ΠΑΝΤΕΙΟΝ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΟΙΝΩΝΙΚΩΝ ΚΑΙ ΠΟΛΙΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
PANTEION UNIVERSITY OF SOCIAL AND POLITICAL SCIENCES

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΟΙΚΟΝΟΜΙΑΣ ΚΑΙ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ

ΤΜΉΜΑ ΟΙΚΟΝΟΜΙΚΗΣ ΚΑΙ ΠΕΡΙΦΕΡΕΙΑΚΗΣ ΑΝΆΠΤΥΞΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΕΦΑΡΜΟΣΜΕΝΩΝ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ»

Regression modelling of an ordinal dependent variable.

Αντασκαλίτσας Μπογδάν

Αθήνα, 2020

Τριμελής Επιτροπή
Σταύρος Ντεγιαννάκης, Αναπληρωτής Καθηγητής Παντείου Πανεπιστημίου
Κλάιβ Ρίτσαρντσον, Ομότιμος Καθηγητής Παντείου Πανεπιστημίου (Επιβλέπων)
Γρηγόρης Σιουρούνης, Επίκουρος Καθηγητής Παντείου Πανεπιστημίου

Περίληψη

Εφαρμογή και σύγκριση διωνυμικής, ταξινομημένης, continuation-ratio, adjacent-categories και πολυωνυμικής λογαριθμικής παλινδρόμησης στα δεδομένα από μεγάλο πολυεθνικό δειγματοληπτικό ερωτηματολόγιο για παιδιά. Η λογιστική παλινδρόμηση χρησιμοποιείται για την ανάλυση της σχέσης μιας διωνυμικής εξαρτώμενης μεταβλητής με ορισμένες επεξηγηματικές μεταβλητές. Όταν η εξαρτημένη μεταβλητή έχει τρεις ή περισσότερες κατηγορίες, υπάρχουν διαφορετικές μέθοδοι ανάλυσης και μοντέλα, όπως ταξινομημένη, continuation-ratio, adjacent-categories και πολυωνυμικής λογαριθμικής παλινδρόμησης. Ο σκοπός αυτής της εργασίας είναι να τις παρουσιάσει και να τις συγκρίνει σε ένα σύνολο δεδομένων από δειγματοληπτικές έρευνες. Η ανάλυση έγινε με R στο RStudio. Ξεκινάμε εξετάζοντας τη βιβλιογραφία για τα logit μοντέλα μας και μετά εξετάζουμε και συγκρίνουμε τα αποτελέσματα. Αποτελέσματα: Αρχικά παρατηρούμε ότι η τέταρτη και σπανιότερη κατηγορία στο δείγμα μας μπορεί να συγχωνευθεί με την τρίτη για να παράγει λιγότερες ακραίες τιμές. Ορισμένες ανεξάρτητες μεταβλητές όπως, οι χώρες της Γερμανίας και της Ισλανδίας, καθώς και οι ηλικίες, το φύλο και οι μεταβλητές επιπέδου εκπαίδευσης των γονέων ήταν στατιστικά ασήμαντες σε ορισμένες περιπτώσεις. Η καλύτερη ικανότητα πρόβλεψης εμφανίζεται από το μοντέλο continuation-ratio, το καλύτερο μοντέλο AIC είναι το Multinomial, με ακόλουθο το continuation-ratio και το καλύτερο μοντέλο log-likelihood είναι το continuation-ratio.

Abstract

Logistic regression is the basic method used to analyze the relationship of a binomial dependent variable to some explanatory variables. When the dependent variable has three or more categories, there are different analysis methods and models, including ordered logistic, continuation ratio, adjacent categories and multinomial logistic regression. The purpose of the present thesis is to review and present these models, and compare the results obtained by applying them to a dataset drawn from a large multinational sample survey among schoolchildren. The analysis was done in R Studio.

The dependent variable had four categories originally. First we notice that the fourth and rarest category in our sample can be merged with the third to produce fewer outliers. Some of the independent variables such as, the countries Germany and Iceland, as well as age, gender and parent's education variables were statistically insignificant in some cases. The best predictive capability was displayed by the continuation ratio model, the best AIC model was the Multinomial, with a close second the Continuation Ratio and the best loglikelihood model was the Continuation Ratio.

Table of Contents

1 Introduction

The purpose of this paper is to compare different ways of logistic regression for categorical variables. Application and comparison of binomial logistic, ordered logistic, cumulative ratio, adjacent-categories ratio and multinomial logistic regression on the same data as Tsitsika et al. (2014). As we know, logistic regression is used to analyze the relationship of a binomial dependent variable to some explanatory variables. When the dependent variable has three or more categories, there are different analysis methods with similar models. The purpose of this proposed work is to present them and compare them in different datasets from sample surveys. The analysis can be done via SPSS. The data-set is from a questionnaire about Internet Addictive Behavior(IAB) and the factors that may have a role in such behavior. Probably the most important activity in statistical modelling of data is fitting regression models that relate the value y of a dependent variable Y to the values of several explanatory variables (predictors). The basic model is

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots + \beta_p x_p + \varepsilon$

where x_1, x_2 …. are the values of the explanatory variables and ε is a random error. The constant term β_0 and the regression coefficients

$\beta_1, \beta_2, \ldots., \beta_p$

which represent the effects of the explanatory variables on the dependent variable are parameters that must be estimated. In the classical regression model, the estimation may be carried out by least squares or, if a statistical distribution has been assumed, by the method of maximum likelihood. There are many variations of this model, most of them differing in the assumptions that are made concerning the error term. However, all of them require Y to be a proper quantitative measurement so that its value can be predicted numerically.

In many applications, the above model is not suitable, because the dependent variable Y is not quantitative but categorical. For example, in most biomedical applications Y is a binary variable, such as Success/Failure for the outcome of the treatment of a patient. Although the values of this variable may be coded numerically in the computer (for example, as 0/1), this is just for our convenience. Any two different values (1/2, -1/+1, 0/100,….) would do the same job. Obviously, numerical prediction of the "value" of Y is meaningless.

For this reason, various alternative regression-type models have been developed. When Y has two categories, the best known are probit and logistic regression. The second of these will be presented in the next chapter. It is a feature of all regression models that the "linear predictor"

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots + \beta_p x_p$

still represents the effect of the explanatory variables and in some way its value predicts the

category of Y.

As stated, logistic regression is used to analyse the relationship of a binary dependent variable to some explanatory variables. It cannot be used directly when the dependent variable has three or more categories. This would happen if the outcome of the patient's treatment was Success/Partial success/Failure. In this case, there are different analysis methods with similar models. For a review, see, for example, Agresti (2002). The purpose of the present dissertation is to present the basic models and compare them in the analysis of a data set concerning Internet Addictive Behaviour among adolescents, obtained from a sample survey and previously analysed by Tsitsika et al. (2014). The methods that will be presented, besides binary logistic regression, are ordered logistic, cumulative ratio, adjacent-category ratio and multinomial logistic regression.

Statistical packages do not necessarily offer all of these methods. For example, for more than two categories of Y, SPSS provides only the ordered logistic and multinomial logistic models. In order to fit the models, the R program was used in the present work. These models are generally fitted by maximum likelihood.

2 Bibliography overview

2.1 Binomial Logit

The first model is the binomial or binary logistic odds regression model,or logit regression, which aims to describe the proportion of successes, $P=Y_i/n$, in each subgroup in terms of factor levels and other explanatory variables which characterize the subgroup. An example of this is the mortality rate which is expressed by death and the opposite or if a new trial medicine achieved to cure a patient or not. In our case, we will not consider grouped data; that is n=1 always. This is done by modeling the probabilities π_i as, $g(\pi_i)=x^T\beta$ where x, is a vector of explanatory variables (dummy variables for factor levels and measured values of covariates), β is a vector of parameters and g is a link function. The general logistic regression model is

$$logit(\pi_i)=\log[\frac{\pi_i}{(1-\pi_i)}]=x^T\beta$$

2.1.1

x , is a vector of continuous measurements corresponding to covariates and dummy variables corresponding to factor levels and β is the parameter vector. The quantity $\pi_i/(1-\pi_i)$ is known as the odds of success. It is very widely used for analyzing multivariate data involving binary responses. It provides a powerful technique analogous to multiple regression and ANOVA for continuous responses. To get the probability of π_i we can use the following,

$$\pi_i=\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}$$

2.1.2

(Dobson,1990)

The estimators are interpreted as the probability of something occurring versus the opposite for a specific subgroup of factors, for example if the weather among other things affects the grip of tires, the $\exp(\beta i)$ or the odds ratio shows (OR) how many times the odds of Y=1 are multiplied when the variable x_i changes by one (while all others are stable). If the estimator is greater than zero or the OR is greater than one then the odds rise, if the estimator is less than zero or the OR is less than one then the odds fall and when the estimator is zero or the OR is one then the odds remain unchanged. If we want to predict the value (category) of Y, we estimate P(Y=1) and say for example

"if P(Y=1) >0,5 then we predict Y=1 for that case", then we assess the success of the model by comparing each prediction with the real values.

2.2 Ordered Logistic

In order to encompass the case of the dependent variable representing multiple groups, we need to expand our logistic regression model, to more than two groups, for example, patients of a disease have several stages of severity, mild, medium and severe. Thus, the second model is ordered logistic with proportional odds assumption. A cumulative probability for Y is the probability that Y falls at or below a particular point. For outcome category j, the cumulative probability is

$$P(Y \le j) = \pi_1 + \pi_2 + ... + \pi_j,\ j = 1,..,J \quad .$$

The cumulative probabilities reflect the order, with

$$P(Y \le 1) \le P(Y \le 2) \le \cdots \le P(Y \le J) = 1 \quad .$$

Models for cumulative probabilities do not use the final one, P(Y≤J), since it necessarily equals 1. The logits of the cumulative probabilities are

$$logit[P(Y \le j)] = \log[\frac{P(Y \le j)}{1-(P(Y \le j))}] = \log[\frac{\pi_1 + ... + \pi_j}{\pi_j + ... + \pi_J}] \quad .$$

2.2.1

These are called cumulative logits. For J=3, for example, models use both

$$logit[P(Y \le 1)] = \log[\frac{\pi_1}{(\pi_2 + \pi_3)}] \quad \text{and} \quad logit[P(Y \le 2)] = \log[\frac{(\pi_1 + \pi_2)}{\pi_3}] \quad .$$

Each cumulative logit uses all the response categories. A model for cumulative logit j looks like a binary logistic regression model in which categories 1–j combine to form a single category and categories j+1 to J form a second category. For an explanatory variable x, the model

$$logit[P(Y \le j)] = aj + \beta x,\ j = 1,..., J-1$$

2.2.2

has parameter β describing the effect of x on the log odds of response in category j or below. In this formula, β does not have a j subscript. Therefore, the model assumes that the effect of x is identical for all J−1 cumulative logits. When this model fits well, it requires a single parameter rather than J−1 parameters to describe the effect of x. Model interpretations can use odds ratios for the cumulative probabilities and their complements. For two values x_1 and x_2 of x, an odds ratio comparing the cumulative probabilities is

$$\frac{P(Y \le j|X=x_2)/P(Y>j|X=x_1)}{P(Y \le j|X=x_1)/P(Y>j|X=x_1)} \quad .$$

2.2.3

The log of this odds ratio is the difference between the cumulative logits at those two values of x. This equals $\beta(x_2-x_1)$, proportional to the distance between the x values. Specifically, for $x_2-x_1=1$, the odds of response below any given category multiply by e^{β} for each unit increase in x. For this log odds ratio $\beta(x_2-x_1)$, the same proportionality constant (β) is true for each cumulative probability. This property is called the proportional odds assumption of model 2.2.2 . (Agresti, 2002)

2.3 Continuation-Ratio logits

In this approach, logits are formed for ordered response categories in a sequential manner. The models apply simultaneously to

$$\log(\frac{\pi_1}{\pi_2}), \log(\frac{\pi_1+\pi_2}{\pi_j}), ..., \log(\frac{\pi_1+...+\pi_{j-1}}{\pi_J}) \quad .$$

2.3.1

These are called continuation-ratio logits. They refer to a binary response that is contrasting each category with a grouping of categories from lower levels of the response scale. A second type of continuation-ratio logit contrasts each category with a grouping of categories from higher levels of the response scale; that is,

$$\log(\frac{\pi_1}{\pi_2+...+\pi_j}), \log(\frac{\pi_2}{\pi_3+...+\pi_j}), ..., \log(\frac{\pi_{j-1}}{\pi_j}),$$

2.3.2

Models using these logits have different parameter estimates and goodness-of-fit statistics than models using the other continuation-ratio logits.(Agresti,2002)

2.4 Adjacent-Categories Logits

Another approach forms logits for all pairs of adjacent categories. The adjacent-categories logits are

$$\log\left(\frac{\pi_{j+1}}{\pi_j}\right), j=1,\ldots,J-1 \quad .$$

2.4.1

For J=3, these logits are $\log(\pi_2/\pi_1)$ and $\log(\pi_3/\pi_2)$.

With a predictor x, the adjacent-categories logit model takes the form

$$\log\left(\frac{\pi_{j+1}}{\pi_j}\right)=\alpha_j+\beta_j x, j=1,\ldots,J-1 \quad .$$

2.4.2

For it, the effects $\beta_j=\beta$ of x on the odds of making the higher instead of the lower response are identical for each pair of adjacent response categories. Like the cumulative logit model of proportional odds form, this model has a single parameter rather than J−1 parameters for the effect of x. This makes it simpler to summarize an effect. The adjacent-categories logits, like the baseline-category logits, determine the logits for all pairs of response categories. For the simpler model, the coefficient of x for the logit, $\log(\pi_a/\pi_b)$, equals $\beta(a-b)$. The effect depends on the distance between categories, so this model recognizes the ordering of the response scale.(Agresti, 2002)

2.5 Baseline-Category Logits

Baseline-Category Logits, logit models for nominal response variables, pair each category with a baseline category. When the last category (J) is the baseline, the baseline-category logits are

$$\log\left(\frac{\pi_j}{\pi_J}\right), j=1,\ldots,J-1 \quad .$$

2.5.1

Given that the response falls in category j or category J, this is the log odds that the response is j. For J=3, for instance, the model uses $\log(\pi_1/\pi_3)$ and $\log(\pi_2/\pi_3)$. The baseline-category logit model with a predictor x is

$$\log\left(\frac{\pi_j}{\pi_J}\right)=aj+b_j x, j=1,\ldots,J-1 \quad .$$

2.5.2

The model has J−1 equations, with separate parameters for each. The effects vary according to the category paired with the baseline. When J=2, this model simplifies to a single equation for $\log(\pi_1/\pi_2)=logit(\pi_1)$, resulting in ordinary logistic regression for binary responses. The

equations for these pairs of categories determine equations for all other pairs of categories. For example, for an arbitrary pair of categories a and b,

$$\log\left(\frac{\pi_j}{\pi_J}\right)=\log\left(\frac{\pi_a/\pi_j}{\pi_b/\pi_J}\right)=\log\left(\frac{\pi_a}{\pi_J}\right)-\log\left(\frac{\pi_b}{\pi_J}\right)=$$
$$=(\alpha_a+\beta_a x)-(\alpha_b+\beta_b x)=(\alpha_a-\alpha_b)+(\beta_a-\beta_b)x$$

2.5.3

The equation for categories a and b has the form $\alpha+\beta x$ with intercept parameter $\alpha=(\alpha_a-\alpha_b)$ and with slope parameter $\beta=(\beta_a-\beta_b)$. Software for multi-category logit models fits all the equations simultaneously. Estimates of the model parameters have smaller standard errors in contrast to when binary logistic regression software fits each component equation in 2.5.2 separately. For simultaneous fitting, the same parameter estimates occur for a pair of categories regardless of which category is the baseline. The choice of the baseline category is arbitrary.(Agresti, 2002)

3 Data

TABLE 1. DEMOGRAPHIC CHARACTERISTICS OF THE SAMPLE BY GENDER, AGE, PARENTAL EDUCATIONAL LEVEL, AND COUNTRY

	Gender		*Age*		*Parental education*	
	Female, N *(%)*	*Male,* N *(%)*	*14–15 years,* N *(%)*	*16–17 years,* N *(%)*	*Low/middle,* N *(%)*	*High,* N *(%)*
All adolescents	7,000 (52.7)	6,284 (47.3)	8,156 (61.4)	5,128 (38.6)	4,165 (37.3)	7,007 (62.7)
Greece	1,010 (51.3)	957 (48.7)	1,375 (69.9)	592 (30.1)	819 (44.4)	1,024 (55.6)
Spain	1,024 (51.7)	956 (48.3)	1,296 (65.5)	684 (34.5)	713 (40.5)	1,046 (59.5)
Romania	1,021 (55.8)	809 (44.2)	486 (26.6)	1,344 (73.4)	775 (48.6)	820 (51.4)
Poland	1,030 (52.1)	948 (47.9)	1,468 (74.2)	510 (25.8)	767 (50.0)	766 (50.0)
Germany	1,281 (54.4)	1,073 (45.6)	1,351 (57.4)	1,003 (42.6)	575 (30.4)	1,319 (69.6)
The Netherlands	625 (50.0)	624 (50.0)	480 (38.4)	769 (61.6)	188 (19.8)	763 (80.2)
Iceland	1,009 (52.4)	917 (47.6)	1,700 (88.3)	226 (11.7)	328 (20.5)	1,269 (79.5)

Illustration 3.1 Table 1 taken from Tsitsika, A., Janikian, M., Schoenmakers, M.T., et al. ,2014 .

We have a sample 13,284 observations, from which 52,7% are Female and 47,3% Male, 61,4% are 14 to 15 years old and 38,6 are 16 to 17 years old and the parent's education level is low/middle (37,3%)and high (62,7%). The Functional Internet behavior category is 86,1% of the sample and the Dysfunctional Internet behavior is 13.9% of the sample. In the dataset we have 9600 usable observations, no missing data, which are categorized as IAT_Cat, which is our dependent variable of 4 categories : 0-No signs of IAB, 1-Mild signs of IAB, 2-At risk for IAB and 3-IAB, then we have our factors, Country (Greece, Spain, Romania, Poland, Germany, Netherlands and Iceland), Age_FINAL_kat is the age groups, mainly 14 - 16 and 16 - 18 years old, Q1 is the sex factor, male and female, Q20 is the age of first contact with the Internet, edu_post_tetr is the education level of the parents, SNS is the amount of usage, >=2hours/day or <=2hours/day, Gamers is whether they play games or not. Cats3 is a new set of categories of the dependent variable IAT_Cat (4 categories) which combines the the two last categories 2 and 3 together into one, thus leaving us with three categories and Cats2 is another new set of categories of the IAT_Cat variable which combines the 0 and 1 as well as 2 and 3 categories together, as a result we have only 2 categories.

In the dataset, category 0-No signs of IAB is 46,38%, category 1-Mild signs of IAB is 39,69%, category 2-At risk for IAB is 12,93% and category 3-IAB is 1,1%. From these percentages we see why we might want to combine the categories to three or two.

TABLE 2. PERCENTAGE OF ADOLESCENTS WITH FUNCTIONAL AND DYSFUNCTIONAL INTERNET BEHAVIOR BY GENDER, AGE, PARENTAL EDUCATIONAL LEVEL, AND COUNTRY

	Functional Internet behavior (N = 11,029)		*Dysfunctional Internet behavior*[a] *(N = 1,778)*		*% Total dysfunctional Internet behavior (95% CI)*
	% No signs of IAB (95% CI)	*% Mild signs of IAB (95% CI)*	*% At risk for IAB (95% CI)*	*% IAB (95% CI)*	
All adolescents	47.2 (46.1–48.3)	38.9 (37.9–39.9)	12.7 (12.0–13.4)	1.2 (1.0–1.5)	13.9 (13.1–14.7)
Gender					
Female	48.5 (47.1–50.0)	38.8 (37.4–40.1)	11.8 (11.0–12.7)	0.9 (0.7–1.2)	12.7 (11.8–13.6)
Male	45.7 (44.3–47.2)	39.0 (37.7–40.3)	13.6 (12.7–14.6)	1.6 (1.3–2.0)	15.2 (14.2–16.3)
Age					
14–15 years	48.0 (46.7–49.4)	39 (37.7–40.2)	12 (11.2–12.8)	1.1 (0.8–1.4)	13.0 (12.2–14.0)
16–17 years	46.0 (44.3–47.6)	38.8 (37.3–40.3)	13.8 (12.7–14.9)	1.5 (1.2–1.9)	15.2 (14.2–16.4)
Parental education					
Low/middle	45.3 (43.4–47.1)	38.4 (36.7–40.1)	14.9 (13.6–16.3)	1.4 (1.1–1.9)	16.3 (14.9–17.8)
High	47.4 (46.0–48.9)	39.9 (38.6–41.3)	11.6 (10.8–12.5)	1.0 (0.8–1.4)	12.6 (11.8–13.5)
Spain	19.3 (17.2–21.6)	57.9 (55.4–60.5)	21.3 (19.2–23.5)	1.5 (0.9–2.3)	22.8 (20.6–25.1)
Romania	45.5 (42.6–48.5)	36.8 (34.3–39.3)	16.0 (13.9–18.4)	1.7 (1.1–2.4)	17.7 (15.5–20.1)
Poland	50.3 (47.9–52.8)	36.4 (34.3–38.7)	12.0 (10.5–13.7)	1.3 (0.8–1.9)	13.2 (11.6–15.0)
Greece	59.0 (56.2–61.7)	28.3 (26.0–30.8)	11.0 (9.4–12.9)	1.7 (1.1–2.5)	12.7 (10.9–14.8)
The Netherlands	45.3 (42.1–48.5)	42.5 (39.5–45.6)	11.4 (9.3–13.9)	0.8 (0.4–1.5)	12.2 (10.0–14.7)
Germany	53.9 (51.1–56.7)	35.4 (33.0–38.0)	9.7 (8.2–11.5)	0.9 (0.6–1.4)	10.6 (9.0–12.5)
Iceland	56.2 (53.5–58.9)	35.9 (33.3–38.5)	7.2 (5.9–8.7)	0.8 (0.4–1.6)	7.9 (6.4–9.7)

Illustration 3.2 Table 2 taken from Tsitsika, A., Janikian, M., Schoenmakers, M.T., et al. ,2014 .

TABLE 3. ODDS RATIOS AND 95% CONFIDENCE INTERVALS FOR RELATIONSHIPS BETWEEN ONLINE ACTIVITIES AND DYSFUNCTIONAL INTERNET BEHAVIOR

	%[a]	OR[b]	95% CI
Gambling	8.4	2.97	2.52–3.49
Social networking sites (e.g., Facebook)	92.5	2.62	1.95–3.51
Monetary prize games	15.5	2.58	2.26–2.95
Chat rooms	60.0	2.45	2.16–2.79
Internet forums	50.4	2.44	2.18–2.74
Searching for sexual information	35.4	2.40	2.16–2.67
Making personal Web sites or blogging	28.6	2.31	2.07–2.59
Instant messaging (e.g., MSN)	85.6	2.29	1.86–2.81
Downloading movies	66.9	2.16	1.87–2.50
Downloading music	87.8	2.15	1.72–2.70
Real-time strategy games	33.8	2.11	1.89–2.36
Downloading games	51.3	2.00	1.79–2.24
Downloading software	61.2	2.00	1.78–2.25
Multiplayer role-playing games	44.8	1.82	1.63–2.04
Searching for medical information	43.2	1.80	1.62–2.00
Shooter games	48.0	1.66	1.49–1.86
E-mail	84.3	1.57	1.31–1.87
News sites	74.5	1.48	1.31–1.69
Hobbies	82.7	1.42	1.22–1.66
Purchasing goods	53.1	1.31	1.17–1.47
Single-player games (e.g., solitaire, backgammon)	64.2	1.26	1.13–1.41
Watching videos or movies	97.5	1.01	0.68–1.48
Doing homework or research	93.3	0.68	0.57–0.83

[a]Proportion of adolescents reporting the specific activity at least weekly.

[b]All ORs are statistically highly significant ($p < 0.001$) except for "watching videos or movies" ($p = 0.98$).

ORs, odds ratios.

Illustration 3.3 Table 3 taken from Tsitsika, A., Janikian, M., Schoenmakers, M.T., et al. ,2014 .

TABLE 4. FACTORS INDEPENDENTLY ASSOCIATED WITH DYSFUNCTIONAL INTERNET BEHAVIOR IN MULTIPLE LOGISTIC REGRESSION: ADJUSTED ODDS RATIO AND 95% CONFIDENCE INTERVAL

	OR (95% CI)	p
Country		
Spain	4.94 (3.71–6.58)	<0.001
Romania	2.68 (1.94–3.70)	<0.001
Poland	1.94 (1.43–2.61)	<0.001
The Netherlands	1.80 (1.25–2.59)	0.002
Greece	1.54 (1.12–2.11)	0.008
Germany	1.44 (1.04–1.99)	0.030
Iceland	1.00[a]	
Gender		
Female	1.00	
Male	0.95 (0.82–1.10)	0.49
Age		
14–15 years	1.00	
16–17 years	1.01 (0.88–1.16)	0.89
Parental education		
Low/middle	1.00	
High	0.85 (0.74–0.98)	0.028
Age at first use of the Internet (years)	0.94 (0.91–0.98)[b]	0.001
Daily use of SNSs		
No use/<2 hours	1.00	
≥2 hours/day	3.47 (3.05–3.94)	<0.001
Average hours of playing games per weekday		
No gaming/<2 hours	1.00	
≥2 hours/day	2.34 (1.99–2.75)	<0.001

[a]Indicates reference category.
[b]For 1-year increase.
SNSs, social networking sites.

Illustration 3.4 Table 4 taken from Tsitsika, A., Janikian, M., Schoenmakers, M.T., et al. ,2014 .

Female, IAT_Cat = No signs of IAB			Male,No signs of IAB		
Country	16-17.9	>=18	16-17.9	>=18	
Greece	364	159	319	122	
Spain	110	48	96	44	
Romania	91	272	48	211	
Poland	248	102	244	66	
Germany	261	224	202	152	
Netherlands	67	106	55	103	Subtotal
Iceland	341	35	323	39	46,38%
, IAT_Cat = Mild signs of IAB			Mild signs of IAB		
Greece	166	61	164	75	
Spain	328	155	315	161	
Romania	81	210	64	157	
Poland	195	73	152	61	
Germany	185	125	145	117	
Netherlands	74	105	61	103	Subtotal
Iceland	233	34	179	31	39,69%
IAT_Cat = At risk for IAB			At risk for IAB		
Greece	59	29	53	23	
Spain	120	70	92	58	
Romania	24	79	32	91	
Poland	50	21	60	24	
Germany	42	17	61	36	
Netherlands	18	26	17	31	Subtotal
Iceland	41	8	48	2	12,83%
IAT_Cat = IAB			IAB		
Greece	5	0	13	4	
Spain	6	7	8	3	
Romania	0	8	2	6	
Poland	4	1	7	2	
Germany	1	2	3	7	
Netherlands	1	2	0	5	Subtotal
Iceland	2	0	4	3	1,10%

Table 3.5 Brief statistics of our dataset.

4 Methods and tools

The tools that were used were R v4.0.2(2020-06-22), R Studio v1.13.1056(2020-07-07) and SPSS 21. The data is from Tsitsika A. et al. (2014). As noted in Chapter 3, preliminary analysis indicated that we could reduce the number of categories of the dependent variable to 3 or even 2, because the fourth category(IAB) was very small.

4.1 Summaries of analyses

4.1.1 Binomial Logit

The binomial Logit model was fitted with glm() from the stats library with the binomial family distribution. The function is Cats2 = Country + Q1 + Q20 + age_FINAL_kat + edu_post_tetr + SNS + gamers. The factors and the reference categories are Cats2 : Category 0 together with category 1 represents “No signs of IAB” and “Mild signs of IAB” make up the reference category 0, category 2 “At risk for IAB “ and category 3 “IAB” added together they are category 1, Country: Greece(reference), Spain, Romania, Poland, Germany, Netherlands and Iceland, gender : female(reference) and male, Q20 (Age at first use of the Internet (years)): the youngest age is the reference, age_FINAL_kat (is the age groups): 14 - 16(reference) and 16 - 18 years old, edu_post_tetr (parent's level of education): low/middle (primary or secondary school) (reference) and high (postsecondary or tertiary education) educational level, SNS: moderate SNS use (<2 hours daily) (reference) and heavier SNS use (>=2 hours daily), gamers (Average hours of playing games per weekday): “No gaming/<2 hours” (reference) and “>2 hours/day”.

FORMULA	Country + Q1 + Q20 + age_FINAL_kat + edu_post_tetr + SNS + gamers							
FAMILY	binomial							
AIC	7075,42							
DEVIANCE	7049,42							
Log Likelihood	-3524.712 (df=13)							
	Estimate	Std. Erro	z value	Pr(>|z|)	exp(b)	2,50%	97,50%	
(Intercept)	-2,1433	0,2062	-10,3966	2,56995140246232e-25	0,1173	0,0781	0,1753	***
CountrySpain	0,8790	0,1054	8,3401	7,42007066491754e-17	2,4084	1,9614	2,9652	***
CountryRomania	0,4695	0,1133	4,1432	3,42443968730127e-05	1,5992	1,2814	1,9983	***
CountryPoland	0,1377	0,1216	1,1322	0,2576	1,1476	0,9039	1,4564	
CountryGermany	-0,2058	0,1193	-1,7252	0,0845	0,8140	0,6439	1,0282	.
CountryNetherlands	0,0355	0,1441	0,2462	0,8056	1,0361	0,7793	1,3717	
CountryIceland	-0,5994	0,1424	-4,2093	2,56130015515506e-05	0,5491	0,4145	0,7247	***
Q1Male	0,2092	0,0688	3,0402	0,0024	1,2326	1,0773	1,4108	**
Q20	-0,0728	0,0146	-4,9745	6,54100722135478e-07	0,9297	0,9035	0,9568	***
age_FINAL_kat16-17.9	0,0146	0,0679	0,2146	0,8301	1,0147	0,8880	1,1587	
edu_post_tetrHigh	-0,1880	0,0644	-2,9175	0,0035	0,8286	0,7304	0,9404	**
SNS>= 2 hours/day	1,2871	0,0635	20,2693	2,40076323362427e-91	3,6221	3,1998	4,1043	***
gamersYes	0,3589	0,0741	4,8467	1,25526261892998e-06	1,4318	1,2389	1,6563	***

Table 4.1.1.1 Binomial Logit

The binomial function is 2.1.1,

$$logit(\pi_i) = \frac{log\pi_i}{(1-\pi_i)} = \chi^T \beta$$

Our p-values show that almost all countries are statistically significantly different from Greece at the 0,05 level, except Poland and Netherlands. The age group 16-18 is significant. The factors with the highest odds are Spain (2,40 times), Romania (1,5 times), SNS (3,62 times) and gamers (1,4 times). The factors with the lowest odds and therefore reduce the possibility of IAB is Germany (0,81 times), Iceland (0,54 times) and parents' level of education. The other factor all have odds ratios close to one so they have a small effect.

Binomial Logistic				
Predicted		0	1	%Correct
cats2	0	8252	10	99,879%
	1	1329	9	0,673%
		Overall	%	86,052%

Table 4.1.1.2Binomial Contingency Table

4.1.2 Multinomial Logit

The multinomial logit model used multinom() from the nnet library, the nnet's package multinom() function did not give the best or desired format of results, since it uses a different estimation method, maybe VGAM's multinomial family argument for vglm() is a better choice.

FORMULA	MULTINOMAL REGRESSION						
Cats3=Country + Q1 + Q20 + age_FINAL_kat + edu_post_tetr + SNS + gamers							
DEVIANCE	17236,31						
Log Likelihood	-8618,15	(df=26)					
AIC:	17288,31				INTERVALS		
	1	2	1 exp(b)	2 exp(b)	X97.5...1	X2.5...2	X97.5...2
(Intercept)	-0,3620	-1,1014	0,6963	0,3324	0,8601	0,2470	0,4473
CountrySpain	2,0381	2,1341	7,6759	8,4491	9,2040	6,6436	10,7453
CountryRomania	0,5722	0,7434	1,7721	2,1031	2,1082	1,6555	2,6716
CountryPoland	0,4030	0,3222	1,4962	1,3802	1,7818	1,0710	1,7787
CountryGermany	0,2275	-0,1189	1,2555	0,8879	1,4808	0,6932	1,1372
CountryNetherlands	0,6543	0,3503	1,9239	1,4195	2,3665	1,0471	1,9244
CountryIceland	0,0225	-0,6231	1,0228	0,5363	1,2346	0,3993	0,7202
Q1Male	-0,0213	0,1827	0,9790	1,2005	1,0874	1,0356	1,3916
Q20	-0,0717	-0,1136	0,9308	0,8926	0,9523	0,8648	0,9214
age_FINAL_kat14-15.9	-0,1596	-0,5481	0,8525	0,5780	0,9531	0,4941	0,6762
age_FINAL_kat16-17.9	-0,2024	-0,5533	0,8168	0,5751	0,9244	0,4834	0,6840
age_FINAL_kat>=18	0,0000	0,0000	1,0000	1,0000	NA	1,0000	1,0000
edu_post_tetrHigh	-0,0003	-0,1903	0,9997	0,8267	1,1051	0,7196	0,9498
SNS>= 2 hours/day	1,0945	1,8923	2,9878	6,6344	3,3020	5,7729	7,6245
gamersYes	0,3180	0,5452	1,3744	1,7250	1,5337	1,4717	2,0220

Table4.1.2.1 Multinomial Regression

Multinomial					
Predicted		0	1	2	%Correct
cats3	0	3341	1106	5	75,045%
	1	1649	2157	4	56,614%
	2	425	909	4	0,299%
				Overall %	57,313%

Table 4.1.2.2 Multinomial Model Contingency Table

The equation for this model is Cats3=Country + Q1 + Q20 + age_FINAL_kat + edu_post_tetr + SNS + gamers. The multinomial logit function is 2.5.2,

$$\log\left(\frac{\pi_j}{\pi_J}\right) = aj + b_j x\,,\ j = 1,\ldots,J-1$$

Except for Cats3 the other reference categories categories are the same, Cats3: Category 0 represents “No signs of IAB” (reference 1), category 1 is “Mild signs of IAB”, category 2 “At risk for IAB “ and category 3 is “IAB” make up a new category 2. Again all countries are statistically significant except Iceland. Parents education level is not statistically significant as well the gender

variable. For the first equation, the factors with the highest odds are Spain (7,67 times), all countries except Iceland (1,02 times) ranging from 1,92 to 1,25 times, gamers (1,37) and SNS (2,98 times), while the lowest or reducing are the age groups from 14 to 18 years old (0,81-0,85 times). The other factors all close to one having little effect. These odds are for the category 1.

In the second equation, the factors with the highest odds are Spain (8,44 times), all countries except Iceland (0,53 times) and Germany (0,88 times) ranging from 2,3 to 1,3 times, gamers (1,72) and SNS (6,63 times), while the lowest or reducing are the age groups from 14 to 18 years old (0,57 times), Q20 (0,89) and parent's education (0,82). The other factors all close to one having little effect. These odds are for the category 2.

4.1.3 Ordinal Logit

The ordinal logit model used polr() from the MASS library. The ordinal logit function is 2.2.1,

$$logit[P(Y \leq j)] = \log[\frac{P(Y \leq j)}{1-(P(Y \leq j))}] = \log[\frac{\pi_1 + ... + \pi_j}{\pi_j + ... + \pi_J}]$$

All countries except Germany are statistically significant. The gender, parent's education and age variables are not statistically significant. The reference categories are the same as stated in the multinomial logit regression section.

FORMULA	ORDiNAL LOGISTIC						
Cats3=Country + Q1 + Q20 + age_FINAL_kat + edu_post_tetr + SNS + gamers							
DEVIANCE		17391,88					
ZETA	0|1	0,1590					
	1|2	2,4293					
AIC:	17419.88						
Log Likelihood	-8695.94	(df=14)			INTERVALS	2.5 %	97.5 %
	Value	Std. Error	t value	exp(b)	(Intercept)	0,0781	0,1753
CountrySpain	1,5812	0,0730	21,6714	4,8608	CountrySpain	1,9614	2,9652
CountryRomania	0,6066	0,0775	7,8308	1,8343	CountryRomania	1,2814	1,9983
CountryPoland	0,3433	0,0792	4,3362	1,4096	CountryPoland	0,9039	1,4564
CountryGermany	0,0803	0,0752	1,0682	1,0837	CountryGermany	0,6439	1,0282
CountryNetherlands	0,4716	0,0926	5,0939	1,6026	CountryNetherlan	0,7793	1,3717
CountryIceland	-0,2010	0,0858	-2,3434	0,8179	CountryIceland	0,4145	0,7247
Q1Male	0,0795	0,0457	1,7412	1,0828	Q1Male	1,0773	1,4108
Q20	-0,0796	0,0099	-8,0387	0,9234	Q20	0,9035	0,9568
age_FINAL_kat16-17.9	0,0206	0,0452	0,4557	1,0208	age_FINAL_kat16	0,8880	1,1587
edu_post_tetrHigh	-0,0861	0,0435	-1,9771	0,9175	edu_post_tetrHigh	0,7304	0,9404
SNS>= 2 hours/day	1,2789	0,0432	29,6109	3,5927	SNS>= 2 hours/d	3,1998	4,1043
gamersYes	0,3433	0,0477	7,1941	1,4096	gamersYes	1,2389	1,6563
0|1	0,1590	0,1434	1,1087	1,1724			
1|2	2,4293	0,1459	16,6475	11,3513			

Table 4.1.3.1 Ordinal Logit

For the first equation, the factors with the highest odds are Spain (4,86 times), Romania (1,83 times), Netherlands (1,62 times) and Poland (1,4 times), gamers (1,4 times) and SNS (3,59 times), while the lowest or reducing are Iceland (0,83 times) and parents' education (0,91 times). The other factors all close to one having little effect.

Ordinal					
Predicted		0	1	2	%Correct
cats3	0	3246	1186	20	72,911%
	1	1539	2116	155	55,538%
	2	373	847	118	8,819%
				Overall %	57,083%

Table 4.1.3.2 Ordinal Model Contingency Table

4.1.4 Adjacent Categories Logits

The adjacent category together with the continuation ratio model used vglm() from the VGAM library and the acat() family. The function is (2.4.2),

$$\log\left(\frac{\pi_{j+1}}{\pi_j}\right)=\alpha_j+\beta_j\chi,\ j=1,\ldots,J-1$$

Our equation is cats3 = Country + Q1 + Q20 + age_FINAL_kat + edu_post_tetr + SNS + gamers. The reference categories are the same as stated in the multinomial logit regression section. All variables are statistically significant except one country, Iceland, and the age, gender and parent's education variables.

Adjacent Categories									
vglm cats3 ~ Country + Q1 + Q20 + age_FINAL_kat + edu_post_tetr + SNS + gamers acat(link = "loglink",									
parallel = FALSE, reverse = FALSE, zero = NULL, whitespace = FAL									
AIC	17288,31								
deviance	17236,31								
Log Likelihood	-8618,15								
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**exp(b)**	**INTERVALS**	**2.5 %**	**97.5 %**	
(Intercept):1	-0,5216	0,1604	-3,2512	0,0011	0,5935	(Intercept):1	0,4334	0,8129	**
(Intercept):2	-1,1278	0,2180	-5,1723	2,3118346175	0,3237	(Intercept):2	0,2111	0,4964	***
CountrySpain:1	2,0381	0,0926	22,0026	2,7198577178	7,6758	CountrySpain:1	6,4015	9,2039	***
CountrySpain:2	0,0959	0,1122	0,8549	0,3926	1,1007	CountrySpain:2	0,8833	1,3715	
CountryRomania:1	0,5722	0,0886	6,4572	1,0666390496	1,7721	CountryRomania:1	1,4896	2,1082	***
CountryRomania:2	0,1712	0,1227	1,3958	0,1628	1,1868	CountryRomania:2	0,9331	1,5094	
CountryPoland:1	0,4030	0,0891	4,5214	6,1423603314	1,4962	CountryPoland:1	1,2564	1,7818	***
CountryPoland:2	-0,0808	0,1311	-0,6162	0,5378	0,9224	CountryPoland:2	0,7134	1,1926	
CountryGermany:1	0,2275	0,0842	2,7018	0,0069	1,2555	CountryGermany:1	1,0645	1,4808	**
CountryGermany:2	-0,3464	0,1281	-2,7046	0,0068	0,7072	CountryGermany:2	0,5502	0,9090	**
CountryNetherlands:1	0,6543	0,1056	6,1940	5,8645796639	1,9239	CountryNetherlands:1	1,5641	2,3665	***
CountryNetherlands:2	-0,3040	0,1528	-1,9903	0,0466	0,7378	CountryNetherlands:2	0,5469	0,9954	*
CountryIceland:1	0,0225	0,0960	0,2347	0,8144	1,0228	CountryIceland:1	0,8473	1,2346	
CountryIceland:2	-0,6456	0,1514	-4,2649	1,9999988455	0,5243	CountryIceland:2	0,3897	0,7055	***
Q1Male:1	-0,0213	0,0536	-0,3966	0,6917	0,9790	Q1Male:1	0,8813	1,0874	
Q1Male:2	0,2040	0,0722	2,8236	0,0047	1,2263	Q1Male:2	1,0644	1,4128	**
Q20:1	-0,0717	0,0117	-6,1444	8,0269239988	0,9308	Q20:1	0,9098	0,9523	***
Q20:2	-0,0419	0,0154	-2,7204	0,0065	0,9590	Q20:2	0,9305	0,9884	**
age_FINAL_kat16-17.9:	-0,0428	0,0533	-0,8032	0,4219	0,9581	age_FINAL_kat16-17	0,8631	1,0636	
age_FINAL_kat16-17.9:1	0,0377	0,0713	0,5283	0,5973	1,0384	age_FINAL_kat16-17	0,9030	1,1940	
edu_post_tetrHigh:1	-0,0003	0,0511	-0,0052	0,9959	0,9997	edu_post_tetrHigh:1	0,9044	1,1051	
edu_post_tetrHigh:2	-0,1900	0,0679	-2,8004	0,0051	0,8269	edu_post_tetrHigh:2	0,7240	0,9446	**
SNS>= 2 hours/day:1	1,0945	0,0510	21,4496	4,6026949459	2,9878	SNS>= 2 hours/day:1	2,7034	3,3020	***
SNS>= 2 hours/day:2	0,7977	0,0673	11,8593	1,9265080860	2,2205	SNS>= 2 hours/day:2	1,9462	2,5334	***
gamersYes:1	0,3180	0,0560	5,6820	1,3316517789	1,3744	gamersYes:1	1,2316	1,5337	***
gamersYes:2	0,2272	0,0776	2,9264	0,0034	1,2551	gamersYes:2	1,0779	1,4615	**

Table 4.1.4.1 Adjacent Categories Logits

For the first equation, the factors with the highest odds are Spain (7,67 times), Romania (1,77 times), Netherlands (1,92 times) and Poland (1,49 times), gamers (1,37 times) and SNS (2,98 times), while the lowest or reducing are age, gender, Q20 and parents' education all ranging from

0,9 to 1 times. Almost all factors for the second equation are lower than the first, except for age, gender and Q20 which are higher.

Adjacent					
Predicted		0	1	2	%Correct
cats3	0	3341	1106	5	75,045%
	1	1649	2157	4	56,614%
	2	425	909	4	0,299%
				Overall %	57,313%

Table 4.1.4.2 Adjacent-categories Model Contigency Table

4.1.5 Continuation Ratio Logit

The continuation ratio model used vglm() from the VGAM and the cratio() family. The function is 2.3.2,

$$\log\left(\frac{\pi_1}{\pi_2+\ldots+\pi_j}\right), \log\left(\frac{\pi_2}{\pi_3+\ldots+\pi_j}\right), \ldots, \log\left(\frac{\pi_{j-1}}{\pi_j}\right),$$

Our equation is cats3 = Country + Q1 + Q20 + age_FINAL_kat + edu_post_tetr + SNS + gamers. The reference categories are the same as stated in the multinomial logit regression section. All variables are statistically significant except two countries, Germany and Iceland, as well as the variables for age, gender and parent's education.

Continuation Ratio									
vglm cats3 ~ Country + Q1 + Q20 + age_FINAL_kat + edu_post_tetr + SNS + gamers cratio(link = "logitlink",									
parallel = FALSE, reverse = FALSE, zero = NULL, whitespace = FALSE)									
AIC	17283,79								
deviance	17231,79								
loglikelihood	-8615,89								
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**exp(b)**	**INTERVALS**	2.5 %	97.5 %	
(Intercept):1	-0,2171	0,1518	-1,4306	0,1525	0,8048	(Intercept):1	0,5977	1,0837	
(Intercept):2	-1,1593	0,2200	-5,2691	1,3706620713475e-07	0,3137	(Intercept):2	0,2038	0,4828	***
CountrySpain:1	2,0534	0,0884	23,2347	2,03305083782931e-11	7,7942	CountrySpain:1	6,5546	9,2682	***
CountrySpain:2	0,1104	0,1133	0,9739	0,3301	1,1167	CountrySpain:2	0,8943	1,3944	
CountryRomania:1	0,6153	0,0821	7,4917	6,80053127593818e-14	1,8502	CountryRomania:1	1,5751	2,1734	***
CountryRomania:2	0,1868	0,1241	1,5051	0,1323	1,2054	CountryRomania:2	0,9451	1,5373	
CountryPoland:1	0,3780	0,0831	4,5497	5,37270249791853e-06	1,4593	CountryPoland:1	1,2400	1,7173	***
CountryPoland:2	-0,0606	0,1324	-0,4580	0,6469	0,9412	CountryPoland:2	0,7260	1,2200	
CountryGermany:1	0,1463	0,0788	1,8569	0,0633	1,1575	CountryGermany:1	0,9919	1,3508	.
CountryGermany:2	-0,3444	0,1296	-2,6564	0,0079	0,7087	CountryGermany:2	0,5497	0,9137	**
CountryNetherlands:1	0,5796	0,1000	5,7983	6,69735091798919e-09	1,7852	CountryNetherlands:	1,4676	2,1716	***
CountryNetherlands:2	-0,2760	0,1537	-1,7953	0,0726	0,7588	CountryNetherlands:	0,5614	1,0256	.
CountryIceland:1	-0,1181	0,0906	-1,3039	0,1923	0,8886	CountryIceland:1	0,7441	1,0612	
CountryIceland:2	-0,6231	0,1526	-4,0844	4,41893955777176e-05	0,5363	CountryIceland:2	0,3977	0,7232	***
Q1Male:1	0,0253	0,0509	0,4969	0,6193	1,0256	Q1Male:1	0,9282	1,1332	
Q1Male:2	0,2337	0,0729	3,2066	0,0013	1,2632	Q1Male:2	1,0951	1,4572	**
Q20:1	-0,0818	0,0111	-7,3806	1,57565652544356e-13	0,9214	Q20:1	0,9016	0,9417	***
Q20:2	-0,0420	0,0155	-2,7035	0,0069	0,9589	Q20:2	0,9302	0,9885	**
age_FINAL_kat16-17.9:	-0,0339	0,0506	-0,6686	0,5037	0,9667	age_FINAL_kat16-17	0,8754	1,0676	
age_FINAL_kat16-17.9:	0,0395	0,0718	0,5507	0,5819	1,0403	age_FINAL_kat16-17	0,9037	1,1976	
edu_post_tetrHigh:1	-0,0446	0,0484	-0,9209	0,3571	0,9564	edu_post_tetrHigh:1	0,8698	1,0516	
edu_post_tetrHigh:2	-0,2020	0,0683	-2,9555	0,0031	0,8171	edu_post_tetrHigh:2	0,7147	0,9342	**
SNS>= 2 hours/day:1	1,2884	0,0483	26,6868	6,69457241574626e-157	3,6270	SNS>= 2 hours/day:	3,2995	3,9870	***
SNS>= 2 hours/day:2	0,8217	0,0677	12,1409	6,40802161317824e-34	2,2744	SNS>= 2 hours/day:	1,9919	2,5971	***
gamersYes:1	0,3670	0,0534	6,8692	6,4561727196591e-12	1,4434	gamersYes:1	1,2999	1,6028	***
gamersYes:2	0,2236	0,0782	2,8585	0,0043	1,2505	gamersYes:2	1,0728	1,4577	**

Table 4.1.5.1 Continuation Ratio Logit

Signif. Codes: 0 '***' ,0.001 '**', 0.01 '*', 0.05 '.', 0.1 '

For the first equation, the factors with the highest odds are Spain (7,79 times), Romania (1,85 times), Netherlands (1,78 times) and Poland (1,45 times), gamers (1,44 times) and SNS (3,62 times), while the lowest or reducing are age, gender, Q20 and parents' education all ranging from 0,8 to 1 times. Almost all factors for the second equation are lower than the first, except for age, gender and Q20 which are higher.

Continuation					
Predicted		0	1	2	%Correct
cats3	0	3351	1095	6	75,270%
	1	1649	2157	4	56,614%
	2	427	905	6	0,448%
				Overall %	57,438%

Table 4.1.5.2 Continuation-Ratio Model Contigency Table

4.2 Comparisons

According to Akaike Information Criterion the order of the models from best to worst is Multinomial (17288,31), Continuation Ratio (17283,79), Adjacent Categories (17288,31) and then Ordinal (17419,88). The Log-Likelihood criterion orders the models as such: Continuation Ratio (-8615,89), Adjacent Categories (-8618,15), Multinomial (-8618,15) and then Ordinal (-8695,94). The binomial logit model can't be compared, the equation is different, but provides a simple reference, overall it has the best criteria. By using classification/contigency tables, we see that only the ordinal model managed to predict category 2 better than the others, while at the same time it has the worst overall percentage of correct predictions (57,083%). The multinomial and the adjacent categories model are exactly the same in terms of overall and individual category predictions (57,313%) ,while the best predictive capability is displayed by the continuation-ratio model (57,438%). The binomial logistic model has the best overall prediction, because it has only 2 categories. Binomial estimations are 86,052% correct.

Binomial Logistic				
Predicted		0	1	%Correct
cats2	0	8252	10	99,879%
	1	1329	9	0,673%
		Overall %		86,052%

Multinomial					
Predicted		0	1	2	%Correct
cats3	0	3341	1106	5	75,045%
	1	1649	2157	4	56,614%
	2	425	909	4	0,299%
				Overall %	57,313%

Continuation					
Predicted		0	1	2	%Correct
cats3	0	3351	1095	6	75,270%
	1	1649	2157	4	56,614%
	2	427	905	6	0,448%
				Overall %	57,438%

Ordinal					
Predicted		0	1	2	%Correct
cats3	0	3246	1186	20	72,911%
	1	1539	2116	155	55,538%
	2	373	847	118	8,819%
				Overall %	57,083%

Adjacent					
Predicted		0	1	2	%Correct
cats3	0	3341	1106	5	75,045%
	1	1649	2157	4	56,614%
	2	425	909	4	0,299%
				Overall %	57,313%

Table 4.2.1 Contingency Table Summary

					Mult 1		Adj 1		Cont 1		Adj 2		Cont 2		Mult 2	
	binomial		order		exp(b)		exp(b)		exp(b)		exp(b)		exp(b)		exp(b)	
(Intercept)	exp(b)		exp(b)		0,6963	***	0,5935	**	0,8048		0,3237	***	0,3137	***	0,3324	
CountrySpain	0,1173	***	4,8608	***	7,6759	***	7,6758	***	7,7942	***	1,1007		1,1167		8,4491	***
CountryRomania	2,4084	***	1,8343	***	1,7721	***	1,7721	***	1,8502	***	1,1868		1,2054		2,1031	***
CountryPoland	1,5992	***	1,4096	***	1,4962	***	1,4962	***	1,4593	***	0,9224		0,9412		1,3802	***
CountryGermany	1,1476		1,0837		1,2555	**	1,2555	**	1,1575	.	0,7072	**	0,7087	**	0,8879	
CountryNetherlands	0,8140	.	1,6026	***	1,9239	***	1,9239	***	1,7852	***	0,7378	*	0,7588	.	1,4195	***
CountryIceland	1,0361		0,8179	*	1,0228		1,0228		0,8886		0,5243	***	0,5363	***	0,5363	***
Q1Male	0,5491	***	1,0828		0,9790		0,9790		1,0256		1,2263	**	1,2632	**	1,2005	***
Q20	1,2326	**	0,9234	***	0,9308	***	0,9308	***	0,9214	***	0,9590	**	0,9589	**	0,8926	***
age_FINAL_kat16-1	0,9297	***	1,0208		0,8120	**	0,9581		0,9667		1,0384		1,0403		0,5780	*
edu_post_tetrHigh	1,0147		0,9175		1,0000	***	0,9997		0,9564		0,8269	**	0,8171	**	0,5751	**
SNS>= 2 hours/day	0,8286	**	3,5927	***	2,9878	***	2,9878	***	3,6270	***	2,2205	***	2,2744	***	1,0000	***
gamersYes	3,6221	***	1,4096	***	1,3744	***	1,3744	***	1,4434	***	1,2551	**	1,2505	**	0,8267	***
Y1	1,4318	***	1,1724												6,6344	
Y2			###												1,7250	

Table 4.2.2 Coefficients Comparison

Signif. Codes: 0 '***' ,0.001 '**', 0.01 '*', 0.05 '.', 0.1 '

The coefficients that are significant in all models are the countries: Spain, Romania, Poland, Netherlands, Q20, SNS >=2 hours a day and being gamers. The dissimilarities are that Germany is significant only in the multinomial, adjacent-categories and continuation-ratio model, Iceland is significant only in the ordinal model, being male was significant only in the binomial model. The age variable is significant only in the binomial and multinomial models. The parent's education level was insignificant in all cases.

In between model comparisons show that, in the second Adjacent-ratio's equation versus the first one, Spain, Romania and Poland become statistically insignificant while Iceland becomes significant and the rest of the countries remain significant. The gender and the parent's education level become significant, while the rest factors remain the same between the two equations. In the Continuation Logit model's second equation we observe Spain, Romania, Poland and Netherlands, although barely, become statistically insignificant, while Iceland becomes significant and the rest of the countries remain significant. The gender and the parent's education level become significant, while the rest factors' significance remains the same between the two equations. In the Multinomial logit model's second equation, Poland becomes statistically insignificant and Iceland becomes significant, all other countries significance is unchanged. The gender and the parent's education level become significant, while the rest factors' significance remains the same between the two equations. As for the effect of the factors in the first equations, starting from Spain which has the highest odds (~7,7 times) and in decreasing order Netherlands, Romania, Poland, Germany and lastly Iceland (~1 times). The other factors don't have a big effect except for the gaming and SNS

factors, ~1,3 times and ~3 times on average respectively bigger odds. In the second equation of the Adjacent-ratio and the continuation logit model, starting from Romania which has the highest odds (~1,1 times) and in decreasing order Spain, Poland Netherlands,Germany and lastly Iceland(~0,53 times).The other factors don't have a big effect (~1 times) except for the gaming and SNS factors, 2,25 times and 1,25 times on average respectively bigger odds. There is a notable difference in comparison of the Multinomial logits second equation, where Spain has a very high odd of 8,45 times, and slightly increased odd of most countries except Iceland's which remains the same. An decreased odd for the age (0,58 times) and parent's education (0,58 times) factor and a much decreased odd of the SNS (1 times) and gamers (0,83 times) factor. The binomial logit model, orders the countries in their effect as such, Romania (2,41 times), Poland, Germany, Iceland, Netherlands and Spain (0,12 times), the gender factor has an effect of 0,55 times and the Q20 factors 1,23 times. Age and parents' education have a small factor effect of 0,93 and 1,04 times, while SNS and gamers have an odd of 0,83 and 3,62 times, SNS and Spain factors have decreased odds compared to all other models, in contrast gamers' factor odds are higher than all other models. The ordered/ordinal logit model, orders the countries in their effect as such, Spain (4,86 times), Romania, Poland, Netherlands, Germany and Iceland (0,82 times), the gender factor has an effect of 1,08 times and the Q20 factor 0,92 times. Age and parents education have a small factor effect of 1,02 and 0,92 times, while SNS and gamers have an odd of 3,59 and 1,41 times.

5 Conclusions

The best predictive capability is displayed by the continuation-ratio model (57,438%), the best AIC model is the Multinomial (17288,31), with a close second the Continuation Ratio (17283,79) and the best log-likelihood model is the Continuation Ratio (-8615,89). Therefore, we can say that the best model is the Continuation Ratio logit, due to it having the best predictive capability and the comparative statistics are also better than most, but secondary in nature. One of the observations made is that SPSS has a limited amount of models that can be used for categorical analysis, binomial ,multinomial and proportional odds, which severely limits us in our work, while R doesn't have this issue.

References

1. Tsitsika, A., Janikian, M., Shoenmakers, T., Tzavela, E.C., Olaffson, K., Wojcik, S., Macarie, G.F., Tzavara C., The EU NET ADB Consortium, and Richardson, C.,(2014). Internet Addictive Behavior in Adolescence: A Cross-Sectional Study in Seven European Countries, Cyberpsychology, behavior, and social networking, 17, 528-535.
2. Agresti, A. (2002) Categorical Data. Second edition. Wiley.
3. Dobson, A. J. (1990) An Introduction to Generalized Linear Models. London: Chapman and Hall
4. Hastie, T. J. and Pregibon, D. (1992) Generalized linear models. Chapter 6 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.
5. McCullagh P. and Nelder, J. A. (1989) Generalized Linear Models, London: Chapman and Hall.
6. Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S, New York: Springer.

Appendix A

```
wd1="E:/My Documents/Πανεπηστημιο/metaptixiako/3ο εξαμηο/8-20/"

setwd(wd1)

library(foreign)

library(nnet)

internet=read.spss("file for analysis.sav", use.value.labels = TRUE)

View(internet)

attach(internet)

library(MASS)##polr

library(stats)##multinom glm confint

library(VGAM)

#invisible(lapply(paste0("package:", names(sessionInfo()$otherPkgs)),detach,character.only =
TRUE, unload = TRUE))    # Unload add-on packages

#missing values?

is.na(internet)

#regressions

LogRegrGlmCats2=glm(cats2~Country+Q1+Q20+age_FINAL_kat+edu_post_tetr+SNS+gamers,fa
mily="binomial")

#LogRegrGlmCats3=glm(cats3~Country+Q1+Q20+age_FINAL_kat+edu_post_tetr+SNS+gamers,f
amily="binomial")

#OrdRegr1Cats2=polr(as.factor(cats2)~Country+Q1+Q20+age_FINAL_kat+edu_post_tetr+SNS+g
amers)

OrdRegr1Cats3=polr(as.factor(cats3)~Country+Q1+Q20+age_FINAL_kat+edu_post_tetr+SNS+ga
mers,Hess=TRUE,model = TRUE)

MultiRegrCats3=multinom(cats3~Country+Q1+Q20+age_FINAL_kat+edu_post_tetr+SNS+gamers
,Hess=TRUE)

#MultiReGLCats3=vglm(cats3~Country+Q1+Q20+age_FINAL_kat+edu_post_tetr+SNS+gamers,f
amily = multinomial)


summary#MultGARegCats3=vgam(cats3~Country+Q1+Q20+age_FINAL_kat+edu_post_tetr+SNS
+gamers,multinomial)

#library(mgcv)##multinomial gam,NEVER LOAD IT!

#MultGARegCats3=gam(cats3~Country+Q1+Q20+age_FINAL_kat+edu_post_tetr+SNS+gamers,f
amily=multinom(K=2),data=internet)

#detach(package:mgcv,unload = TRUE)

ContRatioCats3=vglm(cats3~Country+Q1+Q20+age_FINAL_kat+edu_post_tetr+SNS+gamers,fam
ily=cratio(link="logitlink",parallel=FALSE,reverse = FALSE, zero = NULL,whitespace = FALSE))

AdjucCatsCats3=vglm(cats3~Country+Q1+Q20+age_FINAL_kat+edu_post_tetr+SNS+gamers,fa
```

```
mily=acat(link = "loglink", parallel = FALSE, reverse = FALSE,zero = NULL, whitespace =
FALSE),)

##regression summaries/results
logout1=summary(LogRegrGlmCats2)
#logout2=summary(LogRegrGlmCats3)
#ordout1=summary(OrdRegr1Cats2)
ordout2=summary(OrdRegr1Cats3)
multout=summary(MultiRegrCats3)
#multGAM=summary(MultGARegCats3)#error fixed with VGAM
contout=summary(ContRatioCats3)
adjaout=summary(AdjucCatsCats3)#error fixed somehow idk

##confidence intervals
bL1=exp(confint(LogRegrGlmCats2))
#exp(confint(LogRegrGlmCats3))
bO2=exp(confint.default(OrdRegr1Cats3))
bMu=exp(confint(MultiRegrCats3))
bMv=exp(confint(MultGARegCats3))#error functions hasnt been written yet
bAd=exp(confint(AdjucCatsCats3))
bCo=exp(confint(ContRatioCats3))

#classification table
# real y and predicted y with fitted and then xtabs for table
#fitMu=cbind(round(fitted(MultiRegrCats3)),round(fitted(ContRatioCats3)),round(fitted(OrdRegr1
Cats3)),round(fitted(AdjucCatsCats3)))
fitMu1=cbind(fitted(MultiRegrCats3),fitted(ContRatioCats3),fitted(OrdRegr1Cats3),fitted(AdjucCa
tsCats3))
t1=proc.time()
fitMu0=0
fitCo1=0
fitOr2=0
fitAd3=0

for (i in row(fitMu1) ) {
  fitMu0[i]=which.max(fitMu1[i,1:3])[[1]]-1
```

```
  fitMu0[i]=which.max(fitMu1[i,1:3])[[1]]-1
  fitCo1[i]=which.max(fitMu1[i,4:6])[[1]]-1
  fitOr2[i]=which.max(fitMu1[i,7:9])[[1]]-1
  fitAd3[i]=which.max(fitMu1[i,10:12])[[1]]-1
}

proc.time()-t1
fitAll=cbind(fitMu0,fitCo1,fitOr2,fitAd3)#combine all the fits
#fitAll1=cbind(fitMu01,fitCo11,fitOr21,fitAd31)
xtabs(~fitMu0)
xtabs(~fitOr2)
xtabs(~fitAd3)
xtabs(~fitCo1)

View(round(fitted(MultiRegrCats3)))
View(round(fitted(OrdRegr1Cats3)))
View(round(fitted(ContRatioCats3)))
View(round(fitted(AdjucCatsCats3)))

XLo=xtabs(~cats2 + round(fitted(LogRegrGlmCats2)))
xMu=xtabs(~cats3 + fitAll[,1])
xCo=xtabs(~cats3 + fitAll[,2])
xOr=xtabs(~cats3 + fitAll[,3])
xAd=xtabs(~cats3 + fitAll[,4])
#calcualte p-values from z-test and t-test
#dt(tvalue, df) and comapare,
ordtval=c(21.6714,7.8308,4.3362,1.0682,5.0939,-2.3434,1.7412,-8.0387,0.4557,-
1.9771,29.6109,7.1941)
View(dt(ordtval,14))# ordinal p values
library(broom)
View(tidy(MultiRegrCats3)) #multinomila model p values
#summaries
logout1
#logout2
ordout2
```

```
multout
bL1
bO2
bMu
bMv
bAd
bCo

#logistic binomial
write.csv2(as.character(logout1$call$formula),file="log1 formula.csv")
write.csv2(logout1$call$family,file="log1 family.csv")
write.csv2(logout1$deviance,file="log1 deviance.csv")
write.csv2(logout1$aic,file="log1 aic.csv")
write.csv2(logout1$coefficients,file="log1 coeffs.csv")
write.csv2(bL1,file="log1 confidence intervals.csv")
#ordered polR
write.csv2(as.character(ordout2$call$formula),file="ord2 formula.csv")
write.csv2(ordout2$call$family,file="ord2 family.csv")
write.csv2(ordout2$deviance,file="ord2 deviance.csv")
write.csv2(ordout2$coefficients,file="ord2 coeffs.csv")
write.csv2(ordout2$zeta,file="ord2 zeta.csv")
write.csv2(bO2,file="ord2 confidence intervals.csv")
#nnet multinom
write.csv2(as.character(multout$call$formula),file="mult formula.csv")
write.csv2(multout$deviance,file="mult deviance.csv")
write.csv2(multout$coefficients,file="mult coeffs.csv")
write.csv2(multout$aic,file="mult aic.csv")
write.csv2(bMu,file="mult confidence intervals.csv")
#cratio@
write.csv2(as.character(contout@call)       ,file="cont formula.csv")
write.csv2(          contout@criterion    ,file="cont deviance.csv")
write.csv2(          contout@coef3        ,file="cont coeffs.csv")
write.csv2(          AIC(ContRatioCats3)  ,file="cont aic.csv")
write.csv2(          bCo                  ,file="cont confidence intervals.csv")
#acat
```

```
write.csv2(as.character(adjaout@call)       ,file="adj formula.csv")
write.csv2(          adjaout@criterion    ,file="adj deviance.csv")
write.csv2(          adjaout@coef3        ,file="adj coeffs.csv")
write.csv2(          AIC(AdjucCatsCats3)  ,file="adj aic.csv")
write.csv2(          bAd                  ,file="adj confidence intervals.csv")
#vgam multinomial(no need to write it)
#write.csv2(as.character(multGAM$call$formula),file="multGAM formula.csv")
#write.csv2(multGAM$deviance,file="multGAM deviance.csv")
#write.csv2(multGAM$coefficients,file="multGAM coeffs.csv")
#write.csv2(multGAM$aic,file="multGAM aic.csv")
#write.csv2(bMv,file="multGAM confidence intervals.csv")
```

CV

Αντασκαλίτσας Μπογδάν

E-mail: Adascalitsa.Bogdan@gmail.com
Κιν.Τηλ: 6945487533

'Ετος γέννησης: 1996
Φύλο:'Αρρεν

ΣΠΟΥΔΕΣ

Ημερομηνίες (από-εως): 2014 - 2018:

Ανώτατη εκπαίδευση: Τμήμα Περιφερειακής και Οικονομικής Ανάπτυξης του Πάντειου Πανεπιστημίου Αθηνών.

ΕΠΑΓΕΛΜΑΤΙΚΗ ΕΜΠΕΙΡΙΑ

Μερικής απασχόλησης: Αποθήκη – τιμολόγηση - κατάστημα MANETTI

ΑΤΟΜΙΚΕΣ ΔΕΞΙΟΤΗΤΕΣ ΚΑΙ ΙΚΑΝΟΤΗΤΕΣ

Πληροφορική: MS Office, OpenOffice, ERP(διαχείριση αποθήκης,τιμολόγηση), MQL 4, Eviews, SPSS.

Γλώσσες:

Ελληνικά, Επίπεδο: Άριστο

Ρώσικα, Επίπεδο: Άριστο

Αγγλικά,Επίπεδο: Άριστο Πτυχία: Πανεπιστήμιο του Μίσιγκαν : Proficiency (27 ΜΑΙΟΥ 2012)

ΔΙΠΛΩΜΑ ΑΥΤΟΚΙΝΗΤΟΥ-ΜΟΤΟΣΥΚΛΕΤΑΣ

Δίπλωμα αυτοκινήτου: Ναί

Δίπλωμα μοτοσυκλέτας: Όχι