Multinomial regression in the analysis of associated factors in alcohol consumption in adolescent students

Sandra García-Bustos[1], Ysaí Ronquillo[1], Gema Zambrano[1]
[1] Facultad de Ciencias Naturales y Matemáticas, Escuela Superior Politécnica del Litoral, Km. 30.5 Vía Perimetral, Guayaquil, Ecuador, slgarcia@espol.edu.ec, yronquil@espol.edu.ec, gemazamb@espol.edu.ec

Abstract– ***The present study shows the relevance of the implementation of the multinomial logistic regression model (MLR) for risk analysis and identification of risk factors for a condition or event such as alcohol consumption in students.***

A total of 395 young people residing in Portugal and belonging to two schools in the country were analyzed, considering their mathematics qualifications, information about their family history and inclusive their academic data. This study concludes that the increase in the scale of alcohol consumption is related to the fact of studying in a particular school, whether living under the guardianship or not of the parents influences the probability of having a greater scale of consumption. The likelihood ratio and Akaike criteria have been used to determine the variables that influence alcohol consumption.

Key words: Multinomial Logistic Regression, risk factors, alcohol consumption, likelihood ratio, AIC.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2020.1.1.476
ISBN: 978-958-52071-4-1 ISSN: 2414-6390

Regresión multinomial en el análisis de factores asociados en el consumo de alcohol en estudiantes adolescentes

Sandra García-Bustos[1], Ysaí Ronquillo[1], Gema Zambrano[1]
[1] Facultad de Ciencias Naturales y Matemáticas, Escuela Superior Politécnica del Litoral, Km. 30.5 Vía Perimetral, Guayaquil, Ecuador, slgarcia@espol.edu.ec, yronquil@espol.edu.ec, gemazamb@espol.edu.ec

Abstract–This study allows us to determine the relationship between the predictor variables with the probability of increasing alcohol consumption, through extensions of the Logistic Regression such as Multinomial Logistic Regression and Ordinal Logistic Regression.
Three hundred and ninety-five young people living in Portugal and belonging to two schools in the country were analyzed, taking into account information on their family history, and even their academic data. This study concludes that the increase in the scale of alcohol consumption is influenced by studying in one school or another, in the same way, the association between living with parents or not with the increased probability of having a greater scale of consumption. The probability ratio and Akaike criteria have been used to determine the model selection.

Keywords-- Multinomial Logistic Regression, risk factors, alcohol consumption, likelihood ratio; AIC.

Resumen: Este estudio nos permite determinar la relación entre las variables predictoras con la probabilidad de aumentar el consumo de alcohol, a través de extensiones de la regresión logística, como la regresión logística multinomial y la regresión logística ordinal.
Se analizaron trescientos noventa y cinco jóvenes que vivían en Portugal y pertenecían a dos escuelas del país, teniendo en cuenta la información sobre sus antecedentes familiares e incluso sus datos académicos. Este estudio concluye que el aumento en la escala del consumo de alcohol está influenciado por el estudio en una escuela u otra, de la misma manera, la asociación entre vivir con los padres o no con la mayor probabilidad de tener una mayor escala de consumo.

La razón de probabilidad y los criterios de Akaike se han utilizado para determinar la selección del modelo.

Palabras clave: regresión logística multinomial, factores de riesgo, consumo de alcohol, razón de probabilidad; AIC

I. INTRODUCCIÓN

La adolescencia es un periodo de transición entre la niñez y la adultez, en la cual se desarrolla la personalidad, el autoconocimiento y el establecimiento de metas que servirán de motivación para el futuro en la toma de decisiones de los individuos. Sin embargo, esta etapa es vulnerable y se ve influenciado por factores externos del entorno y las interacciones humanas realizadas, provocando la adquisición de hábitos poco favorables como el consumo de sustancias psicotrópicas y de estupefacientes que aumentan riesgos a contraer enfermedades [1].

Estudios científicos han determinado que durante la adolescencia e inclusive en el periodo de la universidad existen una variedad de fuentes que provocan experiencias desagradables con excesivo estrés, que por consiguiente generan problemas emocionales, trastornos alimenticios, adicciones a sustancias, actitudes temerarias y baja autoestima, originados principalmente por hábitos de autoprotección, la exigencia educacional, aceptación social, entre otros factores [2].

Además, los principales problemas de la salud pública de la sociedad actual se encuentran relacionados con los problemas de salud mental y el consumo de bebidas alcohólicas [3], en donde, el consumo de bebidas alcohólicas es influenciado por cambios de desarrollo, medio ambiente y estilo de vida de los jóvenes [4] y a su vez en el entorno físico y psicológico [5].

Por otra parte, las técnicas estadísticas multivariantes son ampliamente utilizadas para entender causas que influyen en ciertos comportamientos o sucesos. Una de estas metodologías son los modelos lineales generalizados que son técnicas de clasificación y predicción dependiendo del objetivo del estudio. Entre estos modelos se encuentra el modelo de regresión logística multinomial (MLR) el cual es una extensión del modelo de regresión logística binomial.

Generalmente en el análisis multivariante se requiere el cumplimiento de supuestos como la normalidad y la continuidad de los datos, sin embargo, cuando estos supuestos son violados afecta en la validez de los resultados de un estudio [6]. Una de las principales características del MLR además de la fácil interpretación y de no asumir la distribución normal en los errores, es la robustez que posee ante las violaciones de los supuestos de normalidad multivariante y la igualdad de matriz de varianza y covarianza entre grupos [7]. MLR es ampliamente utilizada para la resolución de problemas, particularmente en los campos de psicología, finanzas, ingeniería y medicina, especialmente para el análisis de riesgos e identificación de factores de riesgo para una afección/evento/enfermedad dada [8].

Es por ello por lo que se pretende estudiar a través de modelos lineales generalizados, la relación de los factores y

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2020.1.1.476
ISBN: 978-958-52071-4-1 ISSN: 2414-6390

variables de explicación que inciden en la probabilidad de incrementar el consumo de alcohol en estudiantes adolescentes, dado que se ha establecido que la fase inicial de un proceso de adicción es en la adolescencia y cuyos factores relevantes son las influencias psicológicas, sociales y culturales [9].

II. METODOLOGÍA

Para el estudio en cuestión, se analizó una base de datos proporcionada por Cortez and Silva [10] que contiene información de 395 estudiantes de dos escuelas de Portugal. La base de datos presenta las calificaciones de matemáticas de los alumnos, información de su historia familiar y de hábitos diarios recolectados durante los años 2005-2006. El análisis de los datos se realizó con la herramienta RStudio [11]. Inicialmente se hizo un análisis exploratorio con el cual se obtuvo información básica de los estudiantes de las escuelas Gabriel Pereira (GP) y Mourinho da Silveira (MS).

Posteriormente, se realizaron dos estudios de inferencia estadística para poder determinar el efecto que tienen las variables relacionadas a educación, género y datos sociales de los estudiantes en su nivel de consumo de alcohol. Dado que la variable de respuesta es categórica y además fue medida según una escala de Likert de 1 a 5, se utilizaron las técnicas Regresión Logística Ordinal y luego Regresión Logística Multinomial.

Es importante tener en claro las diferencias entre la Regresión Logística, Regresión Logística Ordinal y Regresión Logística Multinomial; además de conocer qué es el AIC.

Regresión Logística

Según [12] el modelo de regresión Logística tiene el propósito de estudiar el efecto que tienen los predictores, sean numéricos o factores, en la probabilidad de éxito de una variable de respuesta binaria, modelo con el cual se puede realizar clasificaciones dadas nuevas observaciones. El modelo logit, el más utilizado, es expresado como sigue:

$$\ln\left(\frac{p_j}{1-p_j}\right) = \vec{X}^t\vec{\beta} + \vec{\epsilon} \quad (1)$$

Básicamente, el modelo indica el comportamiento de los odds (relación entre la probabilidad de éxito y fracaso, $\frac{p_j}{1-p_j}$) en relación con los valores de $\overrightarrow{X_j}$.

Ahora, si la respuesta no es binaria, sino más de dos categorías, esto indica que, de manera particular la variable de respuesta no es de tipo Bernoulli, entonces se tienen dos posibles situaciones, la variable tiene categorías nominales o categorías ordinales.

Regresión Logística Ordinal.

El modelo supone que la relación entre la probabilidad de pertenecer o no a una categoría o grupo g, son proporcionales, es decir solo los diferencia el valor de β_{0g}, mientras que las variables influyen de la misma manera en cada una de las categorías [12]. El modelo viene dado por:

$$\frac{P\left(Y > g\middle|\overrightarrow{X_j}\right)}{P\left(Y \leq g\middle|\overrightarrow{X_j}\right)} = e^{\beta_{0g}+\beta_1 X_1+\cdots+\beta_j X_j} + \epsilon \quad (2)$$

Donde ϵ es el término aleatorio del modelo y *g*=1, 2, ..., *G-1*, e indica el grupo al que se pertenece.

Regresión logística Multinomial.

Agresti [12] indica que este modelo consiste en ajustar varios modelos logísticos con respuesta binaria con una restricción que consiste en una repara-metrización de la variable de respuesta *Y* con el objetivo de que la suma de probabilidades de pertenecer a cada uno de los grupos sea igual a 1, definiendo G variables dummy de la siguiente manera:

$$Y_{jg} = 1, si\ la\ j-ésima\ observación\ pertenece\ a\ la\ g-ésima\ categoría$$

Y la forma del modelo es:

$$P\left(Y_{jg} = 1\middle|\overrightarrow{X_j}\right) = p_{jg} = \frac{e^{\overrightarrow{X_j}^t\overrightarrow{\beta_g}}}{1+\sum_{g=1}^{G-1} e^{\overrightarrow{X_j}^t\overrightarrow{\beta_g}}} + \epsilon \quad (3)$$

$$g = 1, 2, \ldots, G-1$$

$$P\left(Y_{jG} = 1\middle|\overrightarrow{X_j}\right) = p_{jG} = \frac{1}{1+\sum_{g=1}^{G-1} e^{\overrightarrow{X_j}^t\overrightarrow{\beta_g}}} + \epsilon \quad (4)$$

$$para\ cuando\ g = G$$

AIC, Criterio de Información Akaike

Agresti [12] señala que el AIC puntúa un modelo de acuerdo con la cercanía en que sus valores ajustados tienden a acercarse a los verdaderos valores de la variable respuesta, en términos del error de predicción. El modelo óptimo es el que tiende a tener un ajuste más cercano a la realidad en relación con otros modelos con el mismo conjunto de datos. Se calcula de la siguiente manera:

$$AIC = -2\ln(L) + 2k \quad (5)$$

Donde *L* es la verosimilitud del modelo y *k*, el número de parámetros.

En este estudio se probaron dos modelos utilizando tanto regresión Multinomial como logística ordinal, se utilizó el AIC y la razón de verosimilitud como criterios para seleccionar los predictores más importantes.

A continuación, se detallan las variables y la simbología usada en este estudio.
Variable de repuesta **Walc**: Consumo de alcohol en los fines de semana en Escala de Likert.

Muy bajo-1 2 3 4 5- Muy alto

Para el ajuste del primer modelo se usaron como variables predictoras:
Age: Edad del estudiante medida en años, la cual está entre 15 y 22 años
Pstatus: Variable binaria, T: si vive con sus padres, A: si vive aparte
School: Variable binaria, indica si está en la Escuela GP o MS.

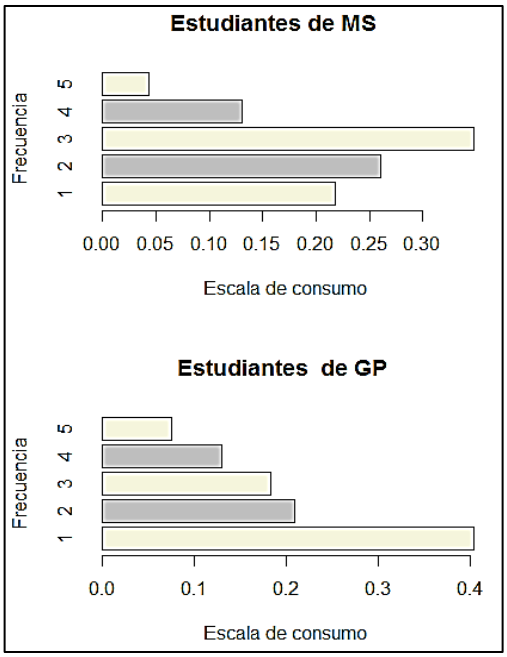
El segundo modelo tiene como variables predictoras:

Goout: Indica la frecuencia con que sale con amigos, en la escala: Muy poco 1 2 3 4 5 Muy seguido
Pstatus: Variable binaria, T: si vive con sus padres, A: si vive aparte.
School: Variable binaria, indica si está en la Escuela GP o MS.
Studytime
Tiempo de estudio semanal, en la escala:
1: menos de dos horas
2: de 2 a 5 horas
3: de 5 a 10 horas
4: más de 10 horas

III. RESULTADOS

Al realizarse el análisis exploratorio de las variables es posible observar que el 11% de la muestra pertenece a la escuela MS y el 89% a la escuela GP. Cerca del 20% de los estudiantes de las instituciones MS y GP tienen un consumo alto y muy alto de alcohol durante los fines de semana (Fig. 1). Por otro lado, la mayor concentración de estudiantes de MS tiene un consumo no considerado ni alto ni bajo de alcohol los fines de semanas; mientras que los estudiantes de GP en mayor proporción consideran que su consumo de alcohol es bajo los fines de semana. Sin embargo, a través del estimador de Tau-b de Kendall (t=0.08), que es una medida de asociación ordinal, no muestra una relación entre el consumo de alcohol en los fines de semanas con respecto a la escuela que asiste el estudiante($p<0.05$).

Fig. 1



Consumo de alcohol en fines de semana

Entre los hallazgos importantes en encuentra que tanto las calificaciones y las relaciones familiares se encuentran altamente correlacionadas positivamente con el ambiente del estudiante.

De la Fig. 2 se observa que hay similitud de las edades en los grupos de estudiantes que tienen una escala de consumo de alcohol entre 2 y 5, a diferencia de los estudiantes que toman una muy baja dosis de alcohol los fines de semana, pues la mediana (línea del centro de las cajas) es menor que en el resto de los casos.

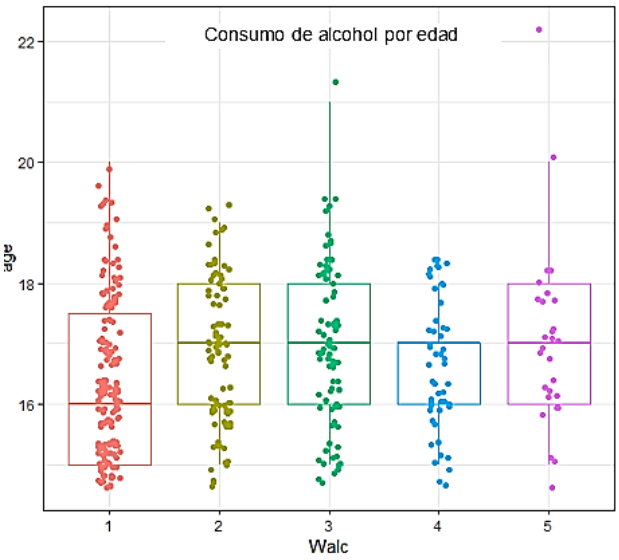


Fig. 2 Diagramas de caja de consumo de alcohol frente a edad

Primeramente, se ajustan los modelos de regresión logística multinomial para posteriormente aplicar los modelos de regresión logística ordinal:

Primer modelo ajustado: Este modelo presenta como variables predictoras: la edad del estudiante, si vive con ambos padres o no y el tipo de escuela al que asiste.

$$Walc \sim Age + Pstatus + school$$

TABLA I
ODDS RATIO DEL MODELO UNO CON AJUSTE DE REGRESIÓN MULTINOMIAL

Walc	Intercepto	Age	Pstatus_T	School_MS
2	0.0088	1.2339	1.9587	1.6613
3:medio	0.0037	1.2427	3.7563	2.4782
4	0.4134	1.062	0.8037	1.7279
5: muy alto	0.0003	1.4769	0.9207	0.6202
AIC	1169.51		Residual Deviance	1137.51

La tabla I muestra los odds ratio, de los cuales se puede interpretar que vivir junto a los padres en lugar de vivir aparte, se relaciona con un aumento en la probabilidad de pasar de una escala de 1 a una de 2 en el consumo de alcohol cada fin de semana (odds ratio mayor a 1). No obstante, la probabilidad de pasar de una escala de 1 a una de 4 o 5 disminuye cuando se compara a alguien que vive junto a sus padres con alguien que vive aparte (odds ratio menores a 1).

La probabilidad de tener una escala de 5 a tener una de 1 disminuye cuando el estudiante es de MS y no de GP.

Segundo modelo ajustado: Este modelo presenta como variables predictoras al tiempo de estudio, si vive con los padres, la escuela en que asiste el estudiante y la frecuencia con la que sale con sus amigos.

$$Walc \sim studytime + Pstatus + school + goout$$

TABLA II
ODDS RATIO DEL MODELO DOS CON AJUSTE DE REGRESIÓN MULTINOMIAL

Walc	(Intercept)	Pstatus_T	goout2	goout3	goout4
2	0.21	1.94	2.52	4.04	1.63
3:medio	0.04	4.11	2.06	7.63	7.89
4	0.28	0.83	1.39	2.08	10.90
5: muy alto	0.00	0.74	227899.70	590202.40	1462432.00
Walc	goout5	studytime2	studytime3	studytime4	school_MS
2	3.18	sa0.51	0.36	0.21	2.13
3:medio	9.63	0.56	0.41	0.20	3.17
4	17.49	0.28	0.07	0.06	1.71
5: muy alto	10508160.00	0.25	0.06	0.18	1.16

De la tabla II se tiene que, la probabilidad de que un estudiante tenga un mayor consumo de alcohol disminuye cuando dedica más de dos horas semanales al estudio en lugar de 2 o menos (odds ratios menores a uno en las variables studytime). La probabilidad de tener una escala de consumo de 2 y no de 1, o una escala de 3 y no de 1, aumenta cuando el estudiante vive con sus padres y no aparte. Sin embargo, la probabilidad de tener una escala de 4 o 5 comparado a tener una escala de 1, disminuye cuando el estudiante vive con sus padres y no aparte. La probabilidad de pasar a tener un mayor consumo de alcohol está influenciado a dedicar más horas a salir con amigos (odds ratio mayor a 1 en las variables *goout*).

TABLA III

PRUEBA DE RAZON DE VEROSIMILITUD PARA COMPARAR LOS MODELOS 1 Y 2

Modelos	Resid. Df	Resid.dev	LR stat.	pvalue
1	1564	1137.51		
2	1540	1010.242	127.2684	0.000

Con la prueba de Razón de verosimilitud presentado en la tabla III, se concluye que existe evidencia estadística para rechazar la hipótesis de que no hay diferencias significativas entre los modelos 1 y 2. Por lo tanto, el mejor modelo con ajuste Multinomial es el modelo 2, pues se obtuvo un menor AIC.

A continuación, se presenta la estimación de los modelos 1 y 2 considerando el modelo de regresión logística ordinal, es decir todas las variables explicativas tienen el mismo coeficiente sin importar la categoría de la variable respuesta, mostrándose la diferencia entre categorías en la estimación del intercepto.

Modelo 1

TABLA IV
ODDS RATIO DEL MODELO UNO CON AJUSTE DE REGRESIÓN LOGISTICA ORDINAL

Age	**Pstatus_T**	**School_MS**	
1.1821	1.1693	1.2119	
Interceptos			
1/2	2/3	3/4	4/5
11.93	29.12	78.98	258.75
AIC	1170.61	Residual Dev.	1156.61

Los resultados muestran que el aumento de la variable edad se vincula con un aumento en la probabilidad de tener una mayor escala de consumo de alcohol. También se observa que vivir junto a los padres se relaciona con el incremento en la probabilidad de tener una mayor dosis de consumo de alcohol los fines de semana. El valor del AIC de este modelo es similar al obtenido con ajuste Multinomial, con la diferencia que en el modelo logístico ordinal se necesitan estimar menos coeficientes.

Modelo 1I

De los resultados mostrados en la tabla V se puede observar que estudiar en MS en lugar de estudiar en GP se asocia a un aumento en la probabilidad de tener una mayor escala de consumo de alcohol los fines de semana. Vivir junto a los padres en lugar de vivir aparte se relaciona con aumento en la probabilidad de tener un mayor consumo de alcohol. Mientras que, si se compara un estudiante que dedica de 2 a 5 horas de estudio semanal con otro estudiante que dedica menos de dos horas y con las mismas condiciones se tiene que disminuye la probabilidad de tener una mayor escala de consumo de alcohol.

TABLA V
ODDS RATIO DEL MODELO DOS CON AJUSTE DE REGRESIÓN LOGISTICA ORDINAL

goout2	goout3	goout4	goout5	studytime2
2.0085	3.7373	8.1717	23.052	0.4658
Pstatus_T	1/2	2/3	3/4	4/5
1.1322	1.3653	3.8751	13.076	52.09
studytime3	studytime4	school_MS	AIC	Resid. Dev.
0.2997	0.1937	1.4418	1086.82	1060.82

TABLA VI
PRUEBA DE RAZON DE VEROSIMILITUD PARA COMPARAR LOS MODELOS 1 Y 2 EN REGRESIÓN LOGISTICA ORDINAL

Modelos	Resid. Df	Resid.dev	LR stat.	pvalue
1	388	1156.61		
2	382	1060.82	95.79068	0.000

Con la prueba se concluye que existe evidencia estadística para rechazar la hipótesis de que no hay diferencias significativas entre los modelos 1 y 2.

Se elige por tanto como mejor modelo, el segundo pues el valor de AIC es relativamente menor.

IV. CONCLUSIONES

Es común que se ajusten modelos de discriminación cuando la variable de respuesta es categórica, sin embargo, los modelos de regresión logística ordinal y el multinomial tienen ventaja en cuanto a que los resultados puede ser interpretados y comprendidos de manera más sencilla. Por eso han sido utilizados por varios autores como en [1], en el cual se usó para determinar los factores que influyen en el consumo de alcohol. Otros estudios han consistido en analizar una asociación el consumo de alcohol y otras variables relacionadas a la salud mental del estudiante [2,4].

Este estudio permitió conocer cómo influye en el aumento de la escala de consumo de alcohol el hecho de estudiar en una escuela o en otra, de igual manera la asociación entre vivir o no con los padres con el aumento en la probabilidad de tener una mayor escala de consumo. Este resultado en particular sugiere a estudios futuros relacionados al consumo de alcohol de estudiantes adolescentes, incluir variables respecto a la escala de consumo de alcohol de los padres y construir modelos considerando esta variable como predictor.

REFERENCES

[1] G. Kapansahim and F. Taner Ersöz, "Investigation of the Alcohol Usage Habits of University Students by Logistic Regression Analysis," *Journal of Institute of Social Science,s,* 9(1), 1-15, 2019.

[2] G. E. Fogle and T. F. Pettijohn, "Stress and Health Habits in College Students," *Open Journal of Medical Psychology*, 2(2) 61-68, 2013.

[3] S. Bell and A. Britton, "An exploration of the dynamic longitudinal relationship between mental health and alcohol consumption: a prospective cohort study, " *BMC Medicine*, 12(91), 2014.

[4] D. Said, "Risk factors for mental disorder among university students in Australia: findings from a web-based cross-sectional survey", *Social Psychiatry and Psychiatric Epidemiology*. 35-44, 2003.

[5] J. Hallett, "Excessive Drinking—An Inescapable Part of University Life?" A Focus Group Study of Australian Undergraduates." *Open Journal of Preventive Medicine*, 04, 616-629, 2014.

[6] B. Tabanick and L. Fidell, *Using multivariate statistics*, 6th ed., Pearson, 2013.

[7] Y. Chan, "Biostatistics 305 Logistic regression Analysis", *Singapore Medical Journal*, 46(6), 2005.

[8] A. Bayaga, "Multinomail Logistic regression : Usage and application in risk analysis", *Journal of Appied Quantitative Methods*, 5(2), 288-297, 2010

[9] D. Abrams and G. Wilson, "Clinical Advances in Treatment of Smoking and Alcohol Addiction", In: Frances Aj, Hales RE, eds. The American Psychiatric Association: annual review, psychiatric update,Vol. 5 Washington, DC: Amerckan Psychiatric press, 1986.
[10]P. Cortez and A. Silva, Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
[11]R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2008. https://www.R-project.org/
[12] A. Agresti, *Categorical Data Analysis*, Universidad de Florida. Florida, USA, 2002