Modeling Binary outcome

Case Study: See example in R

Cedergren's 1974 study of final *s*-deletion in Panama City, Panama. Cedergren had noticed that speakers in Panama City, like in many dialects of Spanish, variably deleted the *s*at the end of words. She undertook a study to find out if there was a change in progress: if final "s" was systematically dropping out of Panamanian Spanish. She performed interviews across the city in several different social classes, to see how the variation was structured in the community. She also investigated the linguistic constraints on deletion, so she coded for a phonetic constraint - whether the following segment was consonant, vowel, or pause and the grammatical category of word that the "s" is part of a: monomorpheme, where the s is part of the free morpheme (e.g.,menos) verb, where the"s"is the second singular inflection (e.g.,tu tienes,el tienes) determiner, where "s"is plural marked on a determiner (e.g.,los,las) adjective, where "s" is a nominal plural agreeing with the noun (e.g.,buenos) noun, wheresmarks a plural noun (e.g.,amigos).

Test of hypothesis

1. Is the effect observed statistically significant or attributable to chance?
2. Three types of hypothesis: a) tests of goodness of fit of the overall model. b) tests of effect of any one risk factor contained within the model. c) tests of the linear effect of ordered categorical risk factors.
3. Deviance is calculated from the likelihood, which is a measure of how likely a particular model is, given the observed data.
4. A measure of the difference between the postulated model and the model that, by definition, is a perfect fit to the data (called full or saturated model).
5. Deviance is given by
$$D = -2\{\log \hat{L} - \log \hat{L}_F\}$$
6. The deviance of the model can be used to test for goodness of fit of the model to the data. The model deviance is compared to chi-square with the model deviance df. The df for a model deviance is calculated as "df = number of data items - number of independent parameters in the fitted model".

7. Number of independent parameters is 1 for the intercept term, 1 for quantitative variable and $l - 1$ for a categorical variable with l levels.

8. In the case of lack of fit, Further explanatory variables may be needed. We may have inadequately modeled the effect of the current variables. Transformations might be needed, important interactions might be missing, Outliers may be in the data. Assumption of binomial variation may be incorrect. It is much more meaningful to test for specific effects.

Effect of a Risk factor

Model nesting: Model A is said to be nested within model B if model B contains all the variables of model A plus at least one other. Constant is thought of as a variable.

Table 1: default

Model A	Model B
constant	constant + social class
constant + SBP	constant + SBP + cholesterol
constant + age +	constant + age + cholesterol+
cholesterol + BMO + smoking	BMO + SBP + smoking+ activity in leisure

When model A is nested within model B, we can test the hypothesis that the extra terms in B have no effect by calculating the difference between the deviance of models A and B, denoted ΔD.

- Ex. Considering the example with the following data

Table 14. Ratio of coronary heart disease (CHD) events to total number by systolic blood pressure (SBP) and cholesterol.

SBP	Serum total cholesterol (mmol/l)				
(mmHg)	≤5.41	5.42–6.01	6.02–6.56	6.57–7.31	>7.31
≤118	1/190	0/183	4/178	8/157	4/132
119–127	2/203	2/175	6/167	10/166	11/137
128–136	5/173	9/176	9/181	8/167	11/164
137–148	5/139	3/156	10/154	13/174	16/174
>148	5/123	8/123	12/144	13/179	23/180

- Four models may be fitted

1. $\hat{\text{logit}} = b_0$
2. $\hat{\text{logit}} = b_0 + b_1^{(1)}x_1^{(1)} + b_1^{(2)}x_1^{(2)} + b_1^{(3)}x_1^{(3)} + b_1^{(4)}x_1^{(4)} + b_1^{(5)}x_1^{(5)}$
3. $\hat{\text{logit}} = b_0 + b_2^{(1)}x_2^{(1)} + b_2^{(2)}x_2^{(2)} + b_2^{(3)}x_2^{(3)} + b_2^{(4)}x_2^{(4)} + b_2^{(5)}x_2^{(5)}$
4. $\hat{\text{logit}} = b_0 + b_1^{(1)}x_1^{(1)} + b_1^{(2)}x_1^{(2)} + b_1^{(3)}x_1^{(3)} + b_1^{(4)}x_1^{(4)} + b_1^{(5)}x_1^{(5)}$
$+ b_2^{(1)}x_2^{(1)} + b_2^{(2)}x_2^{(2)} + b_2^{(3)}x_2^{(3)} + b_2^{(4)}x_2^{(4)} + b_2^{(5)}x_2^{(5)},$

- Analysis of deviance table

Model	D	d.f.
1 Constant	94.58	24
2 Constant + SBP	56.73	20
3 Constant + cholesterol	49.48	20
4 Constant + SBP + cholesterol	18.86	16

Note: D = deviance.

- Compare models 1 and 2 to assess the significance of SBP.
- Models 1 and 3 for cholesterol
- Models 1 and 4 for SBP and cholesterol together
- Models 3 and 4 for SBP over and above cholesterol
- Models 2 and 4 for cholesterol over and above SBP.