MATH 60604A
Statistical modelling
§ 4c - Application of logistic regression

Léo Belzile

HEC Montréal
Department of Decision Sciences

Fixation-intention to buy example

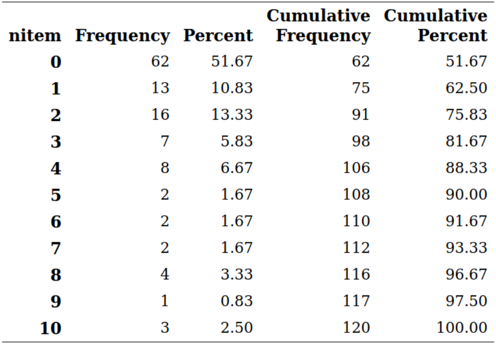
In the study, subjects navigated a website that contained, among other things, an advertisement for candy. Simultaneously, an “eye-tracker” tracked where the subject’s eye were fixated on the screen. We can therefore measure whether the subject saw the ad, and for how long. Moreover, a software was used (FaceReader) to measure the facial expressions and infer the emotions of the subject while viewing the ad.

Suppose that, instead of measuring the intention to buy from the questionnaire, we had contacted the subjects one month later to see if they had bought the product.

Number of items bought

The data includes two variables not considered until now:

- `buy`: a binary variable equal to unity if the subject bought the product and zero otherwise.
- `nitem`: integer giving the number of times the item was bought.



nitem	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	62	51.67	62	51.67
1	13	10.83	75	62.50
2	16	13.33	91	75.83
3	7	5.83	98	81.67
4	8	6.67	106	88.33
5	2	1.67	108	90.00
6	2	1.67	110	91.67
7	2	1.67	112	93.33
8	4	3.33	116	96.67
9	1	0.83	117	97.50
10	3	2.50	120	100.00

The variable `nitem` varies from 0 to 10 with a mean of 1.71 purchases per participant. 51.7% of the participants did not make any purchase.

We focus for now on `buy` as response. The explanatory variables are as before, namely

- `fixation`: total duration of fixation on the ad (in seconds).
- `emotion`: measured during the ad fixation, the ratio of the probabilities of having a positive emotion and of having a negative emotion.
- `sex`: the sex of the subject, zero for male and one for female.
- `age`: the age of the subject (in years).
- `revenue`: annual salary (in dollars), one of
  1. $[0, 20000)$;
  2. $[20000, 60000)$;
  3. 60000 and above.
- `educ`: the subject's level of education, one of
  1. high school or lower;
  2. college;
  3. university degree.
- `marital`: binary variable giving the marital status, zero for single and one for subjects in a relationship.

Logistic model with all the predictor variables

If we let $\pi = \mathsf{P}\left(Y = 1 \mid \mathbf{X}\right)$ denote the probability of making a purchase given the value of the explanatory variables, the model is

$$
\begin{aligned}
\mathrm{logit}(\pi) = \beta_0 &+ \beta_1 \texttt{sex} + \beta_2 \texttt{age} + \beta_3 \texttt{revenue}_1 + \beta_4 \texttt{revenue}_2 \\
&+ \beta_5 \texttt{educ}_1 + \beta_6 \texttt{educ}_2 + \beta_7 \texttt{marital} \\
&+ \beta_8 \texttt{fixation} + \beta_9 \texttt{emotion}.
\end{aligned}
$$

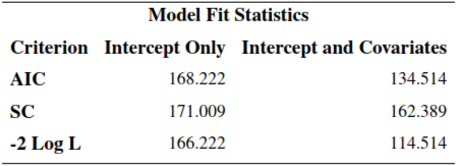
SAS code with `proc logistic`

- The β parameters are really only interpretable on the exponential scale.
- We can use `proc logistic` with `clparm=pl odds=pl expb` to get exponentiated parameters and confidence intervals.
- With `proc logistic`, the default parametrization for categorical variables is obtained through the option `param=glm`.

SAS code with `proc logistic`

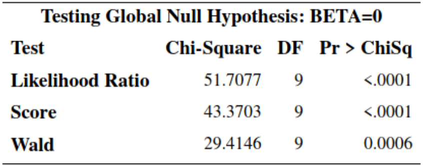
```
proc logistic data=statmod.intention;
class educ revenue / param=glm;
model buy(ref="0")=sex age revenue educ marital
    fixation emotion / clparm=pl odds=pl expb;
run;
```

SAS output for the `logistic` procedure



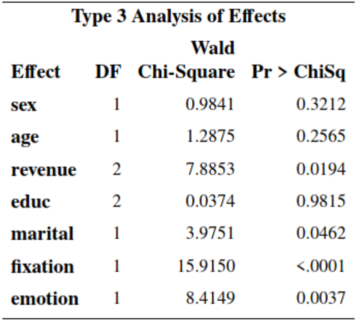
Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	168.222	134.514
SC	171.009	162.389
-2 Log L	166.222	114.514



Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	51.7077	9	<.0001
Score	43.3703	9	<.0001
Wald	29.4146	9	0.0006

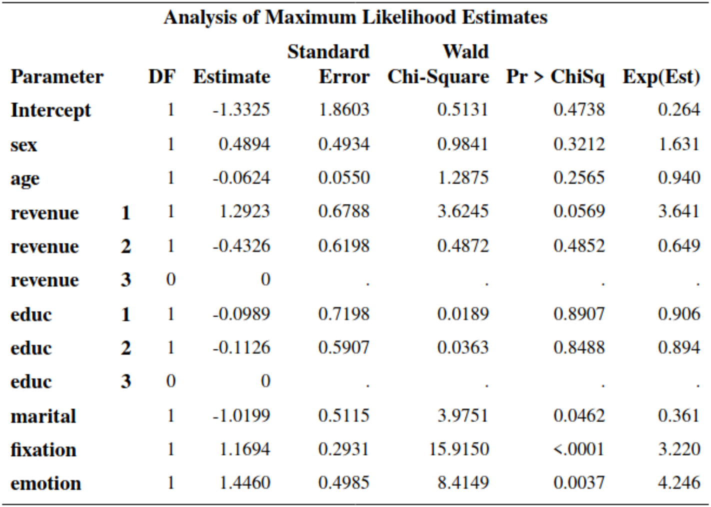


Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	0.9841	0.3212
age	1	1.2875	0.2565
revenue	2	7.8853	0.0194
educ	2	0.0374	0.9815
marital	1	3.9751	0.0462
fixation	1	15.9150	<.0001
emotion	1	8.4149	0.0037

- Goodness of fit diagnostics are given for the fitted model and the null model in which the probability of success is constant (Intercept Only).
- In addition to information criteria, likelihood tests (Wald, score, likelihood ratio) are given for testing the global null hypothesis of significance, $\mathscr{H}_0 : \beta_1 = \cdots = \beta_p = 0$.
- Parameter significance (Type III effects) is based on Wald statistics (for likelihood ratio, use the `genmod` procedure with option `type3`).

Parameter estimates

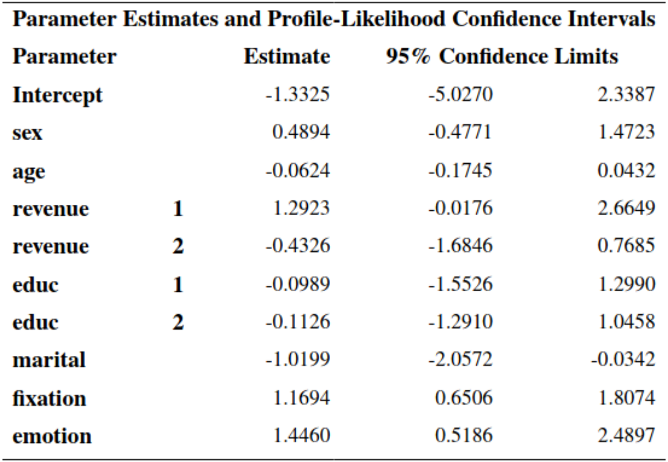


Analysis of Maximum Likelihood Estimates

				Standard	Wald		
Parameter		DF	Estimate	Error	Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	-1.3325	1.8603	0.5131	0.4738	0.264
sex		1	0.4894	0.4934	0.9841	0.3212	1.631
age		1	-0.0624	0.0550	1.2875	0.2565	0.940
revenue	1	1	1.2923	0.6788	3.6245	0.0569	3.641
revenue	2	1	-0.4326	0.6198	0.4872	0.4852	0.649
revenue	3	0	0	.	.	.	.
educ	1	1	-0.0989	0.7198	0.0189	0.8907	0.906
educ	2	1	-0.1126	0.5907	0.0363	0.8488	0.894
educ	3	0	0	.	.	.	.
marital		1	-1.0199	0.5115	3.9751	0.0462	0.361
fixation		1	1.1694	0.2931	15.9150	<.0001	3.220
emotion		1	1.4460	0.4985	8.4149	0.0037	4.246

- The option `expb` will provide each of the $\exp(\widehat{\beta}_j)$ values in the last column.
- The tests of significance are Wald-based.

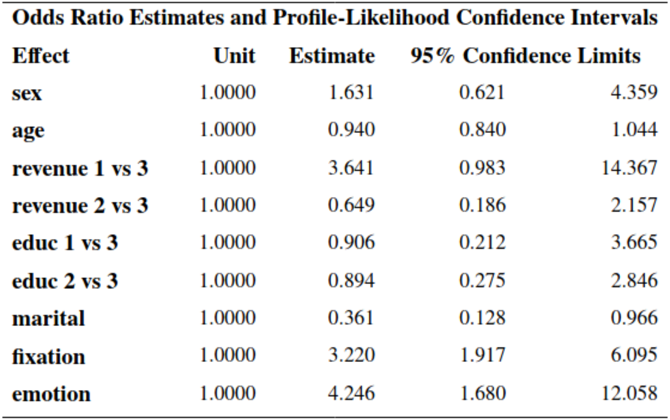
Table of coefficients



Parameter Estimates and Profile-Likelihood Confidence Intervals

Parameter		**Estimate**	**95% Confidence Limits**	
Intercept		-1.3325	-5.0270	2.3387
sex		0.4894	-0.4771	1.4723
age		-0.0624	-0.1745	0.0432
revenue	**1**	1.2923	-0.0176	2.6649
revenue	**2**	-0.4326	-1.6846	0.7685
educ	**1**	-0.0989	-1.5526	1.2990
educ	**2**	-0.1126	-1.2910	1.0458
marital		-1.0199	-2.0572	-0.0342
fixation		1.1694	0.6506	1.8074
emotion		1.4460	0.5186	2.4897

Table of exponentiated coefficients



Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
sex	1.0000	1.631	0.621	4.359
age	1.0000	0.940	0.840	1.044
revenue 1 vs 3	1.0000	3.641	0.983	14.367
revenue 2 vs 3	1.0000	0.649	0.186	2.157
educ 1 vs 3	1.0000	0.906	0.212	3.665
educ 2 vs 3	1.0000	0.894	0.275	2.846
marital	1.0000	0.361	0.128	0.966
fixation	1.0000	3.220	1.917	6.095
emotion	1.0000	4.246	1.680	12.058

- On the exponentiated scale, the parameter is not significant at level 5% if 1 is in the confidence interval.
- To obtain likelihood-ratio based confidence intervals, use option `plrl`.

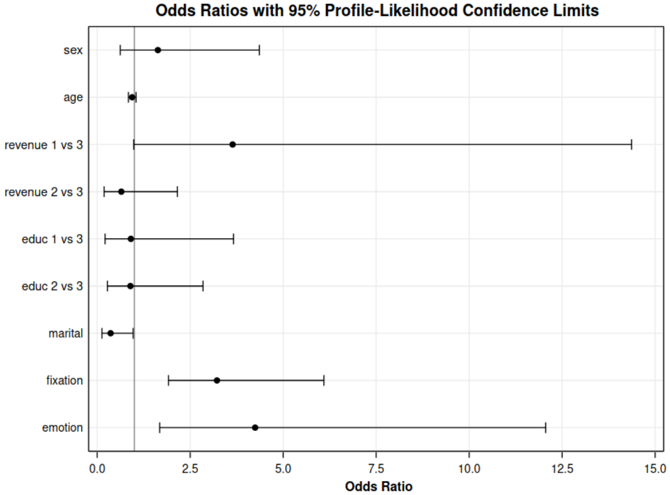
Parameter interpretations

- $\exp(\widehat{\beta}_{\texttt{sex}}) = 1.631$: the odds of buying for women (`sex=1`) is 1.631 times higher than for men, **when all other variables in the model are held constant**. Therefore, women have a higher chance of buying than men, even after adjusting for all the other variables.
- $\exp(\widehat{\beta}_{\texttt{age}}) = 0.94$: when age increases by one year, the odds of buying change by a factor of 0.94, i.e. decreases by 6%, **when all other variables in the model are held constant**.
- $\exp(\widehat{\beta}_{\texttt{fixation}}) = 3.22$: when fixation time increases by one second the odds of buying is multiplied by 3.22 **when all other variables in the model are held constant**.

Comparing the levels of `revenue`

- The coefficient for $\texttt{revenu}_1$ is relative to `revenu=3` and $\exp(\widehat{\beta}_{\texttt{revenu}_1}) = 3.641$: the estimated odds of buying for low-income individuals (`revenue=1`) is 3.641 times the odds of buying for high-income individuals (`revenue=3`), **when all other variables in the model are held constant**.
- To get the odds ratio between levels 1 and 2 of `revenue`, we would need to fit another model, changing the reference category.
- We could also easily do it by hand: $3.641/0.649 = 5.61$ implies that the odds for revenue level 1 are 461% higher than the odds for revenue level 2, when all other variables are held constant.

Visual representation of the odds ratios



The confidence intervals are based on a profile likelihood. They are therefore **invariant to reparametrization** of the model, so you can get a confidence interval for $\exp(\beta_k)$ by exponentiating the limits of the confidence interval for β_k.