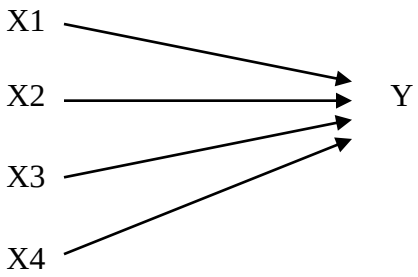
Neuendorf
Logistic Regression

The Model:



Assumptions:

1. Metric (interval/ratio) data for 2+ IVs, and dichotomous (binomial; 2-value), categorical/nominal data for a single DV. . . bear in mind that other types of IVs are allowed when they have been dummy or otherwise coded (SPSS will actually convert categorical IVs to dummies automatically (!) for this procedure only, if the user indicates it in the "Categorical" section of [Binary] Logistic Regression).

2. Predicts the odds of an event occurring (see Addendum 1), which is based on the probability of that event occurring. Precisely, the odds of an event occurring is:

$$\text{Odds} = \frac{\text{probability of event occurring}}{\text{probability of event not occurring}} = \frac{P}{1 - P}$$

3. Assumes a *nonlinear* (specifically an S-shaped, or sigmoidal, curve) relationship between IVs and the DV (expressed as a probability of the occurrence of DV=1); however, this translates to a *linear* relationship between the logit (natural log of the odds of the dependent occurring or not) and the set of IVs. See Addendum 2 for an illustration that compares probabilities, odds, and the logit. See Addendum 3 for an example of the S-shaped curve.

4. Uses a maximum-likelihood rather than least-squares statistical model. In least squares, the procedure selects regression coefficients that result in the smallest sum of squared differences between the observed and the predicted values of the DV (i.e., the smallest sum of squared residuals). In maximum-likelihood, the coefficients that make our observed results "most likely" are selected after a series of iterations.

5. Residuals follow a binomial rather than a normal distribution. Normality of variables is not a stringent requirement.

6. Does not assume homoscedasticity.

7. Assumes that there is little or no multicollinearity (however, SPSS will not assess this in the [binary] Logistic Regression procedure; again, you can run a "fake" Multiple Regression procedure to get TOLs and VIFs if you wish).

Decisions to make:

1. [Hierarchical] Blocks vs. simultaneous model

2. Forced entry ("Enter") vs. stepwise entry of IVs (several options available for both "Forward" and "Backward"--right click on each option in older versions of SPSS for more info.)

Statistics:

1. Goodness-of-fit indicators for overall model at each block's entrance ("omnibus tests")

   A. -2 Log Likelihood (-2LL)--This stat is integral to the notion of the "maximum likelihood" analytical model; it compares the likelihood function of the final model with that for a baseline model (i.e., one in which there is no knowledge about the IVs). The values for -2LL are often large and, much like eigenvalues, don't make objective sense. The -2LL is a cumulative measure across all cases and its size is therefore highly dependent on *n*.

      A small value for -2LL indicates a good fit. A "perfect fit" by the model (called a "saturated model") will result in a likelihood of 1.0, and a -2LL of 0. In SPSS, chi-squares test the significance of -2LL; we like these chi-squares to be significant.

   B. "R^2-like" measures--Cox & Snell, Nagelkerke, pseudo-R^2 (this last one must be hand-calculated; see Hair p. 328). These stats use the log likelihoods of the baseline and final model to construct a "reduction-in-error" proportion, much like R^2.

   C. Hosmer & Lemeshow Chi-square--tests the predictive accuracy of the model by splitting the sample into deciles (ten groups) on the basis of the probability of DV=1 for the purposes of constructing a chi-square table. Here, a non-significant chi-square indicates a good model fit; the actual and predicted values on the DV do not differ significantly, in this case.

2. Goodness-of-fit indicator for each block (in hierarchical order)

   A. Block chi-squares test the significance of changes in -2LL for each block that is added; we like these chi-squares to be significant.

3. Logistic coefficients (B's)--These are unstandardized coefficients that correspond to the b's in multiple regression (i.e., the unstandardized partial regression coefficients). The overall equation is:

   (a) $\text{Logit} = \ln(\text{Odds}) = B_0 + B_1X_1 + B_2X_2 + B_3X_3$

   B_0 is like the intercept/constant "a" in a multiple regression equation.

   A. The Wald statistic tests each B (rather than a t-test, the typical test of a "b" or Beta in multiple regression). Wald has a chi-square distribution. Its formula is:

   $$\text{Wald} = (B/SE_B)^2$$

   B. Exp(B)--exponentiated B, the result of transforming both sides of the above equation (a)

such that the left now is straight odds (note: e = 2.718):

(a) Logit = ln(Odds) = $B_0 + B_1X_1 + B_2X_2 + B_3X_3$

(b) Odds = $e^{B0}\, e^{B1X1}\, e^{B2X2}\, e^{B3X3}$

Each Exp(B) indicates a decrease or increase in the odds of occurrence of the DV. A value of less than 1.0 indicates a lowered odds as a result of that IV, and a value of greater than 1.0 indicates an enhanced odds as a result of that IV. For example, an Exp(B) of 1.20 would indicate that an increase of one unit in that IV would result in a predicted increase in the odds of 20% for the occurrence of the DV. An Exp(B) of .75 would indicate a predicted decrease in the odds of 25%. (Note that the coefficients are partials, indicating that the effect assumes that all other IVs are held constant (controlled for).)

The following formula allows for a fairly straightforward interpretation of an Exp(B) for a given IV:

(Exp(B) – 1) x 100% = the % change in odds of DV being 1 as a result of a unit increase in that IV, when controlling for all other variables in the model at that point

4. Score statistic--this tests the significance of parameter estimates computed via maximum likelihood methods. The test assesses whether a given IV relates significantly to the DV. The test is based on the behavior of the log-likelihood function at the point where the tested parameter is zero. SPSS presents this stat for variables "not in the equation" at that point.

5. Classification analysis--Like in discriminant analysis. . . we can obtain an overall "hit rate" (% of cases correctly classified by the logistic equation). We can also get casewise diagnostics, as in discriminant, and a one-dimensional classification plot, sort of like the territorial map in discriminant.

   As with Discriminant Analysis, you may hand-calculate a Tau or a Press' Q to assess whether the classification outcome is significantly better than chance. (See the Discriminant Analysis handout for formulae.)

References:

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley & Sons, Inc.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage Publications.

Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage Publications.

[NOTE: Even though the Pampel book says it's a primer, it's rather highly mathematical; it is good for understanding odds ratios and probabilities. Menard uses SPSS and SAS examples, and is a bit more applied.]

ADDENDUM 1--Terminology for use with Logistic Regression

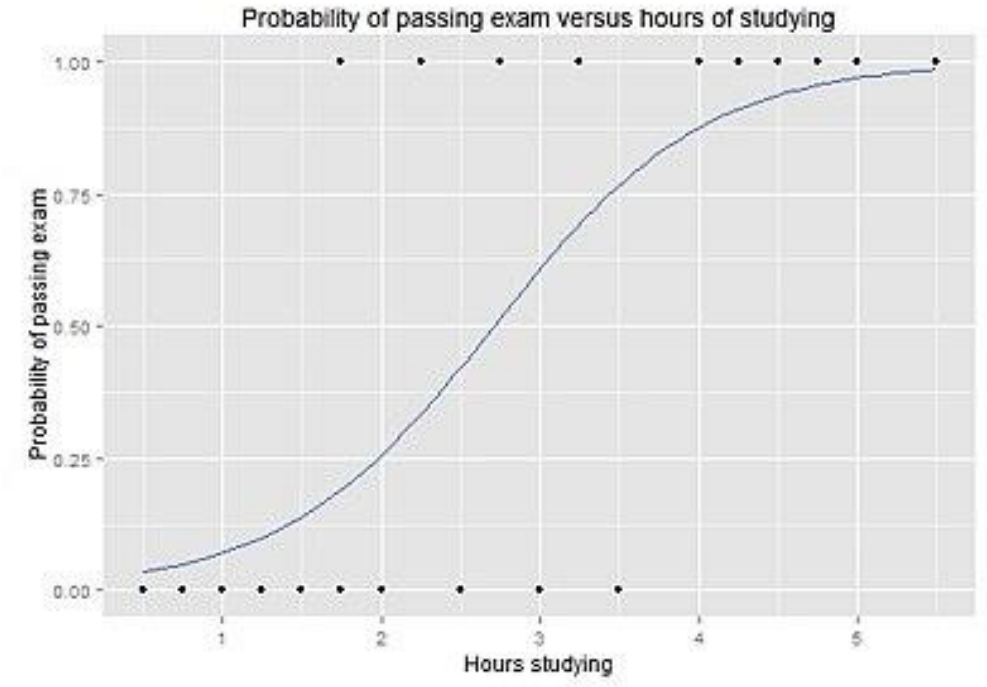
Probability	=	P	=	probability of an event occurring (range of 0 - 1)		
Odds	=	$\frac{P}{1 - P}$	=	ratio of the probability of an event occurring to the probability of the event not occurring (range of 0 - pos. infinity)		
Odds ratio	=	$\frac{Odds_1}{Odds_2}$	=	$\frac{P_1/(1 - P_1)}{P_2/(1 - P_2)}$	=	ratio of two odds
Logit	=	ln(Odds)	=	predicted logged odds (range of neg. infinity - pos. infinity)		

NOTE: The Hair et al. book calls the "odds" the "odds ratio." This runs counter to the use of the terms by Pampel (2000) and by Hosmer and Lemeshow (2000). This handout uses the terms as they are presented by Pampel and by Hosmer and Lemeshow.

ADDENDUM 2--Illustration of the relationship between probabilities, odds, and the Logit (ln(odds)):

P	.01	.1	.2	.3	.4	.5	.6	.7	.8	.9	.99
1 - P	.99	.9	.8	.7	.6	.5	.4	.3	.2	.1	.01
Odds	.01	.111	.25	.429	.667	1	1.5	2.33	4	9	99
Logit	-4.60	-2.20	-1.39	-.847	-.405	0	.405	.847	1.39	2.20	4.60
Exp(Logit)	.01	.111	.25	.429	.667	1	1.5	2.33	4	9	99

ADDENDUM 3—Example of S-shaped (sigmoidal) curve



Sex Roles (2010) 62:747–761
DOI 10.1007/s11199-009-9644-2

ORIGINAL ARTICLE

Shaken and Stirred: A Content Analysis of Women's Portrayals in James Bond Films

Kimberly A. Neuendorf · Thomas D. Gore · Amy Dalessandro · Patricie Janstova · Sharon Snyder-Suhy

Table 6 Prediction of mortality (death) of female characters in Bond films via logistic regression.

	r	Exp(B) at enter	Final Exp (B)	Block Chi-square	–2LL	Cox & Snell R^2	Nagelkerke R^2	Hosmer & Lemeshow Chi-sq test
Block 1				.08	167.53	.00	.00	11.99
Year	.04	1.01	.98					
Block 2				7.42	160.11	.04	.07	16.27*
Age	.01	1.04	1.07					
Body size	−.09	.65	.50					
U.S. accent	−.04	1.05	.99					
Blond	−.05	.64	1.06					
Glasses	.07	3.13	3.96					
Long hair	.03	1.34	.72					
Short hair	.05	.82	.69					
Straight hair	.04	1.55	1.32					
Non-white	−.02	.94	1.41					
Extremely attractive	.07	1.31	.80					
Extremely unattractive	−.03	.73	.34					
Block 3				1.41	158.70	.05	.08	13.95
Major role	.10	1.89	.29					
Block 4				4.98*	153.72	.07	.12	5.25
Sexual Activity	.18*	1.35*	1.61*					
Block 5				30.62**	123.10**	.21	.36	3.78
Good at end of film	−.30**	.15**	.15**					
Attempts to kill Bond	.28**	11.84*	11.84*					
Weapon use	.24**	1.13	1.13					

Each–2LL was tested via chi-square.

*$p<.05$; **$p<.01$