Basic Medical Statistics Course

S7 Logistic Regression

Michael Hauptmann

m.hauptmann@nki.nl

Logistic regression

The concept of a relationship between the distribution of a dependent variable and a number of explanatory variables is also valid when the dependent variable is ***qualitative* (0 or 1)** instead of ***quantitative*** (with an unlimited range).

The relationship is in this case between explanatory variables and ***probability (1)***.

This cannot be a linear relationship, since probabilities have boundaries 0 and 1.

Examples:

dead - alive

side effect - no side effect

disease - no disease

NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

Example

Dataset N=24 (BMI, Blood Pressure, Diabetes)

BMI	BloodPr	BMI	BloodPr
17	120	33	118
22	130	32	170
34	144	37	160
23	122	22	100
43	119	18	101
34	115	23	103
29	132	26	128
19	121	26	110
20	124	33	134
29	140	19	121
25	134	18	123
27	118	25	122

BMI	Diabetes	BMI	Diabetes
17	0	33	0
22	0	32	1
34	1	37	1
23	1	22	0
43	1	18	0
34	0	23	0
29	0	26	0
19	0	26	1
20	0	33	0
29	0	19	0
25	0	18	0
27	0	25	0

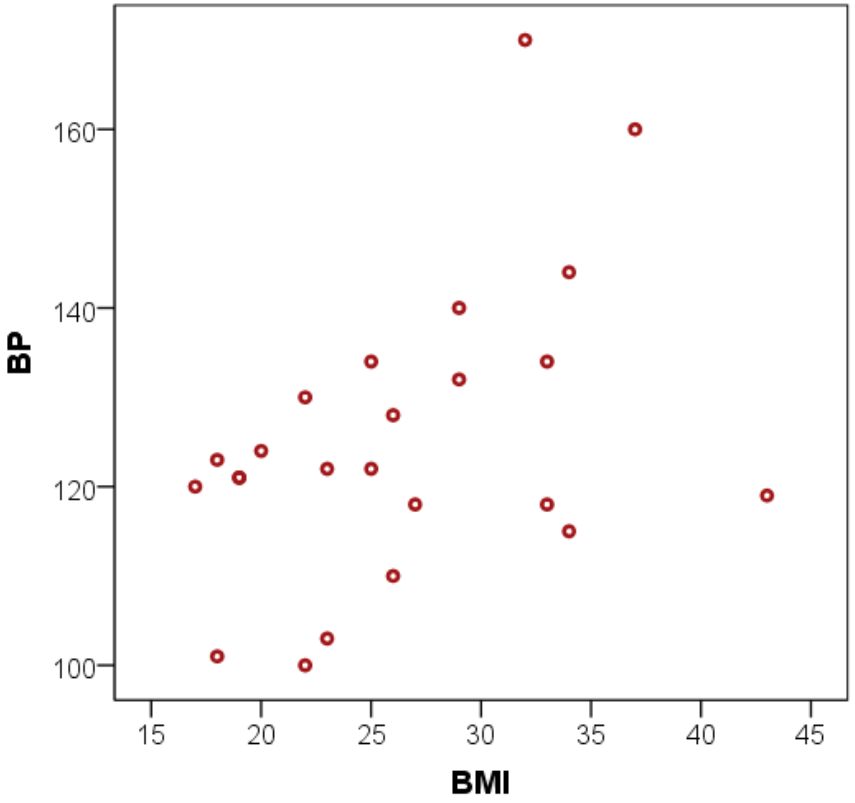
Is BMI (x) predictive for Blood Pressure (y) ?

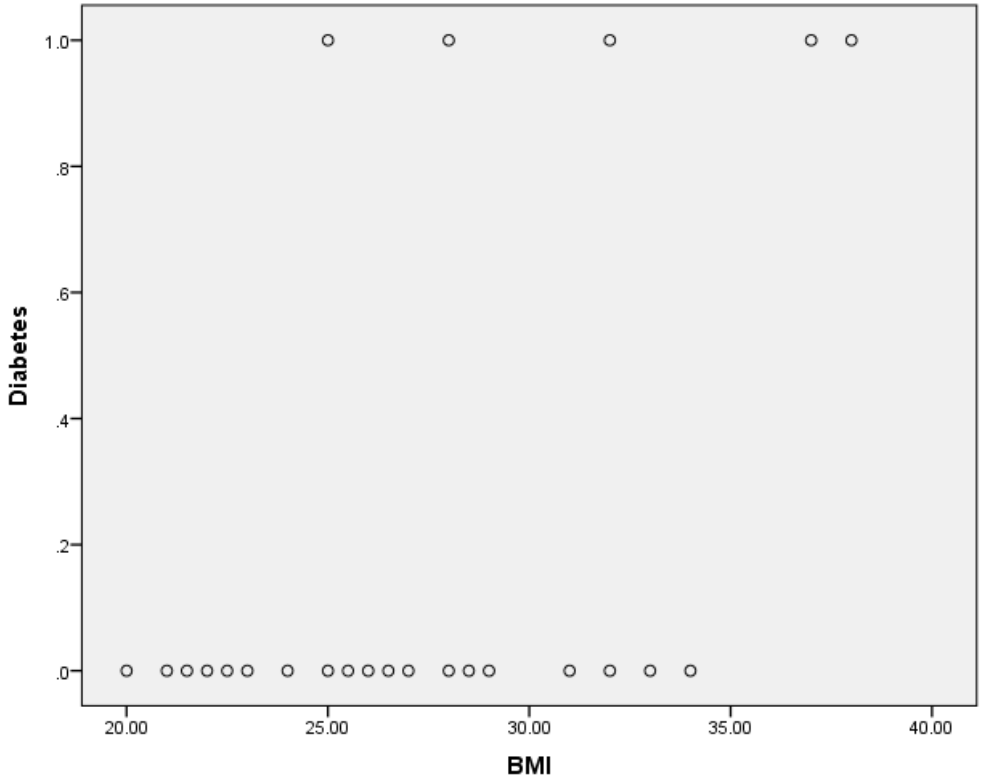
Linear Regression Model

Is BMI (x) predictive for Diabetes (y) ?

Linear Regression Model not appropriate

Scatter plots

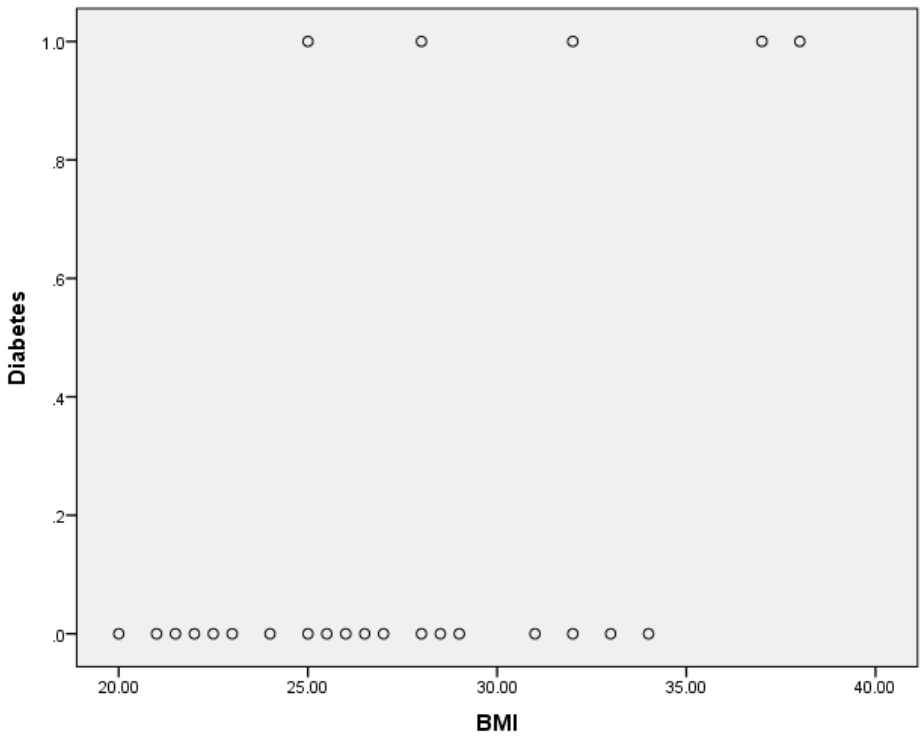


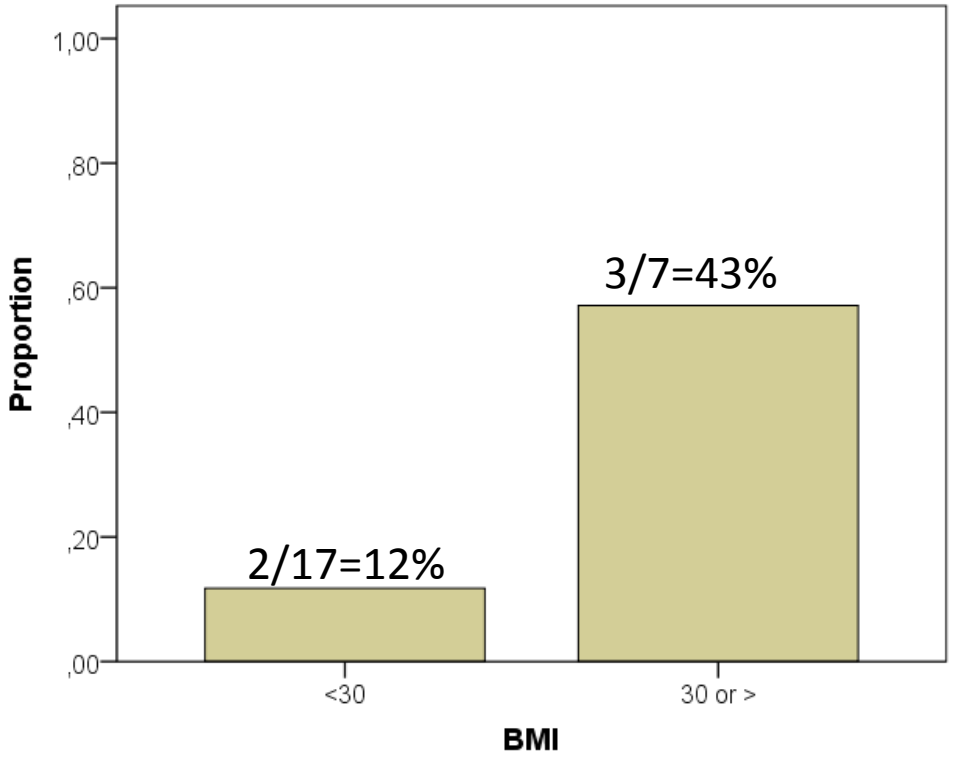


Clear pattern.
We can fit a regression line.

There is a pattern, but this is far from optimal for a linear regression model.

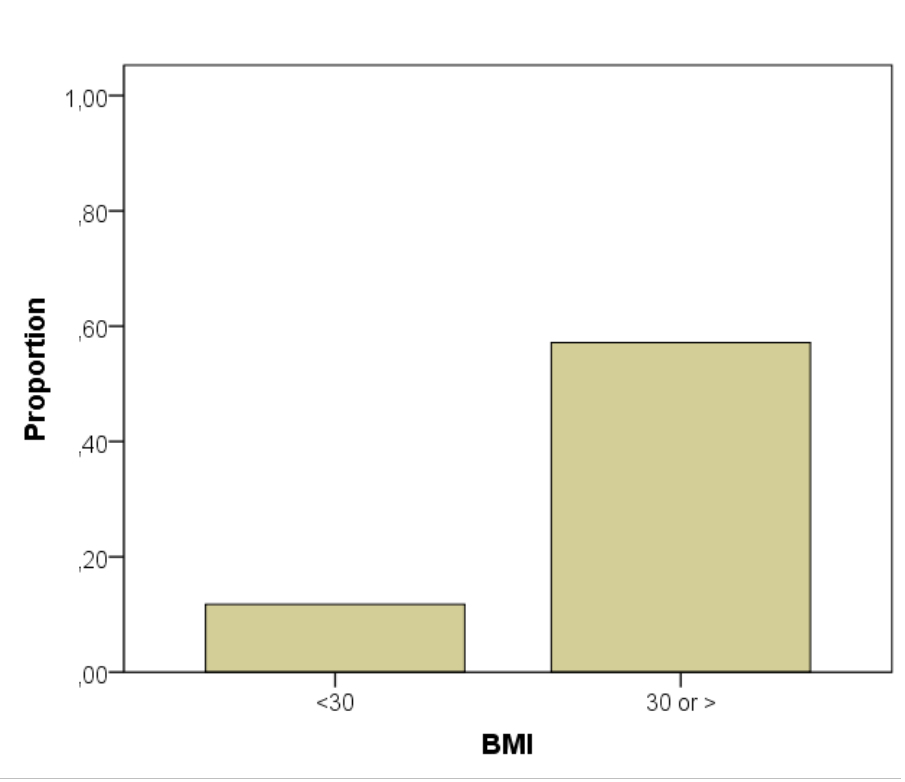
Relationship BMI - Diabetes

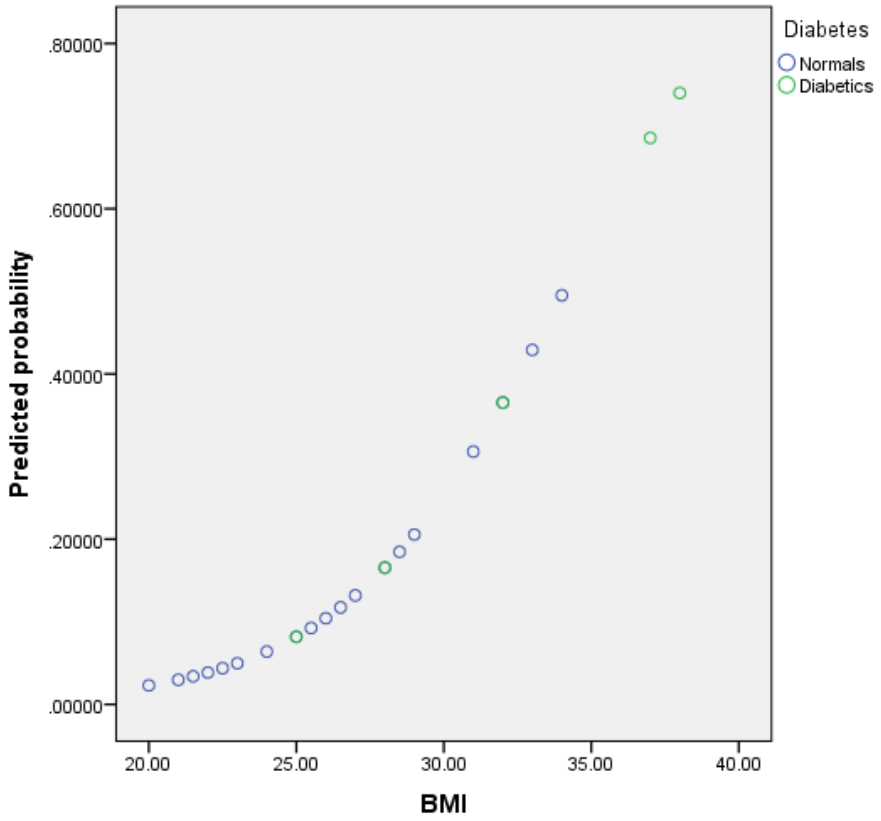




We can make the relationship visible in a simple way, by binning the BMI (e.g. <30, ≥30) and calculate the proportion of diabetes cases within each bin.

Relationship: Logistic regression





The logistic regression model fits the data into a probability (DIAB=1) as a function of BMI.

Linear vs logistic regression

Linear regression:

Outcome Y is a continuous (dependent) variable which we try to predict/explain by an independent variable(s) X.

Binary logistic regression:

Outcome Y is a binary (0,1) (dependent) variable which we try to predict/explain by an independent variable(s) X.

In order to fit this relationship into the regression framework, we need a transformation (the link function).

Logit

In logistic regression, we model the log odds

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots$$

ln odds(p), log odds(p), logit(p)

p	= proportion/probability of Y=1
1 – p	= proportion/probability of Y=0

ln = natural logarithm, e = 2.7183

Odds

Previous data:

Disease/Exposure	BMI ≥30		BMI <30	
Diabetes +	A	n=3	B	n=2
Diabetes -	C	n=4	D	n=15

exposed = BMI ≥**30,** unexposed = BMI <30

Odds (p) = p/(1-p) = A/C for exposed = $\frac{3/7}{1-3/7}$ = ¾ = 0.75

= B/D for unexposed = $\frac{2/17}{1-2/17}$ = $2/15$ = 0.13

Odds ratio = the ratio of odds exposed vs unexposed

= $\frac{A*D}{B*C}$ = $\frac{0.75}{0.13}$ = 5.77

Ratios

$$\text{Odds Ratio (OR)} = \frac{\text{Odds}_{\text{exposed}}}{\text{Odds}_{\text{unexposed}}}$$

OR **= 1** similar odds in both groups

OR **> 1** odds higher in the exposed group

OR **< 1** odds lower in the exposed group

OR ≈ relative risk if outcome is rare (rare disease assumption)

Logistic regression

In logistic regression, we model the 'log odds':

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots$$

ln odds(p), log odds(p), logit(p)

p	= proportion/probability of Y=1
1 – p	= proportion/probability of Y=0

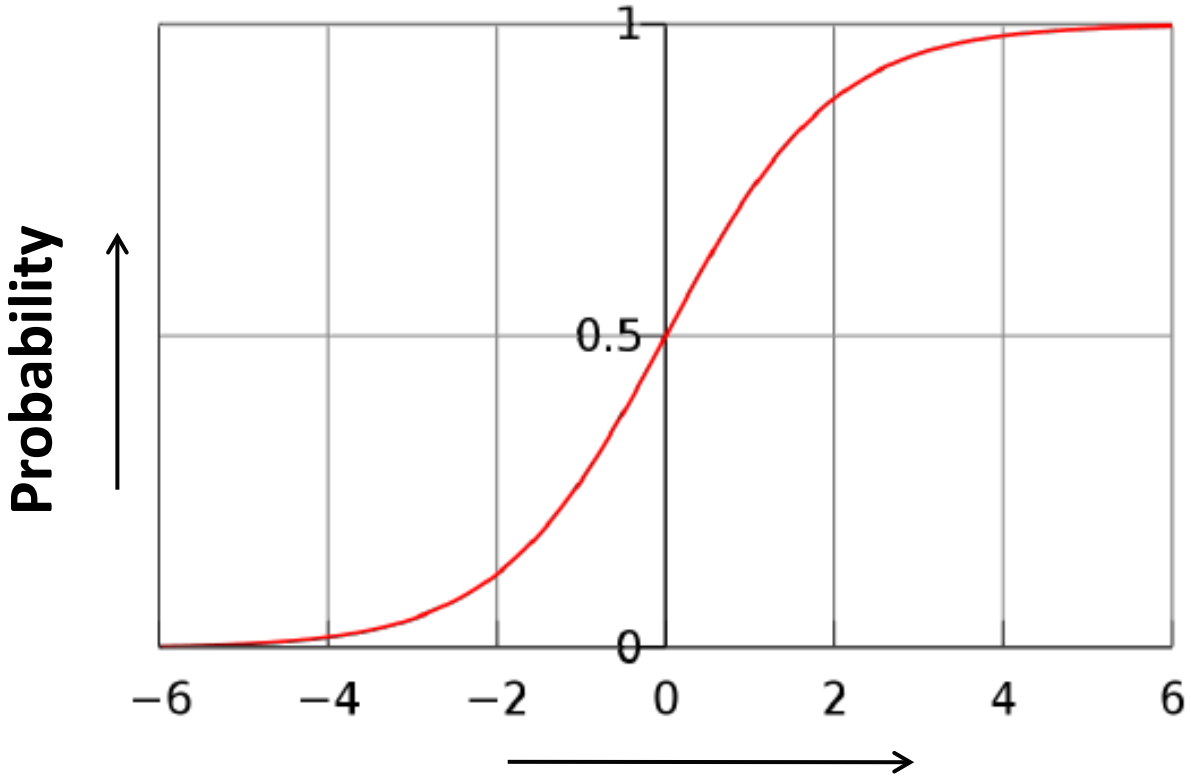
$$\text{Probability}(Y=1 \mid X_1, X_2, \ldots) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots)}$$

ln = natural logarithm, e = 2.7183

Relationship: Logit - Probability

The value of the logit is not restricted.

Probability is restricted (between 0 – 1).



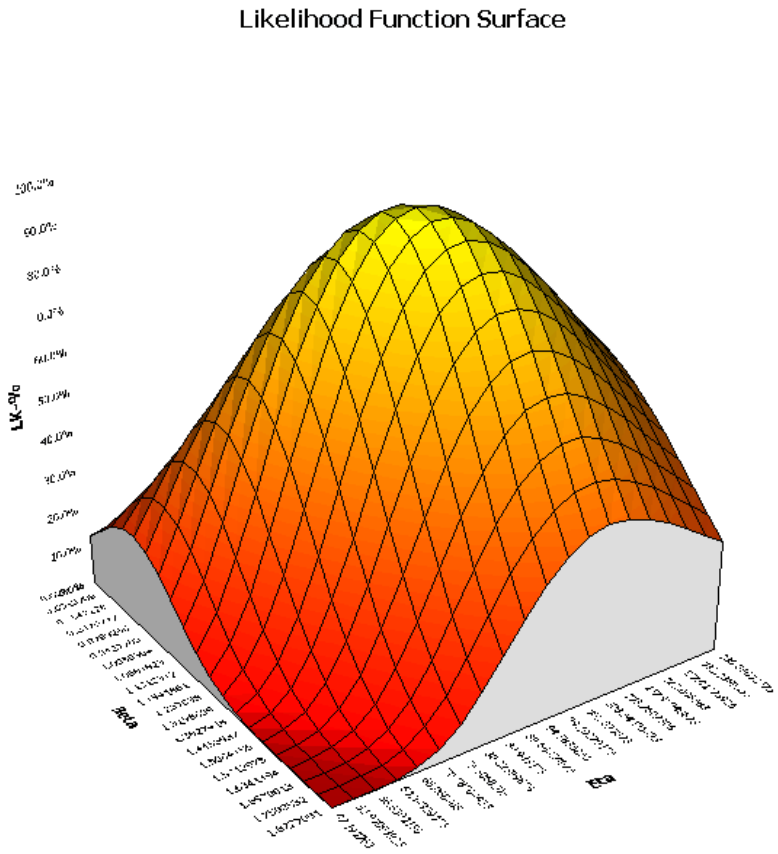
Maximum likelihood method

The maximum likelihood method is a method to estimate model parameters (i.e., regression coefficients and their standard errors), i.e., an alternative to least squares.

Maximum likelihood estimates of regression coefficients maximize the likelihood for the data: the values for which the observed data are most likely.

Likelihood: product of all probabilities over all individuals (in-dependence)

The procedure for fitting a model involves iterative optimization.



Regression coefficients

Linear Regression

BMI as predictor for Blood Pressure: $\mathbf{Y = \beta_0 + \beta_1 * BMI}$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	97,092	12,570		7,724	,000
	BMI	1,071	,461	,444	2,322	,030

a. Dependent Variable: BP

Logistic Regression

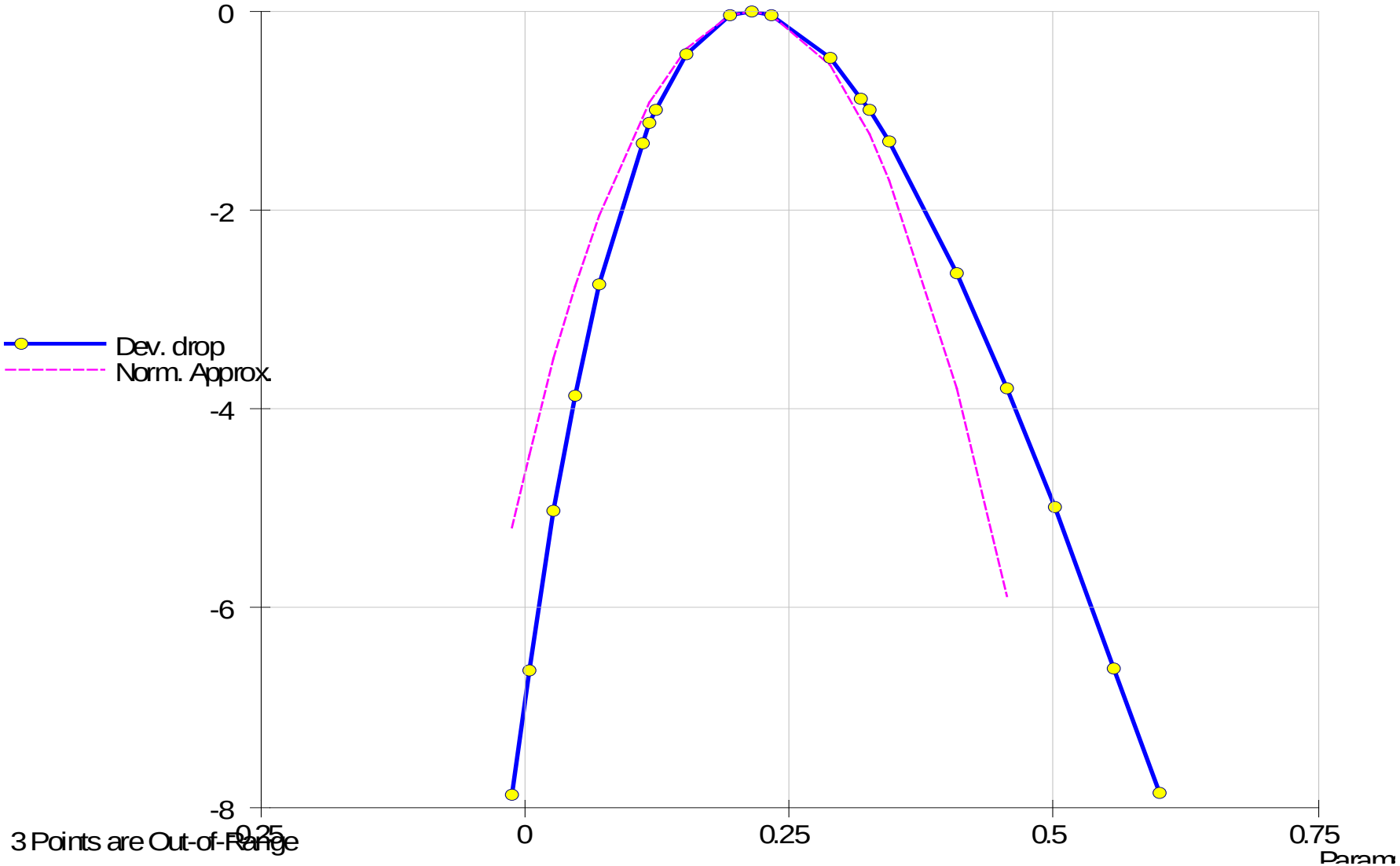
BMI as predictor for Diabetes: $Prob(Y=1) = \dfrac{\exp(\beta_0 + \beta_1 * BMI)}{1+\exp(\beta_0 + \beta_1 * BMI)}$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	BMI	,214	,100	4,624	1	,032	1,239
	Constant	-7,163	3,000	5,702	1	,017	,001

a. Variable(s) entered on step 1: BMI.

Profile likelihood



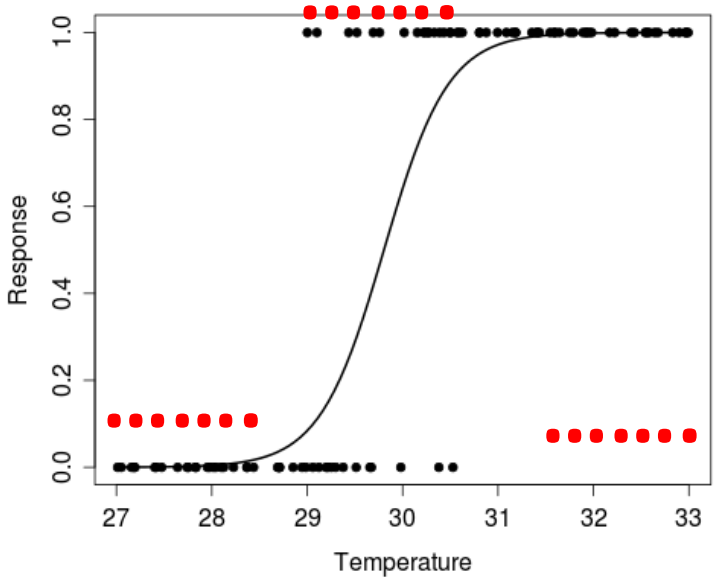
3 Points are Out-of-Range

Goodness of fit evaluation

A Goodness of Fit test measures the discrepancy between ***observed*** values and the values ***expected*** under the model.

- **(Pseudo) R square**: Explained variance. Based on distances between predicted values and outcome (0/1) values.
- **Hosmer & Lemeshow**: Is the model appropriate? Evaluates whether the data show strong deviations from the chosen model.

In case of the red values: Hosmer & Lemeshow test will detect that the model is not appropriate.



Hosmer & Lemeshow test

Test evaluates goodness-of-fit by comparing the observed and the predicted number of cases in deciles of fitted risk (predicted probability).

Test statistic has chi-square distribution.

Null hypothesis: Predicted and observed are the same, i.e., the model is appropriate.

P<0.05: Model prediction differs significantly from observed data, i.e., model is ***not*** appropriate for the data.

Hosmer & Lemeshow test not recommended for small sample sizes

Goodness of fit output

Example output with BMI – Diabetes data

Contingency Table for Hosmer and Lemeshow Test

		Diabetes = 0		Diabetes = yes		
		Observed	Expected	Observed	Expected	Total
Step 1	1	3	2,900	0	,100	3
	2	2	1,913	0	,087	2
	3	3	2,787	0	,213	3
	4	1	1,806	1	,194	2
	5	2	1,716	0	,284	2
	6	1	1,660	1	,340	2
	7	3	2,237	0	,763	3
	8	2	1,617	1	1,383	3
	9	1	,935	1	1,065	2
	10	0	,429	2	1,571	2

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7,777	8	,456

SPSS: Request test under "Options"

(Pseudo) R square

There are several “pseudo R square methods” to assess the explained variance in the model.

Model performance is estimated by measuring the distances between predicted and actual outcome. Higher values of R Square indicate better fit.

Example of SPSS output with Diabetes – BMI Data:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	20,084[a]	,250	,370

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

SPSS: In default output

Variable types

Logistic regression accommodates continuous and categorical covariates.

Continuous: One regression coefficient is estimated. OR is ratio of odds for value x and odds for value x-1. Assumption: Same across values of x. OR=change in risk per one unit change in exposure.

Categorical: Binary indicators for each category are created (dummy variables). One regression coefficient is estimated for each category except reference (has to be chosen, *in SPSS lowest or highest value are possible).*

Binary variables and ordinal variables: Should be handled as categorical variables *(Rule of thumb: ordinal variables with >7 equidistant levels can be handled as continuous*).

In case of (too) many categories, or categories with too few observations (<5-10 cases), categories can be merged.

Example: Dummy variables

Categorical variables are represented by dummy variables. These are automatically created by SPSS. The coding is in the output.

Example: Dummies for tumor stages in study of recurrence of lung cancer

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)	(4)	(5)
tumorstage	1A	16	.000	.000	.000	.000	.000
	1B	18	1.000	.000	.000	.000	.000
	2A	4	.000	1.000	.000	.000	.000
	2B	8	.000	.000	1.000	.000	.000
	3A	27	.000	.000	.000	1.000	.000
	3B	15	.000	.000	.000	.000	1.000

Merging of categories would be better.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	tumorstage			3.584	5	.611	
	tumorstage(1)	.990	.921	1.156	1	.282	2.692
	tumorstage(2)	.847	1.380	.377	1	.539	2.333
	tumorstage(3)	1.435	1.051	1.864	1	.172	4.200
	tumorstage(4)	1.415	.855	2.743	1	.098	4.118
	tumorstage(5)	1.540	.922	2.794	1	.095	4.667
	Constant	-1.946	.756	6.626	1	.010	.143

a. Variable(s) entered on step 1: tumorstage.

Models with >1 covariate

- Purpose of model with >1 covariates:
  - Unbiased estimate of β for the covariate of interest (adjustment for confounding bias)
  - Multivariate model with all relevant predictors

- No. of events is limiting the maximum number of covariates you can include in the model. Rule of thumb: at least **5-10 cases** per parameter (*rough indication below which problems can occur*)

- Stepwise method (forward conditional in SPSS): variables are selected in the order in which they maximize the contribution to the model.

Nested models

Does the goodness of fit of a model significantly improve when covariates are added?

Likelihood ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_1, with the additional covariate) over the maximized value of the likelihood function for the simpler model (L_0, without the additional covariate).

$$\text{Likelihood ratio} = -2 \log (L_1 / L_0)$$

$$-2 (\log L_1 - \log L_0) \sim \chi^2(df)$$

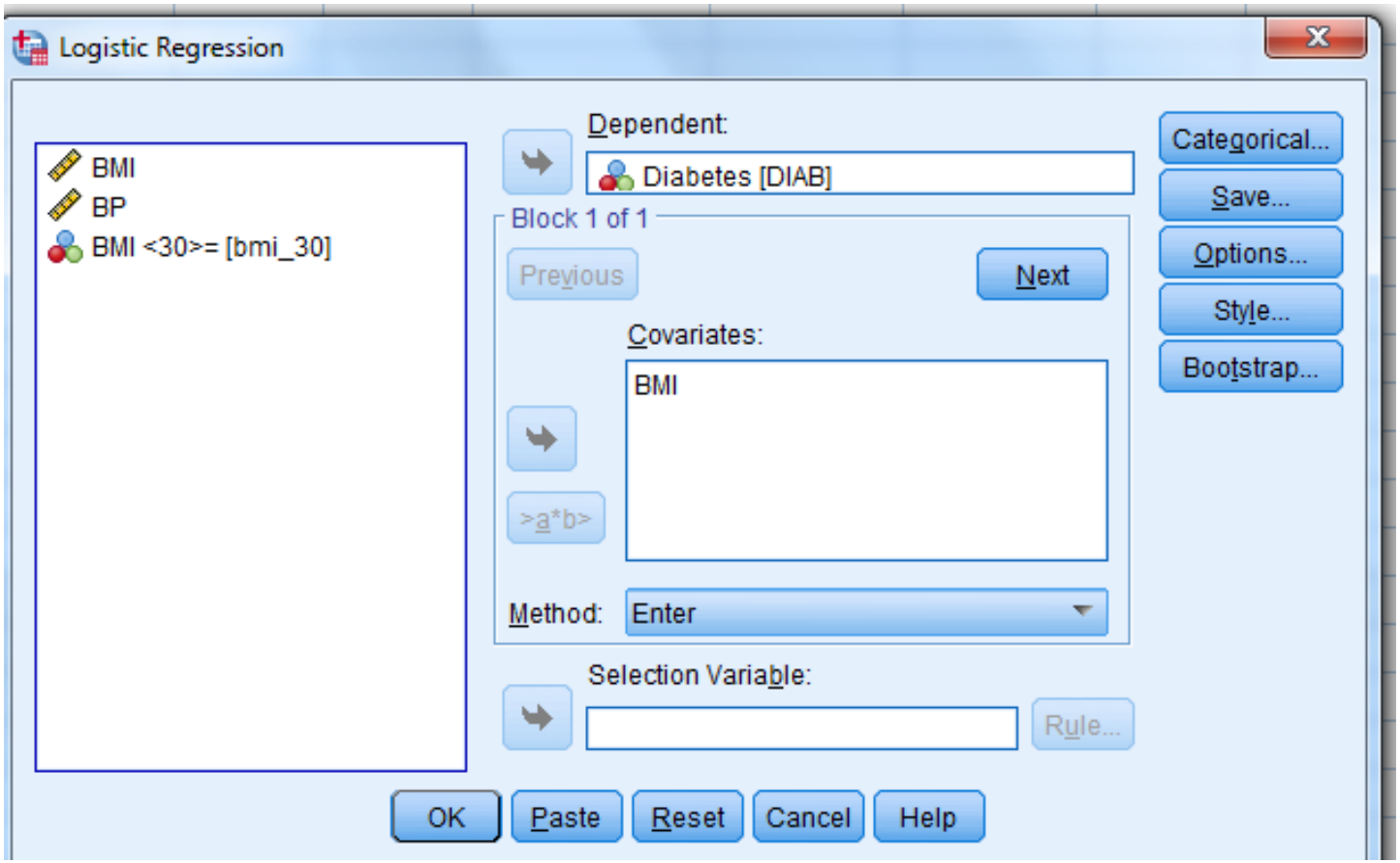
with df = # parameters model 1 - # parameters model 0

Overview linear vs logistic regression

	Linear	Logistic
Outcome Y	continuous & normal	binary
Outliers in Y problematic	yes	no
Outliers in X problematic	yes	yes
Multi-collinearity in X's problematic	yes	yes
Creating dummy variables needed	yes	no (SPSS)
Extrapolation beyond data problematic	yes	yes

SPSS

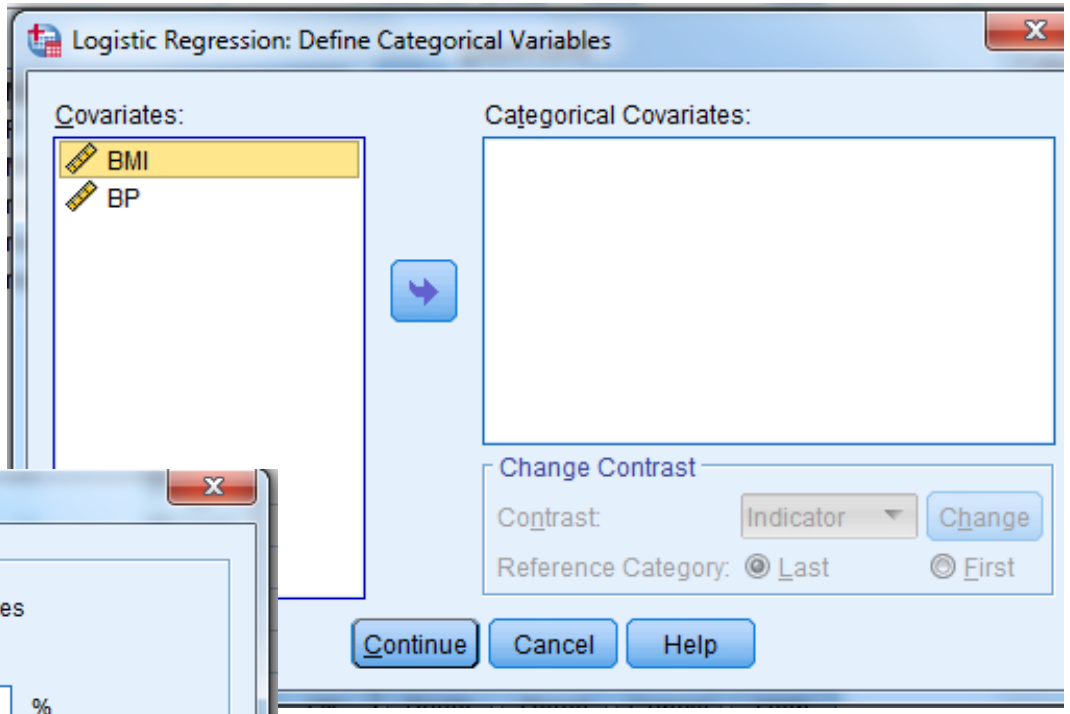
Analyze - Regression – Binary logistic

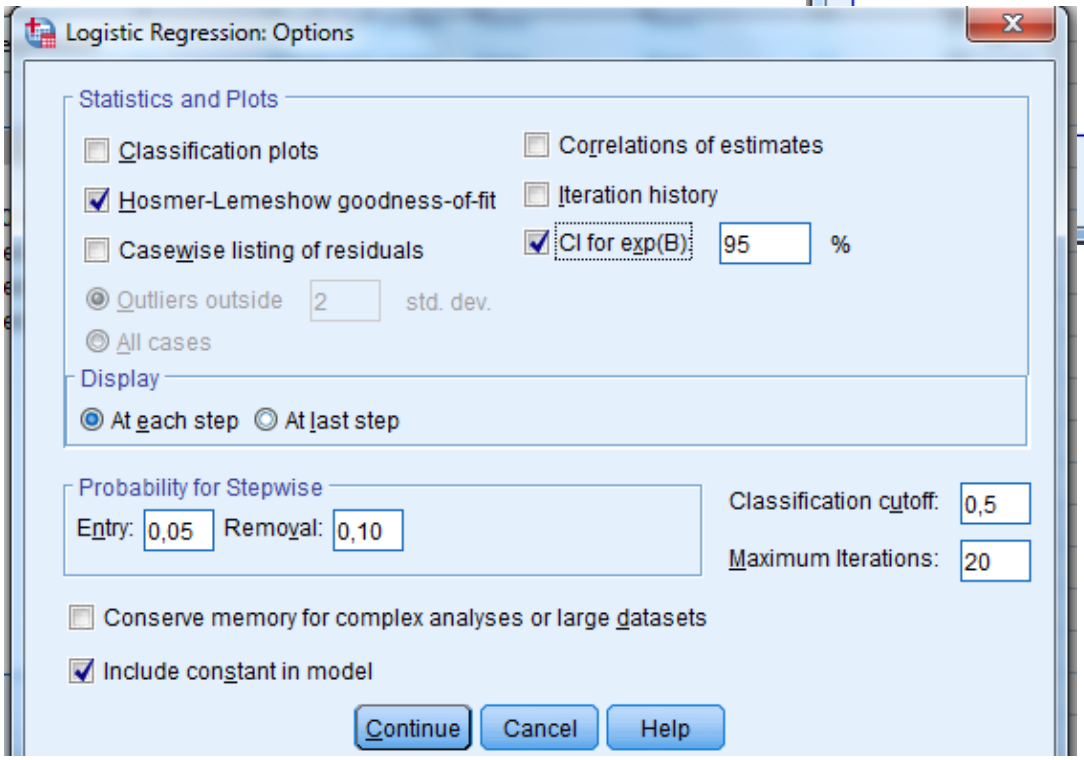


Univariate analysis: 1 covariate in the model

SPSS

Sub-menus





SPSS output

SPSS runs the logistic regression in two steps:

Block 0: Beginning Block
No predictors are included, only the constant (“intercept”)

It includes a table “Variables not in the Equation”, where it is predicted whether an independent predictive variable that is not included yet, would be significant in the model.

SPSS output

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			Diabetes		Percentage Correct
			0	yes	
Step 0	Diabetes	0	18	0	100,0
		yes	6	0	,0
	Overall Percentage				75,0

a. Constant is included in the model.

b. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1,099	,471	5,431	1	,020	,333

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	BMI	6,568	1	,010
	Overall Statistics		6,568	1	,010

SPSS output

Block 1: Method=Enter (default)

Interesting part of the output with results for covariate(s) of interest

Includes (by default) table “Variables in the equation” with estimates for the constant and β

SPSS output

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	6,908	1	,009
	Block	6,908	1	,009
	Model	6,908	1	,009

→ Comparison between model with covariate(s) and model with only constant (**likelihood ratio test**), with H0: "no improvement by adding covariate" (which is rejected).

If stepwise regression is requested, rows would compare newest model with previous one.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	20,084[a]	,250	,370

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	BMI	,214	,100	4,624	1	,032	1,239
	Constant	-7,163	3,000	5,702	1	,017	,001

a. Variable(s) entered on step 1: BMI.

Wald Statistic: test the statistical significance of *each* coefficient (B) in the model, based on Z statistic.

Interpretation

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	BMI	,214	,100	4,624	1	,032	1,239
	Constant	-7,163	3,000	5,702	1	,017	,001

a. Variable(s) entered on step 1: BMI.

'Sig.' indicates the p-value of the Wald statistic: the null hypothesis of 'B=0' is rejected since $p < 0.05$.

Exp(B) is the **Odds Ratio** for a unit increase.

A value of 1.24 implies a relative increase of the odds of +24% per unit increase BMI.

Here, 1 unit = 1 kg/cm^2 (BMI).

- *Increase of BMI from 22 to 23: OR=1.24*
- *Increase of BMI from 33 to 34: OR=1.24 (assumption model)*

Numerical problems

Issues that can affect the accuracy of the estimation of B:

- Multi-collinearity among the independent variables.
- ‘Complete separation' whereby the two outcome groups are perfectly separated by one of the independent variables.
- Zero cells for a dummy-coded independent variable because all of the subjects have the same value for the variable.

Output that indicates numerical problems should not be interpreted.

Multi-collinearity

2 covariates with high correlation (Pearson correlation = 0.99).
We put them separately and together in the Logistic Regression Model.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	vol30gy	.056	.012	22.530	1	.000	1.058
	Constant	-3.449	.620	30.921	1	.000	.032

a. Variable(s) entered on step 1: vol30gy.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	vol30gy2	.056	.012	22.573	1	.000	1.058
	Constant	-3.508	.631	30.874	1	.000	.030

a. Variable(s) entered on step 1: vol30gy2.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	vol30gy	.182	1.056	.030	1	.863	1.200
	vol30gy2	-.126	1.057	.014	1	.905	.881
	Constant	-3.315	1.269	6.826	1	.009	.036

a. Variable(s) entered on step 1: vol30gy, vol30gy2.

We can detect the problem by examining the errors (S.E.): it has become much larger.

vol30gy: volume (%) of esophagus with ≥ 30 Gy

Separation

Covariate perfectly predicts outcome: no results in SPSS

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.688	1.000

a. Estimation terminated at iteration number 18 because a perfect fit is detected. This solution is not unique.

Covariate almost perfectly predicts outcome

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	vol30gy	.052	.045	1.344	1	.246	1.054
	factor_perfect	24.876	6206.669	.000	1	.997	6.363E+10
	Constant	-6.313	2.507	6.342	1	.012	.002

a. Variable(s) entered on step 1: vol30gy, factor_perfect.

We can detect the problem by examining the error (S.E.): it is large (as well as the Exp(B)).

Category with 1 value

Covariate has a category with only non-cases

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	vol30gy	.055	.014	16.731	1	.000	1.057
	cat_1value			.459	4	.977	
	cat_1value(1)	-.593	1.011	.344	1	.558	.553
	cat_1value(2)	.029	.566	.003	1	.959	1.030
	cat_1value(3)	-.018	.750	.001	1	.981	.982
	cat_1value(4)	-18.111	17731.913	.000	1	.999	.000
	Constant	-3.371	.664	25.753	1	.000	.034

a. Variable(s) entered on step 1: vol30gy, cat_1value.

$e^{-18} = 1.5*10^{-8} \approx 0$

We can detect the problem by examining the error (S.E.): it is large.

Example

Esophagitis (inflammation of esophagus) among lung cancer patients who received radiotherapy and chemotherapy.

Research Questions

- Is development of esophagitis associated with dose to the esophagus?
- Is esophagitis associated with chemotherapy?
- Does the model significantly improve when we add chemotherapy to a model with the dose variable 'Vol30Gy'?
  Is the Goodness-of-Fit acceptable?

! SPSS will automatically predict the highest value of the binary outcome variable. So we have to code for example 0=no tox, 1=tox.

Example: results I

Results (univariate) for 2 dose parameters

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	vol30gy	,056	,012	22,530	1	,000	1,058
	Constant	-3,449	,620	30,921	1	,000	,032

a. Variable(s) entered on step 1: vol30gy.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	vol60gy	,048	,012	16,298	1	,000	1,049
	Constant	-1,954	,327	35,725	1	,000	,142

a. Variable(s) entered on step 1: vol60gy.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	146,714[a]	,201	,292

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

V30

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	163,735[a]	,109	,158

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

V60

Example: results II

Results (univariate) for concurrent chemotherapy (0=no, 1=yes)
First table: reference category = first category
Second table: last category as reference

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	conc(1)	1.646	.406	16.465	1	.000	5.187
	Constant	-1.484	.236	39.475	1	.000	.227

a. Variable(s) entered on step 1: conc.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	conc(1)	-1,646	,406	16,465	1	,000	,193
	Constant	,163	,330	,243	1	,622	1,176

a. Variable(s) entered on step 1: conc.

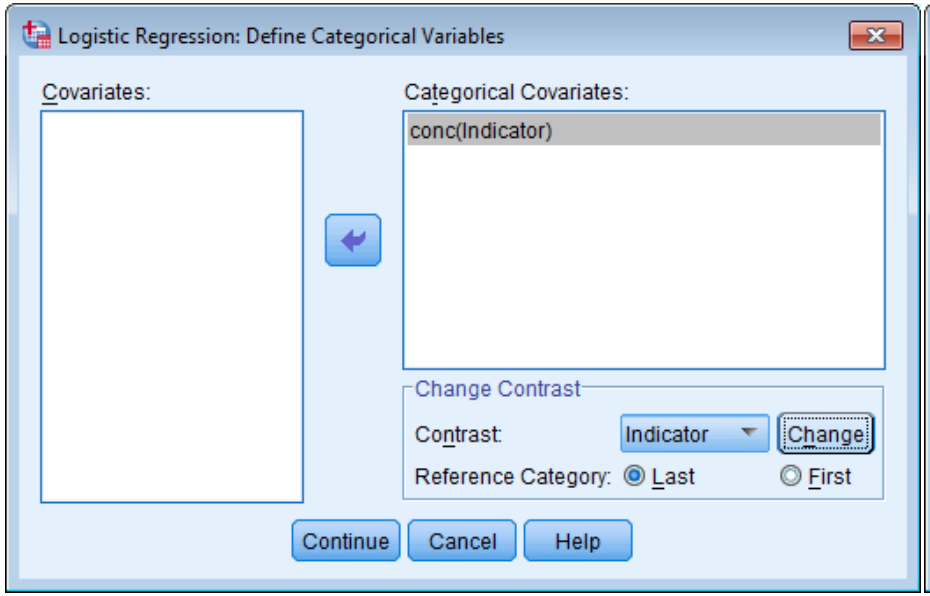
Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	164,981[a]	,102	,148

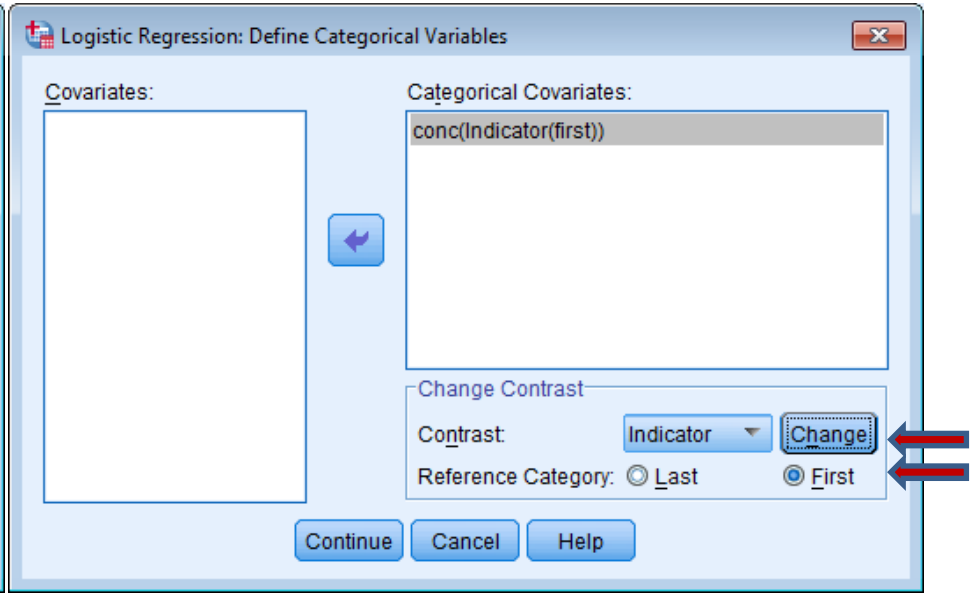
a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

Choice of reference

SPSS default: reference = last category



Reference = first: click first & change



Nested models

Results (nested model) for Vol30Gy and concurrent chemotherapy
Stepwise model (method = Forward Conditional)

The minimum ratio of valid cases to independent variables for stepwise logistic regression is 10 to 1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	35.024	1	.000
	Block	35.024	1	.000
	Model	35.024	1	.000
Step 2	Step	4.678	1	.031
	Block	39.702	2	.000
	Model	39.702	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	146,714[a]	,201	,292
2	142,036[b]	,225	,327

Likelihood ratio test: model significantly improves

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	vol30gy	,056	,012	22,530	1	,000	1,058
	Constant	-3,449	,620	30,921	1	,000	,032
Step 2[b]	conc	,942	,436	4,671	1	,031	2,565
	vol30gy	,051	,012	16,771	1	,000	1,052
	Constant	-3,509	,642	29,858	1	,000	,030

Goodness of fit test

Goodness of Fit (comparing observed and expected no. of cases)

Hosmer and Lemeshow Test

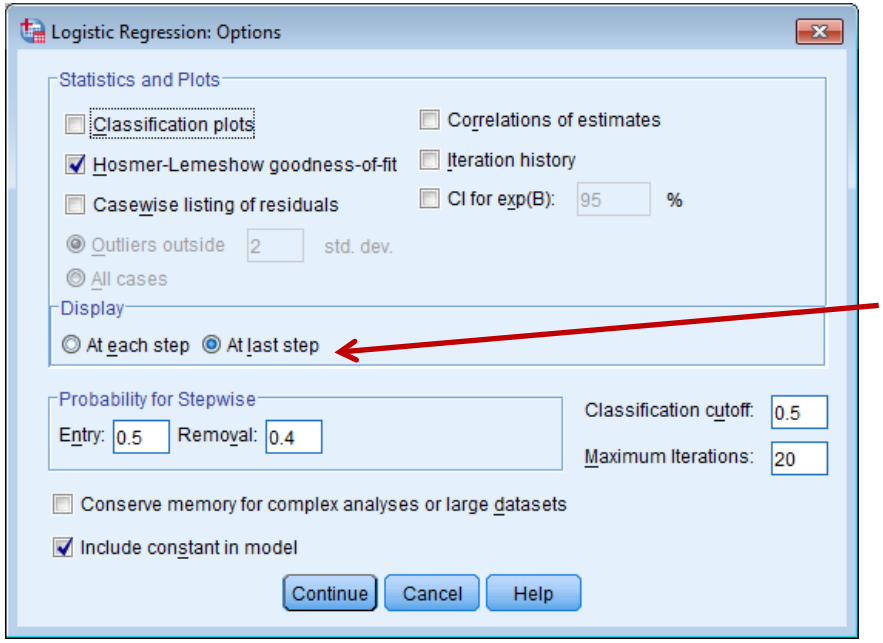
Step	Chi-square	df	Sig.
1	4.419	7	.730

Non-significance indicates that the model is appropriate

Contingency Table for Hosmer and Lemeshow Test

		G0-1 vs G2+ = G0-1		G0-1 vs G2+ = G2+		
		Observed	Expected	Observed	Expected	Total
Step 1	1	24	23.303	0	.697	24
	2	16	15.428	0	.572	16
	3	15	14.651	1	1.349	16
	4	11	13.348	5	2.652	16
	5	12	12.137	4	3.863	16
	6	10	10.542	6	5.458	16
	7	10	9.397	6	6.603	16
	8	8	8.363	8	7.637	16
	9	8	6.832	12	13.168	20

Summary table



You can obtain a summary of all steps by choosing "at last step"

Step Summary[a,b]

Step	Improvement			Model			Correct Class %	Variable
	Chi-square	df	Sig.	Chi-square	df	Sig.		
1	35.024	1	.000	35.024	1	.000	71.2%	IN: vol30gy
2	4.678	1	.031	39.702	2	.000	76.3%	IN: conc

a. No more variables can be deleted from or added to the current model.

b. End block: 1

In the Step Summary table we see which variable was added or removed at each step.

Research questions

Research Questions

- *Is development of esophagitis associated with dose to the esophagus?*

  Yes, we found a statistically sign. relationship with dose.

  → the model significantly improved by adding a dose variable,

  → the estimated β was significant,

  → there were no signs of numerical problems.

- *Is this toxicity associated with chemotherapy*? **Yes**

- *Does the model significantly improve when we add chemotherapy to a model with Vol30Gy*?

  Yes, the likelihood ratio test of the extended model compared to the model with 1 covariate indicates that the model improves significantly. The Goodness of fit was acceptable.

Final Remarks

Logistic Regression is a tool to analyze the effect of covariates on a binary outcome.

In Logistic Regression, we assume that **follow-up time** is constant or not an issue for the studied outcome. It is therefore often used in cross-sectional studies. To incorporate follow-up time, use Cox Regression (*lectures S8 and S9*).

We have looked at ***Unconditional Binary*** Logistic Regression.

Other types of Logistic Regression:

- Multinomial & ordinal Logistic Regression
- Conditional (binary) Logistic Regression