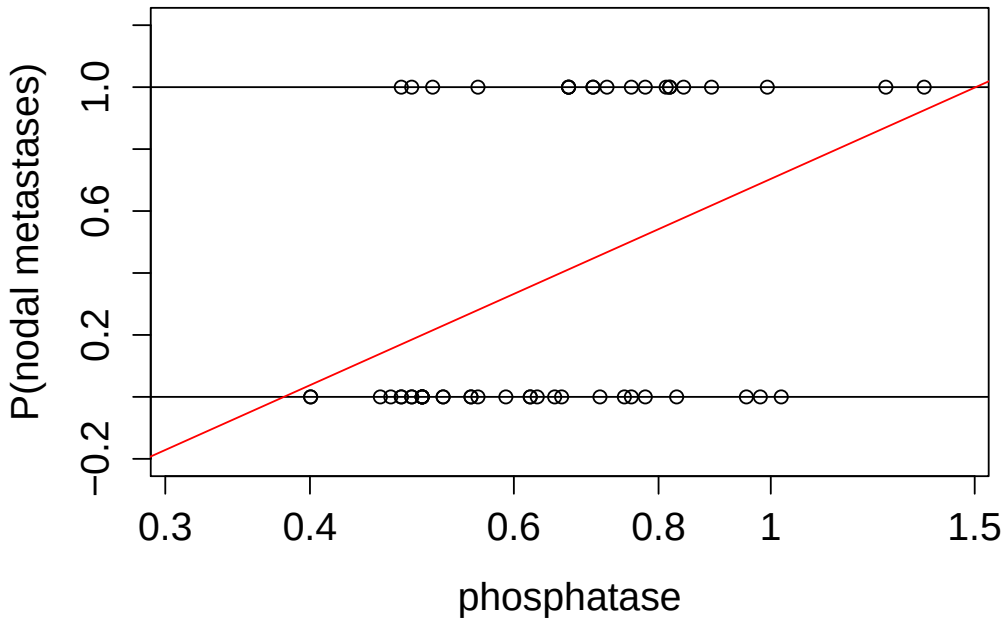
regression function $f(x, \beta)$ = conditional expectation of y, given the value x, i.e.

$$\mathrm{E}(y \mid x) = f(x, \beta)$$

- "outcome" binary event: "success" $(y = 1)$, "failure" $(y = 0)$
- probability for success $p = \mathrm{P}(y = 1)$
- $\mathrm{E}(y) = 0 \times \mathrm{P}(y = 0) + 1 \times \mathrm{P}(y = 1) = p$

Why not use ordinary regression?

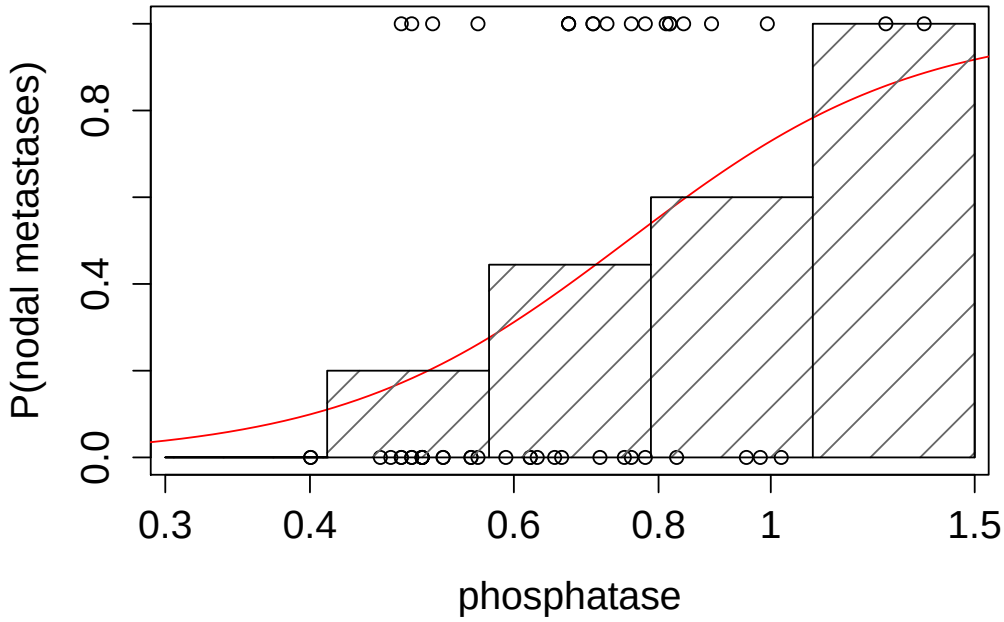
Example: y = presence of nodal metastases
x = phosphatase (logarithmised)



regression = conditional mean of y given $x \longrightarrow \mathrm{E}(y \mid x)$.

Thus: $\mathrm{E}(y \mid x) = \mathrm{P}(y = 1 \mid x) = p(x)$

A probability is modelled — lies between 0 and 1.
$\longrightarrow$ plausible to model $p(x)$ as distribution function.



$y = 0$	2	16	10	4	0
$y = 1$	0	4	8	6	2
OR		∞	3.2	1.9	∞

- OR for [0.58–0.79] vs. [0.41–0.57] $= \frac{16 \times 8}{10 \times 4} = 3.2$
- OR for [0.80–1.09] vs. [0.41–0.57]
  $= \frac{16 \times 6}{4 \times 4} = \frac{16 \times 8}{10 \times 4} \times \frac{10 \times 6}{4 \times 8} = 3.2 \times 1.9 = 6$
- OR for a change of more than one class: multiplicative

Which distribution function to use?

- Assumption: odds ratio for adjoining classes is constant (similar to the assumption of a constant slope of the regression function in linear regression)

As OR multiplicative, log(OR) must be linear.
$\longrightarrow$ for log–odds (logits):

$$\boxed{\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x}$$

(log = natural logarithm = $\log_e$)

$\longrightarrow$ $p(x)$ is logistic distribution function

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

♣ Linearity of the logit–transformation

Assumption:

OR for $x = x_0 + c$ vs $x = x_0$ is constant in x_0 $\quad = \mathrm{OR}(c)$

OR multiplicative $\longrightarrow \mathrm{OR}(c) = \mathrm{OR}(1)^c$

$$\text{Is} \quad g(x) = \log\left(\frac{p(x)}{1-p(x)}\right) \quad \text{linear?}$$

$\mathrm{OR}(c)$: true OR for "$x = c$" vs $x = 0$

$$\log\left(\mathrm{OR}(c)\right) = g(c) - g(0)$$

logarithmise:

$$\begin{aligned} g(x) - g(0) &= \log\left(\mathrm{OR}(1)\right) x \\ g(x) &= g(0) + \log\left(\mathrm{OR}(1)\right) x \\ g(x) &= \beta_0 + \beta_1 x \end{aligned}$$

with $\beta_0 = g(0)$

and $\beta_1 = \log\left(\mathrm{OR}(1)\right)$

Estimation and testing in logistic regression

A. How to estimate β_0, β_1?
B. How to test whether the influence of x on y is not by chance ("significant")?

Scientific hypothesis H_1: $\beta_1 \neq 0$

Example: phosphatase influences presence of nodal metastases

Null hypothesis H_0: $\beta_1 = 0$

Example: phosphatase has no influence

Method: Maximum Likelihood Estimation

Attractive characteristics:

I. maximum likelihood estimates are optimal
($\longrightarrow$ optimal use of data).

II. they are normally distributed with known variance–covariance matrix.
($\longrightarrow$ precision known $\longrightarrow$ statistical tests)

III. tests and confidence intervals are optimal
("likelihood ratio tests")

But:

- iterative procedure, i.e. solution not always correct
- p–values only valid for large n ("asymptotically")
(analogous to χ^2–test)

What is maximum likelihood principle? (informal)

- Probability for event $y_i = 1$ known (Bernoulli), depending on unknown model parameters β_0, β_1 ("likelihood–function")
- Inserting data in model for $p(x)$ yields likelihood function (= function of parameters $\beta_0,\ \beta_1$):

$$\mathrm{P}(y_i = 1|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

- Determine $\hat{\beta}_0,\ \hat{\beta}_1$ by maximising the likelihood, i.e. probabilities to observe these data (x_i, y_i) get maximal.
- Computing: iteratively solve a system of non–linear equations for $\hat{\beta}_0, \hat{\beta}_1$
- variance–covariance matrix for $\hat{\beta}_0, \hat{\beta}_1$ as a byproduct.

$\Rightarrow$ Leads to confidence intervals and tests

Example: prostate cancer

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9919	0.6033	1.64	0.1001
$\log_2$(phosph)	2.4198	0.8778	2.76	0.0058

95% confidence interval for exp(β_1):

	exp(Estimate)	Lower	Upper
$\log_2$(phosph)	11.24	2.01	62.83

Hosmer and Lemeshow test (goodness of fit):

$\chi^2 = 7.245$, $df = 8$, p–value$= 0.510$

Wald test

Test for a single predictor

Wald test statistic

$$W = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)}$$

p–value: use of approximate normal distribution of $\hat{\beta}_1$ and standard error.

Example: Nodal metastases vs. phosphatase

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9919	0.6033	1.64	0.1001
$\log_2$(phosph)	2.4198	0.8778	2.76	0.0058

$$W = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} = \frac{2.42}{0.9} = 2.8$$

- Two-sided approximate *p*–value: $\mathrm{P}(|z| > 2.8) = 0.006$
- Statistically significant, clinically negative influence of an increased phosphatase

Interpretation of coefficients

Linear regression: If x changes by one unit, the mean of y changes by β_1 units.

Relation between $p(x) = \mathrm{P}(y = 1 \mid x)$ and x is linear in logits:

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

Thus: change in x by one unit

$\longrightarrow$ change in logit of $p(x)$ by β_1 units

Interpretation: **binary x variable**

- "odds ratio" OR: ratio of odds for $x = 1$ (pos X-ray) to odds for $x = 0$ (neg X-ray)

$$\begin{aligned} \text{OR} &= \frac{p(1)}{1-p(1)} \Big/ \frac{p(0)}{1-p(0)} \\ \longrightarrow \quad \log(\text{OR}) &= g(1) - g(0) \\ &= (\beta_0 + \beta_1 \times 1) - (\beta_0 + \beta_1 \times 0) \\ &= \beta_1 \\ \text{i.e.} \quad \text{OR} &= \exp(\beta_1) \end{aligned}$$

- OR for neg vs. pos X-ray $= \exp(-\beta_1) = 1/\exp(\beta_1)$

Interpretation: **continuous x variable**

If x changes by one unit, the logit changes by $\log(\text{OR}) = \beta_1$ units.

Thus: odds ratio $= \exp(\beta_1)$ is a measure for an increase in risk (in odds) when x changes by one unit.

logit-increase when x changes by k units:

$$\log(\text{OR}) = (\beta_0 + \beta_1 \times (x + k)) - (\beta_0 + \beta_1 \times x) = k \times \beta_1$$

OR for change of x by k units:

$$\exp(k\,\beta_1) = (\exp(\beta_1))^k = \text{OR}^k$$

Interpretation of coefficients

Example: OR when phosphatase changes by a factor of 2:

$$\text{OR} = \exp(\beta_1) = 11.2$$

OR for a change by a factor of 1.5:

$$1.5 = 2^{0.585} \longrightarrow \text{OR} = 11.2^{0.585} = 4.1$$

Interpretation: categorical or ordinal x variable

One has to introduce binary "design variables",
then interpretation as for binary variables.

Computation of individual risk

Representative sample $\longrightarrow$ $p(x)$ and RR appropriate

Example: y = nodal metastases, $x = \log_2(\text{phosphatase})$

absolute individual risk: $p(x) = \dfrac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$

RR for patients with one unit increase of $\log_2$(phosphatase) compared to mean $\bar{x}$ (i.e. doubling of phosphatase):

$\bar{x} = -0.63$ (corresponds to phosphatase of $2^{-0.63} = 0.64$)

$$p(\bar{x}) = \frac{\exp(1.0 + 2.42 \times (-0.63))}{1 + \exp(1.0 + 2.42 \times (-0.63))} = 0.37$$

$$p(\bar{x} + 1) = \frac{\exp(1.0 + 2.42 \times 0.37)}{1 + \exp(1.0 + 2.42 \times 0.37)} = 0.87$$

$$\longrightarrow \text{RR} = \frac{p(\bar{x} + 1)}{p(\bar{x})} = \frac{0.87}{0.37} = 2.3$$

individual risk with doubled phosphatase is increased by a factor of 2.3. The OR, however, is OR = 11.2 !

Multiple logistic regression

$k > 1$ variables $x_1, \ldots, x_k \longrightarrow$ multiple logistic regression

Reasons as for multiple linear regression:

1. Eliminate potential effects of "confounding" variables in a study with one explanatory variable.
2. Investigate potential prognostic factors of which we are not sure whether they are important or redundant.
3. Develop formulas for a better prediction of individual risk based on explanatory variables

- Problem solved with maximum likelihood principle
- Rule of thumb: at least 20 events and 20 non-events per explanatory variable

Univariate analysis for prostate cancer example

	Estimate	Std. Error	z value	Pr(>|z|)	OR
$\log_2$(phosph)	2.4198	0.8778	2.76	0.0058	11.2
Age	-0.0448	0.0468	-0.96	0.3379	1.0
X-ray	2.1466	0.6984	3.07	0.0021	8.6
Size	1.6094	0.6325	2.54	0.0109	5.0
Grade	1.1389	0.5972	1.91	0.0565	3.1

Multiple logistic regresstion: prostate cancer example

	Estimate	Std. Error	z value	Pr(>|z|)	OR
(Intercept)	-0.5418	0.8298	-0.65	0.5138	
$\log_2$(phosph)	2.3645	1.0267	2.30	0.0213	10.6
X-ray	1.9704	0.8207	2.40	0.0163	7.2
Size	1.6175	0.7534	2.15	0.0318	5.0

Interpretation:

$\hat{\beta}_i$ Influence of x_i when remaining variables are fixed

p–values Does x_i, given the fixed remaining variables, yield additional information about $P(y = 1)$? Significant variables are called "independent risk factors".

$\exp(\hat{\beta}_i)$ OR with fixed remaining variables, i.e. OR of a patient with X-ray = 1, Size = 1, by a factor of 2 decreased phosphatase against a patient with X-ray = 0, Size = 0: $\text{OR} = 5.0 \times 7.2/10.6 = 3.4$

Multiple logistic regression

How to combine the information of several significant explanatory variables?

$$\text{PI} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

is a prognostic index (score).

If PI large ($>$ cut-point), we predict "$y = 1$".

Model choice and model tests

- difficult topic $\longrightarrow$ expert
- similar to linear regression

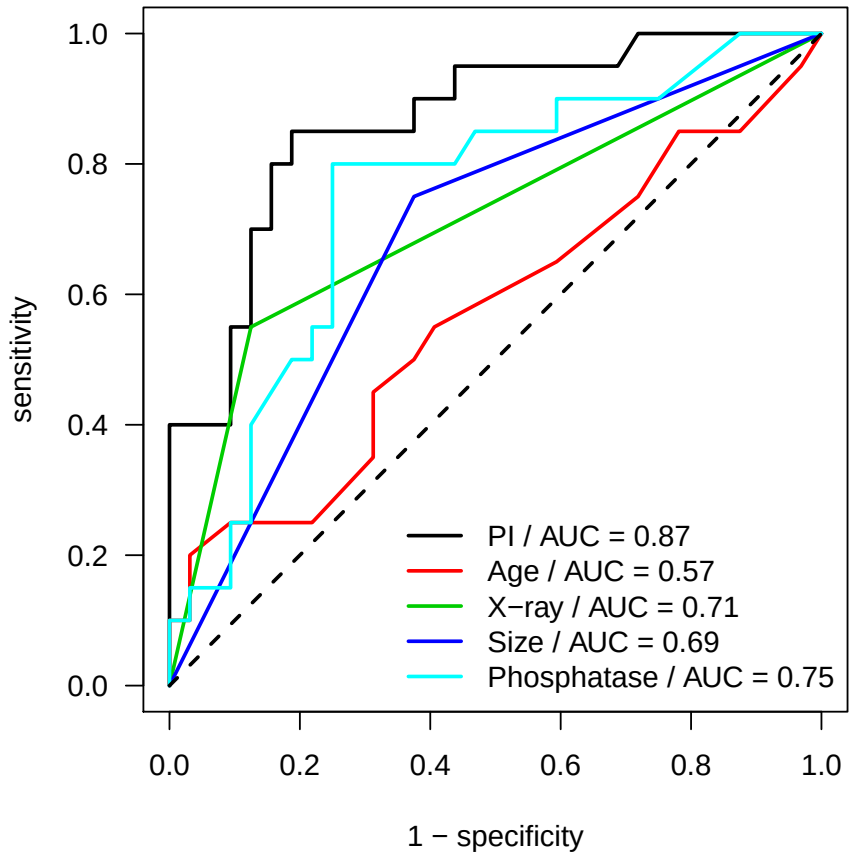
By means of

- statistical tests (comparison of models)
- R^2 often provided, but use is controversial
- judge quality of a model by means of sensitivity and specificity $\longrightarrow$ "ROC analysis"

Goodness of prediction

Example: Nodal metastases with prostate cancer

- ROC (receiver operating characteristic) curve:



$$\text{PI} = 2.4 \times \log_2(\text{phosphatase}) + 2 \times \text{X-ray} + 1.6 \times \text{Size}$$

- area of 0.5 corresponds to complete ignorance.

Choice of variables

Too many: varying parameter, large SEs

Too few: outcome not well explained, bias

$\longrightarrow$ include all variables that are, a priori, of medical interest
$\longrightarrow$ "principle of parsimony"
$\longrightarrow$ if: clear idea $\longrightarrow$ comparative testing of models

If unclear and many predictors:

(i) Compute univariate model for each x variable, eliminate e. g. those with $p > 0.2$

(ii) Build a multiple model with the remaining variables; eliminate clearly non–significant variables

Alternative: stepwise selection

Model building

Linearity of logits in x

- test against nonlinear alternatives (quadratic, Box–Tidwell–test: interaction $x * log(x)$)
- transformation of x to linear relation

Interactions

Example: y = occurrence of a coronary heart disease

x_1 = age, x_2 = gender

Model without interaction ("additive"):

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Meaning: gender related differences are not depending on age.
If gender related differences increase or decrease with age ("specific effect") $\longrightarrow$ modelling including interaction.

$$g(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

OR for gender is then depending on age.

Literature

Matthews, D. E. and Farewell, V. T. (1988). *Using and understanding medical statistics*. 2nd ed., Karger.
- contrary to many other introductions, this book includes logistic regression and survival analysis, 200 pages.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*. 2nd ed., Wiley.
- explain logistic regression using 6 worked out, medical examples, 373 pages.

Ryan, T. P. (1997). *Modern regression methods*. Wiley.
- Chapter 1–8: Linear regression
- Chapter 9: Logistic regression, 59 pages.
- Chapter 10–15: Non-parametric, robust, Ridge-, non-linear regression, experimental design
Contains exercises and further literature. Theoretically demanding, 515 pages.