Contents

1 Model Building and Application in Logistic Regression 2

1.1 Strategies in Model Selection 2

1.1.1 How many predictors? 2

1.1.2 Stepwise variable selection algorithms 4

1.1.3 AIC 5

1.1.4 Classification Tables 6

1.1.5 Correlation 7

1.2 Checking the Model 8

1 Model Building and Application in Logistic Regression

Exploratory versus Confirmatory studies

In confirmatory studies an existing or proposed model shall be confirmed. We do this by fitting the model to a random sample and measuring model fit, checking for interpretation of the parameter estimates etc..
In exploratory studies the intent is to determine a model that explains the observed data the best. Once such a model has been chosen a confirmatory study should be conducted to validate the model. In the exploratory process many different models are considered and should be compared with each other. The final model should be most parsimonious model, which describes the data well.

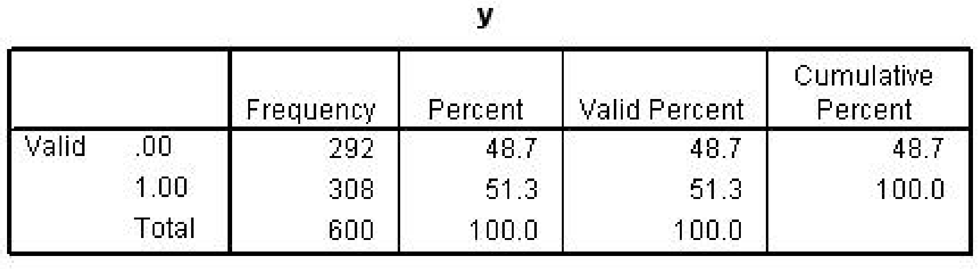
1.1 Strategies in Model Selection

1.1.1 How many predictors?

Guideline: The possible number of predictors to be included with the final model is limited by the sample size. For each parameter in the final model the sample should include at least 10 records of each outcome of the response.

For example, consider a large sample with $n = 500$, including 250 successes and 250 failures, then the final model can include up to 250/10=25 predictors.
But if the sample includes only 50 successes and 450 failures, the final model should not include more than 50/10=5 predictors.

This is a guideline, but the estimates can be very biased if models with too many predictors are fit to data.



y

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	292	48.7	48.7	48.7
	1.00	308	51.3	51.3	100.0
	Total	600	100.0	100.0	

For the HSB model we could use up to 29 parameters according to this guideline, we are lucky because the data is balanced.

Another problem caused by the inclusion of many predictors is multicollinearity.
Which means that some predictors could be linearly dependent (or close to linearly dependent) and therefore the estimates become imprecise because the $SE(\hat{\beta})$ get very large.

For example, if we have two predictors x_1, x_2, where $x_1 = a + bx_2$ (e.g. x_1 is the temperature in centigrade and x_2 is the temperature in Fahrenheit) then because of this linear relationship one of the two variables is redundant. Now assume that the relationship is not perfect, but really close and that maybe there are just some deviations due to lack of precision in the measurements. Then these deviations determine very strongly the estimates for the model parameters and their standard errors. Standard errors are very large in the presence of collinearity, resulting in tests for effect of explanatory variables on the response variable not being significant even when the model utility test ($H_0 : \beta_1 = \cdots = \beta_k = 0$) is significant.

In the presence of many predictors it can happen that one predictor is a linear combination of several other predictors.
$x_1 = c_1x_2 + c_2x_3 + \cdots + c_lx_l$.
This would have the same effect n the estimates and their standard errors as mentioned above.

We call the occurrence of such relationships multicollinearity and the more predictors included with a model the more likely multicollinearity will occur.

To detect predictors that are involved in multicollinearity we should investigate the **variance inflating factor**s (VIFs) for the all predictors. They measure how much larger the standard errors will be in the model with multiple predictors versus the model only including the one predictor and are calculated for the k^{th} predictor

$$VIF_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the coefficient of determination for the regression of all the other predictors on the k^{th} predictor.
VIFs larger 2.5 are suspicious, unless one includes transformations like $x, x^2, \ln(x)$ etc. in the model. It should be noted that VIFs are only based on the predictors and do not consider the response variable.

Example 1
HSB
The variable X was calculated as the total of science writing, civic, reading, and math, therefore
$x =$ math + science + writing + civic + reading
The six variables are collinear and including them all in one model would make the parameter estimates undefined

Principles in model selection:
When choosing predictors for the final model among the variables observed the following principles apply.

1. Predictors deemed essential for the research question or the model should always be included, regardless of them being "significant" or not. Treatment variables are for example always included.

2. When including interaction between predictors with the model the lower order terms for those predictors should be included with the model as well. This is necessary because otherwise the measurement scale of the predictors has an effect on the results.

1.1.2 Stepwise variable selection algorithms

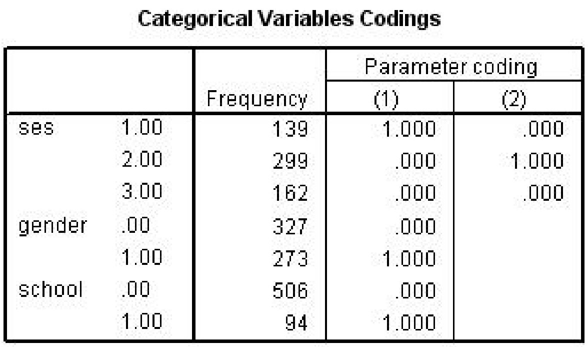
These algorithms can be helpful in building a model but have to be used with caution and the final model has always to be reviewed by the researcher.

In the *backward* variable selection algorithm the process starts with all variables included in the model. Then the model is improved step by step by dropping one variable from the model at each step. The variable dropped is the one demonstrating the "least significant" effect on the response when correcting for the other variables in the model.
The process stops if an additional step does not show a further improvement of the model.

In the *forward* variable selection algorithm the process starts with the intercept-only model. Then the model is improved step by step by adding an extra variable to the model at each step. The variable added is the one demonstrating the "most significant" effect on the response when correcting for the other variables in the model.
The process stops if an additional step does not show a further improvement of the model.

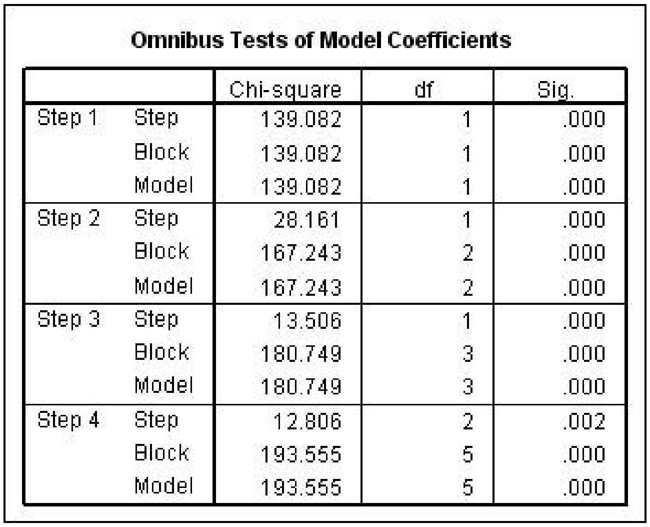
When using these algorithms one should make sure that if an interaction term is included with a model automatically all main effect term have to be included, too.

Example 2
HSB with the following predictors:
SES
school type
gender
x, academic success
reading
writing
math
science
civics



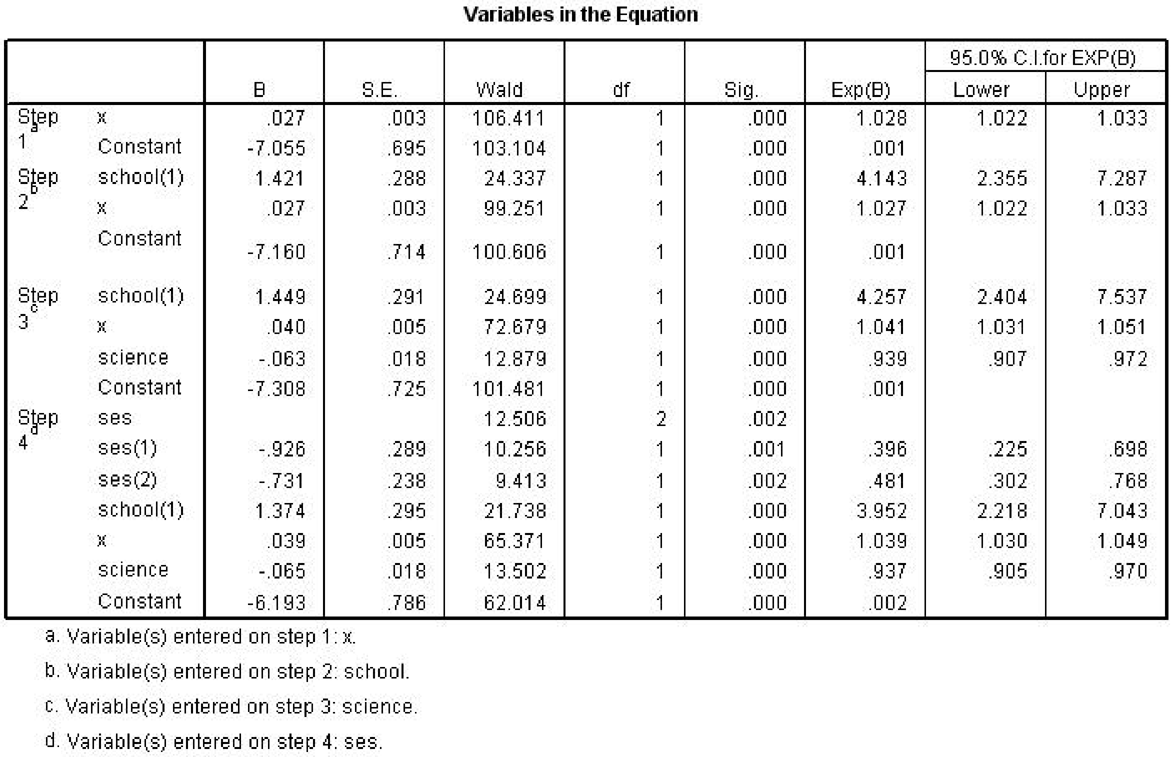
Categorical Variables Codings

		Frequency	Parameter coding	
			(1)	(2)
ses	1.00	139	1.000	.000
	2.00	299	.000	1.000
	3.00	162	.000	.000
gender	.00	327	.000	
	1.00	273	1.000	
school	.00	506	.000	
	1.00	94	1.000	



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	139.082	1	.000
	Block	139.082	1	.000
	Model	139.082	1	.000
Step 2	Step	28.161	1	.000
	Block	167.243	2	.000
	Model	167.243	2	.000
Step 3	Step	13.506	1	.000
	Block	180.749	3	.000
	Model	180.749	3	.000
Step 4	Step	12.806	2	.002
	Block	193.555	5	.000
	Model	193.555	5	.000



Variables in the Equation

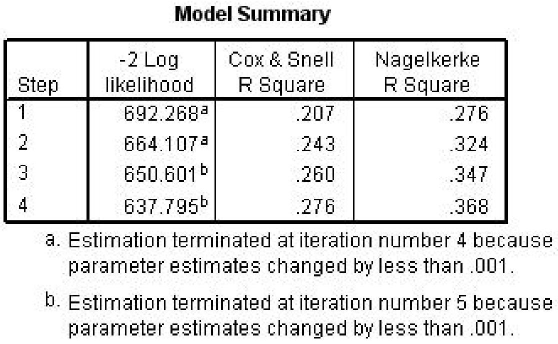
		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	x	.027	.003	106.411	1	.000	1.028	1.022	1.033
	Constant	-7.055	.695	103.104	1	.000	.001		
Step 2[b]	school(1)	1.421	.288	24.337	1	.000	4.143	2.355	7.287
	x	.027	.003	99.251	1	.000	1.027	1.022	1.033
	Constant	-7.160	.714	100.606	1	.000	.001		
Step 3[c]	school(1)	1.449	.291	24.699	1	.000	4.257	2.404	7.537
	x	.040	.005	72.679	1	.000	1.041	1.031	1.051
	science	-.063	.018	12.879	1	.000	.939	.907	.972
	Constant	-7.308	.725	101.481	1	.000	.001		
Step 4[d]	ses			12.506	2	.002			
	ses(1)	-.926	.289	10.256	1	.001	.396	.225	.698
	ses(2)	-.731	.238	9.413	1	.002	.481	.302	.768
	school(1)	1.374	.295	21.738	1	.000	3.952	2.218	7.043
	x	.039	.005	65.371	1	.000	1.039	1.030	1.049
	science	-.065	.018	13.502	1	.000	.937	.905	.970
	Constant	-6.193	.786	62.014	1	.000	.002		

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

b. Variable(s) entered on step 2: school.

c. Variable(s) entered on step 3: science.

d. Variable(s) entered on step 4: ses.



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	692.268[a]	.207	.276
2	664.107[a]	.243	.324
3	650.601[b]	.260	.347
4	637.795[b]	.276	.368

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

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

1.1.3 AIC

Be aware that there is not a correct model and every model is a simplification of reality. But models can explain reality well to different degrees and they can provide insight into relationships between predictors and response.

Measures like AIC (Akaike's Information Criterion) can help in choosing good models. The AIC measures model fit by assessing how close the fitted values based on the model are to the true probabilities. To keep models parsimonious AIC includes a penalty for the number of variables in the model.
The best model is the one which has the smallest

$$AIC = -2(\text{log likelihood} - \text{number of parameters})$$

Example 3
HSB, compare models
M_1 with predictors school type, x, and science and
M_2 with predictors school type, x, science and SES.

Model	AIC
M_1	658.601
M_2	649.795

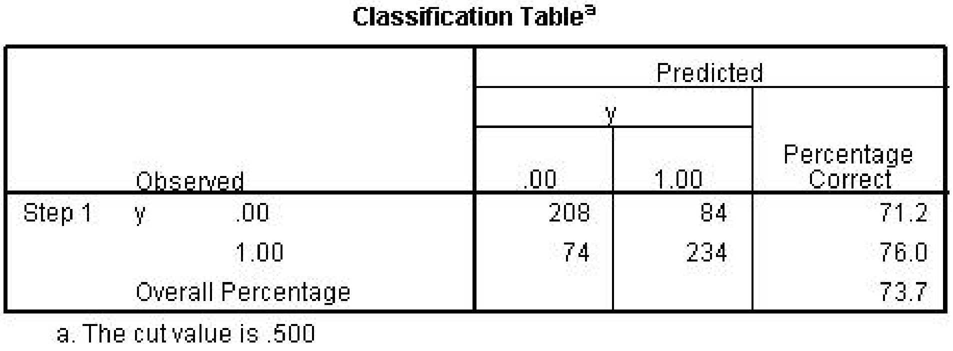
Therefore model M_2 is preferred over model M_1, although it includes more predictors.

1.1.4 Classification Tables

In order to check the predictive power of a model one can use classification tables.
In order to create classification tables we use the model to estimate the probability for success, $\hat{\pi}(x)$, for each individual in the sample. If this probability is greater than 0.5, then it is predicted that a success for this individual is more likely than a failure. Likewise if $\hat{\pi}(x)$ is smaller than 0.5, then one would predict a failure for this individual.
Therefore each individual is classified as success if $\hat{\pi}(x) \geq 0.5$ and as a failure if $\hat{\pi}(x) < 0.5$.
A table showing the actual measurement and the classification for the data is called a classification table.
If the model fits well most individuals from the sample should be predicted to fall in their actual category.
To measure the predictive power of the model sensitivity, specificity, and the overall proportion of correct classifications are used.

Sometimes instead of using a cutpoint of 0.5 for classifying individuals another values π_0 can be used and might result in better sensitivity and specificity.

Example 4
HSB for ”best” model



Classification Table[a]

			Predicted		
			y		Percentage Correct
	Observed		.00	1.00	
Step 1	y	.00	208	84	71.2
		1.00	74	234	76.0
	Overall Percentage				73.7

a. The cut value is .500

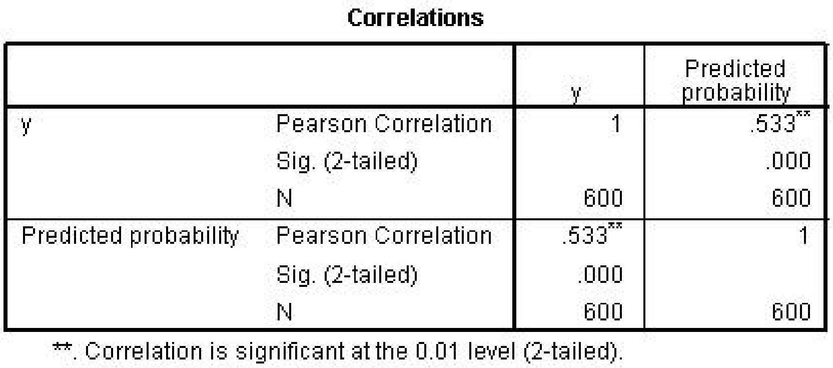
This model gives good predictions with an overall correct classification rate of 73.7%.

1.1.5 Correlation

R the correlation between the observed values y_i and the estimated values $\hat{y}_i$ is also a measure of model fit.
For the logistic regression model this is a correlation between values of 0 and 1 for the response (1=success, 0=failure) and the estimated probabilities for success, $\hat{\pi}_i$. The closer the value to 1, the better the fit between data and model. Because of the discrete response variable the usefulness is limited.

R^2 is again the coefficient of determination, which gives the proportion of variation in the response variable explained by the model.
Cox& Snell R^2 and Nagelkerke R^2 are alternative measures of fit. Cox&Snell can never be 1, for which Nagelkerke is adjusting. The closer Nagelkerke's R^2 to 1 the better the model for predicting the outcomes.
Both measures can by obtained by using SPSS Regression>Binary Logistic Regression.

Example 5
HSB for "best" model

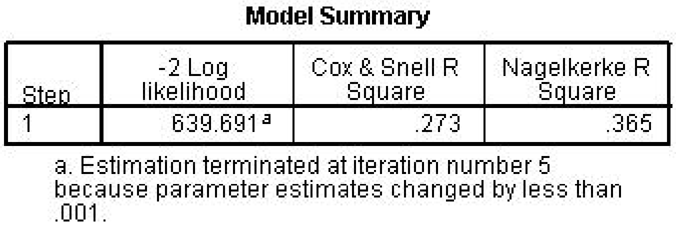


Correlations

		y	Predicted probability
y	Pearson Correlation	1	.533**
	Sig. (2-tailed)		.000
	N	600	600
Predicted probability	Pearson Correlation	.533**	1
	Sig. (2-tailed)	.000	
	N	600	600

**. Correlation is significant at the 0.01 level (2-tailed).

R^2=0.284.



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	639.691[a]	.273	.365

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

Nagelkerke's R^2 = .365 indicates a limited usefulness for predicting the outcomes.

1.2 Checking the Model

As discussed before we analyze measures of model fit

$$G^2 = \text{Deviance} = 2\sum \text{observed}[\log(\text{observed}/\text{fitted})]$$

and

$$X^2 = \text{Pearson} = \sum (\text{observed} - \text{fitted})^2/\text{fitted}$$

and standardized residuals to assess the model fit.

In addition one might want to analyze some *influence diagnostics* to identify observations that are either outliers or highly influential in determining the values of the predictions and estimates of the slopes.
Usually the model fit is assessed for the complete sample. When checking for influential observations estimates and model fit are compared for these based on the complete data set and the data set removing the observation under investigation. If the model fit and/or estimates are *very* different we conclude that the measurement is highly influential, and investigate if it is caused by a data entry error, mis-measurement, or other confounding variables.
Measures of leverage and influence of observations in the sample are:

1. **leverage**: The leverage of an observation measures how far it falls away from the other observation in the predictor space. The leverage will always be between 0 and 1, a cut-off of $\geq 3k/n$ is used to identify potential outliers and influential observations.

2. **Cook's Distance**: Cook's D measures the discrepancy between the mean estimates based on the complete data set and the one excluding the observation investigated. Let $\hat{\mu}_i$ be the estimate of the mean response for observation i based on the complete data set, and $\hat{\mu}_{i(j)}$ the same estimate when leaving out observation j, then

   $$\text{Cook's} \;\; D = \frac{\sum(\hat{\mu}_i - \hat{\mu}_{i(j)})^2}{\text{some standard error}}$$

   For identifying problematic observations a cut-off of > 1 is used.

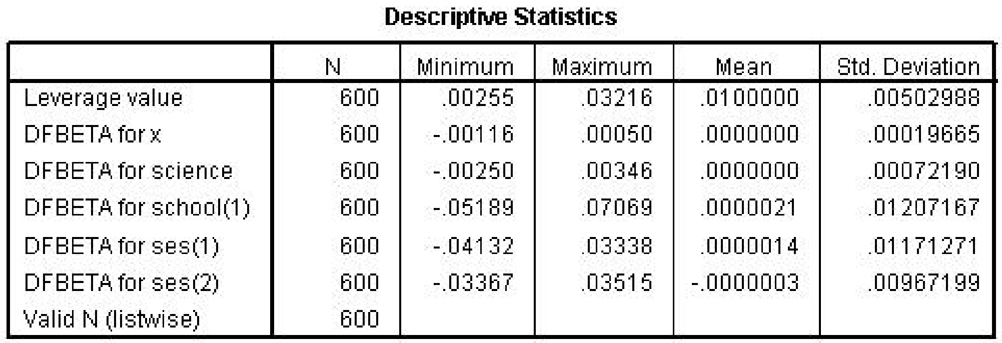
3. ***Dfbeta***: *Dfbeta*s are calculated for every model parameter and every observation. They measure the discrepancy between the parameter estimates based on the complete data set and the ones excluding the observation investigated. Let $\hat{\beta}_l$ be the estimate of the l^{th} parameter based on the complete data set, and $\hat{\beta}_{l(j)}$ the same estimate when leaving out observation j, then

   $$\text{DFbeta}_{l(j)} = \frac{\hat{\beta}_l - \hat{\beta}_{l(j)}}{\text{some standard error}}$$

   For identifying problematic observations a cut-off of $> 2/\sqrt{n}$ is used.

For each of the measures, the larger the measure the greater the influence of the individual on the estimates and model fit.

Example 6
HSB, “best model”



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Leverage value	600	.00255	.03216	.0100000	.00502988
DFBETA for x	600	-.00116	.00050	.0000000	.00019665
DFBETA for science	600	-.00250	.00346	.0000000	.00072190
DFBETA for school(1)	600	-.05189	.07069	.0000021	.01207167
DFBETA for ses(1)	600	-.04132	.03338	.0000014	.01171271
DFBETA for ses(2)	600	-.03367	.03515	-.0000003	.00967199
Valid N (listwise)	600				

Here we show only the summary statistics, but in the analysis one should look at the entire data set to flag observations that seem problematic. Summary statistics are helpful to find out, if observations exist which require further investigation.

The cut-off for the leverage is $> 3k/n = 3(6/600) = 0.03$ (the model has 5 slopes + the intercept). Since maximum leverage is 0.03216, there is at least one measurement that should be investigated.

The cut-off for the Dfbetas is $> 2/\sqrt{n} = 0.0816$. According to the summary statistics none of the observations seems to strongly influence the parameter estimates, since the the minima for all slopes are greater than -0.0816, and the maxima are all smaller than 0.0816.